UNIVERSITY OF TÜBINGEN
Department of Computer Science
Chair for the Methods of Machine Learning

Bachelor Thesis Computer Science

Identifying Sources of Unfairness in Bayesian Logistic Regression

Mila Gorecki

Reviewer: Prof. Dr. Philipp Hennig
Department of Computer Science

Supervisor: Alexandra Gessner
Department of Computer Science

Started: January 22, 2019

Finished: June 6, 2019

Erklärung

Hiermit erkläre ich, dass ich diese schriftliche Abschlussarbeit selbständig verfasst habe, keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe und alle wörtlich oder sinngemäß aus anderen Werken übernommenen Aussagen als solche gekennzeichnet habe.

Tübingen am June 6, 2019

____

Mila Gorecki

Gorecki, Mila:
Fairness in Machine Learning
Bachelor Thesis Computer Science
University of Tübingen
January 22, 2019 - June 6, 2019

Abstract. Machine learning processes have been incorrectly perceived as a neutral way of overcoming human bias in decision-making and promoting fairness in social contexts. Instead, algorithmic decisions have been found to reflect and reinforce existing human biases. Since decisions about people can directly affect their chances in live, the question of fairness is inevitable in this context. In recent years, numerous promising approaches have been proposed and discussed to achieve fairness on a group or individual basis. However, to approach fairness in a holistic way, it is clearly not enough to just work on the symptoms. Instead, further investigations are essential to identify the underlying causes and reveal how they relate to observed biases of algorithmic systems. This should happen within the machine learning community, but also requires interdisciplinary discourse and the allocation of responsibility.

The first part of this thesis explores existing approaches to achieving fairness and discusses open challenges. Using Bayesian logistic regression on an exemplary data set the relationships between data, classifier and outcomes is then analysed in more detail in the second part. The aim is to identify different sources of unfairness and thus contribute to the fairness debate. For this purpose, we propose two methods that are suitable for a more detailed analysis of the classifier's decisions. The first is a pre-processing method that modifies the data with regard to fairness criteria. The second method applies after the training of the classifier and makes it possible to examine the classifications and the associated uncertainty more closely. The methods were evaluated on the basis of three fundamental notations of group fairness, but are not limited to these. Through a better understanding of data set and classifier, we hope to uncover potential sources of unfairness in classification tasks and thus contribute to the development of genuinely fair and transparent algorithmic systems.

Acknowledgements

I would like to thank Alexandra Gessner and Niki Kilbertus for their valuable and constructive suggestions in the development and implementation of this research work. Their willingness to give their time so generously has been very much appreciated. I am also very grateful to Philipp Hennig for his willingness to support a thesis on the topic of fairness, for valuable discussions and for providing the opportunity to meet Moritz Hardt, who gave valuable feedback and shed light on broader visions of the relationship between machine learning, fairness and society. Thanks is also due to Claudia Berger for proofreading my bachelor thesis.

Contents

1 Introduction

With the increasing use of technology in everyday life, the number of decisions that are made or at least influenced by algorithms is growing. By now, algorithmic systems permeate almost every aspect of our lives. This also applies to a vast number of domains that directly affect human lives including education, employment, criminal justice, health care and advertisement. In this context, purpose of machine learning algorithms boils down to finding statistical patterns in the data and using them for decision making. Given a set of examples, the computer is expected to derive decision rules that are general enough to apply for unseen cases. The use of math and the limitation of human intervention in this process has led to the widespread perception that such systems tend to be fair and unbiased. In particular, this gave rise to the hope that they could contribute to overcoming human prejudices and implicit biases. However, analysis of various applications suggests that the idea of neutral, unbiased automatic decision making is a misconception. Instead, algorithmic decisions have been found to be prone to exactly those biases that they were hoped to combat. These systems depend crucially on data which itself incorporates human behaviour and characteristics. Additionally, there are several choices required along the machine learning pipeline at which biases can be introduced into the process. As a consequence, machine learning systems are likely to reflect or reinforce existing human biases or to even introduce new biases when applied in social contexts.

A common example to illustrate the need for fairness is the risk assessment tool COMPAS, that is used across U.S. courts. COMPAS, short for Correctional Offender Management Profiling for Alternative Sanctions, is a risk assessment tool developed by Northpointe, now Equivant. Given six attributes of an offender, the algorithm returns a score corresponding to the estimated probability of recidivism in the future. The risk score is supplemented by two further outputs: a needs scale derived from 137 attributes of the offender that aims to help select appropriate treatment plans and additional interventions, and a gender-specific typology that classifies the offender into one of eight prototype categories. These outputs are used in several states across the US to guide and support judges in their decisions as to whether a defendant should be detained or released while awaiting trial. A detailed analysis conducted by ProPublica using risk scores of defendants in Broward County uncovered substantial racial disparities in the outcomes (Angwin et al., 2016). Based on data of recidivism within two years after trial, they compared the risk scores assigned to more than 7000 people. The authors reported significantly higher false positive rates for black defendants and higher false negative rates for white defendants from which they concluded that black defendants were more likely to be classified as high risk. Responding to that, Equivant argued that the risk scores were calibrated which can be interpreted as another measure of fairness. This example has significantly driven research into fair algorithms in recent years as it shows that automatic decision making can have a significant impact on people's lives and implies that the existing notions of fairness might not be sufficient to prevent unfair practices.

As outlined above, the increasing influence and scope of algorithmic decision making are likely to raise ethical questions and call for detailed examination of existing applications and research on how to get rid of the bias. In a long-term view the increasing use of algorithmic systems and its consequences will profoundly affect our society. Therefore it is of special importance to examine the impact of algorithmic decisions, especially when they are biased, and to get a better understanding where such biases originate from. This aims to ensure algorithmic decision making to be free from discriminatory biases in order to avoid unethical practices. What needs to be highlighted is that biased results can occur even if the developer does not have any bad intentions at all. Angwin et al. (2016) and O'Neil (2016) argue that algorithms inadvertently and potentially inevitably encode human bias. It follows logically from this that it is not enough to regard the development of fair algorithms as a one-time task, but rather as a process that requires continuous monitoring and adaptation as well as the involvement of society through public discussions.

In recent years, the area of fairness in machine learning has become very popular so that, by now, numerous researchers commit their resources to understanding and mitigating existing biases in machine learning systems. Although this is an exciting development, it has also created a proliferation of partially contradictory notions of fairness, and some confusion, as the various approaches can scarcely be separated due to the multitude of perspectives. The area of research is still quite young and some fundamental questions have not yet been resolved: What does it mean for a classifier to be fair? How can we capture fairness by abstract mathematical equations? How to measure the impact of fairness-enhancing methods and ensure fairness in the long term?

Structure. This thesis tries to contribute to answering the existing questions through exploring how unfairness can arise in machine learning and what effects it can have. Its structure is divided into two parts. The first part is devoted to current approaches to achieving fairness and discusses open problems. Following this review is the implementation of an illustrative model. Through the use of Bayesian Logistic Regression on the German Credit Dataset different criteria for fairness are examined and techniques are proposed to reduce the bias of the results. This serves the purpose of analysing the dataset and classifier more closely in order to identify possible sources of bias.

2 Review of Fairness in Machine Learning

This chapter reviews the main concepts that are emerging within the machine learning community to foster fairness and understand the potential sources of unfairness in automated decision making. As a starting point, we will first briefly outline how fairness is addressed from a legal perspective and what it means when an algorithm is biased.

Policing and Discrimination Law. Before we delve into the discussions about how fairness can be accomplished technically, it is worthwhile to get a brief idea of how discrimination is defined and regulated in law. The legal implementation can provide good indications, since law serves as a guideline for the development of future technology, and thus algorithmic decisions are also subject to these conditions. As a preliminary it should be noted that discrimination is no universal concept, but instead crucially depends on the context. While it is discriminatory to distinguish between two persons, for examples on the basis of their sex or their skin colour, without regard to their merits, it can be useful or necessary to make exactly this distinction in order to determine the optimal medical treatment. Determining when observed disparities can be considered discriminatory is a key question for discrimination law and consequently also for the definition of fairness criteria. Barocas et al. (2018) name two characteristics that can also be found in U.S. anti-discrimination law. First, discrimination is domain specific, i.e. certain domains require more monitoring because they relate to important opportunities and directly affect people's chances in life. Barocas et al. (2018) list education, employment, housing and loan granting as domains that are explicitly regulated by U.S. anti-discrimination law. Second, discrimination is feature-specific, i.e. grounded in certain characteristics of people on account of which disadvantageous treatment has been justified in the past, or in societal beliefs about which characteristics should not be used for decision making. U.S. anti-discrimination law defines numerous so-called protected classes that correspond to inherent personal characteristics like race, colour, religion, sex, national origin, age, disability or veteran status. Under German non-discrimination law six classes are defined to be protected. Those are race, ethnicity, gender, religion, disability status, age and sexual identity. Similar to U.S. law the German non-discrimination law explicitly applies to domains like education, employment and housing. What is considered as an objectionable characteristic might change over time, as do those explicitly named categories.

U.S. anti-discrimination law further distinguishes between disparate impact and disparate treatment (Barocas and Selbst, 2016). Similar concepts can be found in German non-discrimination law as well. The latter involves formal or intentional unequal treatment, that is the purposeful attempt to discriminate against a person by explicitly or implicitly considering protected classes for decision making. Disparate impact, on the other hand, comprises unjustified or avoidable forms of discrimination that result in disparate outcomes and thereby disproportionately hurt or benefit a group of people sharing one or more protected attributes. Both concepts approach different aspects of fairness. While disparate treatment aims for equality of opportunity, disparate impact seeks to achieve equality of outcome. Barocas et al. (2018) explain clearly that even at the legislative level there is a discrepancy or tension between the objectives of different notions of fairness. At this point it should be noted that this tension between approaches towards fairness is not a characteristic of decisions made by machines, but is also present in human decisions.

What is Bias? In the given context, the term *bias* describes demographic disparities (disparate treatment or impact) in algorithmic systems, which is questionable for societal reasons (Barocas et al., 2018;

Lipton and Steinhardt, 2018; Mitchell et al., 2018). It has to be distinguished from the concept of a *biased estimator*, which describes a characteristic of a statistical model and refers to the relation of an estimator's expected value and the true value of a parameter. An estimator is unbiased if the expected value corresponds to the true value and thus biased if it under- or overestimates the true value systematically.

Bias induced by algorithmic systems, on the other hand, is commonly used as a collective term that can be attributed to disadvantageous consequences of choices involved in the process of machine learning as well as the properties of training and test data. Referring to the overall machine learning pipeline, the Obama Administration's Big Data Working Group argued discrimination to "be the inadvertent outcome of the way big data technologies are structured and used" (Podesta et al., 2014) and called for "equal opportunity by design" in the follow-up report from 2016 (Executive Office of the President et al., 2016). More recent work by Suresh and Guttag (2019), devoted to identify potential sources of bias, proposes a well-structured framework that separates different sources of consequential damage from each other and thus shows that bias cannot be traced back exclusively to a harmful properties of data.

Structure. This review covers approaches taken within the machine learning community to ensure fairness. Given that this movement is still relatively young, it should be underlined that interdisciplinary exchange is essential in order to include philosophical, ethical and social viewpoints and to comprehensively address fairness. The review highlights the technical approaches, so it makes no claim to completeness, but hopes to further foster interdisciplinary exchange. The first part examines various approaches to mathematically define the concept of fairness. Considerations on operationalizing these definitions in order to create fair algorithmic systems will be explored in the second part. Finally, based on the multitude of available methods, literature will be discussed which deals with the sources of biases in algorithmic systems and further challenges that arise in the development of fair systems.

2.1 Defining Fairness

2.1.1 Common structure

Machine learning methods can be divided into supervised and unsupervised learning. Although fairness plays an important role for both scenarios, when applied in social contexts, this review focuses on supervised learning, i.e. fairness definitions for regression and classification. A discussion of fairness in unsupervised learning tasks falls outside the scope of this review.

In a typical classification scenario the machine learning algorithm has access to a set of training points and labels $(X,Y), X = \{x_1,...,x_n\} \subset \mathcal{X}, Y = \{y_1,...,y_n\} \subset \mathcal{Y}$. The training set is then used to generate a classification rule $f : \mathcal{X} \rightarrow \mathcal{Y}$ that takes new data points as input and predicts an outcome or membership to one of the classes. During the training process the classifier is optimized to maximize accuracy, which is often described as finding regularities in the available data in order to infer a rule. Later, the performance of the machine learning algorithm will be evaluated using a set of test points and labels

Variable	**Meaning**
$X = \{x_1,...,x_n\}$	Data points
$Y = \{y_1,...,y_n\}$	Target variable/label (true label)
$a \in A$	Sensitive attribute
$C = c(X,A) \in \{-1,1\}$	Classifier making predictions
$R = r(X,A) \in [0,1]$	Score function

Table 2.1: Variables used in the context of fairness.

that is independent from the training set and has not been used so far. The objective of this task is to generalize from the training data such that the algorithm performs well on unseen data like the test set, but also future data points.

In the context of fairness it is often assumed that the data set contains one or multiple sensitive attributes *A* (also protected attributes). What qualifies as a sensitive attribute depends on societal norms and has also been defined by law as mentioned above. In general, these are attributes that have served as a basis for discrimination in the past. It is further assumed that fairness or at least some aspects of fairness can be defined and operationalized mathematically. As we will see later, those definitions of fairness can then be used to restrict the space of possible classifiers. Without imposing such a fairness constraint, the classifier is optimized for accuracy during training. However, in many cases the fulfillment of fairness criteria is at odds or at least in parts conflicting with the goal of achieving maximal accuracy which then leads to a fairness-accuracy-trade-off to be made.

2.1.2 Towards Fairness

A naive approach to achieving fairness is to ignore protected attributes like race, sex, disability or cultural membership. This is also referred to as “fairness through blindness” or “fairness through unawareness” where protected attributes are not explicitly used for decision making. However, information about the protected attribute can also be redundantly encoded, which makes it possible to predetermine the value of the protected attribute by unprotected attributes, or used only indirectly via proxies (Pedreshi et al., 2008; Hardt et al., 2016). In contrast to humans, machine learning algorithms are much better at capturing proxies in the data in order to maximize accuracy so that blindness turns out to be ineffective (Narayanan, 2018). In addition, explicitly ignoring the protected attribute can also have the opposite effect, leading to errors occurring mainly in a subgroup of people who share the same value in the protected attribute, which then leads to discrimination against that same subgroup (Dwork et al., 2012). Instead, consciousness about the values of protected attributes has turned out to be more efficient and has led to the development of more advanced approaches.

The available concepts of fairness can be classified according to different aspects. A common division is the distinction between group and individual fairness, which corresponds to the question '*For whom is fairness achieved?*'. Another approach is to subdivide the concepts by their implementation ('*when or how?*'): while pre-processing the data, as a constraint during training, or while post-processing the outcomes. Since there are several sources of bias, it has also been suggested that more emphasis should be given to the type of bias the different definitions are attempting to mitigate ('*what?*'). All three aspects will be covered in the course of this review while focussing on group fairness and individual fairness. In order to illustrate the meaning of the individual criteria or aspects of fairness, I will describe them using the examples of loan granting and hiring throughout the review.

2.1.3 Group Fairness

Approaches to achieve group fairness address systematic disparities in outcomes between demographic groups. Within such a group, members are assumed to share certain attributes and especially the sensitive attributes. Barocas and Hardt (2017) suggest that most notions of fairness that have been proposed so far can be reduced to a set of three fundamental fairness notions: independence, separation and sufficiency. All three criteria can be formulated purely based on statistical measures as summarized in a confusion matrix. Such a matrix is visualized in Figure 2.1. It describes the relationship between true and predicted class membership. Describing the nature of these criteria of group fairness in other words, each criterion is a property of the joint distribution of data, labels, the sensitive attributes and the classifier. Hardt (2017) calls such criteria observational criteria because they result from passively observing the word and do not per se include interventions. This comes with severe inherent limitations which will be discussed later. The following paragraphs are devoted to the three fundamental criteria for group fairness as suggested by Barocas and Hardt (2017).

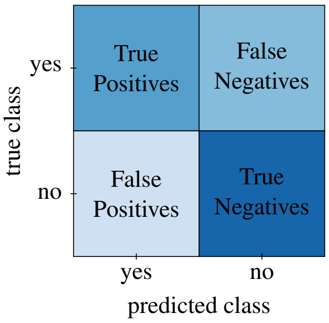


Figure 2.1: For a binary classification task the confusion matrix is comprised of four fields, called true positives, false negatives, false positives and true negative. Depending on the normalisation, however, the exact meaning of those field can vary. In general, high values on the diagonal correspond to a high classification accuracy. Additionally, the confusion matrix visualizes the distribution of different types of error.

Independence. The notion of independence requires the classifications to be independent of the sensitive attribute,

$$C \perp A \;\Leftrightarrow\; P_a\{C = c\} = P_b\{C = c\}. \tag{2.1}$$

For all groups a, b in the population, the probability to receive a positive (or negative) classification has to be equal in order to satisfy independence. As a result the demographics of those receiving a certain classification are identical to the demographics of the population as a whole and, given an individual classification, no inference on the group membership is possible (Dwork et al., 2012). In the case of hiring, independence requires all demographic groups to have equal probability to get hired (or not to get hired), i.e. assigned with the positive (or negative) class. Regarding the task of loan granting, independence requires all groups to be equally likely (or unlikely) to receive a credit loan.

Independence has been proposed several times (Kamiran and Calders, 2009; Calders et al., 2009; Dwork et al., 2012; Zemel et al., 2013; Zliobaite, 2015; Zafar et al., 2017b) and arises from the concept that people should initially have equal chances of being placed in the positive class. It is further motivated by legal support. To determine whether a decision process has led to disparate impact, U.S. anti-discrimination law takes the four-fifths- or 80%-rule as a guideline. This refers to the ratio of the minority group of people with certain sensitive attributes to the majority, or in general to the group with the highest acceptance rate. For example, if a company hired 50% of the men and 10% of the women applying for a job, the ration would be 10:50 and consequently the rate of females accepted for the job is 20% of the acceptance rate of men. Since 20% is significantly lower than the legally legitimized difference of 80%, a person feeling discriminated could call for investigation of gender-related discrimination. Unless it can be demonstrated by the defendant that this imbalance is due to business necessity or due to a lack of alternatives, a ratio below four fifths is regarded by courts as evidence of disparate impact (Barocas and Selbst, 2016).

Equivalent variants of this notion appear under the terms *demographic parity*, *statistical parity*, *equal acceptance rates*, *group fairness* or *disparate impact*. While most definitions agree in that they express the same measurement in different perspectives (as probabilities, expectations or mutual information), Zemel et al. (2013) suggest to come up with a fair representation of the feature space that makes the data independent of the protected attributes and exploiting that representation space for training.

Reformulations as ratio condition or inequality that allow for error of size ε can be viewed as relaxations of independence. These more relaxed definitions are consistent with legal practice, but may not be sufficient to achieve algorithmic fairness, because the classifier might exploit the leeway to maximize accuracy. Following the legal regulations under U.S. law closely, Feldman et al. (2015) formalize such a relaxation, which allows them to certify data sets as free of disparate impact. Corbett-Davies et al. (2017) and Verma and Rubin (2018) present a relaxation called *conditional statistical parity*, which allows for a set of legitimate factors to influence the decision. This approach addresses one of the main limitations of independence: Since this notion excludes correlation between the target labels and the sensitive attribute, it does not allow for the perfect classifier $C = Y$ if there exist such correlations, and prevalence between groups differs. A loss in utility can be the consequence. Moreover independence does not prevent different selection methods for different groups, also referred to as laziness. While candidates in one group might get selected carefully, it is possible to randomly select candidates from the other group regardless of their qualifications. Although the selection process is obviously not fair, equal acceptance rates can be achieved by this strategy such that the notion of independence is satisfied. Such behaviour, although undesirable, occurs automatically when there is no or too little training data available for the second group, which in more general cases is the minority. Hardt (2014) refers to this as sample size disparity, because the uneven distribution of available data across the groups leads the classifier to make more accurate decisions for individuals belonging to the majority for which substantially more data is available. This might result in disparate outcomes for different demographic groups, assuming that those groups are heterogeneous regarding specific attributes.

Separation. The second criterion is separation. It requires the classifications to be independent of the sensitive attribute given the true labels,

$$C \perp A | Y \ \Leftrightarrow P_a\{C = c | Y = y\} = P_b\{C = c | Y = y\}. \tag{2.2}$$

Here, sensitivity (positive prediction given positive true class) and specificity (negative prediction given negative true class) are supposed to be equal for all groups in the population. With regard to the classification task, loan granting separation requires that all applicants receive a positive (or negative) prediction regardless of group membership, given that they would pay back their loan (default) according to ground truth. It applies equally in the case of hiring, where the labels describe whether the applicant has been hired. The notion of separation is appealing because it allows for perfect classification while at the same time avoiding the problem of group-specific sampling strategies by punishing models that perform well only on a subset of demographic groups. Hardt et al. (2016), who propose an equivalent variant of separation called equalized odds, further argue that separation is easier to achieve for more accurate classifiers, which aligns the requirement of fairness with the goal of machine learning to make highly accurate predictions. Identical formulations can also be found in the paper by Chouldechova (2016) termed as *error rate balances* and in the paper by Zafar et al. (2017a) termed as *avoiding disparate mistreatment*, where misclassification rates are measured as fractions over the class distribution in the ground truth labels. Depending on the task, it might be more important to achieve either equal true positive rates or true negative rates across groups. This might be the case when one of the outcomes is associated with substantially higher profit or less harm. Relaxations of this kind have been introduced under the terms *equal opportunity* (Hardt et al., 2016), *balance for the positive or negative class* (Kleinberg et al., 2016) and *predictive equality* (Chouldechova, 2016; Corbett-Davies et al., 2017).

Acknowledging that separation (just like independence) is a purely observational criterion, i.e. fully depends on the joint distribution of data, sensitive attribute and labels, Hardt et al. (2016) also illuminate possible limits of the criterion in their paper. They create two scenarios that both satisfy separation, but lead to fundamentally different interpretations in the context of fairness. Another revealing example comes from Corbett-Davies et al. (2017) in the context of the debate over the COMPAS score. They demonstrate that equality of false positive rates (pendant to true negative rates) is not sufficient to avert unfair practices, because the measurements can be manipulated by external changes of the real-world process. Assume that two groups, a minority and a majority, are given, in which nobody recidivates according to ground truth. Further, suppose a threshold is set to determine whether an arrested person is detained or released prior to trial. People with a COMPAS score below the threshold are released, people with a higher score are detained. In order to achieve equality of false positive rates between those groups, it would be a possible change of the real-world process underlying the data, though unquestionably unfair, to arrest more low-risk individuals belonging to the minority. This leads to a decrease of the false positive rate and the detention rate for the minority, as a greater proportion of members is released. Whilst equality of false positive rates can be achieved thereby, it may in other forms entail disadvantages for the minority, such as increased surveillance or more frequent arrests. This illustrates that the false positive and detention rates themselves do not provide enough information and cannot prevent unfair methods. Although the foregoing argumentation only refers to a relaxation of separation, it could also be shown that the criterion as such does not prevent disparities between the groups in the longer term.

Sufficiency. The third criterion, the notion of sufficiency, requires the true labels to be independent of the sensitive attribute given the classifier's predictions,

$$Y \perp A | C \;\Leftrightarrow P_a\{Y = y | C = c\} = P_b\{Y = y | C = c\}. \tag{2.3}$$

This formalization corresponds to the requirement of equal positive and negative predictive value across all groups: Conditioned on the prediction, the probability that the prediction matches the ground truth label has to be the same regardless of group membership. In the case of loan granting, sufficiency is satisfied when the score already subsumes the sensitive attribute so that the probability to pay back the loan, given a person is granted the loan, i.e. predicted to pay back, and the probability to default, given a negative prediction, is equal across groups. In the context of hiring, people from all groups should have the same probability to be "correctly" hired (or rejected) when they receive a positive (or negative) classification. Here, "correctly" means that the true label matches the predicted one, and that the classifier correctly captures the capabilities of the applicant as measured by the available features.

