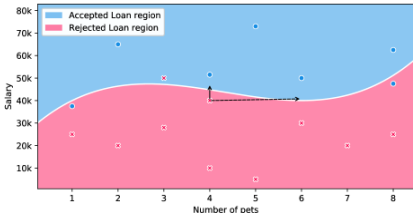
Interpretable Machine Learning

Adversarial Examples and Counterfactual Explanations



Learning goals

- Compare adversarial examples to counterfactual explanations
- See an example where both coincident

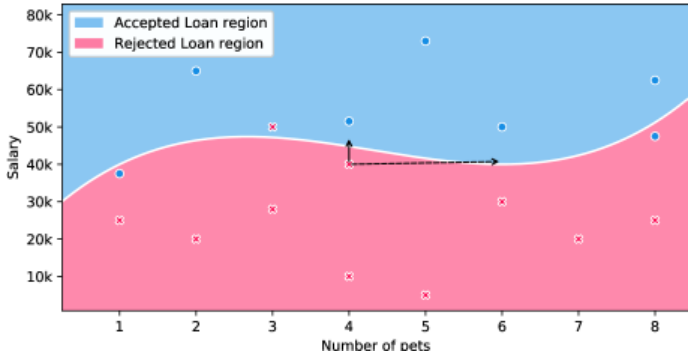
ADE AND COUNTERFACTUAL EXPLANATIONS

It seems as if ADEs and counterfactual explanations (CEs) are defined similarly. Both ADEs and CEs describe inputs close to a given input $\mathbf{x}$ that gets a different assignment. What are their differences?

- Counterfactuals do not have to be misclassified.
- Different notions of distance $\|\cdot\|$ are applied, e.g., $p_{2,\infty}$-norm for ADEs or $p_{0,1}$-norm for CEs.
- Informal difference I: ADEs are mostly considered for high-dimensional data, while CEs are mostly considered in the context of low-dimensional data.
- Informal difference II: ADEs hide changes while CEs highlight them.

SHARED EXAMPLE ▸ Ballet (2019)

- ‘‘If you had two more pets, your loan application would have been granted" is an example of both ADEs and CEs.



Decision boundary of a classifier deciding loan applications. ADE via “number of pets”