Explainability

Machine Learning for Behavioral Data

May 25, 2023

Today's Topic

Week	Lecture/Lab
8	Spring Break
9	Time Series Prediction
10	Unsupervised Learning
11	Unsupervised Learning
12	Fairness
13	**Explainability**
14	Project Presentations
15	Whit Monday

- **What is fairness?**
- **Fairness metrics**
- **Interpreting neural networks**

Getting ready for today's lecture...

- **If not done yet**: clone the repository containing the Jupyter notebook and data for today's lecture into your Noto workspace.
- SpeakUp room for today's lecture:

https://go.epfl.ch/speakup-mlbd

Short quiz about the past...

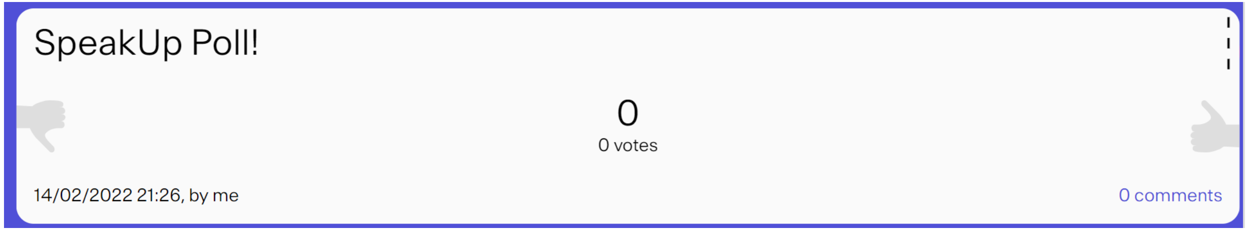
In K-Means Clustering, how should you initialize the cluster centroids?

a) Once, randomly

b) Once, uniformly

c) Visualizing the data and picking appropriate starting points

d) Multiple times randomly and minimizing distortion



Short quiz about the past...

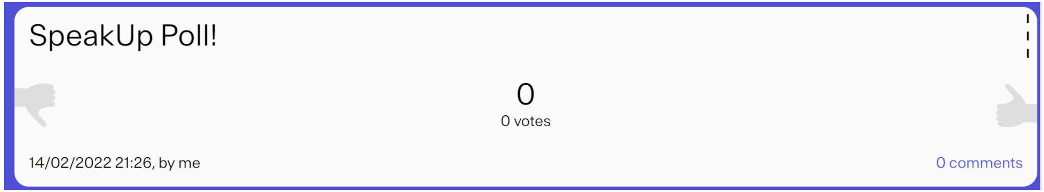
When performing clustering on text data, which distance/similarity metric is appropriate?

a) Silhouette Score

b) Jaccard Similarity

c) Cosine Similarity

d) Euclidean Distance



https://go.epfl.ch/speakup-mlbd

Short quiz about the past…

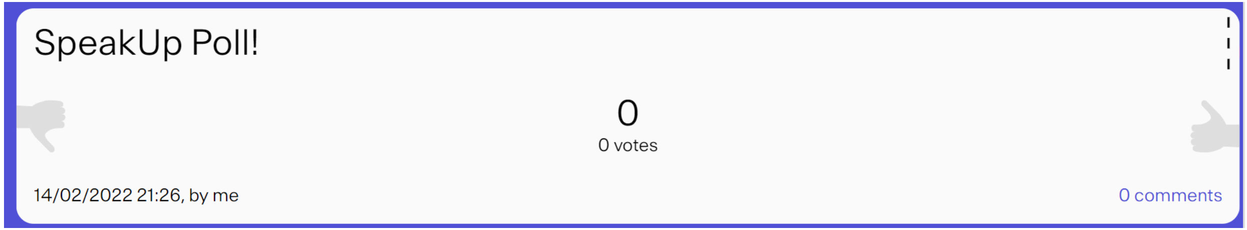
If you use accuracy instead of balanced accuracy for a binary classification task (on an imbalanced data set), this is an example of:

a) Historic Bias

b) Evaluation Bias

c) Measurement Bias

d) Aggregation Bias



https://go.epfl.ch/speakup-mlbd

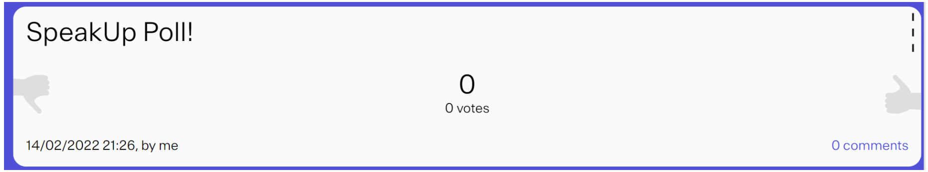
Short quiz about the past…

You are building a model for whether someone will pass a class based on their MOOC clickstream. You are concerned about whether your model's predictions of passing and predictions of failing are equally accurate across demographic groups. Which metric do you use?

a) equalized odds

b) demographic parity

c) predictive (value) parity



https://go.epfl.ch/speakup-mlbd

Agenda

1) **Introduction to Explainability**
   - Taxonomy of interpretability methods
   - Deep Dive: PDP
   - Deep Dive: LIME
2) Course Wrap-Up (project, exam)

Learning Objectives

You should be able to:

- Describe and categorize the explainability methods discussed in class
- Explain their strength and weaknesses
- Interpret their outputs
- Apply the methods (using the APIs) to predictions of a model and discuss the results

Interpretability

Interpretability is the degree to which a human can understand the cause of a decision.