Equivalent notions have been proposed by Chouldechova (2016) under the term *calibration* and by Berk et al. (2018) under the term *conditional use accuracy equality*. Chouldechova (2016) defines calibration in terms of a score, where for every possible value of the score the fraction of positive instances among all instances assigned to this score has to be equal across groups. As briefly outlined in the introduction, calibration was one of the fairness criteria called upon in the COMPAS debate. Equivant claimed their score to be calibrated within groups, which has also been confirmed by Flores et al. (2016). According to Chouldechova (2016), calibration is an established fairness measurement within educational and psychological contexts, which further motivates its use. Since, again, there might be settings in which it is of particular importance to ensure equality for one of the two statistical measures, corresponding relaxations have been formulated. One of those is predictive parity, where only the positive predictive value of the classifier is required to be the same for all groups (Chouldechova, 2016).

Similar to separation, sufficiency is optimality compatible, i.e. it allows the classifier to be perfect. Moreover, the notion realizes a perspective of fairness that might be desirable: Given a person is accepted, i.e. received a positive prediction, he or she is given an equal chance to succeed (true label $Y = 1$). However, as Barocas et al. (2018) remark, sufficiency often follows from unconstrained training and is satisfied without any additional constraints. That unfair outcomes have been reported for existing applications suggests that sufficiency is not effective enough to prevent unfair practices. In particular this is the case, when the sensitive and non-sensitive attributes are correlated, which is the case in many real-world scenarios. This emphasizes that sufficiency might not be enough for real-world applications.

Impossibility Theorem. The three presented notions of fairness provide different perspectives on the same general goal. Each criterion comes with different desirable properties, prompting efforts to combine their advantages to achieve more general fairness. However, Kleinberg et al. (2016) and Chouldechova (2016) have shown that they are mutually exclusive, i.e. no two of them can be satisfied simultaneously, except for degenerate cases like the perfect classifier or equal prevalence for all groups. See also the explanation provided by Barocas and Hardt (2017) for this proof. The incompatibility also holds for approximations of the criteria and does not rely on the choice of algorithms for decision making (Kleinberg et al., 2016). As a consequence, trade-offs are necessary, which involve contextual consider-

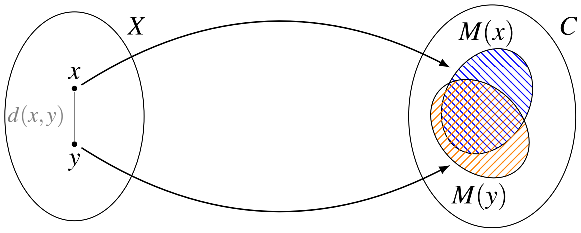


Figure 2.2: Mapping two individuals x,y to similar distributions over outcomes as proposed by Dwork et al. (2012).

ations and are likely to generate tensions between different stakeholders as they pursue different goals (Narayanan, 2018).

2.1.4 Individual Fairness

As opposed to group fairness, individual fairness pursues the goal of treating similar people similarly. The first approach into this direction has been published by Dwork et al. (2012). In their paper "Fairness through Awareness" they formalize the idea that a classification is only fair, when people who are regarded as similar for the classification task at hand receive a similar outcome. Presupposing that there is such a similarity metric, they propose to constrain the algorithm during training with a Lipschitz condition over the outcomes:

$$M : X \to \Delta(C),\ D\left(M(x), M(y)\right) \leq d(x,y). \tag{2.4}$$

As visualizes in Figure 2.2, any two individuals x,y at distance $d(x,y) \in [0,1]$ have to be mapped to distributions $M(x), M(y)$ over outcomes, so that the statistical distance $D(M(x),M(y))$ is at most $d(x,y)$. A classifier that minimizes the expected utility loss is considered to be fair if it fulfils the fairness constraint. Dwork et al. (2012) further explore their notion and its relation to independence, which is not sufficient itself, but can be implied by individual fairness. To show this, the authors introduce the Earthmover distance $d_{EM}(A,B)$, which describes the distance between two distributions A and B. They derive that statistical parity is implied by the Lipschitz constraint if and only if the two distributions are already similar, i.e. the prevalence to receive a certain outcome is similar across groups. Because prevalence is often different between groups in reality, authors examine the possibility of affirmative action. Affirmative actions describe the systematic preferential treatment of demographic groups who have previously suffered from discrimination in order to establish equality. This concept is also controversially discussed in reality as it can lead to reversed discrimination. Using a comprehensible example in this context, Dwork et al. (2012) come to the conclusion that statistical parity and Lipschitz conditions are difficult to combine if the demographics of the groups considered differ significantly from one another. While imposing parity does not do justice to the loss of the decision maker, using only the Lipschitz constraint may not be sufficient to completely prevent unfair practices. In this context, they propose a combination of statistical parity and a relaxed Lipschitz condition that allows affirmative actions. In conclusion, the authors argue that further efforts are needed to ensure fairness when multiple sensitive attributes are present and the demographic groups are not mutually disjoint.

Barocas and Hardt (2017) motivate considerations as to what extent measured similarities between people are informative. Individuals can also appear dissimilar from the perspective of a current measurement, but should be treated similarly assuming that their dissimilarity is caused by injustices they faced in the past. Comparable reasoning encourages the Dwork et al. (2012) to implement affirmative actions and propose an alternative optimization problem, which unites a relaxed Lipschitz condition and statistical parity. However, the primary criticism remains: The authors presuppose the existence of a similarity metric and thereby shift the problem of achieving fairness to finding a suitable metric which allows to determine the degree to which people can be deemed similar in a given task. Dwork et al.

(2012) acknowledge this problem, and refer to a existing realization for a health care setting, but miss to give further guidelines on the nature of suited formalization of similarity. Finding such a function and justifying that it is suited for a task at hand therefore remains a major challenge when aiming to implement this notion of individual fairness.

As Rothblum and Yona (2018) points out, the generalizability of the classifiers from the training data to new data, poses another problem in the work of Dwork et al. (2012). While they still presuppose a similarity measure, Rothblum and Yona (2018) address the generalization from training data to the underlying population by proposing a relaxation of the idea to treat similar people similarly. This allows them to give generalization guarantees using linear and logistic regression techniques.

Zemel et al. (2013) also build on the approach by Dwork et al. (2012), but formulate fairness as an optimization problem that requires to come up with an intermediate representation space $\mathcal{Z}$ of the input space $\mathcal{X}$. This representation is designed to maximize the mutual information of the data set X and representation Z, while simultaneously minimizing the mutual information between A and Z. This means that the representation encodes as much information as possible about the non-sensitive attributes and as little information as possible about membership in the protected class. Consequently, a classifier trained on the new representation Z satisfies independence. In order to create Z, Zemel et al. (2013) define a mapping from individuals $x \in X$ to probability distributions in the representation space Z. The authors further allow the input features of the data set X to have varying impact on the representation, i.e. individual weights. They argue that the use of the same mapping for individuals in the data set to the representation spaces supports individual fairness as similar inputs are mapped to similar representations (Zemel et al., 2013). To allow for more flexibility of the model, they suggest to adapt a weight per group, which would allow to encode similarity within group. This might support individual fairness by rebalancing differences between demographic groups, but it can also cause distortions due to different (unsuited) weighting of the attributes in different demographic groups. Extending the idea of fair representation learning, Edwards and Storkey (2015) and Madras et al. (2018) approach fairness through adversarial learning.

Further approaches to turn the concept of individual fairness into concrete algorithms has been put forward by Joseph et al. (2016a,b), who address individual fairness in a multi-armed bandit problem using the chaining of confidence intervals. They overcome the problem of finding a suitable similarity metric by introducing a notion of regret, that is experienced by the algorithms for each choice.

2.1.5 Causal Reasoning

Another promising approach is causal reasoning (Kilbertus et al., 2017; Kusner et al., 2017), which describes the given task as a causal structural equation model and tries to determine conditions for the corresponding causal graph in order to ensure fairness. One approach would be to demand, for example, that no causal path can be found from the sensitive attribute to the label. This type of analysis is also called counterfactual fairness. Its aim is to assess the influence of the sensitive attribute by changing the values of this very attribute (asking *'what if?'*). Pearl et al. (2016) provides an introduction to central concepts and methods of casual reasoning.

2.2 Achieving Fairness

Several authors have defined fairness criteria and examined existing automated decision making systems based on them. Probably the most prominent analysis of this kind was conducted by ProPublica, who report racial disparity to be present for the risk assessment tool COMPAS Angwin et al. (2016). Beyond the diagnosis of fairness or rather unfairness, numerous papers have been published proposing techniques to meet the existing definitions. They can be grouped according to where in the machine learning pipeline they are applied: as a pre-processing step, during training of the classifier or afterwards as a post-processing step. These approaches are outlined in the following paragraphs.

2.2.1 Pre-Processing

Achieving fairness by pre-processing the data follows the assumption that the available training data is already biased, and aims to prevent that this bias is propagated through the decision making process. The proposed techniques describe transformations of the feature space into a representation that satisfies the desired fairness criterion. They are also referred to as data-based methods.

A straight-forward strategy is to modify the training labels so that positive labels are equally distributed across groups. Comparable approaches to remove dependencies between labels and the sensitive attribute can be found in Pedreshi et al. (2008), Kamiran and Calders (2009) and Calders et al. (2009). The latter two describe this as '*massaging the data set*', which involves an initial ranking of the training points using a Bayes classifier. Based on the resulting ranking, the labels of the negatively labelled data points from the protected group holding the highest ranks are flipped. Alternatively, Calders et al. (2009) propose to reweight the training points based on frequency counts as this does not directly modify the data itself but rather its impact. Hajian and Domingo-Ferrer (2013) criticize that this strategy does only attempt to overcome direct discrimination based solely on the sensitive attribute. They further argue that discrimination might also arise from a combination of different attributes, making it necessary to come up with a data transformation that removes direct and indirect discrimination. They provide several such algorithms for data pre-processing using classification rules, and successfully apply them to two data sets.

Another strategy has been used by Feldman et al. (2015) and Lum and Johndrow (2016). Both apply transformations on the data set using the conditional cumulative distribution of X given the sensitive attribute A. In the first part of their paper Feldman et al. (2015) focus on the diagnosis of fairness, but given a data set is certified to hold disparate impact, they also propose a technique to repair the data set in order to remove the dependency of classifications and group membership. The repairing only work for numerical attributes, but allows to preserve ranks within subgroups. Lum and Johndrow (2016) attempt to overcome this limitation and provide a method using chained conditional models. They successfully apply their method to the COMPAS data set which comprises binary, continuous and discrete attributes. Aiming for both independence and individual fairness, Zemel et al. (2013) propose to come up with an intermediate representation of the feature space $\mathcal{X}$ that is independent of the sensitive attributes and additionally addresses individual fairness (see subsection 2.1.4). The authors successfully apply their pre-processing technique to three data sets, demonstrating that the 'fair and rich' representation Z helps to achieve independence, but also allows to adapt different similarity functions when awareness of the group membership prevents disparate treatment.

To sum it up, there are basically two strategies present in the group of data-based techniques. The first strategy attempts to achieve fairness by changing the labels or the values of the sensitive attribute. The second strategy involves the application of transformations that map the training data into a representational space which enables independence from the sensitive attribute. In the papers discussed here, the sensitive attribute is typically assumed to be a single, binary attribute.

Modifying solely the training data can be both an advantage and a disadvantage of this approach. On the one hand, no modification of the learning algorithm is necessary and it seems sufficient to ensure fair data without having to consider the further process. In addition, the sensitive attribute is no longer necessary during testing. On the other hand, pre-processing treats the learning algorithm in a black box fashion, which can lead to unpredictable losses in accuracy (Zafar et al., 2017b), i.e. they cannot give guarantees about the degree of discrimination in the final classifier. Moreover, since the classifier is not available at this stage, the only criterion that can be achieved by modifying the data is independence from the sensitive attribute or individual fairness with an appropriate metric.

2.2.2 During Training

The second group of methods aims to constraint the training in order to modify the model directly. Those methods are applied during training and also referred to as model-based techniques. An algorithmic system is optimized for maximal accuracy (or minimal error) during training in order to maximize its utility. The idea of methods that are applied during training is therefore quite intuitive: express the desired fairness criterion as a constraint and add it to the optimization objective of the algorithm.

Zafar et al. (2017b) use convex margin-based classifiers to realize a method that is in compliance with both disparate treatment and disparate impact, including a formulation of the constraint that handles the case of business necessity. It is important to emphasize that their approach builds on margin-based classifiers, i.e. the goal of training is to find a decision boundary that separates the training points according to their label with maximal accuracy. Arguing that the 80%-rule cannot be formulated as a convex function, which hinders its incorporation into the objective of the algorithm, they propose a proxy for the 80%-rule: the covariance between the sensitive attributes and the signed distance from the feature vectors to the decision boundary. The decision boundary covariance is a convex-function and can therefore be directly incorporated as a constraint into the optimization problem. Zafar et al. (2017b) successfully validate their method by applying the resulting algorithms to three datasets, one of which is synthetic. This paper is concerned with the notion of independence. In an accompanying paper, Zafar et al. (2017a) propose the notion of avoiding disparate mistreatment, which is equivalent to separation, and present ways to incorporate this notion into the optimization problem as convex-concave constraints. Again, the authors propose a convex proxy, because the differences between misclassification rates are in general non-convex. Following the definition of decision boundary covariance, they propose to compute the covariance between the sensitive attributes and the signed distance between the feature vectors of misclassified users and the decision boundary. Converting the resulting constraints into a disciplined convex-concave program allows them to finally incorporate a notion of separation as a set of constraints into the optimization problem. The authors achieve good results on both a synthetic data set and a real-world data set, but name challenges of the application of their method that require further investigations. There is a considerable amount of further literature that examines how to constrain classifiers during training. Among them are Calders and Verwer (2010), who describe approaches to modify naive Bayes classifiers so that the outcomes are independent of the sensitive attribute, and Kamishima et al. (2012), who use a regularization term to penalize discrimination for logistic regression.

A prerequisite for these model-based techniques is access to the raw data, since the criterion always relates to the available data, and the training process, which in many cases limits the practicability. Additionally, as Zafar et al. (2017b, 2019) acknowledge, the available methods suffer from other limitations. Many formulations are restricted to a narrow range of classifiers. Especially less transparent algorithms like neural networks are rarely considered. Moreover, many techniques rely on binary sensitive attributes, which does not transfer well to more complex real-world challenges. Nevertheless, the idea of constraining the training is appealing, not least, because it offers, depending on the classifier, flexibility in the fairness-accuracy-trade-off. In addition, the sensitive attribute should no longer be needed at test time. Another consideration deals with the aspect that responsibility and also associated obligation are transferred to developers of decision making algorithms, when fairness can be achieved during training.

2.2.3 Post-Processing

The third category of methods, known as post-processing, seeks to achieve fairness by modifying the outcomes of the trained classifier to comply with the desired fairness criterion. It comes with the main advantage that it can be applied to any classifier and only requires access to the sensitive attribute and the corresponding outcome. Furthermore, post-processing does not require any retraining, which makes it applicable at low cost. Since many algorithms are proprietary and the training processes are opaque, post-processing is often the only way to analyse the behaviour of a classifier and adjust the outcomes respectively. As a drawback, post-processing often leads to a significant utility loss of the classifier. Moreover, it requires information about the sensitive attribute for all test points.

Hardt et al. (2016) developed a post-processing step that can be applied to any classifier in order to achieve separation or equalized odds as they call it. They motivate their solution geometrically using the receiver operator characteristic (ROC) curve, which plots the true positive rate against the false positive rate at different thresholds. They further reason that a score function satisfies equalized odds without restriction if the ROC curves of all groups match. Since this is often not the case, the intersections of the group-specific ROC curves must be considered, as they obey equalized odds for the given threshold. Points in the intersection below the ROC curves correspond to trade-offs that only give a suboptimal result. Their method is relatively simple and effective as they give guarantees on optimality preservation. Additionally, the derived predictor minimizes loss and is therefore aligned with the general machine learning goal of maximizing accuracy. Hardt et al. (2016) conclude with considerations when to apply post-processing and encourage further steps to invest in better features and more data.

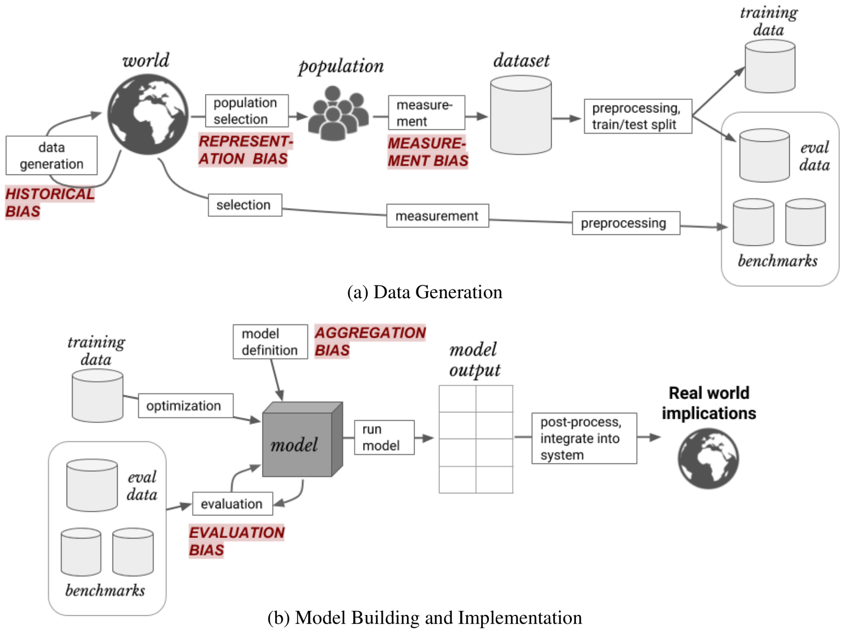
A recent work by Woodworth et al. (2017) proposes a two-step framework which constrains the training in the first step and utilizes post-hoc corrections after training, which is similar to the post-processing step formulated by Hardt et al. (2016).

2.3 Sources of Bias

The previous section discussed options how to achieve fairness, given a data set and a specific classification task. The presented techniques modify the machine learning pipeline while applied either before, during or after training of an algorithm, pursuing to satisfy a specific fairness criterion of interest. To address fairness in machine learning, it is clearly not enough to fight the symptoms, but it is essential to identify the underlying causes and reveal how they relate to observed biases of algorithmic systems. Hence, it is crucial to consider which types of bias are counteracted by the different available techniques. As indicated earlier, biased outcomes are not simply the result of biased data sets and rarely originate from bad intentions on the part of the developers. Instead, the causes are more complex and non-trivial to solve.

Already in the 1990s, Friedman and Nissenbaum (1996) recognized that undesirable dynamics could emerge from the interaction between technology and society, which impedes the identification and elimination of systematically unfair outcomes for certain groups. In their paper they distinguish three categories of bias –pre-existing, technical and emergent bias– that facilitate the diagnosis of causes and the allocation of appropriate countermeasures. In order to minimize bias in algorithmic systems, they suggest several approaches once the type of bias has been determined. However, these approaches remain relatively vague and can also be understood as a general demand to act proactively and anticipate potential problems.

In the light of the increasing use of machine learning in social contexts, several papers have focused on the explanation and systematization of potential sources of bias. Barocas and Selbst (2016) explore how discrimination can arise as an artefact of the data mining process, and discuss the relation to legislative resolutions under U.S. anti-discrimination law. Driven by the observation that the term 'bias' refers to both neutral descriptions and negatively connotated properties of algorithms, Danks and London (2017) propose a taxonomy for algorithmic bias in autonomous systems in order to disentangle different meanings of the same term. They differentiate between training data bias, algorithmic focus bias, algorithmic processing bias, transfer context bias and interpretation bias as possible sources. Similar to Friedman and Nissenbaum (1996), they then propose a two-step approach to mitigate problematic algorithmic biases, including the injection of compensatory biases, which in my view could exacerbate the problem even further as long as feedback loops and interactions between algorithmic systems and the world are not fully understood. Mainly concerned with racial and ethnic disparities, Silva and Kenney (2018) extend this work with a detailed categorization of nine types of bias that can enter the 'ecosystem' of algorithmic decision making and the world at five stages: the input data, algorithmic operations, outcomes, users and the feedback itself. Suresh and Guttag (2019) attempt to overcome the formulation of a general training data bias, and identify five well-defined categories of biases that are strongly aligned with a typical machine learning pipeline including data generation. The categories are historical bias, representation bias, measurement bias, aggregation bias and evaluation bias, which are visualized in Figure 2.3. The first three categories refer to biases that arise during data generation



(a) Data Generation

(b) Model Building and Implementation

Figure 2.3: Sources of bias as described by Suresh and Guttag (2019)

and might therefore be summarized as data bias, while the remaining two take place during modelling. Similar to previous work Suresh and Guttag (2019) highlight the dynamics and interdependencies that arise from interactions between algorithmic decisions and society. They further describe how different types of biases motivate different application-specific solutions.

The data set forms the foundation of learning algorithms and greatly predetermines their performance (Eaglin, 2017; Barocas et al., 2017; Danks and London, 2017). Silva and Kenney (2018) distinguish between two types of bias that can be traced back to the data: the training bias which is caused by deviations in the data and the algorithmic focus bias that arises from the selected attributes and differential usage of information. However, there is no clear separation between the two types and in particular the training data bias refers to a variety of subproblems such as the data set size, over- or underrepresentation of certain demographic groups, potential discrepancies between measurements, skewed samples, unreliable labels and the proprietary nature of many data sets, which further hinders the detection of biases. Building on the fact that data is always a product of many factors, Suresh and Guttag (2019) try to counter this collective term by defining three types of bias in relation to data generation. The historical, representational and measurement biases refer to explicit steps during data generation, allowing Suresh and Guttag (2019) to formulate specific solutions. Historical data often captures biases that are deeply integrated into the state of the world and might get amplified by predictive algorithms potentially leading to discriminatory outcomes. Even if the algorithmic system reflects the world perfectly, it can expose the population to harm. This makes historical bias a fundamental, structural issue, that is non-trivial to solve. Suresh and Guttag (2019) provide Google image search results for 'CEO' as an example. For a long time the results displayed mainly white men, who in the real world make up the majority of managers. Although at odds with reality, Google decided to display more diverse results which could be understood as a step towards the world as it should be. Historical bias raises the inevitable question to what extent the world should be represented as it is. Representational bias occurs when parts of the input space, i.e. particular demographic groups, are underrepresented in the data set. When optimizing a classifier for maximal average accuracy, the underrepresented groups will have less influence on the resulting decision rule and a disproportionate number of errors in this group will be less significant. Suresh and Guttag (2019) give reasons for representation bias. First, sampling methods might reach parts of the population differently well because of underlying societal structures. This has also been described as a *signal problem* by Crawford (2013). Second, changes in the population of interest, or differences from the population used for training can cause representational bias. A common example is the lack of diversity in image data sets like ImageNet (Lohr, 2018). The third bias Suresh and Guttag (2019) identify as part of the "data bias" is the measurement bias which arises from the choice of features, the measurement method used as well as varying quality and granularity between groups. The use of proxies for features like *success* or *stamina* that cannot be measured directly has a compounding effect. In psychology and social science, there are criteria for measurement variables to ensure their quality and validity, but some of those concepts haven't arrived in the broader machine learning community yet (see also Barocas and Hardt (2017)).

Apart from challenges with the data itself, further biases can emerge when training an algorithm. Given a dataset, the data is usually split into training and test data, the latter of which is used to evaluate the model after training. Additionally, benchmarks are often used to compare the performance across models and applications. Regarding the learning process and the design choices made to create the model, Silva and Kenney (2018) identify the algorithmic processing bias that might be introduced to varying weights of the data points or the statistical bias of the algorithms. Again, Suresh and Guttag (2019) break down the process of learning more precisely and introduce aggregation bias which arises from flawed assumptions during model definition, and evaluation bias which can be caused by misaligned or misrepresentative benchmarks. They further emphasize that, similar to the effects of other ranking systems, optimization guided by benchmark results encourages the development of models that only perform well on a limited set of data sets that will, when misrepresentative, disproportionately hurt a subset of the population.

