EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MATHEMATISCH-
NATURWISSENSCHAFTLICHE
FAKULTÄT

DATA SCIENCE & ANALYTICS RESEARCH GROUP
(DR. GJERGJI KASNECI)

Master Thesis

EXPLAINABLE ML METHODS

Learning Disentangled and Diverse **Counterfactual Explanations** with Generative Models (VAEs & GANs)

About the topic

As Machine Learning (ML) techniques grow more and more popular, they have started to support and execute decisions that have primarily been done by humans. For example, learning algorithms are being used by banks to evaluate how likely individuals are to pay back their loans. Likewise, in the US legal system As the influence of these risk assessments increase, decision makers would like both to **understand how complex ML algorithms make decisions** and to enable those with negative predictive outcomes to **change their score** and receive a positive outcome in the future.

Your task

Based on our previous work (see our website), you will use deep learning techniques to improve baseline counterfactual explanation methods. You will closely collaborate with Martin Pawelczyk and other students – currently working on related projects – to build an **explanation model** which enables users to obtain **interpretable and controllable counterfactual explanations**. This thesis attempts to make first steps to close this gap in the literature and ideally results in a (NeurIPS or ICML) workshop paper.

The main part of the thesis is to develop a fair and transparent prediction framework. This includes:

- an experimental replication of basic **counterfactual explanation** algorithms and their evaluation with respect to common metrics.
- an implementation of a new **counterfactual explanation** algorithm (provided by us) and its experimental evaluation

Requirements

Ideally you are motivated to work on **explainable ML methods**. Moreover, you ideally, but, not necessarily,

- have strong programming skills in Python;
- are familiar with TensorFlow or PyTorch.
- Have a sound background in machine learning (and statistics).

Contact

If you are interested, do not hesitate to approach us:

Martin Pawelczyk
Sand 14, C216
martin.pawelczyk(at)uni-tuebingen.de

Dr. Gjergji Kasneci
Sand 14, C221
gjergji.kasneci(at)uni-tuebingen.de

- October 2020