Interpretability

Interpretability is the degree to which a human can understand the cause of a decision.

The higher the interpretability of a machine learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made

Interpretability in Education

Taxonomy of Interpretability Methods

Taxonomy of Interpretability Methods

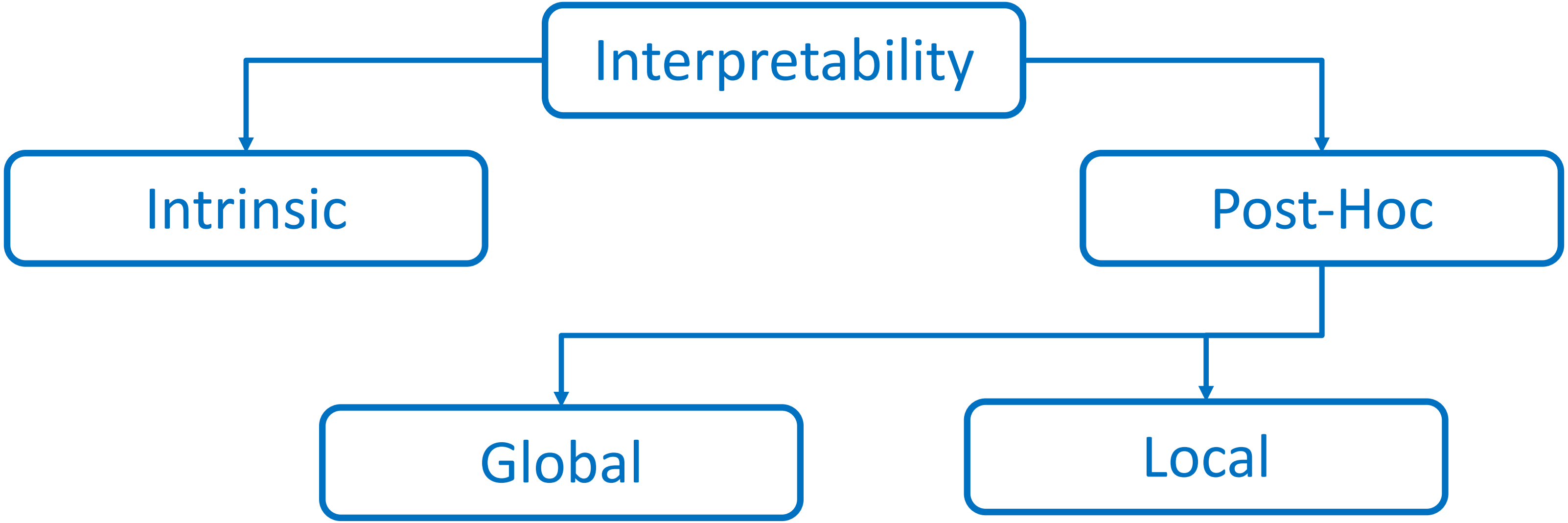
Interpretability

Intrinsic

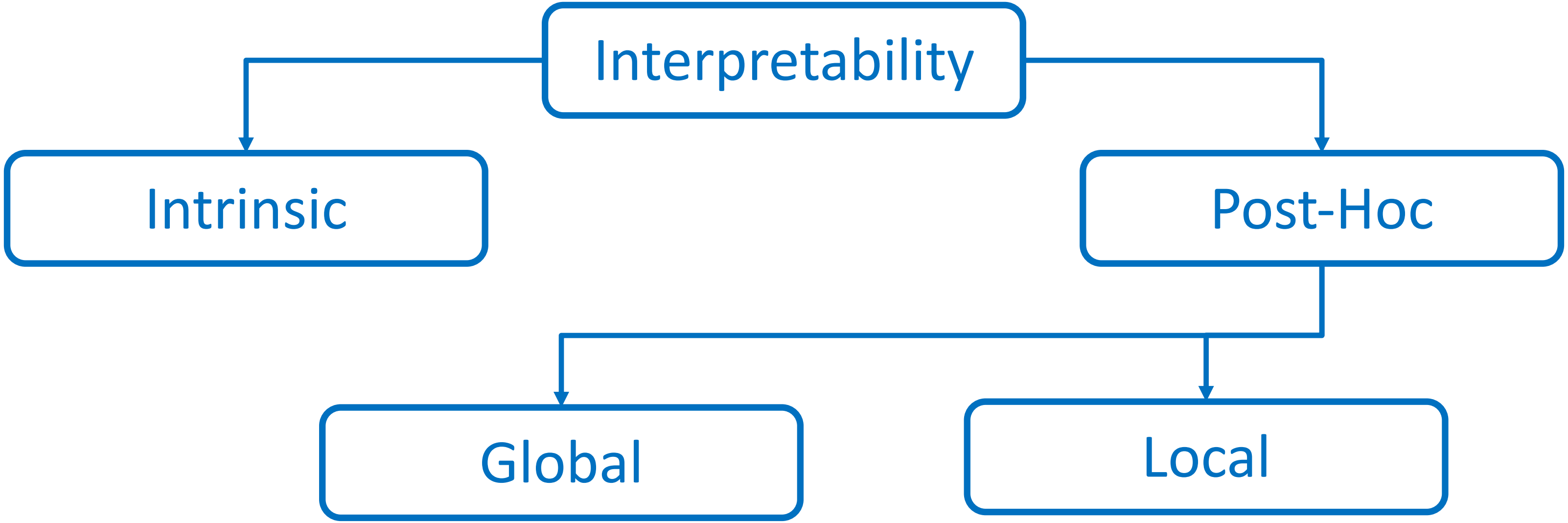
Post-Hoc

- Linear Regression
- Generalized Linear Models (e.g., logistic regression)
- Decision Trees
- (k-Nearest Neighbors)

Taxonomy of Interpretability Methods



Taxonomy of Interpretability Methods



- **Partial Dependency Plot (PDP)**
- Accumulated Local Effects (ALE) Plots
- Global Surrogate

- **Local Surrogate (LIME)**
- Counterfactual Explanations
- SHAP (SHapley Additive exPlanations)

Interpretable Machine Learning Book

Global Method: Partial Dependency Plot (PDP)

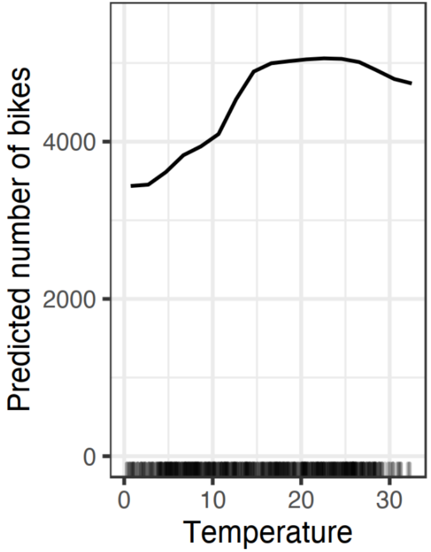