Suresh and Guttag (2019) end with the integration of the algorithm into systems that interact with the world and point to a closed-loop nature, where real-world implications of applied systems in turn affect future measurements. Silva and Kenney (2018), on the other hand, propose six further types of bias that only occur in the interaction between algorithm and user, and its effects. Transfer context bias, interpretation bias and non-transparancy of outcome bias arise from the direct interaction between technology and its users. They arise when the algorithm is used in an inappropriate context, when users interpret the outcome on the basis of their own internalized biases, especially if the output is a probability or score or result from a lack of transparency and explainability. The authors assign three more types of bias to the behaviour of the users themselves. To attribute more credibility to algorithmic decisions than to human decisions and to regard the outcomes as factual is captured by the concept of automation bias. Zarsky (2016) argues that the lack of intuition and empathy of algorithms can become a trap if people classified as negative are always classified negatively, so that positive counter-examples cannot emerge. Consumer bias includes the deliberate manipulation of algorithms through consumers, as was observed with the chatbot Tay (Vincent, 2016), as well as the power imbalance, which is particularly reinforced by the large number of online profiles whose personal information can only be accessed by few powerful people. Finally, Silva and Kenney (2018) mention the existence of feedback loop biases that can have self-reinforcing effects. Although the authors recognize a variety of problems that arise from the resulting dynamics between technology and society, some types of bias are rather to be regarded as formulated fears or tensions. For a clear typification, they require, in my view, a somewhat stronger differentiation and a clearer classification according to the cause-and-effect principle.

In summary, it can be said that there are numerous points at which bias can be introduced into the machine learning process, potentially resulting in an arbitrary combination of multiple types of bias. Algorithms cannot be viewed in isolation, but always interact with their target population, the users. This results in a closed ecosystem of technology and society in which numerous interdependencies, dynamics and feedback loops can be observed, that also entail potential harms. This already indicates that fairness does not offer a general solution, but need to be supplemented by domain knowledge based on the context in which the algorithm at hand is applied to. Beyond identifying sources, Suresh and Guttag (2019) suggest ways to combat the different types of bias. To encounter representational bias, the sampling function or method must be modified in order to reach and evenly cover all groups of the target population. Some of the pre-processing techniques discussed above attempt to counteract the problem of underrepresentation by weighting samples differently. However, this has only limited effectiveness and does not solve the underlying problem. Historical and measurement bias can be addressed by adjusting the projections from the sampled population to features and labels. In order to mitigate aggregation bias, the learned function has to be modified in order to better model the data complexity, e.g. using coupled learning methods, or to transform the data in a way that better suits the learned function. An example for the latter option is representation learning as proposed by Zemel et al. (2013). Addressing evaluation bias requires the machine learning community to carefully analyse weaknesses of common benchmarks and agree on alternative measures of success for algorithms. Buolamwini and Gebru (2018) propose to perform subgroup evaluations additionally to the averaged measures, which provides insights into how the error is distributed within the groups. Further ideas include the use of multiple metrics and transparency about the confidence or uncertainty of the algorithms' decisions. Looking closer at the dynamics between applications and users, Silva and Kenney (2018) point to a tension between the interest of companies that want to protect their methods and data, and current efforts to make algorithms more transparent and results traceable. They further argue that, in order to diagnose and ensure fairness, it is often necessary to collect information about protected attributes for monitoring purposes. However, the use of protected attributes in the algorithm is restricted by law, thus making companies unable to act. It is therefore necessary to adjust regulations in order to allow for fairness-enhancing interventions and simultaneously ensure fairness during the process.

2.4 Limitations and Challenges

Besides the limitations and drawbacks described above that apply for specific criteria, there are some limitations and challenges that have not been taken into account so far. Especially, some aspect of fairness cannot be achieved solely by looking at statistical measures of data attributes, but require serious interventions. The previous section discussed in detail how bias can be propagated through data generation and machine learning pipelines. Within this discussion, numerous challenges have already been addressed. They will briefly be outlined in the following overview. Additionally, the subsequent section attempt to outline some important challenges of ensuring fairness that have hardly been considered so far.

Quality of Data. Ensuring the quality of data and understanding in the context it originates from are essential for successful machine learning. Barocas et al. (2018) point out that evidence-based decision making is only as reliable as the evidence it was based on and thereby stress the concept of 'garbage in, garbage out'. Barocas and Hardt (2017) list several problems to be aware of that can arise in the learning process, some of which have already been mentioned in the previous section. Data sets can contain skewed samples that display a reality that is different from the actual ground truth, or unreliable labels. It is important to emphasize that labels themselves can be the product of human decisions, potentially incorporating biases. In this context, it is further important to note that labels can be of a different nature. If, for example, they describe whether a person has repaid a loan, it is more likely that the label reflects the real truth, as the payback of the loan is measurable. With hiring, on the other hand, the labels reflect whether a person is considered suitable for the job and has therefore been hired. This decision is not absolutely measurable, but includes human considerations and decisions that can be strongly influenced by the setting in which the decision was made. I will refer to labels that come from tasks like loan granting as *objective labels* while labels that involve explicit human judgements which cannot be reconstructed afterwards will be referred to as *subjective labels*. Regarding the actual measurements that are contained in a data set, Barocas and Hardt (2017) identify limited features and the use of proxies as additional problems. These have been discussed in detail in the previous section.

In a nutshell, the quality of data is crucial for the performance of the algorithm. The data should therefore be diligently selected and scrutinized to ensure that it is sufficiently large, diverse and well-annotated. Collecting more data will be for the benefit of the whole population, if there are no distinguishing attributes between groups. Assuming that demographic groups are heterogeneous with regard to attributes relevant for the task at hand, more data might just increase the gap, when it amplifies the differences in available data between groups. Further attention has to be drawn to the validity and reliability of features and annotations. A corresponding concept to foster transparency and facilitate data set selection has been proposed by Gebru et al. (2018).

Feedback Loops. When discussing approaches to achieve fairness in machine learning systems, we focussed on the data and how it is used to create a model by some learning technique. However, as revealed in the previous section, these systems should rather be understood as being embedded in a large socio-technological system (Barocas et al., 2018). First, a data set is generated by projecting and measuring the state of the world on a small number of attributes. The resulting data set is then used to build a model by generalizing from patterns found in the training data. The model can then be used to predict outcomes for new, unseen data. Individuals receiving a prediction might respond to the outcome, e.g. by clicking on the displayed ad. Many systems can use this response as a feedback signal to refine and adapt their model. Collective responses of individuals to the predictions of decision making systems will alter the state of the world, and thereby change the patterns that these systems try to model (Barocas et al., 2018). As Moritz Hardt underlined, consequential decision making is always enfolded in feedback loops between models and the state of the world (Hardt, 2019). The feedback loops pose a big challenge, as their effects are tricky to interpret and can hardly be measured. As a step in this direction, recent work investigated the sources at which bias can be introduced into data generation and the machine learning pipeline, some of which has been discussed in the previous section. They have mostly remained with the acknowledgement that technology and society influence and shape each other. Further work should investigate the impact of these interdependencies.

Inherent Limitations of Observational Criteria. The definitions of group fairness discussed in section 2.1.3 suffer from inherent limitations due to the fact that they rely on statistical measures that are derived from properties of the joint distribution of data, labels, the sensitive attributes and the classifier. This property is what constitutes observational criteria according to Hardt (2017) and Barocas et al. (2018): solely passive observations of the world without any interventions. They come with the implicit assumption that notions of fairness can be expressed in mathematical terms and used to create fair systems (Suresh and Guttag, 2019). However, observational criteria are subject to serious inherent limitations that need to be considered. Barocas and Hardt (2017) design two scenarios with identical joint distribution, which cannot be distinguished by observational criteria but lead to different interpretations. In both cases, the aim is to predict whether a person is a software engineer in order to finally display a job advertisement – a scenario potentially having a significant impact on the person's life. This example stresses the fact that observational criteria are not sufficient to address substantial social questions. Barocas and Hardt (2017) further claim that the problem will not change when more data is available because the collected data will always come from the same distribution. Even more, the dataset forces us to take a particular perspective on the world that may differ from the real world, leaving certain structures unrecognised and alternatives unexplored. Through this argumentation Barocas and Hardt (2017) motivate causal reasoning. However, this approach also comes with the limitation that changing individual attributes can lead to false conclusions, since a person's life would have developed differently and consequently the other attributes might also differ if the person had been white, black, female, male, etc.

Implicit Bias and Representational Harm. What goes beyond the scope of this review is the discussion of different kinds of harm that can be done to an individual. Crawford (2017) proposes a distinction between allocative harms and representational harms. The former arise when a system withholds opportunities to individuals. It concerns questions of the allocation of resources, which is frequently the case in classification tasks. This type of damage is associated with short-term effects and economic consequences for individuals and has been the focus of research so far, which is why the review has focused on these issues (Crawford, 2017; Narayanan, 2018). Representational harm occurs when algorithmic systems reinforce the subordination of some groups along the line of identity. This type of harm is associated with the perpetuation and amplification of stereotypes and comes with social harms, the effects of which are more diffuse and long-lasting. Common examples of this type include the mistagging of a picture of two Afro-Americans as gorillas by Google Photos (Grush, 2015), the lightening of faces as a feature of a 'hot' filter provided by FaceApp (Cresci, 2017), the use of stereotyped pronouns when translating from English to a language with gender-neutral pronouns and back (Narayanan, 2018) and the failure of automated facial analysis systems to recognize black people (Boulamwini, 2017; Lohr, 2018). Those examples indicate that representational harms are rather difficult to formalize as they concern harmful representation of human identity, which can hardly be quantified. Recent work by Sweeney (2013),Kay et al. (2015),Bolukbasi et al. (2016) and Zhao et al. (2017) analyses algorithmic systems in the light of representational harms and investigate how to mitigate these effects in the context of natural language processing, image recognition, search engines and advertising. Crawford (2017) describe five different types of representational harm among which are stereotyping, recognition and denigration that have been reported for several applied algorithms. However, further research is needed to not only achieve fairness in terms of resource allocation, but also in terms of how algorithmic systems represent the world. This requires an understanding of human culture and history and calls for interdisciplinary discourse and collaboration.

2.5 Conclusions

The existing approaches and discussions illustrate that fairness in machine learning is a complex problem in which, in addition to developers of algorithms and curators of data sets, policy makers, humanists, users and representatives of affected minorities, i.e. various stakeholders, have to be involved (Zarsky, 2016; Narayanan, 2018; Barocas and Hardt, 2017; Barocas et al., 2018). The community is still at a relatively early stage, so that many fundamental questions still need to be solved and some structures still need to be developed. However, it also seems that we are on a good track to recognize and address the potential and dangers arising from automated decision making.

Referring to O'Neil (2016) and Narayanan (2018), the goal of algorithmic systems should always be to further and obey human values. Complex concepts such as fairness cannot be reduced to a single formula and cannot be regarded as a one-time or one-size problem. Instead, they require careful consideration of the purpose, scope and context of such systems, application-specific implementation and continuous monitoring. In order to address social problems with machine learning, the dynamics emerging from the interaction between technology and society must be included and further investigated (Hardt, 2019).

In order to bring transparency into the decisions for certain models or data sets, Gebru et al. (2018) and Mitchell et al. (2019) propose to document the strengths and weaknesses, the context and the potential, including possible dangers of data sets and pre-trained models. Confirming the need for more detailed analysis of data and models and hence Crawford (2013) proposes to complement data sources and algorithmic systems with rigorous qualitative research. Going one step further, Crawford (2017) further encourages the self-reflection of the machine learning community and calls for scrutiny of one's own methods and their influence, for scrutiny of goals, and for exploration of the internalized bias that algorithm developers bring to the process and interpretation. Hardt (2019) concludes by stating that fundamental normative questions need to be solved in order to determine in which contexts algorithmic solutions should actually be considered and in which humans, through their intuition, empathy and ability to assess problems and formulate compromises, are superior to machine decisions.

3 Theoretical Background

This chapter deals with fundamental concepts on which the practical part of this work is based. The first section outlines basic concepts of Bayesian inference. The following sections describe the machine learning procedures that were relevant for the implementation.

3.1 Bayesian Inference

Bayesian inference is a methodology to draw conclusions from data based on a probabilistic model. It builds on Bayes' Theorem to update the beliefs about hypotheses given new data. The main characteristic of Bayesian inference is that it is explicit about model uncertainties. This approach of modelling uncertainty is helpful to reason about beliefs and cope with limited data resources. Unless otherwise stated, the introduction is based on this introductions to probability and inference by MacKay (2003) and Murphy (2012).

Before outlining the basic ideas behind Bayesian inference, I will briefly recap basic concepts of probability theory that are essential for inference. The two most fundamental concepts are the joint probability of two random variables and the definition of a marginal likelihood. The joint probability of two random variables, also known as product rule, is defined as

$$P(A,B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A). \tag{3.1}$$

where $P(A|B)$ is called conditional probability. A further concept is called conditional probability. It describes the probability of an event A given another event B has occurred. The marginal probability of A, also known as sum rule or rule of total probability, can be derived from the joint probability by summing over all possible values of B (marginalizing out B):

$$P(A) = \sum_{b \in B} P(A, B = b) = \sum_{b \in B} P(A|B = b) \cdot P(B = b). \tag{3.2}$$

Apart from these essential concepts, the definitions of independence and conditional independence should be mentioned, because both have been used extensively in the review of fairness literature in chapter 2. Two random variables A, B are independent if their joint probability is equivalent to the product of the marginal probabilities:

$$A \perp B \Leftrightarrow P(A,B) = P(A)P(B). \tag{3.3}$$

The conditional independence extends this notion by introducing a third random variable C, conditioned on which independence holds:

$$A \perp B | C \Leftrightarrow P(A,B|C) = P(A|C)P(B|C). \tag{3.4}$$

Bayes' Theorem. Closely related to conditional probability, Bayes' Theorem is obtained by combining the product and sum rule with the definition of conditional probabilities:

$$P(\theta|D,\mathcal{H}) = \frac{P(D|\theta,\mathcal{H}) \cdot P(\theta|\mathcal{H})}{P(D|\mathcal{H})}, \tag{3.5}$$

where $P(\theta|D,\mathcal{H})$ and $P(D|\theta,\mathcal{H})$ can be obtained from equation 3.1 and $P(D|\mathcal{H})$ follows from equation 3.2 when marginalizing over θ. Adopting a Bayesian viewpoint, probabilities can be used as measures of uncertainty or to describe degrees of belief in propositions. Following Cox axioms, probabilities are further suited to describe assumptions and draw conclusions, namely doing inference, given those assumptions (MacKay, 2003). Bayes' Theorem can then be understood as a way to incorporate prior knowledge into computing the probability of an event given the occurrence of another event. In the context of Bayesian inference, the variables are commonly described as parameters of the model θ and data D, which are described in the light of some hypothesis space $\mathcal{H}$.

Bayesian inference follows a fundamental procedure: Data is generated by a latent function and when observed, it might be subject to noise. Initially, assumptions or beliefs about the nature of the latent function are expressed as a prior distribution, $P(\theta|\mathcal{H})$, that is hypothesis $\mathcal{H}$ being parametrized by θ. The prior incorporates prior knowledge by assigning probabilities to all possible parameter values θ. It is important to remark at this point that the choice of prior affects the resulting posterior, i.e. while an informed prior might improve the quality of the posterior, the prior can also be misleading when the underlying assumptions are wrong. The prior is often subject to controversial debates, because it incorporates "subjective" knowledge of the designer. However, it is important to note, that inference is only possible when making assumptions, which is also known as inductive bias (von Luxburg, 2018). Those assumptions are explicitly encoded through the prior in the Bayesian framework. Besides that, inference is also data-driven. This aspect is addressed with the formulation of a likelihood $P(D|\theta,\mathcal{H})$, which connects observations of data to the model.

The posterior probability of the model parameters θ is normalized by the marginal likelihood $P(D|\mathcal{H})$, which is also known as evidence. The normalization constant functions as scaling factor and will not affect the shape of the distribution. However, it is essential to make statements about probabilities. The fact that the marginal is usually hard to compute or not computable at all leads to approximative inference, which aims to estimate properties of the posterior distribution. The posterior is often described as internal belief state of the world which gets updated with new data. In order to make prediction, the posterior has to be applied to new, unseen data. This leads to a posterior predictive distribution, which allows to predict the distribution over outcomes for a new, unseen data point. Given a new data point, the weighted average of possible outcome values weighted by their posterior probability is computed.

In the following, two fundamental methods to perform regression and classification are explained, which utilize the approach of Bayesian inference: Bayesian linear regression and Bayesian logistic regression. Both approaches have the advantage that weights are assigned to the attributes of the data, so that the influence of the individual attributes can be comprehended. This allows a detailed analysis of the decisions of the resulting predictors.

3.2 Bayesian Linear Regression

Linear regression is one of the most basic and most widely used techniques in machine learning. It basically assumes that the output is generated as a linear combination of inputs and attempts to fit a curve to the observed data. As opposed to using point estimates, Bayesian linear regression provides a probabilistic approach by finding a distribution over parameters that gets updated with every new observed data point.

3.2.1 Parametric Linear Regression

Assume, we have data $\mathcal{D} = \{X,Y\}, X = \{x_1,...,x_N\} \subset \mathcal{X} = \mathbb{R}^D, Y = \{y_1,...,y_N\} \subset \mathcal{Y} = \mathbb{R}$ representing pairs of inputs and outputs. The goal is to find the linear function

$$f(x) = \sum_{d=1}^{D} w_d x^{(d)} = x^T w \tag{3.6}$$

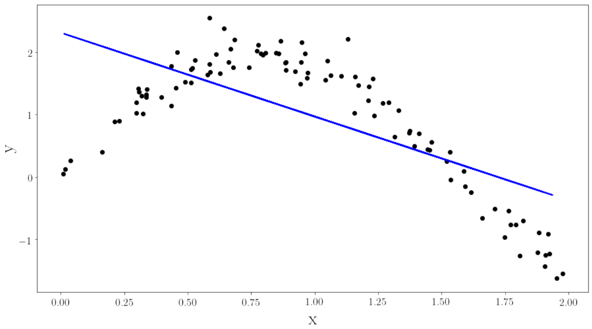


Figure 3.1: Simple example for linear regression, where the underlying data distribution is approximated with a linear function.

with weights $w \in \mathbb{R}^D$ that minimize the loss (von Luxburg, 2018). We further assume that the observations are subject to noise so that the observed target values have the form

$$y = f(x) + \varepsilon \tag{3.7}$$

where $\varepsilon \in \mathcal{N}(0, \sigma^2)$ is additive, independent and identically distributed Gaussian noise, also described as residual error between the linear predictions and the true response. This model can then be rewritten as

$$p(Y|X,w) = \prod_{n=1}^{N} p(y_n|x_n,w) \tag{3.8}$$

$$p(Y|X,w) = \mathcal{N}(Y;\, \mu(x) = X^T w, \sigma^2 I) \tag{3.9}$$

with w being the weights or parameters of the model, μ the mean and σ^2 the variance, which is assumed to be fixed (Murphy, 2012).

Since linear functions are severely restricted in their expressiveness (see Figure 3.1), the model of linear regression can be extended so that the learned function can behave non-linearly in x (von Luxburg, 2018). For this purpose use is made of potentially non-linear mappings of the data points, so called feature functions (or basis functions) $\phi(x) : \mathcal{X}^D \rightarrow \mathcal{X}^N$, short ϕ_x. Aggregating $\phi(x)$ for all $x \in X$ leads to $\Phi(X)$ or short Φ. It should be noted here that the function $f(x)$ remains linear in the parameters w. Rasmussen and Williams (2006) describe the use of feature functions as a projection into a high dimensional space in which the linear model is then applied. The extended parametric model is then takes the following form:

$$f(x) = w^T \phi(x) \tag{3.10}$$

$$p(Y|\Phi,w) = \mathcal{N}(Y;\, \Phi^T w, \sigma^2 I). \tag{3.11}$$

Note, that the weights w are now used in function space, i.e. $w \in \mathbb{R}^N$. The feature functions have to be fixed before seeing the data. In order to select suitable feature functions it is therefore helpful to have prior knowledge about the data (von Luxburg, 2018). A simple example of a basis function is $\phi(x) = \left[1, x, x^2, x^3, x^4, \ldots\right]$, which is used for polynomial regression.

A common strategy to estimate the parameters of a statistical model is to compute the maximum likelihood estimate, also known as least squares because deviations from the fitted line are penalized quadratically. However, when we assume the observations to be subject to noise and outliers to be likely to appear in the data, the maximum likelihood estimate can result in a poor fit (see Murphy, 2012, for a detailed explanantion). To prevent the parameters from becoming too large, which would result in high variance, regularization terms are introduced that encourage the parameters of the model to stay small. Such techniques include ridge regression or Lasso (von Luxburg, 2018). In a Bayesian framework, adding a prior adopts the function of regularization. Using a Gaussian prior for linear regression is equivalent to a quadratic regularization term as used in ridge regression. The prior expresses the beliefs about the weights before observing any data:

$$p(w) = \mathcal{N}(w; \mu, \Sigma). \tag{3.12}$$

Given the prior $p(w)$ and the likelihood $p(Y|w, \Phi_x)$, we can compute the posterior distribution over weights

$$p(w|Y,\Phi) = \frac{p(Y|w,\Phi) \cdot p(w|\Phi)}{p(Y|\Phi)} \tag{3.13}$$

$$p(w|Y,\Phi) = \mathcal{N}\big(w; \mu + \Sigma\Phi(\Phi^T\Sigma\Phi)^{-1}(Y - \Phi^T\mu),\ \Sigma - \Sigma\Phi(\Phi^T\Sigma\Phi + \sigma^2 I)^{-1}\Phi^T\Sigma\big) \tag{3.14}$$

$$= \mathcal{N}\big(w; (\Sigma^{-1} + \sigma^{-2}\Phi^T\Phi)^{-1}(\Sigma^{-1}\mu + \sigma^{-2}\Phi Y),\ (\Sigma^{-1} + \sigma^{-2}\Phi^T\Phi)^{-1}\big) \tag{3.15}$$

as described in the lecture slides by Hennig (2019). Finding the weight that maximizes the posterior, results in the maximum a posteriori (MAP) estimate, which equals the mode of the posterior distribution, i.e. chooses the parameters which maximize the posterior probability in the light of the observed evidence. The MAP can be used as a point estimate of w.

Based on the posterior distribution over weights, predictions for new, unseen data points can be made. Therefore we use the predictive distribution, which weights all possible parameter values by their posterior probability and allows to not only reason about the most likely parameter value, but to also take uncertainties of the model into account. The distribution over function values for a new data point x_* (with corresponding feature function ϕ_*) is given by

$$p(f_*|x_*, \Phi, Y) = \int \underbrace{p(f_*|w, x_*)}_{\text{likelihood}} \ \underbrace{p(w|Y,\Phi)}_{\text{posterior probability}} \ dw \tag{3.16}$$

$$\begin{aligned} p(f_*|x_*, \Phi, Y) = \mathcal{N}\big(&\sigma^{-2}\phi(x_*)^T(\Sigma^{-1} + \sigma^{-2}\Phi^T\Phi)^{-1}\Phi Y, \\ &\phi(x_*)^T(\Sigma^{-1} + \sigma^{-2}\Phi^T\Phi)^{-1}\phi(x_*)\big). \end{aligned} \tag{3.17}$$

The posterior distribution can be understood as an update of the prior beliefs. The posterior predictive distribution builds on this, as it makes predictions using the updated beliefs.

3.2.2 Gaussian Process Regression

The previous paragraph was concerned with inference in the weight space. In order to allow for more flexibility of the model, we used explicit feature functions that map the inputs to a high-dimensional feature space. Going one step further, we can replace all inner products of feature functions by a kernel of the form $k(x, x') = \phi(x)^T\Sigma\phi(x')$ and work with the algorithm in feature space. Inference can then be considered in function space, where we use a Gaussian process to describe probability distributions over functions (Rasmussen and Williams, 2006).

Bailey (2016a) gives a comprehensive explanation on the motivation behind Gaussian processes. She provides an example similar to Figure 3.1, for which linear regression does not appear capable of modelling the latent function reasonably well. Here, instead of explicitly deciding on the number of parameters to use for modelling, as it is done with parametric regression, we would rather like to consider all suited functions regardless of the number of parameters they use to model the data. Since considering all possible functions would be too exhaustive, the formulation of a prior helps to prioritize specific functions. Gaussian processes follow this intuition.

