NTNU | Norwegian University of Science and Technology

PROBABILISTIC TABULAR DIFFUSION FOR EXPLAINABLE AI

Alexander J Ohrt

Supervisor: Kjersti Aas

04 May

Overview

- Explainable AI (XAI)
- Counterfactual Explanations (CEs)
- Generative Modelling
- Diffusion Models
- Selected Results from Thesis Work

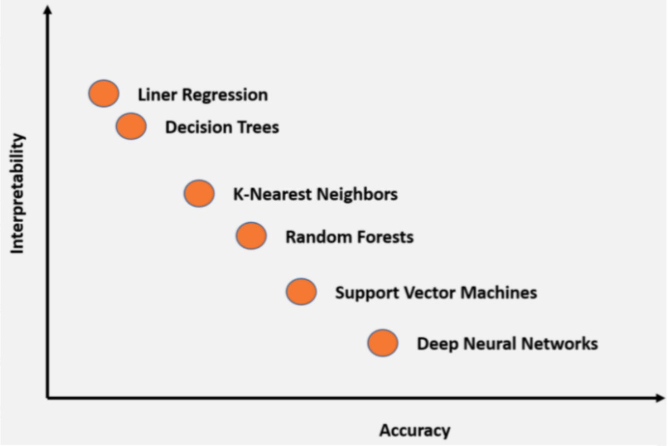
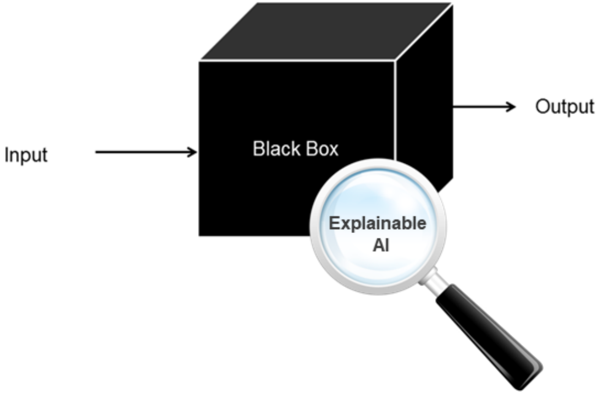


Figure: Qualitative illustration of the tradeoff between accuracy and interpretability in a set of common ML models.

XAI — Why?

- Open the black box
- Understand reasoning
- Debug systems
- Reduce risk
- Guarantee positive effects
- Legislative Reasons, e.g. GDPR, AI Act



XAI — How?

"Meta-Models": Build AI/ML models to explain other black box AI/ML models (post-hoc).

Figure: Comic borrowed from `https://xkcd.com/1838/`

XAI — How?

Examples of Methods:

- Shapley Values and SHAP
- LIME
- Counterfactual Explanations (CEs)

CEs — Example

Table 2.1: The customer that solicits a mortgage. The attributes of the customer are *age, sex, nationality, salary (yearly), work sector, marital status, years as customer at the bank* and *postal code (ZIP)*.

Age	Sex	Nat.	Sal.	Work Sect.	Mar. Stat.	Cust. Years.	ZIP
22	M	Norway	350K	Public	Single	2	7051

Table 2.2: A possible successfully generated counterfactual.

Age	Sex	Nat.	Sal.	Work Sect.	Mar. Stat.	Cust. Years.	ZIP
22	M	Norway	420K	Private	Single	2	7051

Table 2.3: A possible unsuccessfully generated counterfactual.

Age	Sex	Nat.	Sal.	Work Sect.	Mar. Stat.	Cust. Years.	ZIP
200	F	Sweden	2200K	Private	Single	10	7051

CEs — General Criteria

1. *on-manifold*: The counterfactual should lie on the data-manifold ("resemble the training data").
2. *actionable*: The counterfactual should not change fixed features ("not very informative to change *age* e.g.").
3. *valid*: The counterfactual should flip the prediction ("from mortgage denied to mortgage granted").
4. *low cost*: The counterfactual should be as similar as possible to input ("don't change unnecessary characteristics of client").

CEs — Algorithms

- Optimization-based

$$\min_{\boldsymbol{x}'}\{d_1(f(\boldsymbol{x}'), y') + \lambda d_2(\boldsymbol{x}, \boldsymbol{x}')\} \tag{1}$$

- On-Manifold

MCCE — Our Inspiration for Calculating CEs

MCCE: **M**onte **C**arlo sampling of realistic **C**ounterfactual **E**xplanations

Redelmeier, A., Jullum, M., Aas, K., & Løland, A. (2021). MCCE: Monte Carlo sampling of realistic counterfactual explanations. arXiv preprint arXiv:2111.09790.