- PDP is model-agnostic
- PDP show the marginal effects a subset of features have on the predicted outcome of a model
- The subset of features usually consists of one feature (resulting in a 2D-Plot) or two features (resulting in a 3D-Plot)

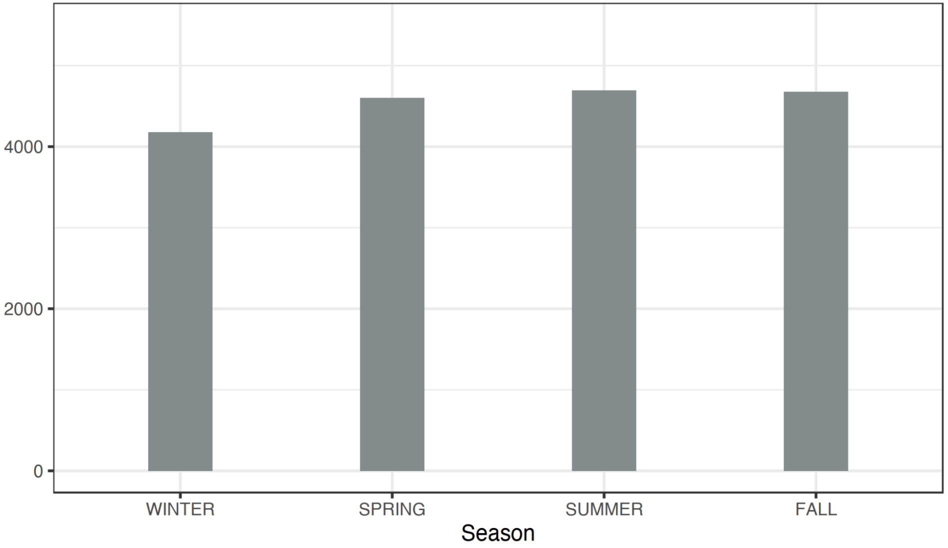
Example – Bike Rental Shop

- Y denotes the number of bikes that will be rented on a given day
- Features (X): season, work day, temperature, humidity, …
- Given: model f such that $y = f(x)$

Example – Bike Rental Shop

- Y denotes the number of bikes that will be rented on a given day
- Features (X): season, work day, temperature, humidity, …
- Given: model f such that $y = f(x)$





Bike Rental Data Set

Partial Function - Regression

$$\boxed{\hat{f}_S(x_S) = E_{X_C}\left[\hat{f}_S(x_S, X_C)\right] = \int_{X_C} \hat{f}_S(x_S, X_C)}$$

Partial Function - Regression

$$\hat{f}_S(x_S) = E_{X_C}\left[\hat{f}_S(x_S, X_C)\right] = \int_{X_C} \hat{f}_S(x_S, X_C)$$