In a more formal way, a Gaussian process $p(f) = \mathcal{GP}(f; m, k)$ is a probability distribution over function values of functions $f : \mathcal{X} \to \mathbb{R}$, so that every finite restriction to function values $f_X : [f_{x_1}, ..., f_{x_n}]$

is a Gaussian distribution (Hennig, 2013). The model is specified solely on the basis of a mean function $m(x) = \mathbb{E}(f(x))$ and a covariance function or kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The covariance function encodes properties of the functions like smoothness. Because it specification implies a distribution over functions, the learning process can be described as finding suitable properties for the covariance function (Rasmussen and Williams, 2006), also called hyperparameter tuning. Gaussian processes are non-parametric models, because they are capable of considering functions with infinitely many parameters. The overall goal is to learn the underlying distribution of the data by finding a distribution over possible functions $f(x)$ that is consistent with the observed data points. Given a training data set, there are potentially infinitely many functions that fit the data. Similar to the procedure described in the introduction to Bayesian inference, Gaussian process regression is comprised of three steps. First, prior believes about the function space are expressed as a prior. Second, the space of possible functions gets then restricted with every new data point we observe, governed to the likelihood over outcomes. Third, using the prior and the likelihood, the Gaussian process results in a posterior over functions that appear plausible to explain the data. Gaussian processes thereby offer an elegant solution to fitting a function to the data by incorporating prior knowledge and assigning a probability to each of the resulting functions (Rasmussen and Williams, 2006).

3.2.3 Links to Frequentist View

As opposed to the Bayesian approach of understanding probabilities as uncertainties, there is also the so-called frequentist interpretation of probabilities. According to this view probabilities can be understood as frequencies of an event (Murphy, 2012). Since both approaches coincide in many respects and often arrive at the same results from different perspectives, the links of the Bayesian view to the frequentist view will be briefly mentioned here.

In the case of linear regression, the noise model assumed for observed target values corresponds to the choice of a loss function (empirical risk) in the frequentist view, while the assumption of a prior distribution in the Bayesian approach corresponds to a particular choice of regularizer in the frequentist approach (von Luxburg, 2018). For example, it can be shown that assuming a Gaussian prior leads to the same result as performing ridge regression, which is characterized by L_2-loss and the squared norm of the weight vector as regularizer (see Bailey (2016b) or von Luxburg (2018) for further details). Equivalently Gaussian process regression is closely related to kernel ridge regression (Hennig, 2019). However, the Bayesian approach allows to incorporate prior knowledge and to further quantify the uncertainty of the model.

3.3 Bayesian Logistic Regression

Contrary to what the name suggests, logistic regression is one of the most widely used methods for classification. Here, the setting is the following: Given a set of training points $X \in \mathcal{X} = \mathbb{R}^D$, assign them to one of C classes, i.e. find linear functions to separate the classes. In the binary case, which we assume here for simplicity, the class labels are usually defined as $Y = \{-1, 1\}$ and the goal is then to find a

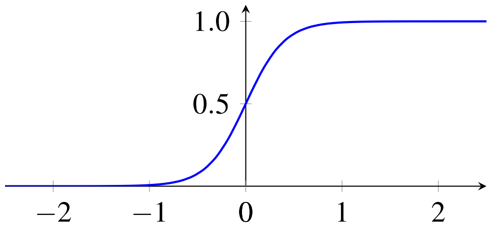


Figure 3.2: The logistic function (compressed along the x-axis)

hyperplane separating the two classes. Logistic regression builds on linear regression in that it turns the outputs of the regression model into class probabilities by mapping them to the interval $[0,1]$ using the logistic function

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{3.18}$$

which is also visualized in Figure 3.2. This step is also known as ”squashing the outputs” and allows to interpret the outputs as probabilities for class membership (Rasmussen and Williams, 2006). Adopting a Bayesian viewpoint for classification leads to predictions in form of class probabilities, which inherently involves a notion of uncertainty about the assignments to one of the classes.

3.3.1 Parametric Logistic Regression

As explained above, parametric logistic regression builds on linear regression and can be phrased in weight or function space. In the parametric case, the prior over weights remains the same:

$$p(w) = \mathcal{N}(w; \mu, \Sigma). \tag{3.19}$$

As with linear regression, explicit feature functions can be used to provide more flexibility. We try to model the latent function which is assumed to be of the form

$$f(x) = w^T \phi(x) \tag{3.20}$$

with $\phi(x) : \mathcal{X}^D \to \mathcal{X}^N$ being the explicit feature functions and $w \in \mathcal{X}^D$ the weights. For classification tasks it is inadequate to assume a Gaussian likelihood, because the target values are discrete class labels. Instead a smoothed step function is used for the likelihood. Therefore, the outputs of the corresponding linear model are squashed through the logistic function to obtain values between 0 and 1. This results in the non-Gaussian likelihood

$$p(y = 1|X, w) = \prod_{n=1}^{N} \sigma(\phi(x_n)^T w). \tag{3.21}$$

Note, that the sum over possible outcomes must add to 1, i.e. $p(y = -1|X, w) = 1 - p(y = 1|X, w)$. The likelihood is thus symmetric and can be denoted more compactly as

$$p(Y|X, w) = \prod_{n=1}^{N} \sigma(y_n \phi(x_n)^T w). \tag{3.22}$$

Using the prior and the likelihood we obtain the unnormalized log posterior

$$\log p(w|\Phi, Y) = -\frac{1}{2} w^T \Sigma^{-1} w + \sum_{n=1}^{N} \log \sigma(y_n \phi(x_n)^T w) + \text{const.} \tag{3.23}$$

As we have seen before, it is not appropriate to assume a Gaussian likelihood because the class labels are discrete. Instead, a non-Gaussian likelihood has been used, which makes the integral analytically not tractable. As a consequence, the posterior is non-Gaussian either and cannot be expressed in a closed analytic form. However, the posterior distribution is still concave and thus also unimodal, i.e. it has a unique maximum. A standard algorithm to find the maximum is Newton’s method, which, starting from an initial guess, takes the curvature of the space into account to approach the maximum of the posterior. For predicting class membership, the maximum of the posterior distribution is sufficient, as it can be used as maximum a posteriori (MAP) estimate of weights w. Beyond the maximum, the variance of the posterior is usually of interest. For this purpose, techniques including expectation propagation, variational inference or the Laplace approximation can be used to approximate the posterior distribution

(Murphy, 2012). The most straight-forward method is probably the Laplace approximation. It approximates the non-Gaussian posterior by a Gaussian one by placing a second order Taylor expansion of the posterior in log space around the maximum of the posterior (Rasmussen and Williams, 2006):

$$p(w|\Phi,Y) \approx q(w|\Phi,Y) = \mathcal{N}\left(w; \arg\max_{w} p(w|\Phi,Y), \left(-\nabla\nabla^T \log p(w|\Phi,Y)\right)^{-1}\right). \tag{3.24}$$

Since the Laplace approximation is a local approximation, it should be noted that it can be arbitrarily wrong. The resulting approximated posterior can be understood as an update of believes after observing the training data. It will then contribute to the predictive distribution by weighting the likelihood of possible outcomes

$$p(y_*|x_*,\Phi,Y) = \int p(y_*|x_*,w)p(w|\Phi,Y)dw = \int \sigma(y_* x_*^T w)p(w|\Phi,Y)dw \tag{3.25}$$

(Rasmussen and Williams, 2006). In order to predict a class for a new data point, a decision boundary in output space is used. Assuming the loss to be symmetric, the data point is assigned to class 1, if resulting probability is greater than 0.5 and assigned to the negative class else with 0.5 defining the decision boundary.

$$y_* = \begin{cases} 1, & \text{if } p(y_* = 1|x_*,\Phi,Y) > 0.5 \\ -1, & \text{if } p(y_* = 1|x_*,\Phi,Y) \leq 0.5 \end{cases} \tag{3.26}$$

In many settings, however, it makes sense to assume asymmetric loss, since one type of misclassification, false positives or false negatives, might entail greater harm than the other. Both the decision maker and the classified persons may be affected by this harm. Likewise, the correct classification can be of varying importance. Varying the impact of particular types of (mis)classification will shift the decision boundary to the left or the right. Accordingly, the threshold to receive a positive classification will be higher or lower. In order to formalize the value of the decision boundary in a one-dimensional case, we will look at the expected loss. A person will be classified as belonging to the class which has smaller expected loss $\mathbb{E}(L|y_*)$. The decision boundary is then derived by

$$\begin{aligned} \mathbb{E}(L|y_* = 1) &< \mathbb{E}(L|y_* = -1) \qquad (3.27)\\ a \cdot p(y_* = -1|x_*,\Phi,Y) &< b \cdot p(y_* = 1|x_*,\Phi,Y) \qquad (3.28)\\ a \cdot (1 - p(y_* = 1|x_*,\Phi,Y)) &< b \cdot p(y_* = 1|x_*,\Phi,Y) \qquad (3.29)\\ p(y_* = 1|x_*,\Phi,Y) &> \frac{a}{a+b}. \qquad (3.30) \end{aligned}$$

In order to assign a person to the positive class ($y = 1$), the predicted probability that this person truly belongs to the positive class has to be higher than $\frac{a}{a+b}$.

3.3.2 Gaussian Process Classification

In accordance with the distinction between parametric linear regression and Gaussian process regression, Gaussian process classification can be differentiated from parametric logistic regression as inference is considered in function space. Moreover, the use of Gaussian processes for classification allows to predict without presupposing an explicit parametrization of the latent function that generated the data. Instead, a Gaussian process prior over the latent function space is assumed. In order to relate outputs of the latent function to class probabilities, it is then squashed through the logistic function to obtain values within the interval $[0,1]$, which can be interpreted as probabilities for class membership in a binary classification task. The use of the logistic function results in $\pi(x) = p(y = 1|x) = \sigma(f(x))$ (Rasmussen and Williams, 2006). Given a data set, the likelihood over functions can then be formulated.

Unfortunately, performing inference and adapting Gaussian process models for classification tasks is not as straight-forward as in regression tasks (MacKay, 1998). For the latter, the computation of the posterior and the predictions involve solely Gaussian distributions making the computation of the integral analytically tractable. In the case of classification, however, the likelihood and therefore also the posterior is non-Gaussian, which has already been highlighted in the previous section. Therefore the evidence is analytically intractable and approximations of the posterior distribution are needed. Common approximation techniques to overcome this problem include the Laplace approximation, expectation propagation and Markov chain Monte Carlo approximations.

3.3.3 Links to Frequentist View

Due to the fact that logistic regression builds on linear regression by using its outputs to compute class probabilities, it follows logically that the Bayesian view and the frequentist view lead to equivalent results in the case of logistic regression. Again, the formulation of a prior corresponds to the choice of a particular regularization, while the loss function can be mapped to the link between model and data. Regarding the predictions, the Bayesian approach allows to take model uncertainty into account, which might be of special importance when reasoning about assignments to a specific outcome that might seriously affect a person's life.

4 Methods

The following chapter documents the methods used in this work. In addition to the data set used and a brief description of the machine learning algorithm applied, this also includes different approaches to test a classifier against specific fairness criteria and for achieving fairness.

4.1 German Credit Data Set

As discussed in chapter 2, the analysis of data sets and applications in relation to fairness focuses primarily on settings like hiring, loan granting, university admission, or risk assessment since these decisions involve the classification of people. In order to illustrate the previously discussed concepts of fairness in an exemplary way, a data set for classification tasks was selected that has been used for the experiments in many of the papers outlined earlier: the German credit data set.

The German credit data set consists of 1000 samples of people who received for a credit loan as well as their corresponding label, encoding whether they paid back the loan. It is an open-access data set from 1994 that is provided within the UCI Machine Learning Repository. In the original data set each instance is described by 20 attributes, 13 of which are categorical and 7 numerical. The repository provides an additional data set which contains only numerical attributes and indicator variables that encode the membership to such a category as 0 (no) or 1 (yes). Since there is no documentation given for the meaning of those attributes, we did not consider the additional data set. The German credit data set is suitable for binary classification tasks as it provides a binary output, i.e. whether a loan applicant paid the loan back or not. The task is then to decide whether a new person should receive a loan.

The data set was chosen because it is commonly used in the literature dealing with fairness, as it illustrates a typical scenario where the introduction of a notion of fairness is rather intuitive. Furthermore the task of loan granting comes with the advantage that the labels in the train and test data set can be assumed as absolute truth as described earlier in this review. This is in contrast to labels which, for example, are available in a hiring task. The latter involves additional human decisions and potential biases so that the labels themselves can be questioned. When granting a loan, on the other hand, it is measurable whether someone has paid the loan back or not. I will refer to the labels that come from tasks like loan granting as *objective labels* as opposed to *subjective labels*. Note that *objective labels* are not the same as objective data, because the selection of the attribute in the data set and also the decision of who gets a loan involves human decisions.

In the original data set categorical attributes are encoded as a string of the form *'A + ⟨# attribute⟩ + ⟨category⟩'*, e.g *A151* refers to the first possible value of attribute 15 with the meaning 'rented housing'. Before training a classifier we pre-processed the data. Pre-processing comprised of changing the original ordering of the categories in a meaningful way and adding additional attributes that serve as indicators of whether a given instance belongs to a category or not. The resulting categorical attributes were realized with integer values centred around zero. The numerical attributes were scaled such that their values were between 0 and 10. The rearrangement and reordering served the purpose of adjusting the scaling of all attributes, because varying scales effect the performance of a classifier significantly. The final pre-processed data set describes each instance by 42 attributes which are listed in detail in Table 4.1 and further visualized in Figure 4.1.

The original data set suffers from several shortcomings. The data set was first published in 1994, but is estimated to reflect the conditions of earlier times. As a result, some of the criteria are no longer appropriate or applicable for today's standards. In addition, the data set is poorly documented, which made it difficult to fully understand the meaning of some attributes. It further remained unclear how and in what context the data were collected. Other obstacles are the imprecise definition of variables and a strong imbalance with regard to some particular attributes. For example, the original attribute 9 represents the combination of personal status and sex, using the five categories "*male:divorced/separated*", "*female:divorced/separated/married*", "*male:single*", "*male:married/widowed*" and "*female:single*". During data preparation we tried to separate personal status and sex, also because the latter is a frequently used sensitive attribute. However, the preparation of attribute 9 turned out to be difficult in that the information on personal status in particular cannot be clearly separated. For men there are three categories divorced/separated, married/widowed and single, while women fall into two categories, namely divorced/separated/married or single. Due to partial overlap of the categories pre-processing has introduced some inaccuracies at this point which could have been prevented by profound documentation and better attribute selection. The unbalancedness of the data set can most clearly be seen in the original attribute 20, the categorical division into foreign workers and others. The data set contains 96.3% foreign workers, which raises the question of whether the labels might have been swapped or whether the selected population might not have been representative of German credit data. Further unbalanced attributes include the strongly right-skewed distribution of age, 84.6% samples of real estate ownership and the lack of examples of requested loans to finance holidays.

attribute	meaning	values	remarks
0	existing account	-1 (no), 1 (yes)	
1	account status	-1 ... $<$ 0 DM 1 0 $\leq$ ... $<$ 200 DM 2 ... $\geq$ 200 DM 0 no checking account	
2	duration	*numeric (in 10 months)*	
3	credit history	-2 critical account/ other credits existing, -1 delay in paying off in the past, 0 existing credits paid back duly till now, 1 all credits at this bank paid back duly, 2 no credits taken/all credits paid back duly	
4	purpose: new car	-1 (no), 1 (yes)	
5	purpose: used car	-1 (no), 1 (yes)	
6	purpose: furniture/equipment	-1 (no), 1 (yes)	
7	purpose: radio/television	-1 (no), 1 (yes)	
8	purpose: domestic appliances	-1 (no), 1 (yes)	
9	purpose: repairs	-1 (no), 1 (yes)	
10	purpose: education	-1 (no), 1 (yes)	
-	purpose: vacation	-1 (no), 1 (yes)	not existent, ignored
11	purpose: retraining	-1 (no), 1 (yes)	
12	purpose: business	-1 (no), 1 (yes)	
13	purpose: others	-1 (no), 1 (yes)	
14	credit amount	*numeric (in 1000 DM)*	
15	savings account/bonds	-1 (no), 1 (yes)	

attribute	meaning	values	remarks
16	present employment since	-2 unemployed, -1 ... < 1 year, 0 1 ≤ ... < 4 years, 1 4 ≤ ... < 7 years, 2 ... ≥ 7 years	
17	employed	-1 (no), 1 (yes)	
18	instalment rate in percentage of disposable income	*numeric*	all values in [1,4]
19	sex	-1 (male), 1 (female)	
20	divorced/separated	-1 (no), 1 (yes)	inaccurate for females due to original attribute
21	single	-1 (no), 1 (yes)	
22	married/widowed	-1 (no), 1 (yes)	inaccurate for females due to original attribute
23	other debtors/guarantors: co-applicant	-1 (no), 1 (yes)	
24	other debtors/guarantors: guarantor	-1 (no), 1 (yes)	
25	present residence	*numeric (in years?)*	no units given, max. residence 4 years
26	property: real estate	-1 (no), 1 (yes)	
27	property: life insurance or building society savings agreement	-1 (no), 1 (yes)	inaccurate due to original attribute
28	property: car or other	-1 (no), 1 (yes)	inaccurate due to original attribute
29	age	*numeric (in 10 years)*	
30	other instalment plans: bank	-1 (no), 1 (yes)	
31	other instalment plans: stores	-1 (no), 1 (yes)	
32	housing: rent	-1 (no), 1 (yes)	living for free: 33 and 34 both -1
33	housing: own	-1 (no), 1 (yes)	living for free: 33 and 34 both -1
34	number of existing credits at this bank	*numeric*	
35	job: unemployed/unskilled (non-resident)	-1 (no), 1 (yes)	
36	job: unskilled (resident)	-1 (no), 1 (yes)	
37	job: skilled employee/official	-1 (no), 1 (yes)	
38	job: highly skilled employee/ management/ self-employed/ officer	-1 (no), 1 (yes)	
39	number of people being liable to provide maintenance for	*numeric*	
40	telephone	-1 (no), 1 (yes)	
41	foreign worker	-1 (no), 1 (yes)	

Table 4.1: Attributes of the pre-processed data set.

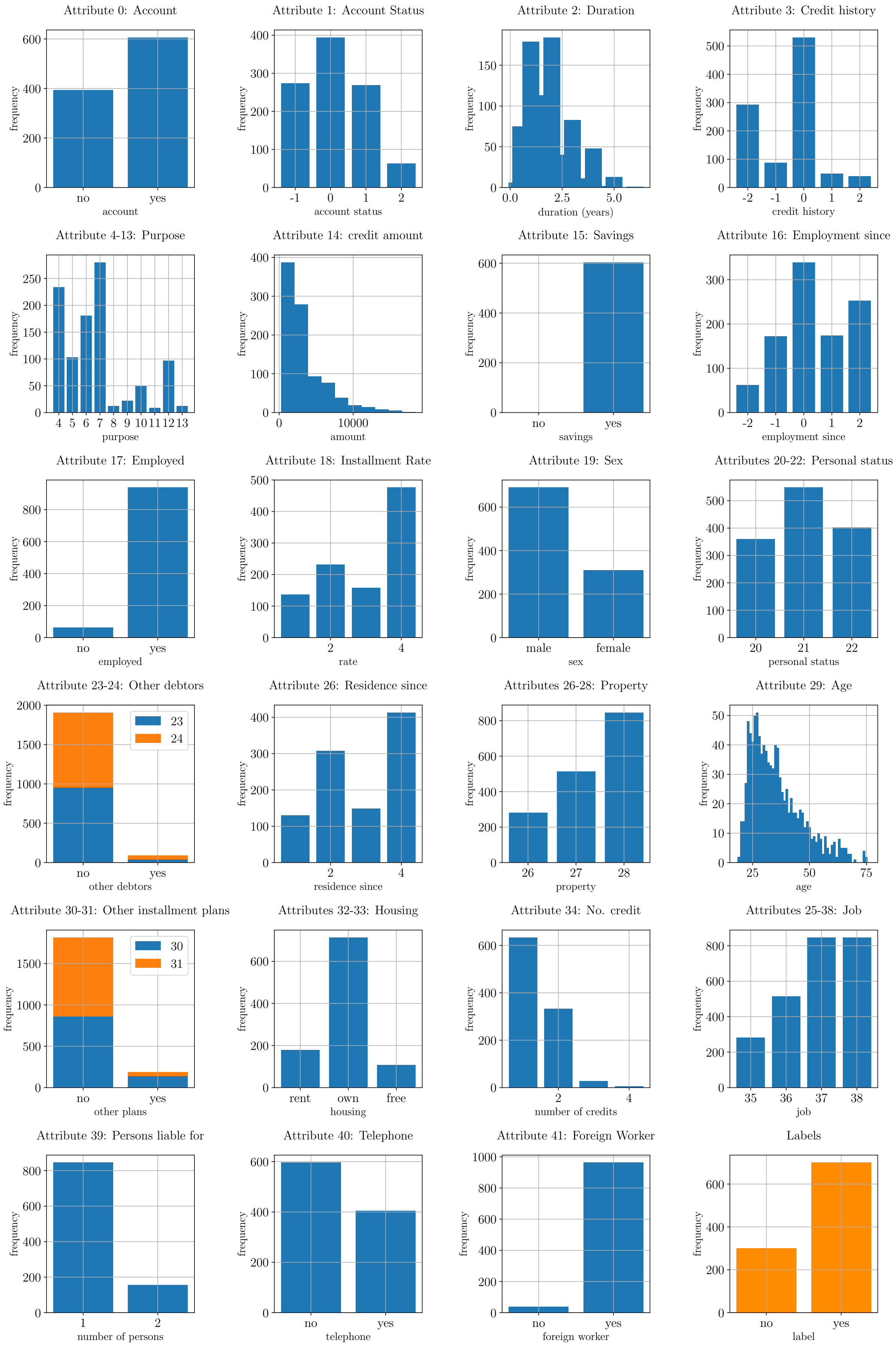


Figure 4.1: Distribution of attributes in the German credit data set.

4.2 Classifier

As already suggested in chapter 3 on theoretical background, we use Bayesian parametric logistic regression to approach the classification task. This comes with two major advantages. First, using a Bayesian approach allows taking uncertainty of the model into account. In the fairness literature this aspect has been neglected so far, although quantifying the uncertainty might enrich decision making. Second, logistic regression provides access to the weights of individual attributes before and after training, because it remains linear in the weights, and thus allows evaluating the influence that certain attributes have on the decision. The weights are directly interpretable. With binary attributes encoded as -1 and 1, an attribute has a positive effect on the decision if the sign of the respective value of the attribute coincides with the sign of its weight. When the sign of the weight and the value of the attribute do not match, this attribute will have cause a reduction of the predictive probability for the class being $Y = 1$.

Implementation. For the application of the Bayesian logistic regression an implementation was used as provided for the lecture 'Probabilistic Inference and Learning' (Gessner, 2019; Hennig, 2019). The model uses explicit polynomial feature functions. Categorical attributes of the data set are mapped to identity, while for numerical attributes the quadratic form is also taken into account.

$$\phi(x) = \begin{cases} [x], & \text{if } x \text{ is categorical} \\ [x, x^2], & \text{if } x \text{ is numerical} \end{cases} \tag{4.1}$$

The aggregation Φ then contains the constant 1, followed by all 42 linearly mapped attributes and finally the squared attributes for the numerical features. This results in a column vector with 54 entries representing each data point.

Training. For training, the data set was randomly shuffled and then split into a training and a test data set, with a ratio of 3:1. Thus the classifier was trained with 750 samples, and later evaluated on the basis of the remaining 250 samples. By default, a prior with zero mean and diagonal covariance is assumed over the weights. Then, the likelihood is computed on the basis of the training data. Finally, the approximate posterior distribution over weights is computed using a Laplace approximation (cf. chapter 3). The resulting classifier was then used to predict class membership.

Predictions. The predictions are two-fold. Initially, the classifier outputs a probability with which the person of interest belongs to the positive class. This probability then leads to the assignment to a class by comparison with a decision boundary. For the prediction of class membership, an asymmetric loss was applied, which penalizes the types of misclassification to varying degrees. This is justified as follows: The loss caused by granting a loan to a person who does not pay it back (*false positive*) is usually higher for both the bank than the loss caused by refusing a loan even though the person would have paid it back (*false negative*). In the context of loan granting, this applies similarly to the people decisions are made upon. The corresponding values for the loss were taken from the data set documentation to define a decision boundary. Since the loss due to false positives is assumed to be higher, the decision boundary shifts to the right. The classifier must be very sure that a person will actually repay the loan before granting it. Let $a = 5$ be the loss caused by false positives and $b = 1$ be the loss due to false negative classification, and $a > b$. As described in chapter 3, a person will be classified as belonging to the class which has smaller expected loss $\mathbb{E}(L|y_*)$. The decision boundary for predictions on the German credit data set is placed at a probability of $100\% \cdot \frac{a}{a+b} = 100\% \cdot \frac{5}{5+1} = 83,33\%$. Consequently, only persons predicted to pay the loan back with a probability higher than $83,33\%$ will be granted a loan.

