Diverse Counterfactual Explanations

An Introduction into the DiCE Package

Daniel Saggau

LMU Munich

2022

What is a DiCE?

- Many XAI methods work with feature importance
- Problem: they don't provide guidance on how to take action
- Example: Credit gets rejected and Income was most important factor
- One may be Interested in what one can do to change the outcome
- Deal with questions like: what would have happened if?
- Provides varying explanations for end-user

Requirements: Training Data, CF Generation Method, Model

What do we need to do?

Process

1. Convert Training data to DiCE data
2. Instantiate pre-trained model (Black Box Model)
3. Combine Model and Data
4. Specify CF Method (model agnostic method only)
5. Specify Constraints

Model Agnostic Method vs. Gradient Based Methods

Model agnostic: Optimization via sampling nearby points to an input + optimizing loss based on proximity (or sparsity, diversity + feasibility)

- *random:* independent random sampling (baseline)
- *kdtree:* k-distance tree (feasibility)
- *genetic algorithm:* (diversity + fast convergence)

Gradient Based: works with differentiable models based on an explicit loss minimization(proximity, diversity and feasibility)

- Classic DL Models (explicit loss minimization)
- VAE Models (Pytorch only)

Set-up of DiCE

Balancing multiple objectives to create cfs

$$\text{Objectives} = \text{Actionability} + \text{Feasability} + \text{Diversity}$$

$$CF = argmin \underbrace{\frac{1}{k}\sum_{t=1}^{k} yloss(f(c_i), y)}_{\text{loss to get desirable outcome}} + \underbrace{\frac{\lambda_1}{k}\sum_{t=1}^{k} dist(c_i, x)}_{\text{distance to original input}} - \underbrace{\lambda_2 dpp_{-}\ \text{diversity}\,(c_1, \ldots, c_k)}_{\text{Loss to provide diverse explanations}}$$

- loss balancing hyperparameters:λ_1, λ_2
- Number of cfs: k

Dataset For Coding Illustration

- Adult income Dataset (Reference Code in Official Documentation)
- Data on demographics, census data and educational background
- Task: classify whether the income of an adult is above $50.000
- Second Dataset to illustrate Regression: California Housing Dataset (new example)
- Task: Estimate the median house value for California districts

Illustrated Features of the DiCE Package

- Model-agnostic and gradient based CF explanation methods
- Estimating Local and Global Feature Importance
- Regression/Classification based CFs
- Modifying Feature weights (rigid features) and Hyper parameter-weights (proximity, diversity)