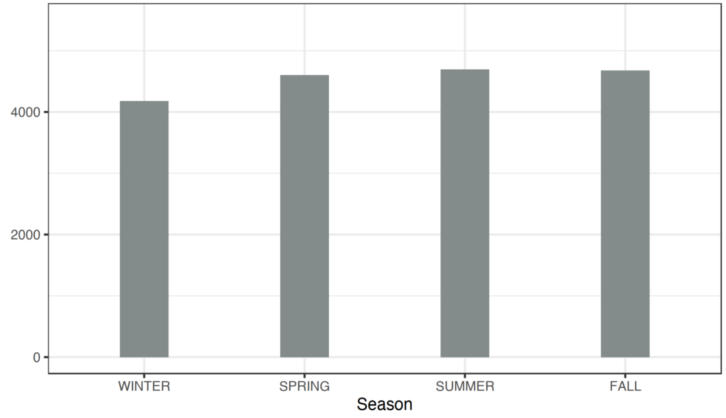
$$\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_S\left(x_S, x_C^{(i)}\right)$$

Partial Function - Regression

$$\widehat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}_S\left(x_S, x_C^{(i)}\right)$$

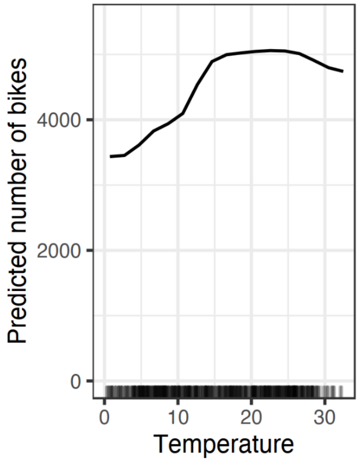
Partial Function - Regression

$$\widehat{f_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n} \widehat{f_S}\left(x_S, x_C^{(i)}\right)$$



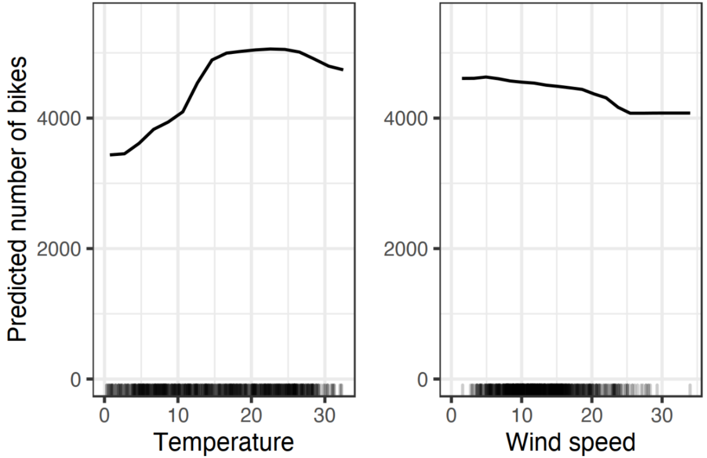
Partial Function - Regression

$$\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_S\left(x_S, x_C^{(i)}\right)$$



Partial Function - Regression

$$\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_S\left(x_S, x_C^{(i)}\right)$$



Partial Function - Classification

$$\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_S\left(x_S, x_C^{(i)}\right)$$

- If classifier outputs a probability, the PDP displays the probability for a certain class given different values for feature(s) in S
- Dealing with multiple classes: draw one line or plot per class