4.3 Fairness-Enhancing Approaches

The primary objective of the practical part of this thesis is to analyse the data set and classifier using the available fairness criteria, and to implement further approaches to establish and reason about fairness. This section will cover the implementation of fairness criteria, which allows to test for fairness in the first place. Then, I will describe two approaches that could foster the reasoning about fairness-related aspects of the data set and the classifier. They can also be used to establish fairness by either pre-processing the training data or post-processing the decisions made by the classifier.

4.3.1 Testing for Fairness

As discussed in chapter 2, the available definitions of group fairness can be reduced to three fundamental criteria: independence, separation and sufficiency. In order to be fulfilled, all three criteria require parity between two different statistical measures, which can be summarized in a confusion matrix. In order to have a measure of how far the outputs of the classifier deviate from a particular fairness criterion, the average deviation is taken from the statistical measures the criterion is based on. Independence requires the probability to receive a particular classification to be equal across groups so that the classifications are independent of the sensitive attribute. In the binary case, in which there are only two possible classifications (usually positive and negative), it is sufficient to consider the acceptance rate only, since the rejection rate and the acceptance rate add up to 1. Parity of one thus implies parity of the other rate. In the implementation difference between the acceptance rates of the groups is computed,

$$\varepsilon_{ind} = P_a\{C=1\} - P_b\{C=1\}. \tag{4.2}$$

Adopting the notation from chapter 2, $C = c$ denotes the decision of the classifier, while P_a and P_b refer to the probability observed for the group with sensitive attribute $A = a$ or $A = b$ respectively. Limiting the difference ε_{ind} to a fixed value corresponds to a relaxation of the criterion, as it was used in some of the previously described papers. The notion of separation requires sensitivity, also known as true positive rate, and specificity, also known a true negative rate, to be equal across groups. To measure the extent to which the classifications deviate from the criterion, I used the average difference of both, the true positive rates and the true negative rates between groups,

$$\varepsilon_{sep} = \frac{(P_a\{C=1|Y=1\} - P_b\{C=1|Y=1\}) + (P_a\{C=0|Y=0\} - P_b\{C=0|Y=0\})}{2}. \tag{4.3}$$

Note, that this formulation is a relaxation of the stricter notion of separation, because differences in true positive and true negative rates can balance out when considering the averaged difference. Moreover, since one parity might be of greater importance to prevent one demographic group from being placed at disproportionally greater harm, it might be useful to compute a weighted average instead. The implementation of sufficiency is analogous to that of separation, except that the statistical measures considered are the positive and negative predictive value, which must be equal across the groups to satisfy the criterion. Consequently, the average difference, which again describes a relaxation of the criterion, is

$$\varepsilon_{suf} = \frac{(P_a\{Y=1|C=1\} - P_b\{Y=1|C=1\}) + (P_a\{Y=0|C=0\} - P_b\{Y=0|C=0\})}{2}. \tag{4.4}$$

Also in the case of sufficiency, it might be useful to assign more weight to one of the two parity requirements when its violation is associated with greater harm.

Summing it up, these three measurements of group fairness have been implemented to assess fairness in the classification algorithm and to measure the effects of pre- and post-processing steps on group fairness. Apart from the fairness criteria themselves, we have also considered at which points fairness of the data set or the classifier can be tested by exploiting the possibilities Bayesian inference provides. Four categories of bias have emerged from this:

1. **Model bias** $f_0 : \mathcal{X} \rightarrow \mathcal{Y} = \{-1, 1\}$
   The model bias related to the relation between input space and predictions. It is characterized by the definition of a prior over weights of the attributes in input spaces and thereby specifies to what extent individual attributes contribute to the final prediction.

2. **Population-inherent bias** $f_0(X)$
   Given a data set that represents the population, an population-inherent bias can be observed when applying the untrained classifier to training data set.

3. **Learnt bias** $f_n : \mathcal{X} \rightarrow \{-1, 1\}$
   After training, the classifier might have modified the weights of specific attributes of the input space. Analysing the classifier at this stage, might therefore reveal a learnt bias.

4. **Compound bias** $f_n(X)$
   When the trained classifier is then applied to a data set, the bias might get amplified, resulting in the last category of bias, the compound bias.

4.3.2 Modifying the Training Data

Algorithm 4.1 Find m most influential training points

```
Input: $\mathcal{D}_{\text{train}} = (X_{\text{train}}, Y_{\text{train}})$ - training data, $m$ number of points to modify, $c$ - fairness criterion
Output: $m$ most influential data points, $clf_{\tilde{X}}$ - modified trained classifier
function FindMostInfluentials($\mathcal{D}_{\text{train}}, m, c$)
    Initialization:
      - Instantiate the prior over weights $p(w)$
      - $\min_c = c(clf_X)$                ▷ Initialize minimal deviation of criterion $c$ with the classifier.
    for all $i = 1$ to $m$ do
        for all $(x, y) \in \mathcal{D}_{\text{train}}$ do                ▷ Loop over training points.
            $\tilde{X} = X_{\text{train}} \setminus x$                ▷ Temporarily remove $x$ from data set.
            $\tilde{Y} = Y_{\text{train}} \setminus y$
            $\Phi_{\tilde{X}} = \Phi(\tilde{X})$
            Retrain with $\tilde{X}$ to obtain $clf_{\tilde{X}}$
              - Instantiate the likelihood $p(\tilde{Y} | \Phi_{\tilde{X}}, w)$
              - Compute the posterior $p(w | \Phi_{\tilde{X}}, \tilde{Y})$
              - Compute predictions on training set
            Compute fairness measure $c(clf_{\tilde{X}})$
            if $c(clf_{\tilde{X}}) <= c(\min_c)$ then                ▷ Find data point with maximal influence.
                $\min_c = c(clf_{\tilde{X}})$
                $x_{\text{max\_imp}} = x$
                $y_{\text{max\_imp}} = y$
            end if
        end for
        $X_{\text{train}} = X_{\text{train}} \setminus x_{\text{max\_imp}}$                ▷ Remove $x_{\text{max\_imp}}$ from data set.
        $Y_{\text{train}} = Y_{\text{train}} \setminus y_{\text{max\_imp}}$
    end for
    return $m$ most influential points, resulting classifier $clf_{\tilde{X}}$
end function
```

The first consideration of how fairness could be further investigated or even achieved focuses on modifications of the training data. The idea is to be able to track back the influence of individual training points on the weight of the sensitive attribute or a selected fairness criterion. This might be helpful to separate data and model bias and reason about potential problems in the data. More generally, the problem can be formulated as finding a set of training points of fixed size that minimizes a selected fairness criterion. It is

$$\text{argmin}_{\{x_i \in X\}_{i=1,\dots,m}} c(clf_{X \setminus \{x_i\}}) \tag{4.5}$$

for a set of fixed size of m data points and a fairness measure c observed for a classifier training on a reduced training data set. By removing this set of data points from the training data set, the extent to which the resulting classifier deviates from the fairness criterion should significantly decrease. Ideally, a combinatorial problem should be solved that identifies the m-tuple whose removal from the training data leads to a strong alignment with the desired fairness criterion. However, since this greatly increases the search space with increasing m, the problem was addressed using a greedy algorithm (cf. submodularity). As outlined in the pseudocode for Algorithm 4.1, it is comprised of m loops over the training data in which the greedy algorithm searches for the data point with the greatest impact on the respective fairness criterion c. It is assumed that the greedy approach approximates the ideal set sufficiently well, especially since it takes into account possible mutual interferences between the data points by sequentially expanding the set.

Alternatively, it could be considered to flip the labels of the data point with the highest impact instead of removing them. This comes with the advantage that the size of the training data set remains the same and thus it might be less intrusive. A further approach is to find the most influential training data points that affect the weight of the sensitive attribute directly.

The motivation for the described approach is to stimulate the in-depth analysis of the data set by identifying and analysing particularly influential data points. The hope would be that this approach can facilitate the identification of potential problems in the data set.

4.3.3 Margin around Decision Boundary

The second technique is implemented after the training and examines the area near the decision boundary more closely. This area is particularly important because close to the boundary there is a potential for classifications that can be inverted simply by changing the sensitive attribute. In other words, data points that are close to the decision boundary are more prone to get discriminated due to their sensitive attributes. We therefore define a margin δ around the decision boundary, in which individual decisions and the influence of the sensitive attribute are examined more carefully. As outlined in the pseudocode for Algorithm 4.2, the margin can be approximated using the test set. Note, that we assume the sensitive attribute to be binary, i.e. $x_a \in \{-1, 1\}$, for the proposed technique. Since the test set may not contain enough "extreme cases", it is furthermore useful to take a closer look at the relation between the margin and the weight of the sensitive attribute. Formalizing the relation potentially offers the advantage of being able to estimate the margin independently of the test set and defining an upper bound for δ. The derivation of the δ and the upper bound of the margin, $\delta_{\max}$, begins with the likelihood of outcomes,

$$p(y|\Phi, \hat{w}) = \sigma(\underbrace{y_i}_{\pm 1} \phi(x)^T \hat{w}) \tag{4.6}$$

$$p(y = 1|\Phi, \hat{w}) = \sigma(\phi(x)^T \hat{w}). \tag{4.7}$$

where $\hat{w}$ is the MAP estimator of the weights which is used to make predictions. We can then exploit the fact that logistic regression is based on the weighted sum of inputs. Thus the individual weights of the attributes are accessible and the influence of the sensitive attribute can be separated from that of the

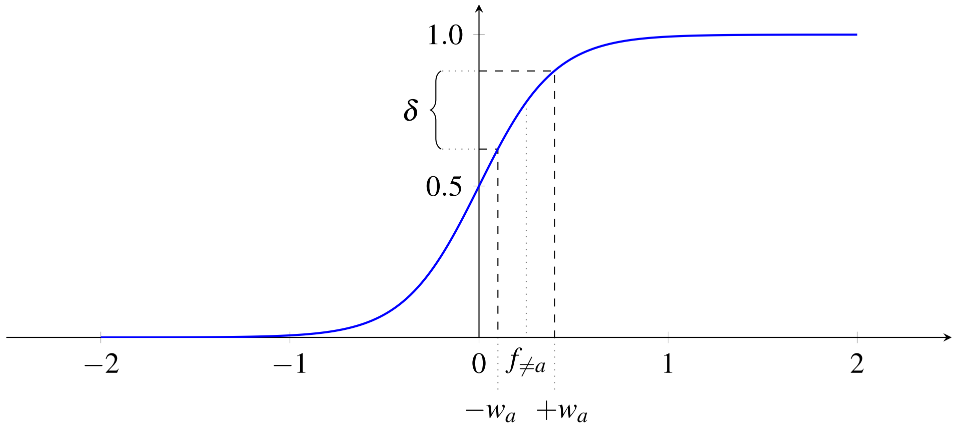


Figure 4.2: Deriving the maximal margin δ based on the weight of the sensitive attribute w_a

non-sensitive attribute:

$$\begin{align}
\phi(x)^T \hat{w} &= \sum_{n=1}^{N} \phi_n(x) w_n \tag{4.8}\\
&= \sum_{n \neq a} \phi_n(x) w_n + \underbrace{\phi_a(x)}_{\pm 1} w_a \tag{4.9}\\
&= \underbrace{\sum_{n \neq a} \phi_n(x) w_n}_{:= f_{\neq a}} \pm w_a \tag{4.10}\\
&= f_{\neq a} \pm w_a \tag{4.11}\\
p(y=1|\Phi, \hat{w}) &= \sigma(f_{\neq a} \pm w_a) \tag{4.12}
\end{align}$$

To obtain δ, it is sufficient to consider the functional values on which an individual is mapped with and without having the sensitive attribute using the logistic function, i.e.

$$\delta = \sigma(f_{\neq a} + w_a) - \sigma(f_{\neq a} - w_a). \tag{4.13}$$

This is also visualized in Figure 4.2. It is important to note that delta depends not only on the sensitive attribute, but also on the values of the other attributes. Its value value therefore varies between individuals, what motivates the search for an upper bound of δ. The upper bound can be determined by considering the point at which the logistic function has the steepest slope. At this point, the weight of the sensitive attribute has maximal impact because deviations to the left or right along the input space cause a huge difference in function values. The logistic function shows maximum slope at $x = 0$. We therefore set $f_{\neg a} = 0$, which results in the maximal δ:

$$\begin{align}
\delta_{\max} &= \sigma(w_a) - \sigma(-w_a) \tag{4.14}\\
&= \sigma(w_a) - (1 - \sigma(w_a)) \tag{4.15}\\
&= 2\sigma(w_a) - 1 \tag{4.16}
\end{align}$$

Note, that the upper bound does only depend on the weight of the sensitive attribute. Thus is can be estimated without looking at the test set. Further notice that the width of the margin around the decision boundary is closely related to the notion of independence, as the margin depends directly on the weight of the sensitive attribute a. δ shrinks with decreasing weight of w_a and vanishes if w_a has no influence on the decision at all.

Algorithm 4.2 Find points within margin around decision boundary

```
Input: $\mathcal{D}_{\text{test}} = (X_{\text{test}}, y_{\text{test}})$ - test data, $clf_X$ - trained classifier, $b$ - decision boundary
Output: margin $\delta_{\max}$, points within margin that have changed by flipping $x[a]$
function FindMaxMargin($\mathcal{D}_{\text{test}}, clf_X, b$)
    Initialization
      - $\delta_{\max} = 0$
    for all $x \in X_{\text{test}}$ do                                ▷ Loop over test points
        Predict class $\hat{y}_x$ and class probability for $x$
        Flip sensitive attribute $x[a] = -x[a]$
        Predict class $\hat{y}_{\tilde{x}}$ and class probability for $\tilde{x}$
        if $\hat{y}_{\tilde{x}} \neq \hat{y}_x$ then
            Compute $\delta = b -$ class probability of $x$
            if $\delta > \delta_{\max}$ then
                $\delta_{\max} = \delta$
            end if
        end if
    end for
    return $\delta_{\max}$, points within margin that have changed by flipping $x[a]$
end function
```

5 Results

This chapter is concerned with the analysis of the outcomes using the implemented fairness criteria and methods to mitigate potential biases. For the implementation, the data set was divided into a training data set and a test data set as described above. To further evaluate the robustness of the proposed approaches, the results are averaged over ten random test-train splits unless otherwise stated. In each split, the data of 750 persons was used for training the classifier. Thus, the analysis is based on the remaining 250 people per test-train split for whom attributes and labels are available.

5.1 Unconstrained Classifier

The results for the unaltered classifier serve as a baseline for the comparison and evaluation of the proposed methods. In order to identify potential problems, it is common practice to start by carefully analysing the properties of the data and the behaviour of the classifier. In the context of fairness, such analyses have significantly contributed to the debate. However, they usually focus on the behaviour of an existing method because the software remains proprietary, which comes with limited scope.

As a brief recap, Bayesian logistic regression is used for the classification task in this thesis. Before the data was considered, a prior with zero mean and diagonal covariance was assumed over the weights. Related to the four types of bias described in 4.3, the prior over weights is connected to model bias that specifies the relationship between input space and predictions. Since the prior was initially set to zero for all weights, no bias against certain attributes in the input space is present and a priori all attributes contribute equally to the predictions. The use of a prior with zero mean entails that the population-inherent bias cannot be explicitly observed. This is the case because the data points are summed up weighted in the framework of logistic regression. If weights of zero are assumed, inconsistencies in the data cannot be expressed. We assume that population bias can be validated by using a more elaborate prior if certain properties of the data have different effects on the different demographic groups using the prior. This type of bias manifests itself when the untrained classifier is used on the data set and imbalances of the predicted labels can be observed for demographic groups.

For predictions on the German credit data set, a decision boundary was placed at a threshold of 83.33%, i.e. the classifier had to assign a loan applicant to the positive class with a relatively high degree of certainty in order to grant the loan.

Accuracy and Loss. Using logistic regression, an average accuracy of 60% can be observed across all test-train splits (standard deviation $1.11 \cdot 10^{-14}$). When using asymmetric loss, however, the accuracy is not particularly meaningful, because the shifted decision boundary skews accuracy. In particular, avoiding expensive false positives is linked to rejecting more applicants, who would actually paid their loan back, and thus the overall accuracy decreases. Another measure to assess the performance of the classifier is the loss incurred. As indicated in the documentation of the data set, false positives were assigned a loss of 5 and false negatives a loss of 1. On the test set an average loss of 69.95 (standard deviation 2.18) could be observed, which is much smaller than the average maximum loss of 834.4 (standard deviation 34.98). The maximum loss would be incurred when the classifier misclassified all test samples.

Posterior Weights and Covariance. Logistic regression comes with the advantage that the weights are available for individual features, which allows to draw direct conclusions about the influence of each

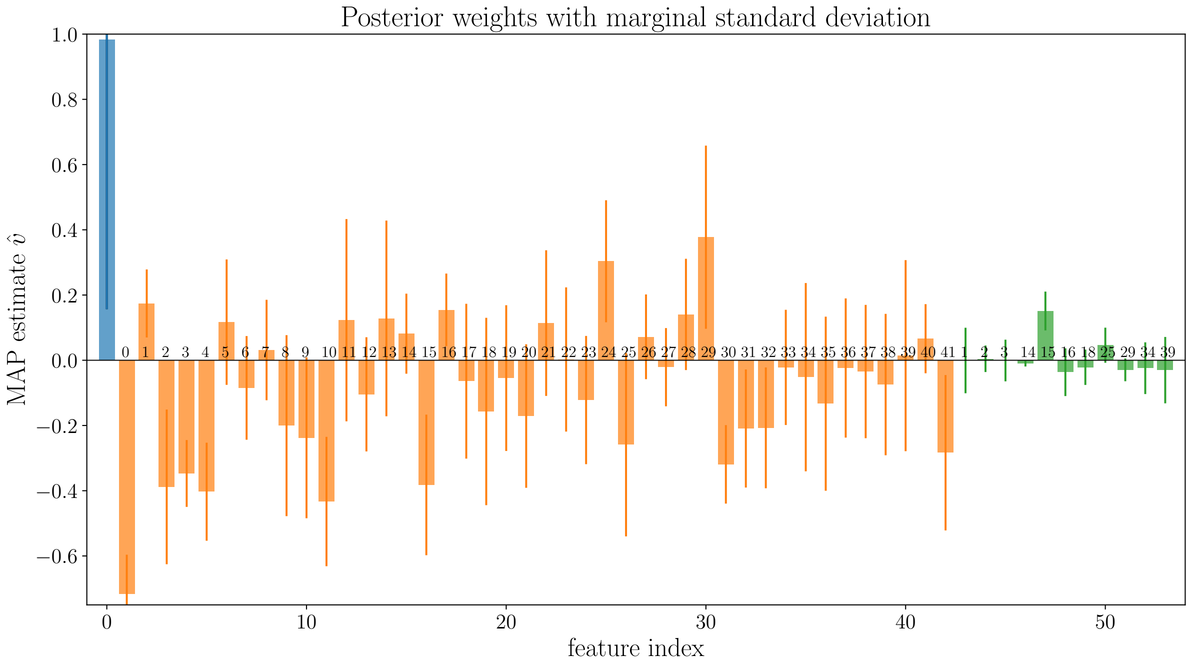


Figure 5.1: Posterior weights of the individual features and their marginal standard deviation as obtained after training of the unconstrained classifier. The weights coloured in blue correspond to the offset, orange weights to linear features. Numerical attributes were additionally mapped to their quadratic form, which are coloured in green.

attribute on the final decision. The posterior weights of the classifier are adjusted based on the training data set. Biases due to unusually high values of posterior weights can thus be traced back to the training phase. This type of bias was previously introduced as learnt bias, since weights are altered according to the characteristics of individuals in the training data set. Since the posterior is composed of likelihood and prior, there is a risk that model and population-inherent bias intermingle as a result of training.

This part of the analysis is based exclusively on the first split, since taking the average balances out different weightings of attributes and hence leads to blurred results. The maximum a posteriori estimates of the weights obtained after training are visualized in Figure 5.1 together with their marginal standard deviation. The blue coloured feature corresponds to the weight of the offset. The weights in orange correspond to the linear features, i.e. the attributes themselves, and the weights of the squared numerical attributes are plotted in green. To assess the weights, it is necessary to consider the nature of the attributes (see table 4.1 for details). For the numeric attributes (1, 2, 3, 14, 15, 16, 18, 25, 29, 34, 39) and their squared features, the sign of the weights corresponds to their direction of influence. The remaining attributes are encoded binary with 1 and -1, which corresponds to having or not having a particular attribute. For these, a negative weight counteracts the granting of credit if an applicant has the corresponding attribute, while a positive weight works towards the approval. If the corresponding attribute is not present, the effect of weights is reversed accordingly. Regarding the displayed MAP estimates of the weights, it is noticeable that the squared features as a whole have relatively little influence on the credit decision. Attribute 0, on the other hand, which describes whether the applicant has a bank account or not, has a great influence. Consequently, it is beneficial for the decision not to have a bank account because the weight is negative. Attributes 1 and 29, which correspond to the income and age of the applicant are positively related to loan granting, while being a foreign worker is negatively weighted for the decision. The latter might serve as indicator to reflected human biases. However, it must also be taken into account that 96.3% of the individuals in the data set are foreign workers. Examining the weights further, it can be observed and probably argued logically that the purpose of the credit, encoded in attributes 4 to 13, plays a role in the allocation. For example, it is positively weighted to buy a loan for a used car instead of a new one. Noteworthy in this context is that an educational loan is valued negatively, while retraining is weighted positively.

In addition to the MAP estimator $\hat{W}$, the standard deviation of the weights should also be considered, as it gives an indication of how uncertain the classifier is about the influence of individual features. Keep in mind that the standard deviations are approximated using the Laplace approximation because the posterior distribution is analytically intractable in the case of logistic regression. As mentioned before, the standard deviation of the individual weights is shown in Figure 5.1. In general, the observed standard deviation of the weights is relatively high, especially for linear features and the offset. Also in view of the circumstance that only the first test-train split is illustrated here, it becomes clear that, depending on the samples in the training data set, strong deviations between observed posterior weights are probable. Comparing the displayed weights with the averaged posterior weights, a tendency toward more positive weights is observable, i.e. the impact of the majority of negative weights decreases, while most positive weights gain in influence. The standard deviation in the first test-train split corresponds approximately to the standard deviation that can also be observed in the averaged values. For numerous attributes such as sex, personal status or job, the indicated standard deviation intersects the x-axis. Thus, positive and negative weighting of the attributes falls within a normal range, which suggests that the classification is highly uncertain. Incorporating this uncertainty in the assessment of the weights, the extent to which some attributes favour or hinder the decision remains not clearly apparent.

It can further be informative to examine the covariances between features. The sign of the covariance of two features describes how the linear relationship between these features behaves. It is positive when larger values of one feature are accompanied by larger values of the other and negative when larger values of one feature are associated with smaller values of the other feature. Looking at the covariance matrix in Figure 5.2, allows conclusions to be drawn about which attributes tend to appear together more frequently and consequently which combinations contribute to credit allocation. The left subfigure visualises the covariance matrix. Since the variance of the features, located on the diagonal of the matrix, outweighs the covariances, the right figure visualizes the off-diagonal entries of the covariance matrix. Along the diagonal, several clusters are visible within which particularly high covariances of the corresponding features can be observed. The first cluster, spanned from feature 5 to feature 14, contains the attributes encoding the purpose of the credit. The second cluster, which clearly stands out from its surroundings, is composed of the attributes sex and personal status. Finally, the rows and columns that belong to features 36-39 show strong covariances between job-related attributes.