MCCE — Our Inspiration for Calculating CEs

Steps:

1. Model underlying data distribution
2. Sample from data distribution
3. Post-process the samples, i.e. filter out counterfactuals

MCCE — Our Inspiration for Calculating CEs

Steps:

1. **Model underlying data distribution**
2. **Sample from data distribution**
3. Post-process the samples, i.e. filter out counterfactuals

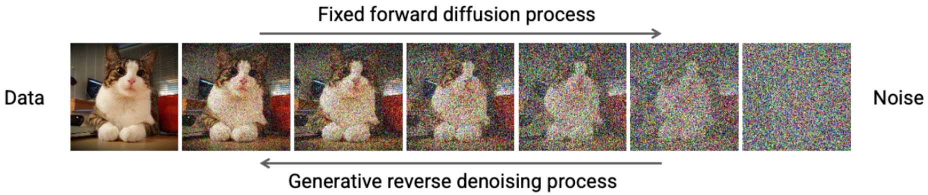
Generative Modelling with Diffusion Models

A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.

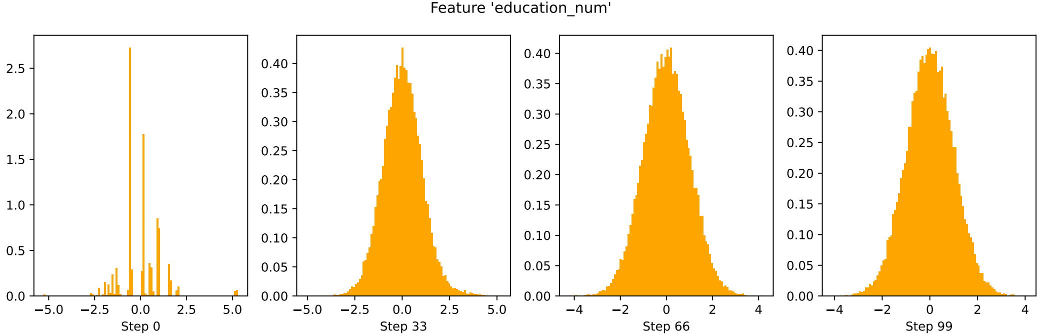
A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.

Figure: Examples of photorealistic images created with Imagen. Borrowed from Imagen paper (May 2022).

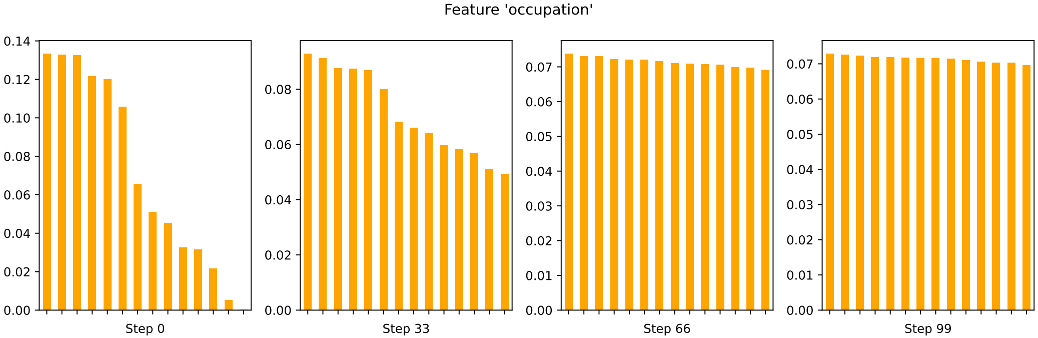
Diffusion Models — How?



Diffusion Models — How?



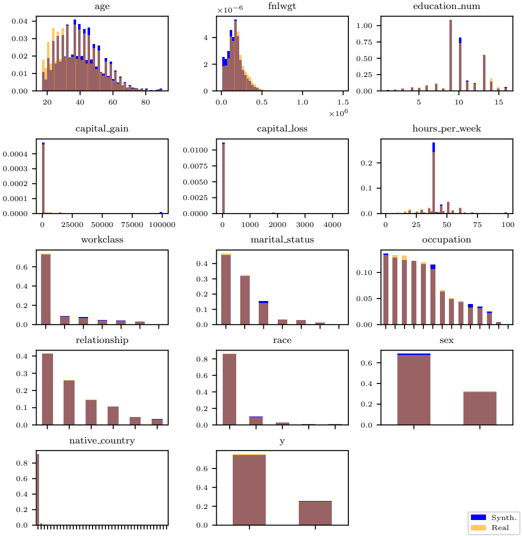
Diffusion Models — How?



Thesis Objectives

- Develop an accessible introduction to diffusion models.
- Implement a model specifically for *tabular data*.
- Perform experiments.

Some Results



Some Results

Table: Factual h and generated counterfactual for CatBoost classifier.

	h	Diffusion
`age`	25	25
`fnlwgt`	188767	218210
`ed_num`	12	13
`cap_gain`	0	0
`cap_loss`	0	0
`h_p_w`	45	52
`workcl.`	Private	Private
`mar_stat`	Never-married	Married
`occup.`	Exec.	Exec.
`rel.`	Not-family	Not-family
`race`	White	White
`sex`	Male	Male
`country`	US	US
`Prediction`	0	1