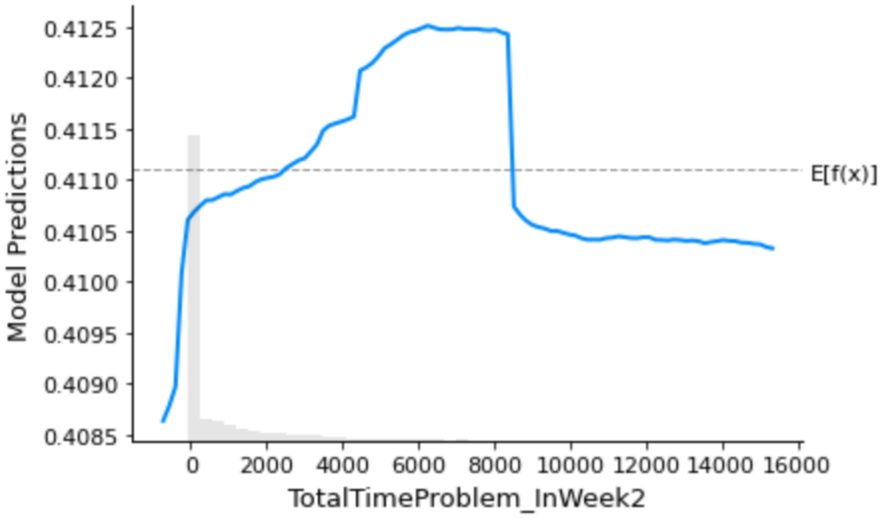
PDP – Strength & Weaknesses

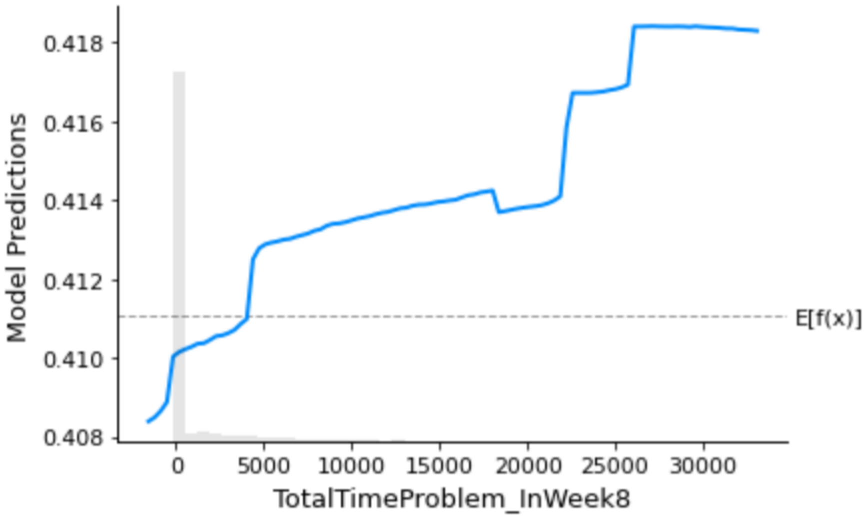
- \+ Model-agnostic
- \+ Computation is intuitive, interpretation is clear
- \+ Easy to implement
- \+ Causal interpretation
- – Maximum number of features in a PDP is two
- – Assumption of independence
- – Some PDP do not show feature distribution

PDP – Your Turn

- Participants: 8679 students of a of an EPFL MOOC with a duration of 10 weeks
- We have trained a classifier to predict whether a student will pass or fail the course based on their clickstream data
- Your Task:
    1. Investigate the PDPs for *TotalTimeProblem* in week 2 and week 8
    2. Discuss: how does this feature influence predictions? Is there a difference between week 2 and week 8? What about the distribution of feature values?

PDP Example – EPFL MOOC





Local interpretable model-agnostic explanations (LIME)

- Idea: use a local surrogate model (interpretable) to explain individual predictions of a black-box model

$$\boxed{\text{explanation}(x) = \underset{g \epsilon G}{\operatorname{argmin}}\, L(f, g, \pi_x) + \Omega(g)}$$

LIME - Recipe

$$\text{explanation}(x) = \underset{g \epsilon G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g)$$

1. Select your instance (sample) of interest

LIME - Recipe

$$\text{explanation}(x) = \underset{g \epsilon G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g)$$

2. Perturb your data set: generate new samples that are variations of the selected sample

LIME - Recipe

$$\text{explanation}(x) = \underset{g \epsilon G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g)$$

3. Get the black-box model predictions for the new samples

LIME - Recipe

$$\text{explanation}(x) = \underset{g \epsilon G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g)$$

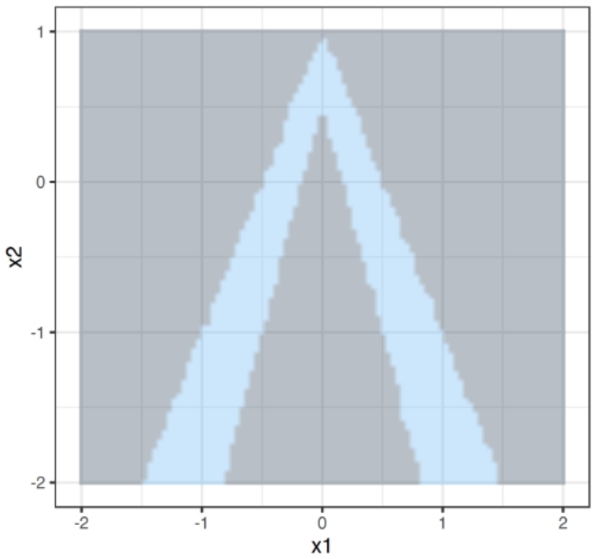
4. Train a weighted, interpretable model on the data set with variations

LIME - Recipe

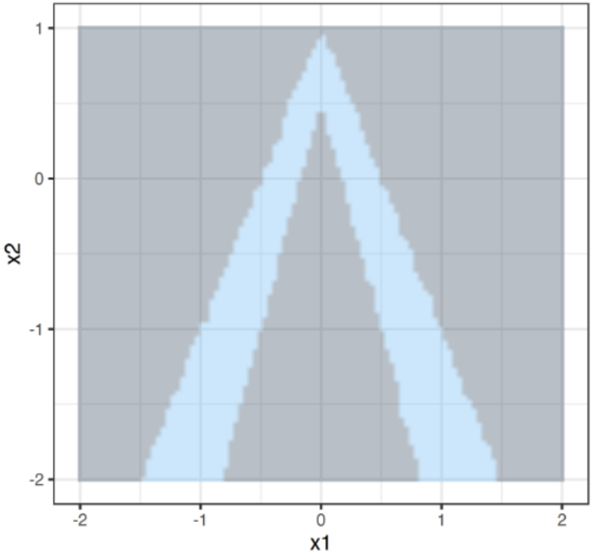
$$\text{explanation}(x) = \underset{g \in G}{\operatorname{argmin}}\, L(f, g, \pi_x) + \Omega(g)$$