Sensitive Attribute. In the context of fairness, the covariance or correlation between the sensitive attribute and other features is of particular interest, as it can reveal potential structures in the data and dependencies between features. To examine the covariance of a single attribute, it is useful to take a closer look at the corresponding row of the covariance matrix. In the context of loan granting and with the given data set, several attributes could be considered worthy of protection because they are either legally protected or their usage for the classification is at least questionable. These include sex, personal status and the age of an applicant as well as the last attribute encoding whether he or she is a foreign worker. Figure 5.3 shows a cross-section of the covariance matrix along the corresponding attributes. The covariance between sex and the attributes encoding the personal status stands out, whereby it should also be noted that these attributes were a joint attribute in the original data set and separated from each other in the preprocessing. As described above, this may have led to inconsistencies that affect the observable effects. The covariances for age and foreign worker show less pronounced effects. However, both attributes are positively related to attributes 4-13, which describe the purpose of the credit request, and attributes 35-37, which refer to the applicant's work.

Although multiple attributes in the data set can be considered sensitive, the subsequent analysis of the trained classifier and the methods described in chapter 4 is limited to an examination of possible gender-related biases. The test data set of the first split contains 181 samples of men and 69 samples of

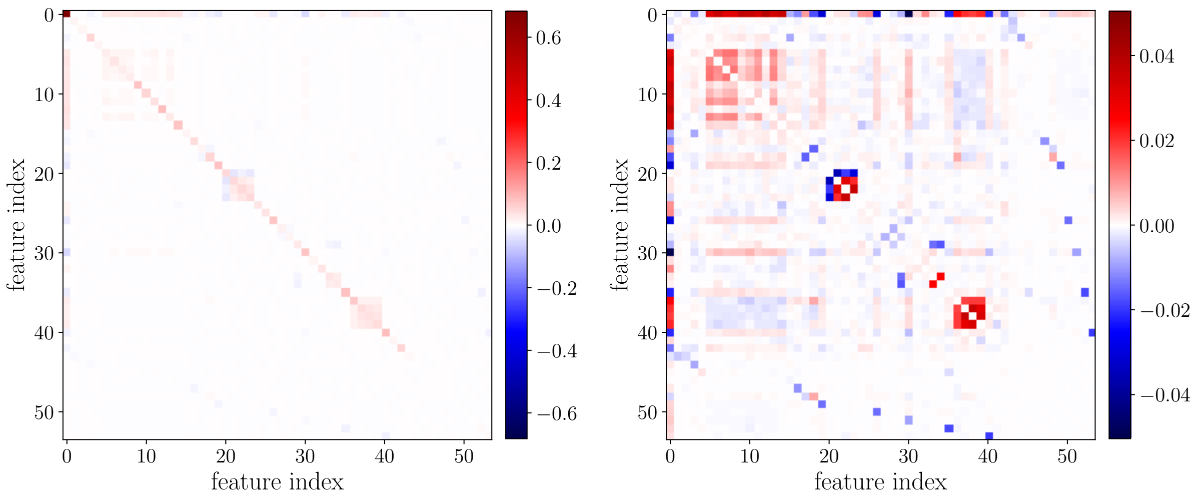


Figure 5.2: Covariance matrix of the features. The left figure shows the unmodified covariance matrix. As the variances of individual features considerably surpass the covariances between features, they were excluded in the right figure for a more detailed visualization of the covariances.

women, being representative of the overall distribution of sex in the German credit data set.

Fairness Criteria. As described before, group fairness criteria are solely based on statistical measures that can be summarised in a confusion matrix (see section 2.1.3 for more details). We will therefore examine the confusion matrix of the classifier for each group, which describes the relation between true and predicted class membership. Figures 5.4a to 5.4c visualize this relationship for the entire test set as well as for men and women in specific using the first test-train split. Note that the entries are normalised. Looking at the confusion matrix for the classifier in general, more accurate performance can be observed for individuals who actually defaulted (true negatives) whereas the performance for individuals belonging to the true positive class seem to be classified close to random. Dividing the test data by groups, i.e. males and females, varying distributions of outcomes can be observed. In particular, the true positives and false negatives vary considerably between groups. While a larger proportion of men is correctly granted credit, a larger proportion of women is falsely rejected. This indicates possible gender-specific discrimination by the classifier. To inform this analysis further, figures 5.4d and 5.4e show the conditional probabilities of paying back the loan or being predicted to do so, given the sex of the applicant. It can be seen that men and women in the test data set are more likely to repay the loan, while the classifier after training shows a tendency not to grant loans. This pattern is consistent with the shifted decision boundary. In addition, it can be observed that women, although those in the test data are more likely to have paid back their loan, are less likely to receive a corresponding prediction.

The implemented fairness criteria can be used to further quantify the imbalance between groups. For the unconstrained classifier, the acceptance rates between men and women differ by 20.11% and thereby showing a lack of independence from the sensitive attribute. Although it is desirable that acceptance rates are independent of group membership, it may be of greater interest, especially for the decision maker, in this case a bank, to consider the probabilities of correctly approving or rejecting a credit application. This corresponds to the idea of separation, which requires equal true positive rates and true negative across groups. For the unconstrained classifier, an average difference of 19.8% was observed between groups, so separation is not satisfied either. The third notion, sufficiency, requires equal positive and negative predictive values across groups. This measure comes closest to the interest of the decision maker, as it examines the probability with which applicants who have received a loan will actually pay the it back and the probability with which applicants who have been rejected would not have paid

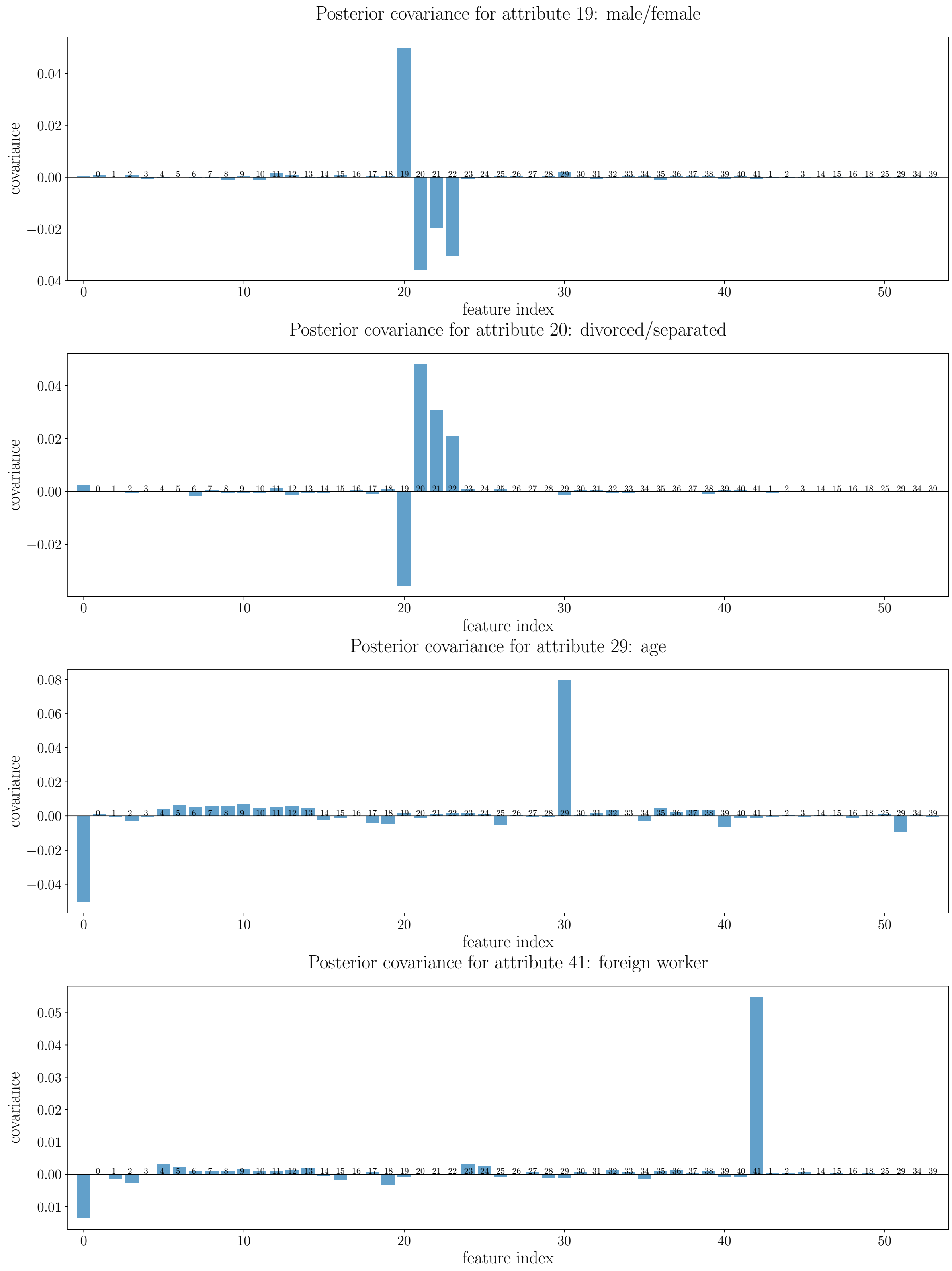
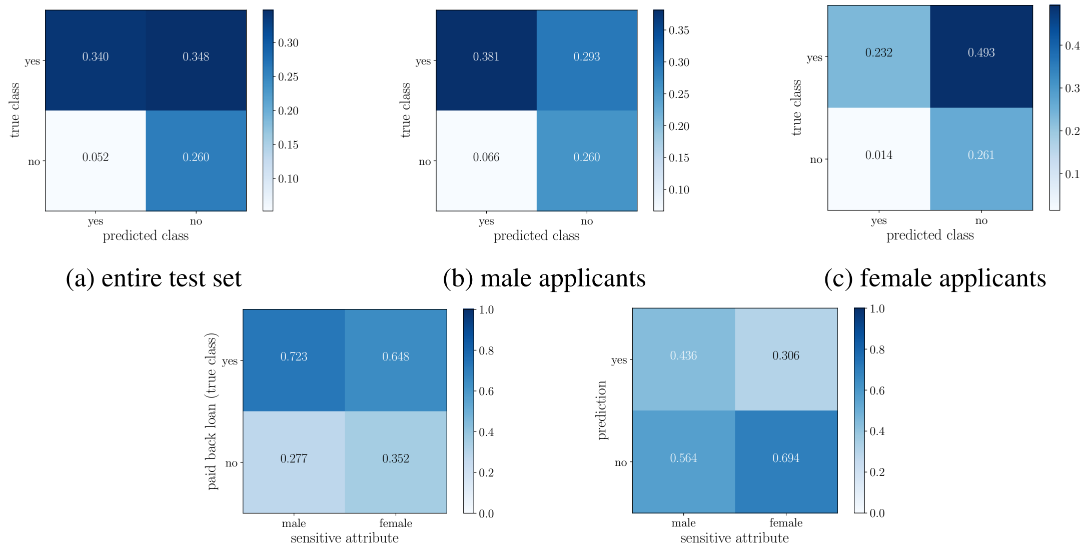


Figure 5.3: Posterior covariance for selected features that could be considered sensitive for the task of loan granting.



(a) entire test set

(b) male applicants

(c) female applicants

(d) true class

(e) predictions

Figure 5.4: Conditional probabilities for the entire test set as well as group-specific for men and women based on a single test-train split. (a)-(c) The confusion matrices visualize the relationship between the true and the predicted class membership. The entries are conditioned on the true class membership and therefore describe the observed true positive, false negative, false positive and true negative rates of the specified population. (d)-(e) The figures visualize the conditional probabilities of (d) true class membership (e) and predicted class membership given the sex of the applicant.

it back. With regard to sufficiency, the unconstrained classifier shows the smallest average deviation between groups with 10.65%, but does not fulfil it either.

5.2 Modification of the Training Data Set

As described in 4.3.2, the modification of the training data provides an opportunity to gain further insights into the data set and possible implications for the classifier. Therefore, those training points are removed in a sequential manner, for which maximal impact of the fairness criterion of interest was observed. This process was repeated for ten different test-train splits in order to estimate its robustness. In the following sections, the results for all three measures of group fairness are presented. In the hope to uncover potential pitfalls of the data set, the ten most influential training points from the first test-train split were additionally subjected to a detailed examination.

5.2.1 Independence

Modifying the training data regarding independence pursues the goal of obtaining a classifier whose predictions are independent of the sensitive attribute. The difference between the acceptance rates of the classifier for the individual demographic groups is used to quantify the deviation from the notion of independence. In an iterative process, those training points that have the greatest influence on the criterion are removed from the training data. Re-training on the modified data set then hopefully leads to a fairer classifier.

Fairness criterion. The development of the deviation by the number of excluded points is shown in Figure 5.5. Here, the training data set was modified to satisfy independence. In orange, the mean

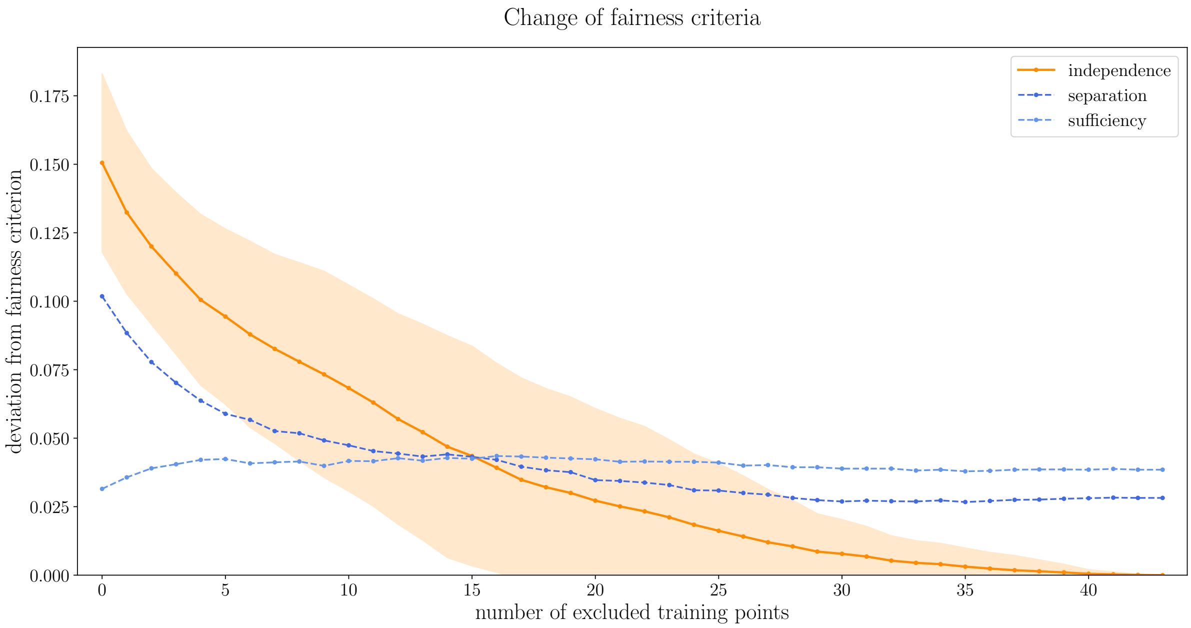


Figure 5.5: Development of the deviation from the measure of independence with number of most influential points removed from the training data set. Alongside the orange line, which is assigned to independence, the average development of the measures of separation and sufficiency is plotted in blue. The standard deviation observed across ten train-test splits is visualized in light orange. On average, independence was achieved after the removal of 28 samples.

deviation from the notion of independence and its standard deviation are illustrated. On average, 28 samples had to be excluded in order to obtain a classifier satisfying independence. The blue lines indicate the deviation from separation and sufficiency for which no optimization has been performed. While the deviation from sufficiency increases slightly, the measure of separation gets improved, but both curves plateau after a small number of iterations. The overall results are compliant with previous findings by Chouldechova (2016) and Kleinberg et al. (2016) that the three fundamental group fairness criteria cannot be achieved simultaneously. Obviously the exact number of excluded points depends on the prevalence of the demographic groups in question as well as on the test-train split. Hence, the development presented in Figure 5.5 should rather be considered as an estimate of the effort required to obtain a fairer classifier. In particular, it is not sufficient to study only this development. Instead, a detailed examination of the characteristics of the excluded data points and the ”fair” classifier resulting from modifications of the training data set is necessary.

Posterior weights. In analogy to the analysis of the unconstrained classifier, the weights of the “independent” classifier and its behaviour on the test data set can be examined. The latter will be discussed in section 5.2.4. The changes in the weights as observed in the first test-train split are shown in Figure 5.6. In comparison to the original classifier, the standard deviation shows a tendency to decrease. Moreover, the absolute value of the majority of negative weights has decreased which consequently have less impact on the final decision. Two weights, in particular the weight assigned to attribute 19 (sex) and attribute 22 (being married or widowed), have even changed their direction of impact. Interestingly, while the greatest difference can be observed for the sensitive attribute sex, the weight did not change to zero, as one could have expected under the notion of independence. However, since the sensitive attribute sex is correlated with other attributes in the data set, the weighted sum over those attributes should add up to zero.

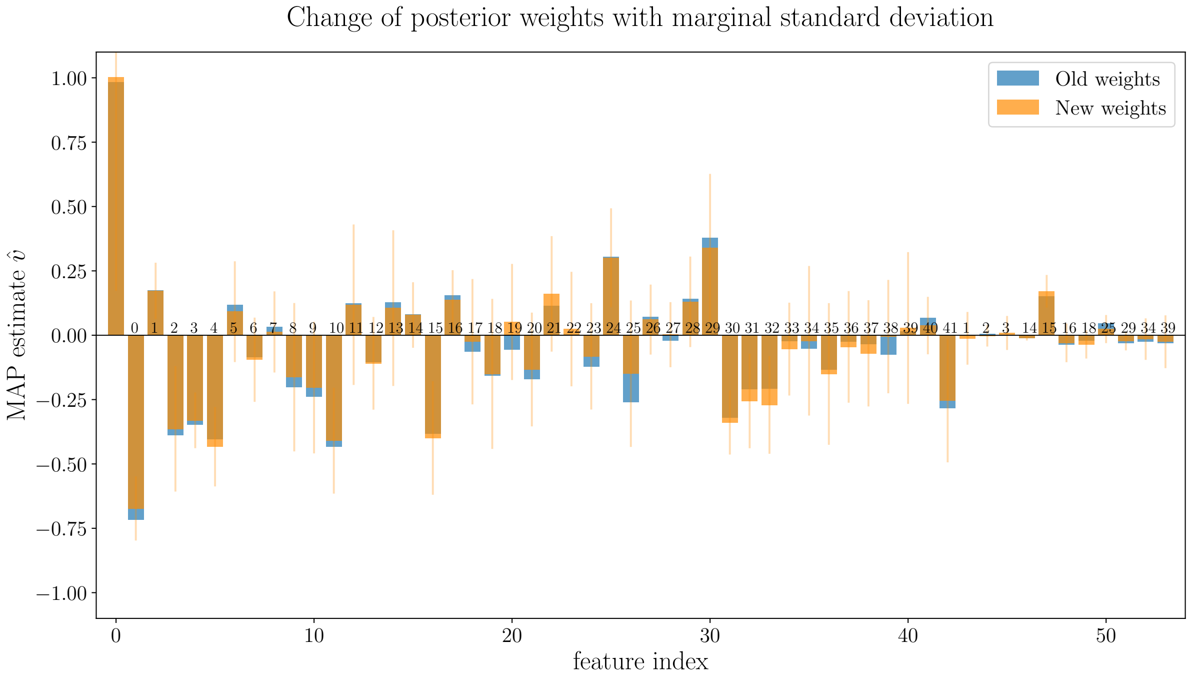


Figure 5.6: Independence: Change of posterior weights after training the classifier with the modified data set as observed for the first test-train split. In blue the posterior weights of the unmodified classifier are displayed, in orange the posterior weights obtained for the "independent" classifier.

Most influential training points. The properties of the removed data points may indicate problems in the data set, as these data points have led to a more unfair classification on the training data. Which points were discarded strongly depends on the test-train split. Figure 5.7 shows the characteristics of the ten most influential data points that were removed in the first split to achieve independence. In the first split, the weight of the sensitive attribute is negative ($w_a = -0.055$), i.e. it is beneficial for the decision to be a man. Interestingly, the majority of excluded samples were women. As described by the attribute 1, the majority of those people have no or only little money on their bank accounts, although they are mostly skilled or highly skilled employees (attributes 25-28). Regarding the labels of those ten points, the majority of them did not pay back loan. That might suggest, that the characteristics of the people excluded from the training data set led to an reduction of the predictive probability for the class being $Y = 1$, influencing the predictive probability for the whole group. When comparing the true class membership to the predicted class membership as predicted by the original classifier, the majority of the excluded data points would be correctly classified as negative. This behaviour can probably be explained with the purpose of modifying the data: achieving independence, i.e. equal acceptance rates across groups. By reducing the number of true negatives in one group, the acceptance rate of this group increases, which is conducive to the goal of independence.

5.2.2 Separation

As discussed earlier, separation can be of special interest for both the decision maker and the individuals decided upon, because it requires parity between sensitivity and specificity across groups. Both measures describe the probability that a customer has been correctly allocated a loan or not and thereby avoid a major loss in utility for the decision maker and harm for the individuals of a particular demographic group. For the proposed data modification approach, deviation from the notion of separation is quantified as mean difference between true positive and true negative rates across groups. Sequentially,

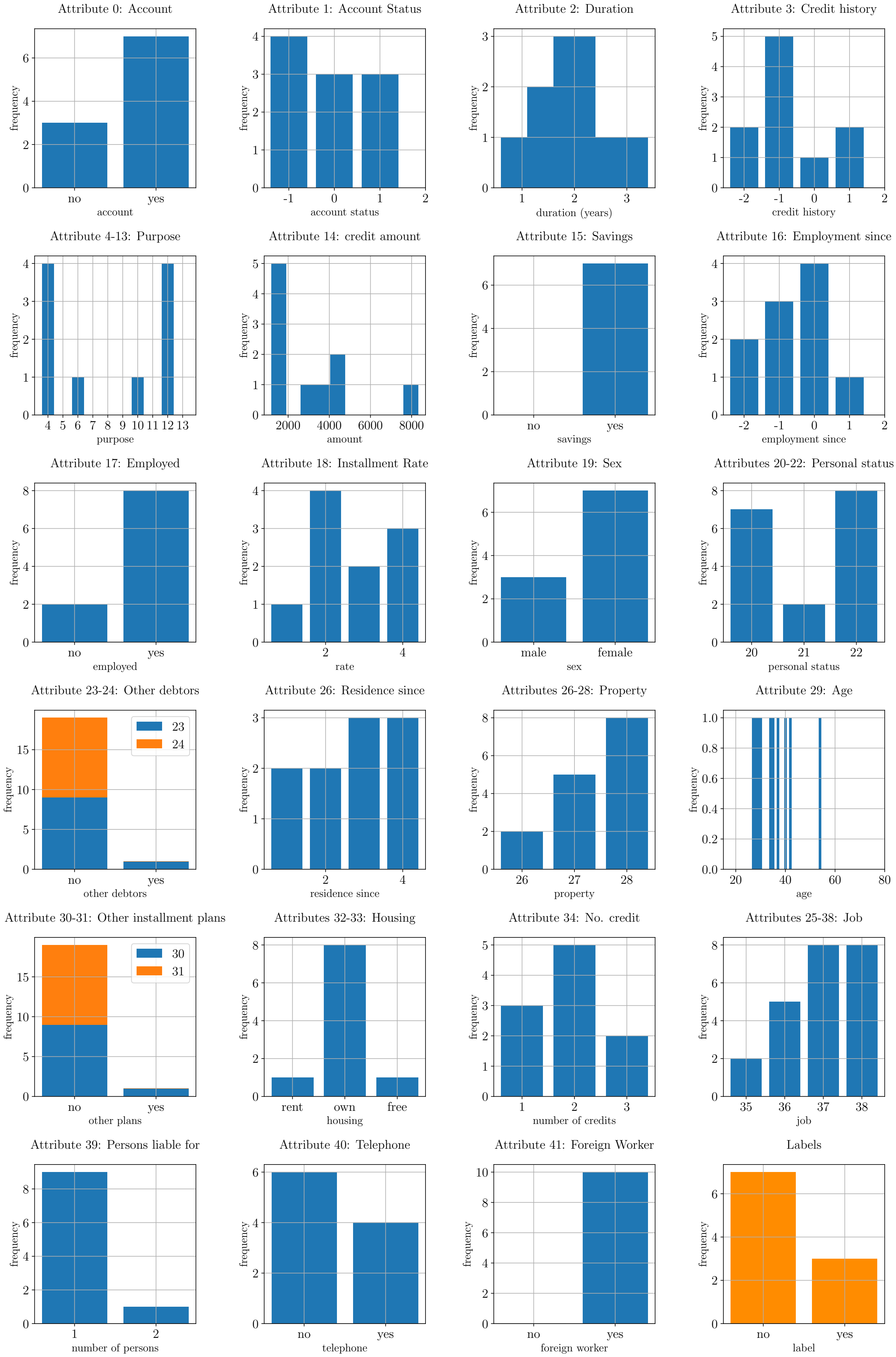


Figure 5.7: Distribution of attributes for the ten most influential data points that were removed in the first split to achieve independence.

