Lecture 12

Interpretability, Explainability, and Fairness

UNIVERSITETET
I OSLO

Interpretability

- High model transparency
- Understand exactly **why** and **how** the model is generating predictions
- Need to observe the **inner mechanics** of the AI/ML method
- Interpreting the model's **weights and features** to determine the given output.

Explainability

- Take an ML model and explain the **behavior in human terms**
- With **complex models** one cannot fully understand how and why the inner mechanics impact the prediction
- Use model agnostic methods (for example, partial dependence plots, SHapley Additive exPlanations (SHAP))

UNIVERSITETET I OSLO

Why is interpretability and explainability important?

- Trust from business
- Avoiding errors
- Fairness
- Compliance
- …

UNIVERSITETET I OSLO

Interpretability

- Usually ‘simple’ models
  - Linear models (log. regression, linear regression)
  - Trees
  - (KNN)
- The more complex a model the harder it is to understand inner workings
- Challenging examples: Boosted models, random forests, NN

UNIVERSITETET
I OSLO

Partial dependence

- Shows marginal effect of one or two features on predicted outcome
- Shows if dependence is linear, monotonic, or more complex
- Can help model trouble-shooting
- Formally defined as

$$\hat{f}_S(x_S) = \mathrm{E}_{X_C}\left[\hat{f}(x_s, X_C)\right]$$

- Can be estimated from data (Monte Carlo)

UNIVERSITETET I OSLO

Shapley Values (SHAP)

- Formal definition

$$\phi_j(v) = \sum_{S \subset \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!(p - |S| - 1)!}{p!} \{v(S \cup \{j\}) - v(S)\}$$

- Computationally expensive
- Efficient methods exists
  - Some exact for specific model classes
  - Some approximate
- Only explanation method with solid theory

UNIVERSITETET
I OSLO

Fairness - how to measure?

- Regression case

$$fairness = \left| \frac{1}{|Z_1|} \sum_{i \in Z_1} \hat{y}_i - \frac{1}{|Z_2|} \sum_{i \in Z_2} \hat{y}_i \right|$$

- Classification case

$$f_{\text{equal outcome}} = \min \left(\frac{P(\hat{y}=1|z=1)}{P(\hat{y}=1|z=0)}, \frac{P(\hat{y}=0|z=1)}{P(\hat{y}=0|z=0)} \right)$$

$$f_{\text{eq. opp}} = \min \left(\frac{P(\hat{y}=1|z=1, y=1)}{P(\hat{y}=1|z=0, y=1)}, \frac{P(\hat{y}=0|z=1, y=0)}{P(\hat{y}=0|z=0, y=0)} \right)$$

UNIVERSITETET
I OSLO

Fairness - what to do about it?

- Pre-processing
  - Remove sensitive variable (still implicit bias possible)
  - Project out sensitive variable (Gram-Schmidt, similar to PCA)
- At training time: Model constraints
  - $\underset{\theta}{\operatorname{argmin}} L(y, f_\theta(x)) + \operatorname{fairness}(f_\theta)$
- At prediction time
  - E.g. through different prediction thresholds in classification

UNIVERSITETET
I OSLO