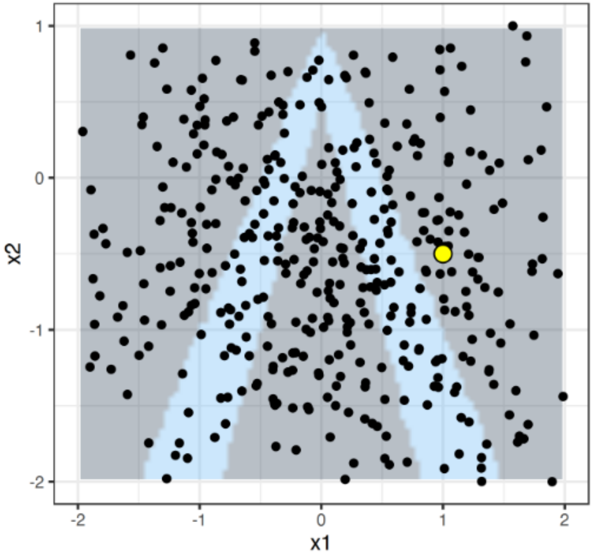
5. Explain the prediction by interpreting the local model

LIME – Perturbation of Sample

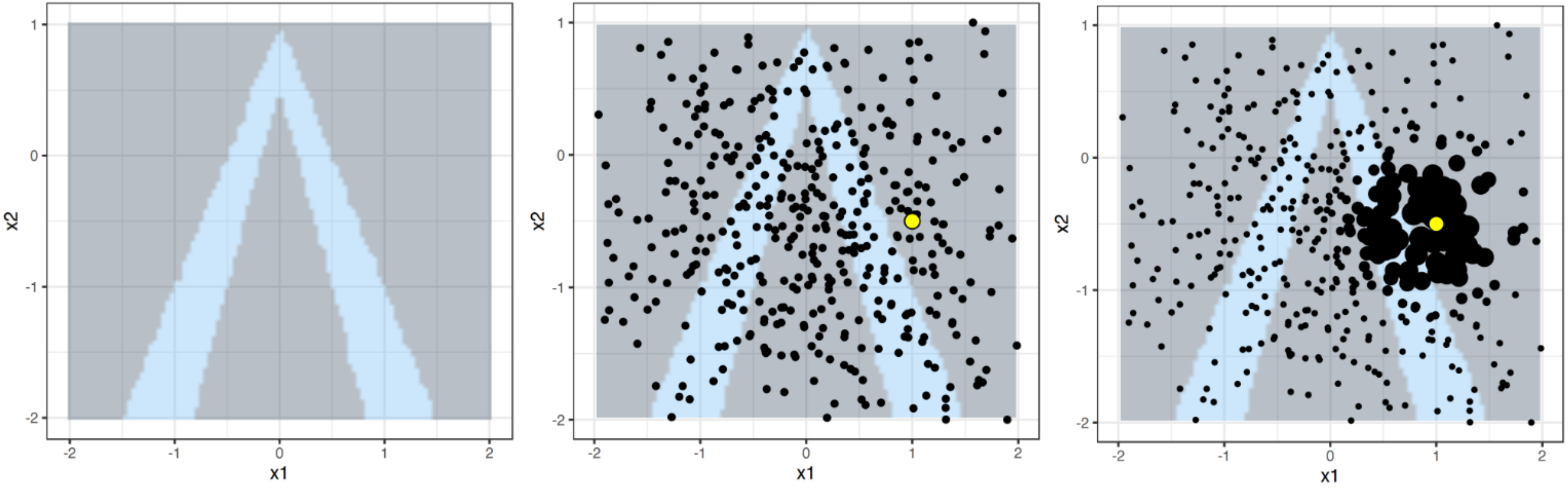


LIME – Perturbation of Sample





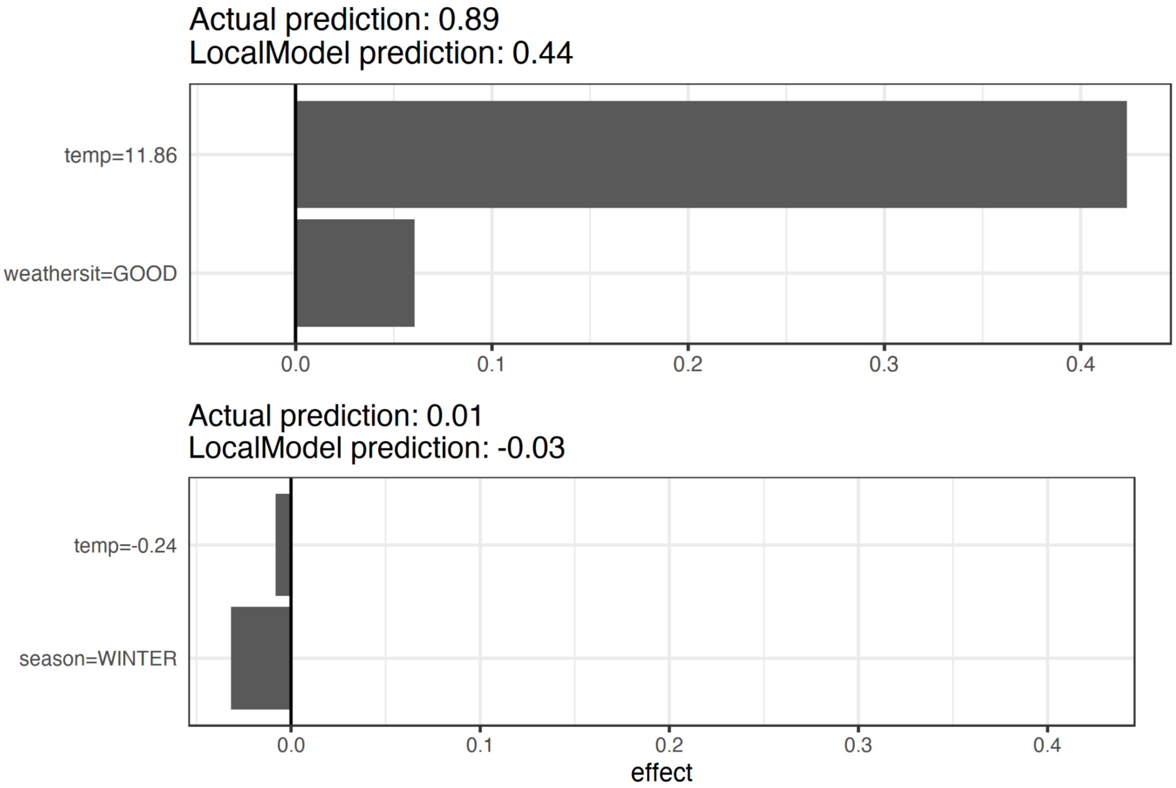
LIME – Perturbation of Sample



Example – Bike Rental Shop

- Y is binary and indicates, whether the number of bikes rented on a given day will be **above average** $(y = 1)$
- Features (X): season, work day, temperature, humidity, …
- Given: model f such that $y = f(x)$

Example – Bike Rental Shop



Bike Rental Data Set

LIME – Strengths and Weaknesses

+ Model-agnostic (we can replace the underlying model and still use the same surrogate model)
+ When using for example Lasso regression, explanations are short (= selective)
+ Benefit from literature on training and interpreting interpretable models
+ Fidelity measure gives us an idea of reliability
– Definition of local neighborhood unsolved problem
– Sampling ignores correlation between features (-> unlikely data points)
– Instability of explanations

LIME – Your Turn

- Your Task:
  1. Run LIME on two instances of your choice
  2. Share the plots for the two instances with us as well as your observations (Are the same features important for both instances? Can you interpret the feature effects?)

Summary

- Interpretability is important (not only for education)
- We can use intrinsic interpretable models or post-hoc methods to get interpretable predictions