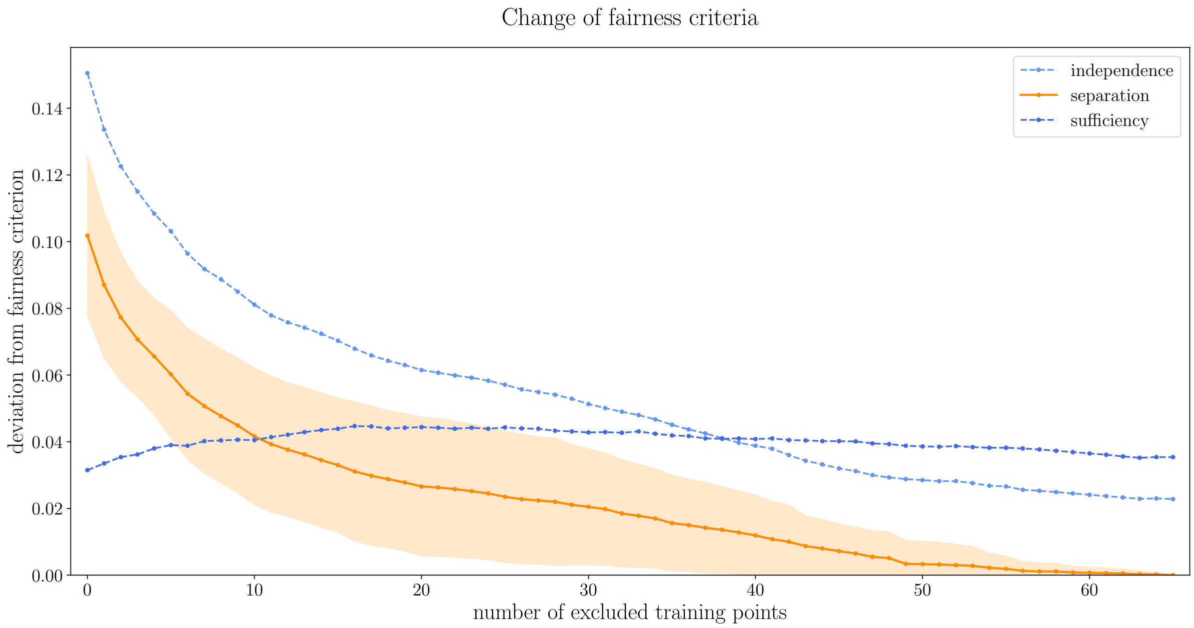


Figure 5.8: Development of the deviation from the criterion of separation with number of points removed from the training data set, averaged over ten test-train splits. The standard deviation is illustrated in light orange. The average development of independence and sufficiency is plotted in blue. On average, a classifier satisfying separation was obtained by removing 45 samples.

those training points that have the greatest influence on the criterion are removed from the training data. Re-training on the modified data set then leads to a fairer classifier in terms of separation.

Fairness criterion. The development of the deviation by the number of excluded points is shown in Figure 5.8 as observed for ten different test-train splits. Here, the training data set was modified to satisfy separation. In orange, the mean deviation from the notion of separation and its standard deviation are illustrated. On average, 45 samples had to be excluded to get a classifier satisfying separation. This is considerably more than was necessary for independence, which also manifests in the more rapid flattening of the curve assigned to separation. Moreover, the standard deviation is slightly smaller and increased less with increasing number of excluded points in the case of separation. The blue lines indicate the deviation from independence and sufficiency for which no optimization has been performed. Similar to the observations in the previous section, the deviation from independence decreases with the number of points which has been excluded to approach separation. The deviation from sufficiency, on the other hand, increases slightly and plateaus at approximately 0.05. Again, the overall results are consistent with previous findings by Chouldechova (2016) and Kleinberg et al. (2016) that the three fundamental group fairness criteria cannot be achieved simultaneously.

Posterior weights. Comparing the posterior weights of the classifier when trained on the original or the modified data set, could give further insights. The differences are illustrated in Figure 5.9. No uniform trend can be identified, but the weight of the sensitive attribute decreases significantly and thus loses influence on the final decision. The largest changes can be observed for attribute 10, 18 and 25, which describe the educational purpose of the credit, the instalment rate and the present resident time of the applicant respectively. Since these attributes are not obviously related to the sensitive attribute, no concrete implications can be derived from them. In the case of separation, none of the linear features flips its direction of influence as it was the case for independence.

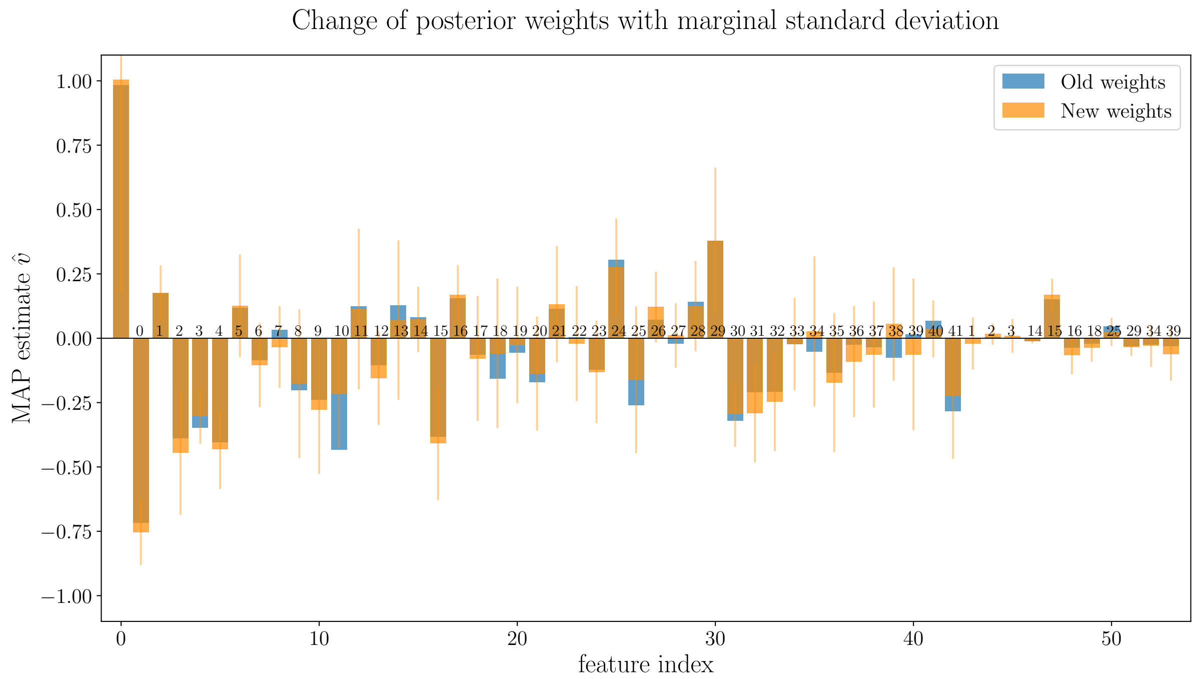


Figure 5.9: Separation: Change of posterior weights after training the classifier with the modified data set from the first test-train split. In blue the posterior weights of the unmodified classifier are displayed, in orange the posterior weights obtained for the "separated" classifier.

Most influential training points. Similar to the case of independence, the ten most influential training points from the first test-train split were selected for closer examination. Their properties may indicate problems in the data set, as the removal of these data points have led to fairer classification on the training data. Which points were discarded strongly depends on the test-train split. Figure 5.10 shows the characteristics of the ten most influential data points that were removed in the first split to achieve separation on the training data. The corresponding weight of the sensitive attribute $w_a = -0.055$, thus being a woman counteracts a positive classification. In contrast to the most influential points discussed in the context of independence, the majority of samples, that were excluded from the training set to achieve separation, were people who paid back their loan. Almost all of them do not have other debtors. The distribution over attributes 4-13 of the excluded persons differs considerably from the corresponding distribution of the population. In particular the most influential persons applied disproportionately often for educational and business-related credits. This might be reflected in the change of the posterior weight for attribute 10 and point to an gender-related imbalance regarding the purpose of the credit. When comparing the true labels and the predicted labels of the ten most influential training points, the majority would have been wrongly classified as creditworthy. Reducing the proportion of false positives is aligned with the goal of sufficiency in that it increases the true positive rate on the training set. Since the majority of excluded training points are female, this behaviour could indicate an increase in the true positive rate of women.

5.2.3 Sufficiency

The procedure for modifying the data set in terms of sufficiency is the same as described before. Those training points that show the greatest influence on the criterion are removed from the training data set in a sequential manner. The classifier trained on the modified training data set then produces fairer predictions on the training data set in terms of sufficiency. It should be noted that the unprocessed classifier has already shown the smallest average deviation for this criterion. This it is not particularly

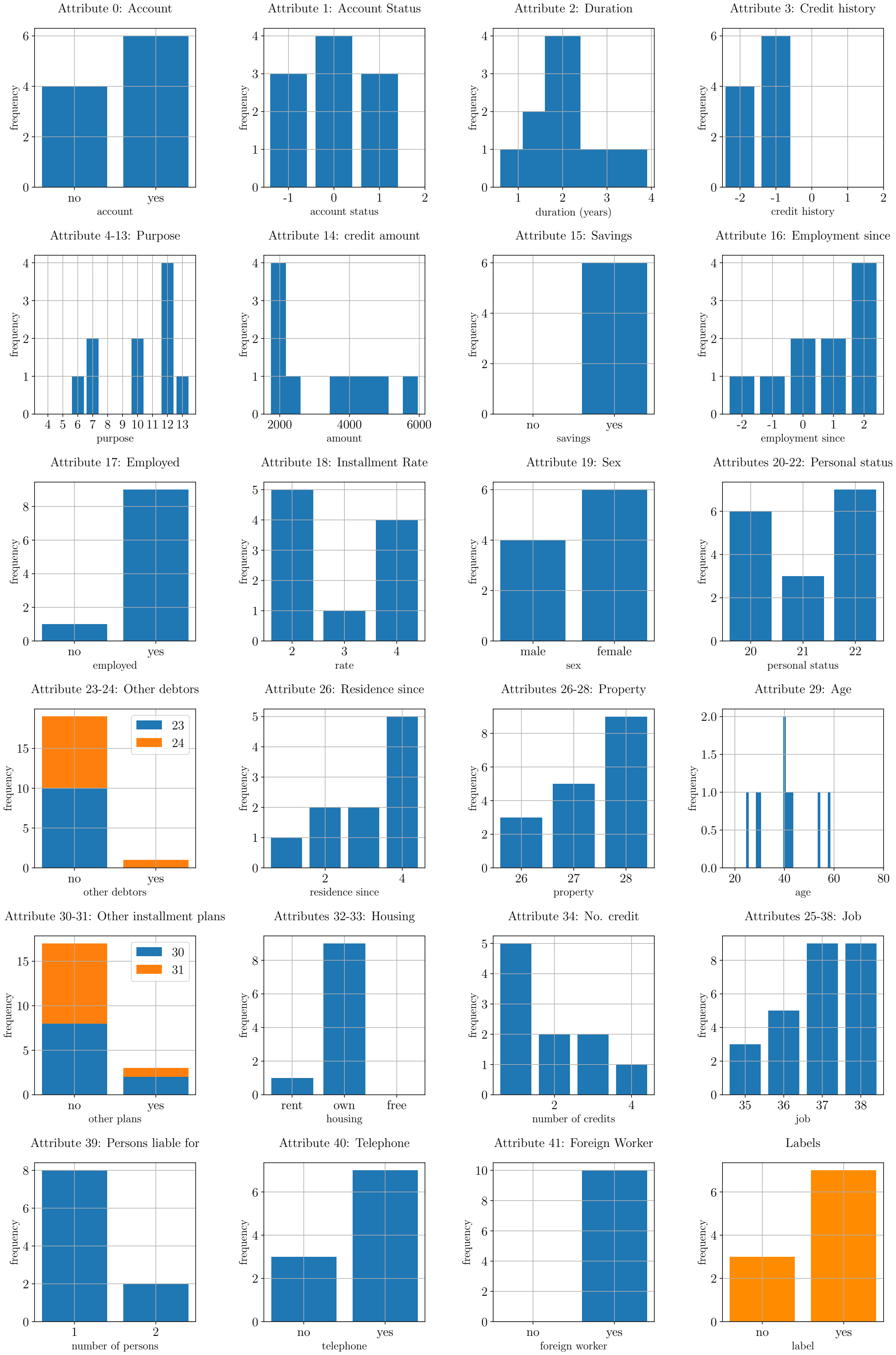


Figure 5.10: Distribution of attributes for the ten most influential data points that were removed in the first split to achieve separation.

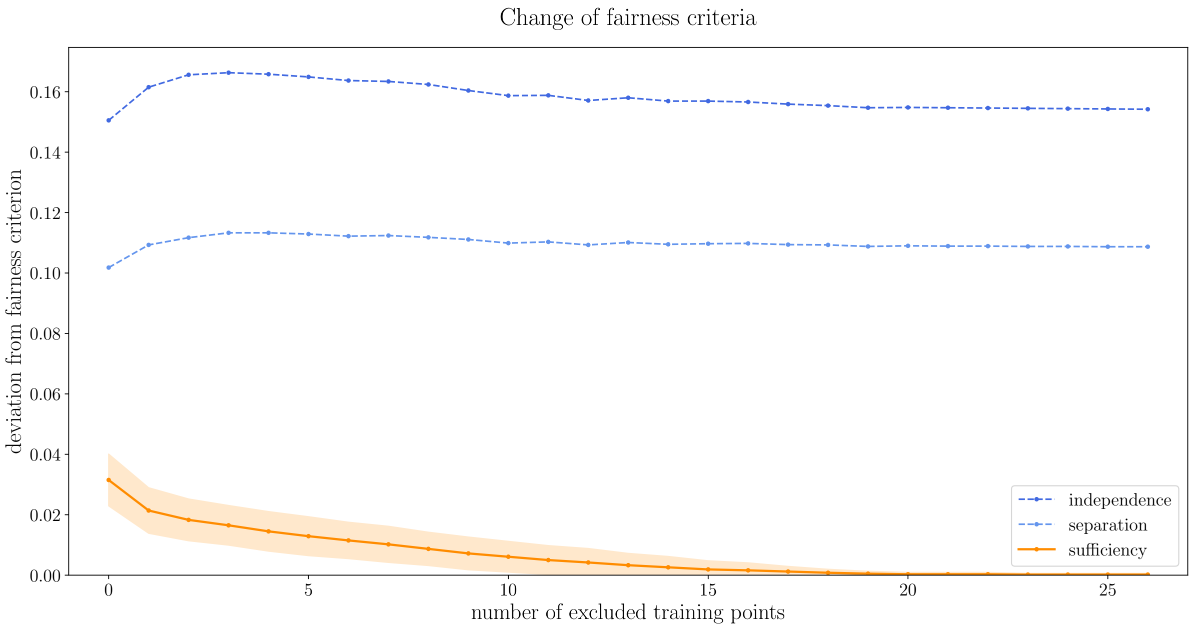


Figure 5.11: Average development of the deviation from the measure of sufficiency (orange) with number of data points removed from the training set. The standard deviation of the criterion of interest is depicted in light orange. The average development of independence and separation over ten test-test splits is plotted in blue. On average, a classifier satisfying sufficiency was obtained by removing 17 samples.

surprising, because among the group fairness criteria, sufficiency is most aligned with the goal of utility maximization (see Barocas et al., 2018).

Fairness criterion. The development of the deviation from sufficiency with the number of excluded training points is illustrated in Figure 5.11 as measured over ten different test-train splits. The orange line describes the mean deviation. The standard deviation in visualized in light orange. In contrast to the previously discussed developments, the optimization with regard to sufficiency exhibits significantly lower standard deviations. On average, 17 samples had to be excluded to get a classifier that satisfies sufficiency. This is considerably less than was necessary for independence and separation. The blue lines indicate the deviation of independence and separation for which no optimization has been performed. In line with previous observation, the deviation from independence increases slightly and remains almost constant at approximately 0.16. The same development can be observed for the deviation from separation which plateaus at approximately 0.11. These results also align with the finding from Chouldechova (2016) and Kleinberg et al. (2016) in that the three fairness criteria are not simultaneously satisfied.

Posterior weights. When looking at the changes in the posterior weights in Figure 5.12, only minor changes can be observed overall. Similar to the finding in the the context of separation, the weight of the sensitive attribute approaches zero and thereby looses influence on the final decision. Further decreases in absolute weight are visible for the credit history (attribute 3) and purpose-related weights of attributes 10 and 11. Opposite developments of attributes 4 and 5 can be found, which refer to whether the applicant intends to pay a new or old car from the loan. Due to the changed weights, it is now even more beneficial for the final decision to choose an used car.

Most influential training points. Analogous to the procedure described above, the ten training data points from the first test-train split that had the greatest influence on sufficiency were studied in more

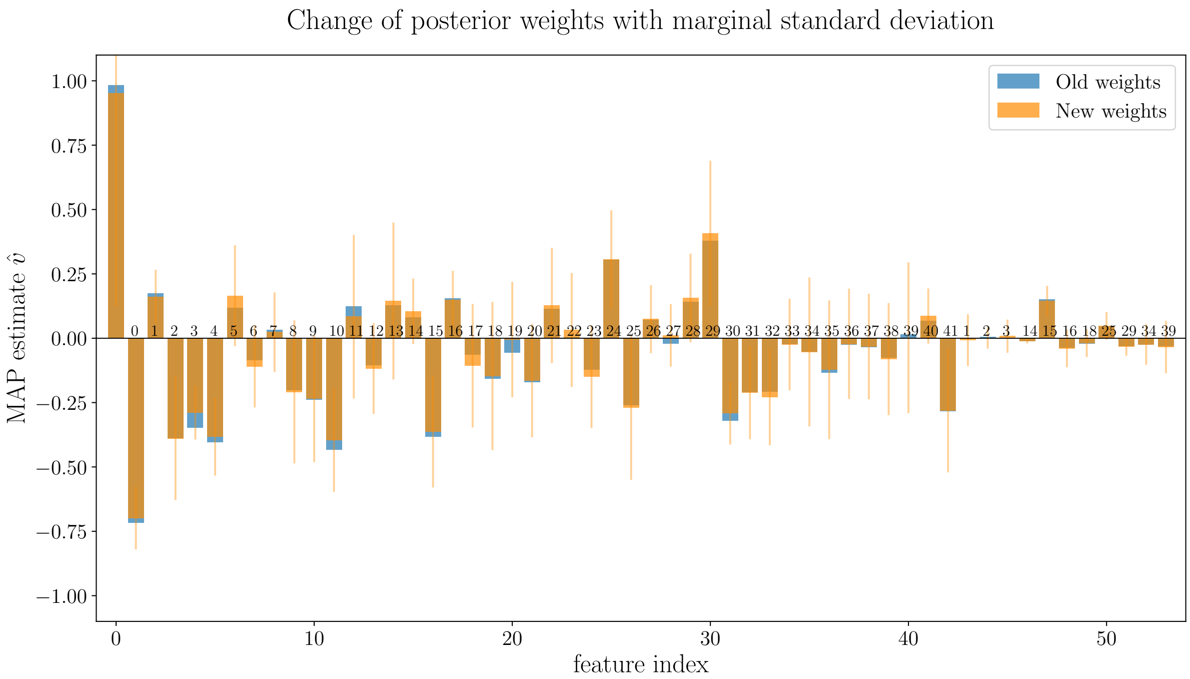


Figure 5.12: Sufficiency: Change of posterior weights after training the classifier with the modified data set as observed for the first test-train split. In blue the posterior weights of the unmodified classifier are displayed, in orange the posterior weights obtained for the "sufficient" classifier.

detail. Their properties may indicate problems in the data set, as the removal of these data points have led to fairer classification on the training data. Which points were discarded , again, strongly depends on the test-train split. Figure 5.13 shows the characteristics of the ten most influential data points that were removed in the first split to achieve separation on the training data. The corresponding weight of the sensitive attribute $w_a = -0.055$, thus being a woman counteracts a positive classification. In contrast the distribution over labels that can be observed for the entire population, the majority of samples, that were excluded from the training set to achieve sufficiency, were people who defaulted. This finding is further underlined by comparing the true and predicted labels. All excluded points would be rejected a loan when classified with the "sufficient" classifier. Since the majority indeed defaulted, the data modification reduces the number of true negatives. This leads to an decrease of the negative predictive value or increase of the false omission rate. Looking at the other attributes, the majority of were female and the purpose of the credits either car- (attributes 4 and 5) or knowledge-related (attributes 10-12). Decreasing the number of true negatives for females mainly, is consistent with the observations in 5.1 that mainly females were rejected from the unconstrained classifier.