- Methods can be categorized into global and local
- PDP is easy to interpret, but has an independence assumption and is limited to a low number of features
- LIME leads to short explanations, but also ignores correlation between features and might lead to instable explanations

Agenda

1) Introduction to Explainability
   – Taxonomy of interpretability method
   – Deep Dive: PDP
   – Deep Dive: LIME
2) **Course Wrap-Up (project, exam)**

In-Depth Evaluation

- The school of IC performs an in-depth evaluation of each course
- The in-depth evaluation helps us to get more detailed feedback from you on the course
- Student evaluations are also a criterion for evaluating the professors’ teaching
- For MLBD, the in-depth evaluation will take place during the poster session on May 22 (on paper)

Project – Poster Presentations

- Poster Presentations on May 22 in the BC atrium, starting at 15:00
- Send us your posters by May 16 at 23:59 (Google Form) or print them yourselves
- Each team will get a presentation slot assigned – if you don't sign up for the slot, we will assign you to a slot: Sign Up Link
- You will have 5-6 minutes to present and 3-4 minutes for questions
- There will be prizes by the start-ups as well as the teaching team

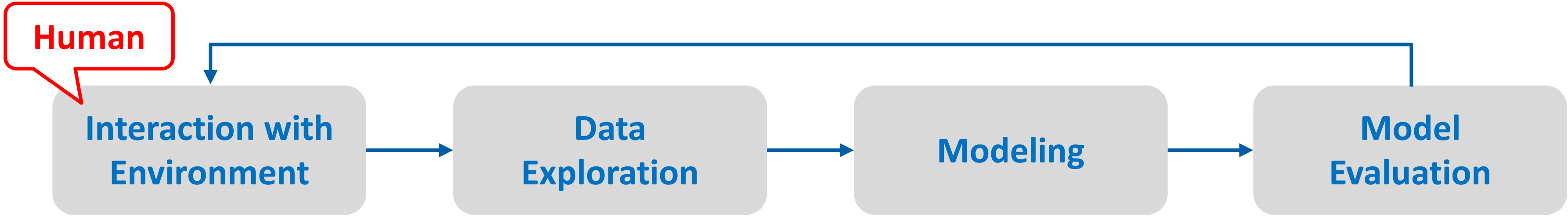
Project – Final Milestone

- Final project (Code + Report) to be delivered by June 9, 2023 23:59 CET
- Detailed guidelines (template and structure of report) have been posted on Moodle

Final Exam - Content

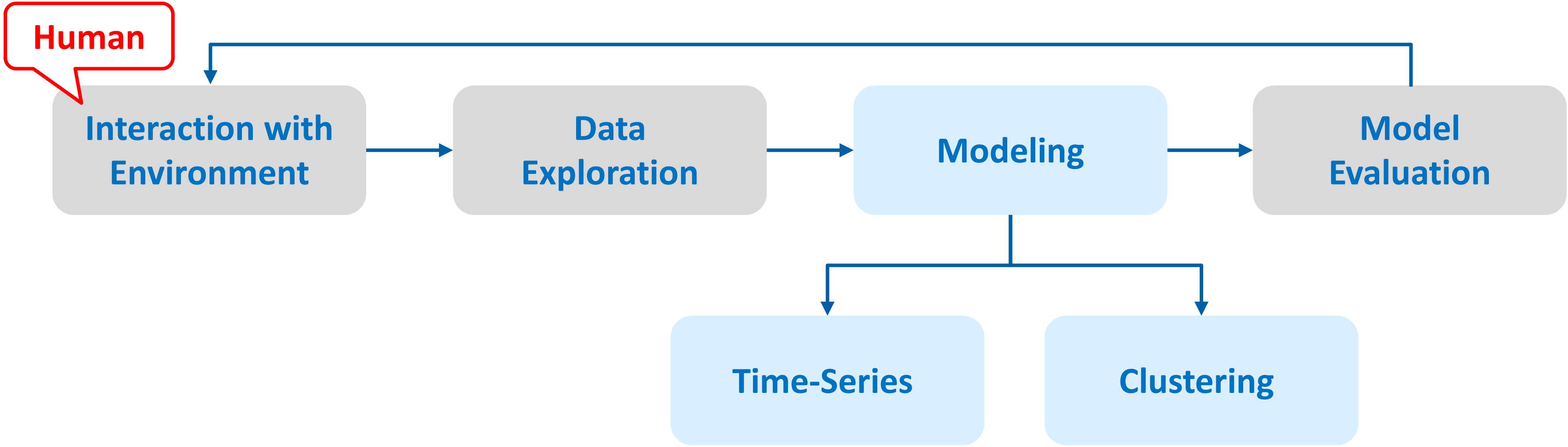
- Mix of conceptual and coding questions
- In the exam: all topics covered in the lecture and tutorials until (including) May 22

Final Exam - Content



- Regression
- Mixed-Effect Models
- Decision Tree
- Random Forests
- K-Nearest Neighbors

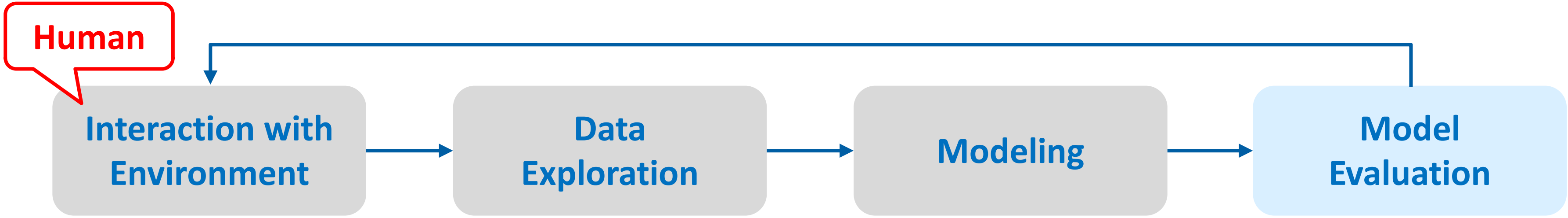
Final Exam - Content



- BKT, AFM, PFA
- Vanilla RNN
- GRU
- LSTM

- K-Means
- Spectral Clustering
- Choosing k*
- DTW

Final Exam - Content



- Fairness
- Explainability

Final Exam - Content

Design/choose an appropriate learning algorithm and features

↓

Select evaluation method

↓

Choose appropriate performance metrics

↓

Select baseline approaches for comparison

↓

Report your results providing error bars

There are many ways to solve a given task (e.g., predicting student performance). It is important that:

- You provide a clean and complete evaluation of your solution
- You are able to justify your decisions for each step

Final Exam - Administrative

- 50% of the final grade
- Saturday, July 1, 9:15-12:15 (CO020 and CO021)
- On campus:
  - Conceptual questions: on paper, 1 hour, counts 50% of the exam grade
  - Coding questions: at the computer, 2 hours, counts 50% of the exam grade
- Environment:
  - Using EPFL NOTO
  - Packages will be pre-installed for you

Final Exam - Administrative

- For both the coding and conceptual questions, you are allowed to use the lecture slides, the lecture and lab notebooks, the internet, …
- You are not allowed to communicate with other people (and we count posting on forums like Stack Overflow as communicating with other people)
- You are not allowed to use ChatGPT (or any other language model)

MOCK Exam

- We have posted the exam of last year
- **Lab session on May 31 (Wednesday)**:
  - A TA will explain and discuss the solutions with you
  - If you plan to attend, try to solve the exam beforehand
- We will post the solutions of last year's exam on Moodle in the last week of the semester

Any Questions?