5.2.4 Behaviour on the Test Set

In order to check to what extent the modification of the training data set contributes to fairer classifications on new, unseen data, the behaviour on the test data sets from ten test-train splits was finally examined. Regarding the four types of bias that were introduced in section 4.3, studying the behaviour in the test set corresponds to investigating potential compound bias exhibited by the trained model on the population. Analogous to the description of the unconstrained classifier, the deviation from the implemented fairness criteria was calculated. The values obtained after averaging over the test-train splits are listed in Table 5.1. For independence, a significant improvement can be observed on the test set, although the average acceptance rate for men at 0.46 is still higher than that for women (0.32). Note also

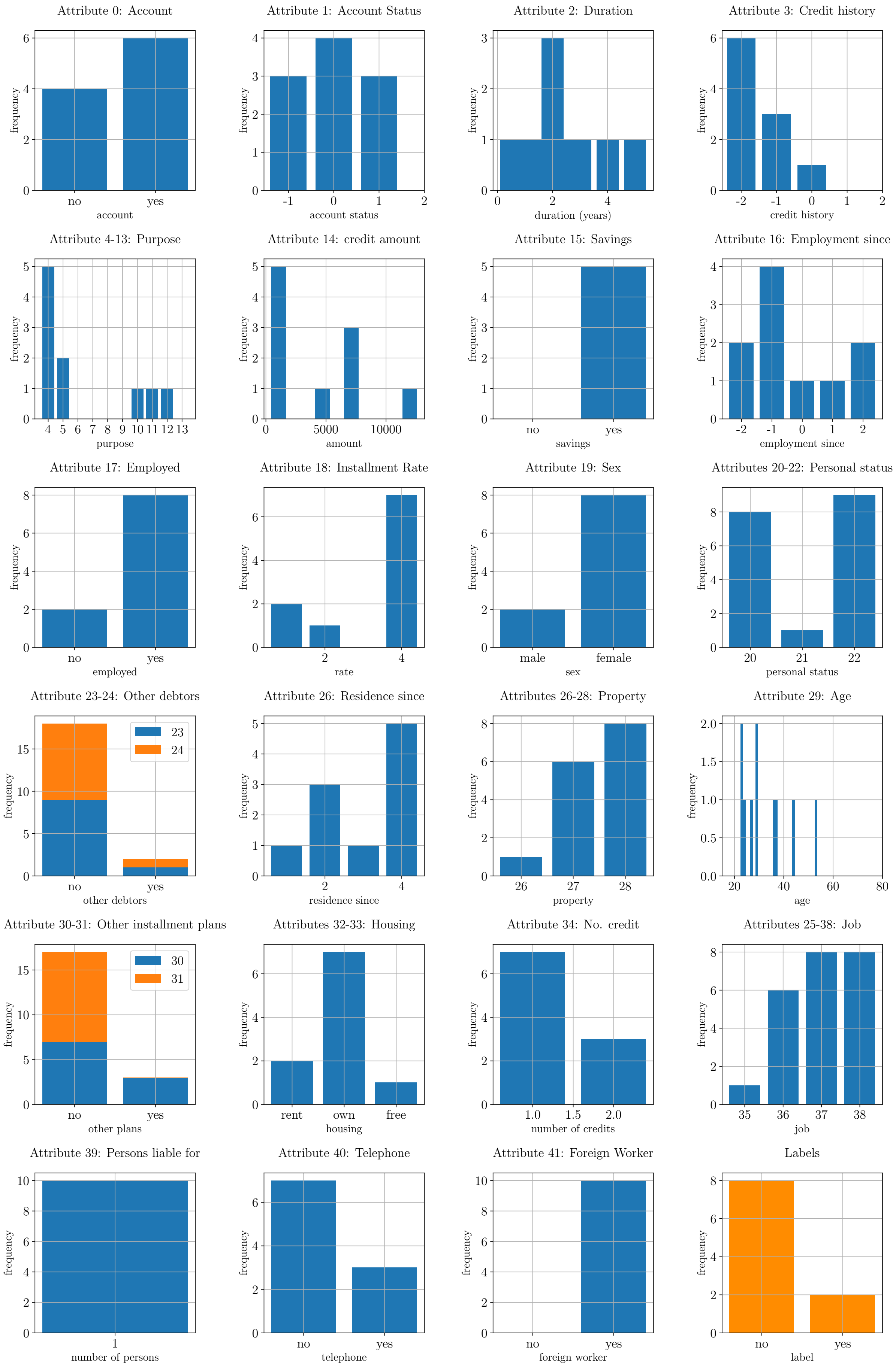


Figure 5.13: Distribution of attributes for the ten most influential data points that were removed in the first split to achieve sufficiency.

training data set modified for	**independence**	**separation**	**sufficiency**
-	0.201	0.198	0.106
independence	**0.091**	0.096	0.071
separation	0.116	**0.110**	0.060
sufficiency	0.134	0.123	**0.065**

Table 5.1: Average deviation from the fairness criteria as observed on the test sets. The first row shows the fairness criteria obtained with the unconstrained classifier. The subsequent rows list the average deviations observed after training the classifier on the maximally modified training data sets.

that the measured deviations for separation and sufficiency also improved for the classifier trained on the data set modified for independence. Similar observations can be made for the classifiers whose training data have been modified for separation and sufficiency. Interestingly, for the classifier trained to satisfy sufficiency, a decrease is also observable for the deviations from independence and separation. This is in contrast to the findings on the training data set, where diverging developments of independence and separation were found. Furthermore, higher accuracies and lower losses were observed for all optimized classifiers. For the ‘independent” classifier the highest accuracy of 64.28% was observed with an asymmetric loss of 66.44. Both values are slightly better than those reported for the unconstrained classifier (60% accuracy and an average loss of 69.95). For the “separated” classifier an accuracy of 63.08% and an asymmetric loss of 68.77 was observed. The “sufficient” classifier performed with 62.04% accuracy and an average loss of 67.11. Those results are in contrast to the idea that achieving fairness has to limit accuracy as it has often been argued.

The results show that changing the training data has an effect on the learned mapping of the classifier and thus also on the predictions for new, unseen data points. The fact that the fairness values also improve considerably on the test data is a pleasing result which demonstrates the lasting effect of the preprocessing. However, it needs to be further investigated why optimizing for sufficiency has a positive effect on the notions of independence when considering the generalization on unseen data points.

5.3 Decision Boundary Margin

The second approach to gaining insights into a classifier’s decisions is to put a margin δ around the decision boundary, within which classifications are examined with more care. In particular, it is checked

split	**weight** w_a	**maximal observed** δ	**upper bound** $\delta_{\max}$
0	-0.055	0.013	0.027
1	0.055	0.015	0.028
2	0.08	0.021	0.04
3	0.006	0.0	0.003
4	0.094	0.021	0.047
5	-0.026	0.006	0.013
6	-0.021	0.004	0.010
7	-0.064	0.017	0.032
8	0.069	0.013	0.035
9	0.016	0.003	0.008

Table 5.2: Empirical results for the weight of the sensitive attribute sex and the corresponding decision boundary as observed in ten different test-train splits

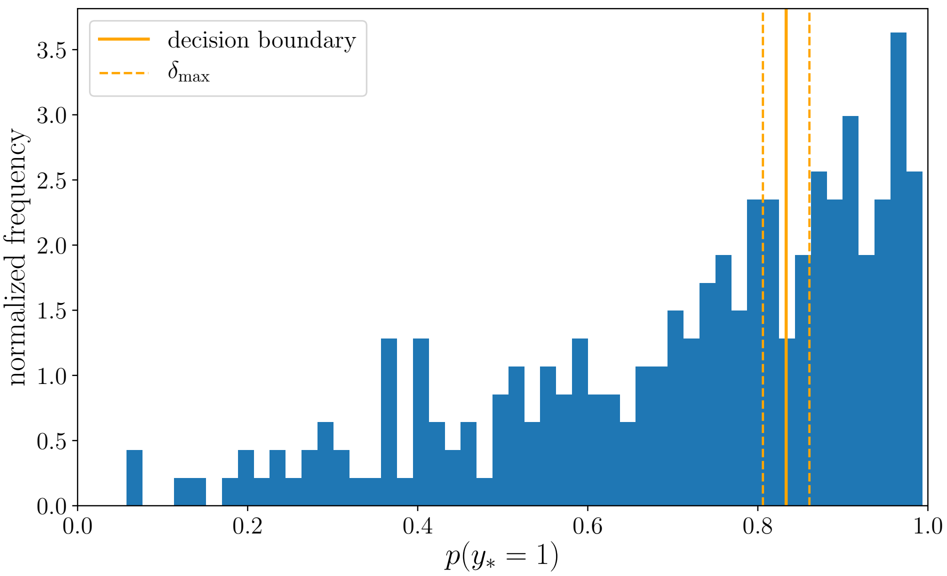


Figure 5.14: Decision boundary with maximal margin $\delta_{\max}$ on both sides of the boundary.

for all affected individuals within this margin whether flipping the sensitive attribute changes the classification. As described in 4.3.3, the exact width of the margin depends not only on the weight of the sensitive attribute w_a, but also on the individual characteristics of the other attributes, which when added together correspond to $f_{\neq a}$. In this way, a separate margin could be determined for each test person individually. However, the aim of this approach is to find a margin in which all persons fall who were potentially misclassified due to their sensitive attribute. Therefore, an upper bound $\delta_{\max}$ was derived in 4.3.3 which is visualized for the first test-train split in Figure 5.14.

Averaged over the ten test-train splits, the classification changes for approximately five people by fictitious flipping of the sex. Approximately three people per split were affected positively, i.e. they received a credit after flipping their sex. Who experiences a positive change depends on the sign of the weight of the sensitive attribute. If this is negative, women experience positive classifications through flipping. If, on the other hand, the weight of the sensitive attribute is positive, men experience a positive change in the classification through fictitious flipping of the sex. The average weight of the sensitive attribute is $\overline{w}_a = 0.015$, i.e. mainly men experienced a positive classification through fictitious flipping to women. By predicting the outcome for every person in the test set, once with the original sex and once with the flipped sex, an empirical estimate of $\delta_{\max_w}$ can be obtained. Averaged over the test-train splits the mean observed margin is $\overline{\delta} = 0.011$. Table 5.2 gives the weight of the sensitive attribute w_a, the observed maximum δ and the corresponding theoretical upper bound $\delta_{\max}$ per split.

5.4 Collective Observations

Looking at the results in their entirety, there are different possibilities to investigate the causes of unfairness more closely. While it is relatively easy in the studied example to define a model that a priori does not make any potentially disadvantageous assumptions about attributes of the population, it is difficult to distinguish neatly between learnt bias and the bias already inherently encoded in the population.

Considering the results of the data modification, the proposed procedure seems appropriate to provide further information on the data set and classifiers and thus contribute to the debate on fairness. In particular, it can be observed that the classifiers trained on the modified training data sets generalize to new data to some extent and thus achieve better values regarding the fairness criteria on the test data set. A close examination of the most influential data points gives insights into structures of the data set and thereby allows to reason about potential problems arising with the data. The second technique is a practical approach to question the decisions of the classifier and to uncover uncertainties of the model. In addition, however, it must be determined to what extent observed changes within the margin lead to new decisions. By fictive flipping of the sex, samples “are created”, which did not necessarily occur like this in the underlying population. If the person had been born with a different sex, other attributes such as income, job or the number of persons for which the person is responsible might have been different at the time of measurement, so the meaningfulness is limited in this respect.

In addition, however, it should be noted that the observed effects strongly depend on the division of the data set into training and test data sets. Varying observations can be made in different splits. The documented results are therefore subject to great uncertainty, which is further increased by the uncertainty of the model itself, reflected in the high standard deviation of the posterior weights.

6 Conclusions and Outlook

This thesis aimed to explore different sources of unfairness that can arise within machine learning pipeline and how those can further be mitigated. It was driven by the idea that fairness is an essential issue for the application of machine learning techniques in social contexts, which needs to be further explored and understood. Multiple different approaches have been proposed that present the emergence of unfairness as a complex problem for that no quick fix is available. Instead, fairness appears to depend strongly on the context and involved stakeholders. In this respect, we can clearly conclude that fairness in machine learning cannot be solved exclusively by the machine learning community, but requires interdisciplinary cooperation and the involvement of different stakeholder in order to create not only the technical, but also the social and political conditions that make fairness possible (Zarsky, 2016; Barocas and Hardt, 2017; Narayanan, 2018; Barocas et al., 2018; Suresh and Guttag, 2019; Hardt, 2019). Ways to make the decisions of algorithms more transparent and comprehensible would facilitate communication between different stakeholders. The conclusions drawn from this thesis will be structured in a similar way to the review. Beginning with data-related aspects, further considerations regarding models, predictions and broader concerns will be addressed.

Data. Data is fundamental to machine learning processes. Consequently, its quality can greatly affect the resulting models and their predictions (see also chapter 2.4). Understanding where the data that is used for training and evaluation originated from, and what methods were used for its generation is not only crucial for good modelling, but also helps to identify sources of bias. Driven by this argumentation, Gebru et al. (2018) encourage to establish the use of datasheets for data sets. Such sheets are supposed to contain detailed information about the purpose, composition, collection process of attributes and label, proposed usage, weaknesses or associated risks and further meta-data of the provided data set. Establishing such a profound documentation of data sets will facilitate the communication between data set curators and users and enforce transparency potentially entailing a positive impact on fairness-related investigations.

In order to foster the analysis of the data and study how it contributes to the performance of the model in terms of fairness, we proposed a pre-processing technique that allows to find those data points in the training data set that have the greatest impact on the fairness criterion of interest. For this work the most influential data points were excluded from the training data set and the modified was then used to create a fairer classifier. Alternatively, flipping the labels of the most influential data points was considered, which is a similar approach to what Calders et al. (2009) describe as *massaging the data set*. Here, the authors propose to rank the training points and use this ranking to select negative labels from the protected group and positive labels from the unprotected group to be flipped. Arguing that the massage is rather intrusive, as it involves explicit modifications of the samples, Calders et al. (2009) further propose to reweight the data points to achieve fairness. Kamiran and Calders (2010) extend this idea by proposing a preferential sampling strategy that allows to change the distribution of the data set without explicitly changing its attributes or labels. Their approach comes closest to the concept of data modification proposed in this work, in that they first train an unconstrained classifier and then remove or duplicate data points until the classification outcomes become free of discrimination. Since the authors involve the decision boundary in determining which data points are affected, their work is an interesting way to combine the two techniques proposed in this thesis and incorporate them into the training of the classifier. What distinguishes data modification from the work mentioned above is primarily its motivation. While we wanted to use this technique in the first place to gain a better understanding of the relationship between data and classification, the work by Calders et al. (2009); Kamiran and Calders

(2010) has focused on the generation of fair classifiers. We think, however, that the use for the exact analysis is an important aspect, which is particularly well-suited in the context of linear or logistic regression where the weights of specific features can be examined and interpreted. Our method has been successfully tested with three fundamental concepts of fairness, but it should also be applicable to other notions. It has enabled detailed analysis of the data and classifiers, enabling to identify potential sources of unfairness.

This thesis further contributed to analysis of data sets by examining the German credit data set and outlining potential issues that arise from its insufficient documentation. Remarkably, this dataset is widely used in the literature on fairness to assess the quality of proposed fairness criteria. Although it provides data for a typical scenario that concerns fairness, the data set is poorly documented and contains several inconsistencies as also discussed with Hardt (2019). As future work, I would therefore like to test the proposed methods on another, less trivial dataset which is closer to real-world conditions.

Models and Predictions. The selection of machine learning methods used for the learning task is the second key aspect, alongside data, to ensure fairness. First of all, the different approaches to ensuring fairness are often only suitable for a limited range of classifiers (Zafar et al., 2019). For this work, we used Bayesian logistic regression which allows to explicitly argue about the assumptions build into the model, incorporate uncertainties into the decision-making process and examine how the classifier changes in relation to modifications of the data set. Our idea of data modification builds on the possibility to track the influence of individual data points on the learnt classification rule. In this respect, it is therefore limited to logistic regression. The second technique proposed in this work is to find a margin around the decision boundary within which people are more vulnerable to discrimination. In the case of logistic regression, it was possible to define an upper bound on the margin such that the margin could be determined independent of the test set. This method, however, is not restricted to the application in logistic regression. Since the implementation is based solely on the test data, a empirical upper margin can be used when the influence of individual attributes is not tractable. Usually, the test set is much smaller than the training data set. In contrast to the described preprocessing method, the determination of a margin is therefore not associated with higher computational costs, which makes it suitable for use with any classifier, including neural networks.

In addition to the applicability of techniques to approach fairness, the available machine learning methods also differ in the extent to which result can be analysed. While (deep) neural networks, for example, are often referred to as black boxes because the decision making process can hardly be explained, simpler models such as logistic regression allow decisions to be traced. In connection with debates about fairness, the issue of interpretability of models and their decisions has therefore become a increasingly important issue in recent years. Transferring the ideas by Gebru et al. (2018) from data to models, Mitchell et al. (2019) encourage the use of model cards to document key characteristics of trained machine learning algorithms. The authors propose such model cards to be comprised of detailed information about the intended application and context, used training and evaluation procedures as well as benchmark evaluation in a variety of conditions including performance for cultural and demographic groups. This would significantly facilitate the transparency of models as well as the debate on fairness and possibilities for intervention by developers of algorithms, decision makers and regulators.

Naturally, companies are also subject to legal regulations when choosing a suitable method, which attributes a special importance to legislation. In particular, legislation can lay the foundations to promote transparency, explainability and accountability in algorithmic decision. In the European Union, data protection law already enables individuals to have their personal data rectified and deleted and to challenge decisions. Institutional decisions must therefore be explainable, which somewhat restricts the use of complex machine learning. Logistic regression, as opposed to neural networks, should be mentioned here as an example of a practically relevant and comprehensible statistical classification model. Although the EU already provides individuals with means to control the use of their personal data, regulations in the context of inference are not yet fully established. With the increasing use of automated decision making, however, it can also be observed that an assessment of humans occurs in areas in which it has not been carried out before (Gesellschaft für Informatik, 2018). In these areas, inferences appear often opaque in that they cannot be understood or refuted by individuals. Comprehensive regulatory guidelines are often still lacking. Wachter and Mittelstadt (2018) therefore demand that the data protection law be extended to include a *right to reasonable inference*, which enables automated decision making processes to be audited and outcomes to be scrutinized with legal effect. This would close the gap of accountability and foster the use of transparent algorithms for decision making in social contexts.

Sources of Bias. Regarding different sources of bias, we attempted to distinguish between different types of bias – model bias, population-inherent bias, learnt bias and compound bias – that can be studied through the use of Bayesian logistic regression. However, as pointed out in chapter 5, population-inherent bias is hard to quantify when using a prior with zero mean and diagonal covariance. Further work should therefore focus on investigating this type of bias, as it allows more precise allocation of responsibilities between data set and algorithm. In addition, it is important to examine in more detail how the interaction between the model, the classified persons and the state of the world is structured. In this respect, Suresh and Guttag (2019) and Silva and Kenney (2018) have conducted important preliminary work that underlines the multifaceted nature of bias. As Hardt (2019) pointed out,the dynamics emerging from the interaction between technology and society must be included to effectively address social problems with machine learning.

Outlook. In summary, the increasing relevance of fairness in machine learning brings exciting developments that promote not only fairness itself, but also the transparency of developed methods and interdisciplinary exchange. Especially in social contexts, the primary goal of algorithmic systems should always align with human values (O'Neil, 2016). Complex concepts such as fairness can neither be reduced to a single formula nor can they be regarded as a one-time or one-size problem. Referring to Narayanan (2018), there is no right definition of fairness, because different metrics of fairness matter to different stakeholders. Consequently, approaching fairness will always include trade-offs between group and individual fairness as well as fairness and utility. In this respect, application of machine learning in social systems requires careful considerations of the purpose, scope and context as well as application-specific implementations and continuous monitoring. The application of machine learning models inevitably affects the population in that it shapes the behaviour of people and thereby changes the exact state of the world those model are trying to predict. In order to respond to this, an informed interdisciplinary exchange is necessary. One option would be to exploit the knowledge about data generation from psychology and social sciences. Being further explicit about uncertainties and assumptions made for modelling as it is done in Bayesian inference, appears to be a promising approach to improve the interdisciplinary exchange, overcome miscalibrated expectations and move closer towards fairness. This is also linked to the call from Crawford (2013) to improve the self-reflection of the machine learning community, scrutinize applied methods and their influence on society and complement technical implementation with rigorous qualitative research.

Further, considerations should be made, to what extent technical implementations align with social notions of fairness (Mitchell et al., 2018) and in which context we, as society, want predictions to be made by algorithms (O'Neil, 2016). In many cases, profound empirical-grounded tools and low-tech solutions can provide promising alternatives to automated decision making, suggesting that the solution to fairness can also be not to predict, but to investigate how to intervene on social problems individually and systematically (Barabas et al., 2017; Hardt, 2019)

Bibliography

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. *ProPublica, May*, 23, 2016.

K. Bailey. Gaussian Processes for Dummies, 2016a. URL `https://katbailey.github.io/post/gaussian-processes-for-dummies/`. Accessed: 18.02.2019.

K. Bailey. From both sides now: the math of linear regression, 2016b. URL `https://katbailey.github.io/post/from-both-sides-now-the-math-of-linear-regression/`. Accessed: 18.02.2019.

C. Barabas, K. Dinakar, J. Ito, M. Virza, and J. Zittrain. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. *arXiv preprint arXiv:1712.08238*, 2017.

S. Barocas and M. Hardt. Fairness in Machine Learning, 2017. URL `https://vimeo.com/248490141`. Accessed: 08.01.2019.

S. Barocas and A. D. Selbst. Big data's disparate impact. *Cal. L. Rev.*, 104:671, 2016.

S. Barocas, K. Crawford, A. Shapiro, and H. Wallach. The problem with bias: from allocative to representational harms in machine learning. special interest group for computing. *Information and Society (SIGCIS)*, 2017.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2018. `http://www.fairmlbook.org`.

R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 0049124118782533, 2018.

T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

J. Boulamwini. How I'm fighting bias in algorithms, 2017. URL `https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms`. Accessed: 30.01.2019.

J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

T. Calders and S. Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2016.

S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

K. Crawford. The hidden biases in big data. *Harvard Business Review*, 1, 2013.

K. Crawford. The trouble with bias, 2017. URL `https://www.youtube.com/watch?v=fMym_BKWQzk`. Accessed: 22.05.2019.

E. Cresci. FaceApp apologises for 'racist' filter that lightens users' skintone. *The Guardian*, Apr 2017. URL `https://www.theguardian.com/technology/2017/apr/25/faceapp-apologises-for-racist-filter-which-lightens-users-skintone`.

D. Danks and A. J. London. Algorithmic Bias in Autonomous Systems. In *IJCAI*, pages 4691–4697, 2017.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

J. M. Eaglin. Constructing recidivism risk. *Emory LJ*, 67:59, 2017.

H. Edwards and A. Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

Executive Office of the President, C. Munoz, D. P. C. Director, M. U. C. T. O. S. O. of Science, T. Policy)), D. D. C. T. O. for Data Policy, C. D. S. P. O. of Science, and T. Policy)). *Big data: A report on algorithmic systems, opportunity, and civil rights*. Executive Office of the President, 2016.

M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

A. W. Flores, K. Bechtel, and C. T. Lowenkamp. False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. *Fed. Probation*, 80:38, 2016.

B. Friedman and H. Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347, 1996.

T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumeé III, and K. Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

Gesellschaft für Informatik. Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren. 2018. *Studien und Gutachten im Auftrag des Sachverständigenrats für Verbraucherfragen.*

A. Gessner. Probabilistic Inference and Learning. Code for lecture on Gaussian Process Classification, personal communication, 2019.

L. Grush. Google engineer apologizes after Photos app tags two black people as gorillas. *The Verge*, July 2015. URL `https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas`.

S. Hajian and J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, 25(7):1445–1459, 2013.

M. Hardt. How big data is unfair. *Understanding sources of unfairness in data driven decision making.[Medium serial on the Internet]*, 2014.

M. Hardt. Biases beyond observation, 2017. URL `https://ainowinstitute.org/symposia/videos/biases-beyond-observation.html`. Accessed: 24.04.2019.

M. Hardt. Machine Learning in Social Systems. Presentation at the lecture series 'Machine Learning' at the University of Tübingen, 2019.

M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

P. Hennig. Gaussian Processes, 2013. URL `https://www.youtube.com/watch?v=50Vgw11qn0o`. Accessed: 19.02.2019.

P. Hennig. Probabilistic Inference and Learning. Lecture Slides, Feb 2019.

M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559*, 1(2), 2016a.

M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016b.

F. Kamiran and T. Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6. IEEE, 2009.

F. Kamiran and T. Calders. Classification with no discrimination by preferential sampling. In *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*, pages 1–6. Citeseer, 2010.

T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.

M. Kay, C. Matuszek, and S. A. Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3819–3828. ACM, 2015.

N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

Z. C. Lipton and J. Steinhardt. Troubling trends in machine learning scholarship. *arXiv preprint arXiv:1807.03341*, 2018.

S. Lohr. Facial recognition is accurate, if you're a white guy. *The New York Times*, 9, 2018.

K. Lum and J. Johndrow. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*, 2016.

D. J. MacKay. Introduction to Gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, 168: 133–166, 1998.

D. J. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229. ACM, 2019.

S. Mitchell, E. Potash, and S. Barocas. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

A. Narayanan. 21 definitions of fairness and their policies, 2018. URL `https://fairmlbook.org/tutorial2.html`. Accessed: 14.01.2019.

C. O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Books, 2016.

J. Pearl, M. Glymour, and N. P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568. ACM, 2008.

J. Podesta, P. Pritzker, E. Moniz, J. Holdren, and J. Zients. Big data: seizing opportunities, preserving values (Executive Office of the President), 2014.

C. E. Rasmussen and C. K. Williams. *Gaussian processes for machine learning*. MIT Press Cambridge, MA, 2006.

G. N. Rothblum and G. Yona. Probably approximately metric-fair learning. *arXiv preprint arXiv:1803.03242*, 2018.

S. Silva and M. Kenney. Algorithms, platforms, and ethnic bias: An integrative essay. *Phylon (1960-)*, 55(1 & 2): 9–37, 2018.

H. Suresh and J. V. Guttag. A Framework for Understanding Unintended Consequences of Machine Learning. *arXiv preprint arXiv:1901.10002*, 2019.

L. Sweeney. Discrimination in online ad delivery. *arXiv preprint arXiv:1301.6822*, 2013.

S. Verma and J. Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

J. Vincent. Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge*, 24, 2016.

U. von Luxburg. Machine Learning: Algorithms and Theory. Lecture Slides, June 2018.

S. Wachter and B. Mittelstadt. A right to reasonable inferences: re-thinking data protection law in the age of Big Data and AI. *Columbia Business Law Review*, 2018.

B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, 2017.

M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017a.

M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2017b.

M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

T. Zarsky. The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, 41(1):118–132, 2016.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

I. Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.