

Explainability

Machine Learning for Behavioral Data
May 23, 2022

Today's Topic

Week	Lecture/Lab
9	Spring Break
10	Guest Lecture: Neuroscience
11	Unsupervised Learning
12	Unsupervised Learning
13	Ethical Machine Learning
14	Ethical Machine Learning
15	Project Presentations

- 
- Fairness
 - Explainability

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/mlbd-lecture>

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

SpeakUp Poll!

0
0 votes

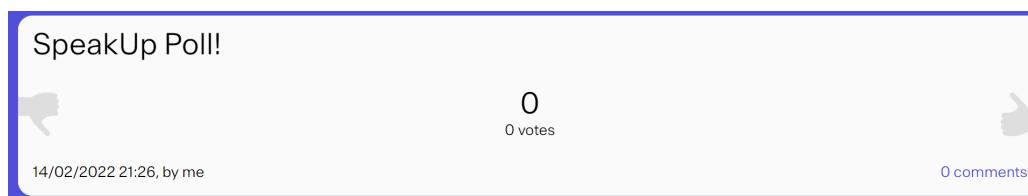
14/02/2022 21:26, by me

0 comments

Short quiz about the past...

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

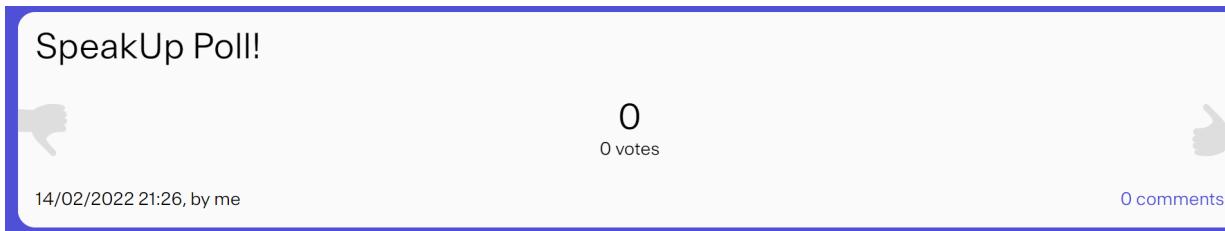
- a) Silhouette Score
- b) Jaccard Similarity
- c) Cosine Similarity
- d) Euclidean Distance



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



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



Agenda

1) Introduction to Explainability

- Taxonomy of interpretability methods
- Deep Dive: PDP
- Deep Dive: LIME

2) Course Wrap-Up (feedback, 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

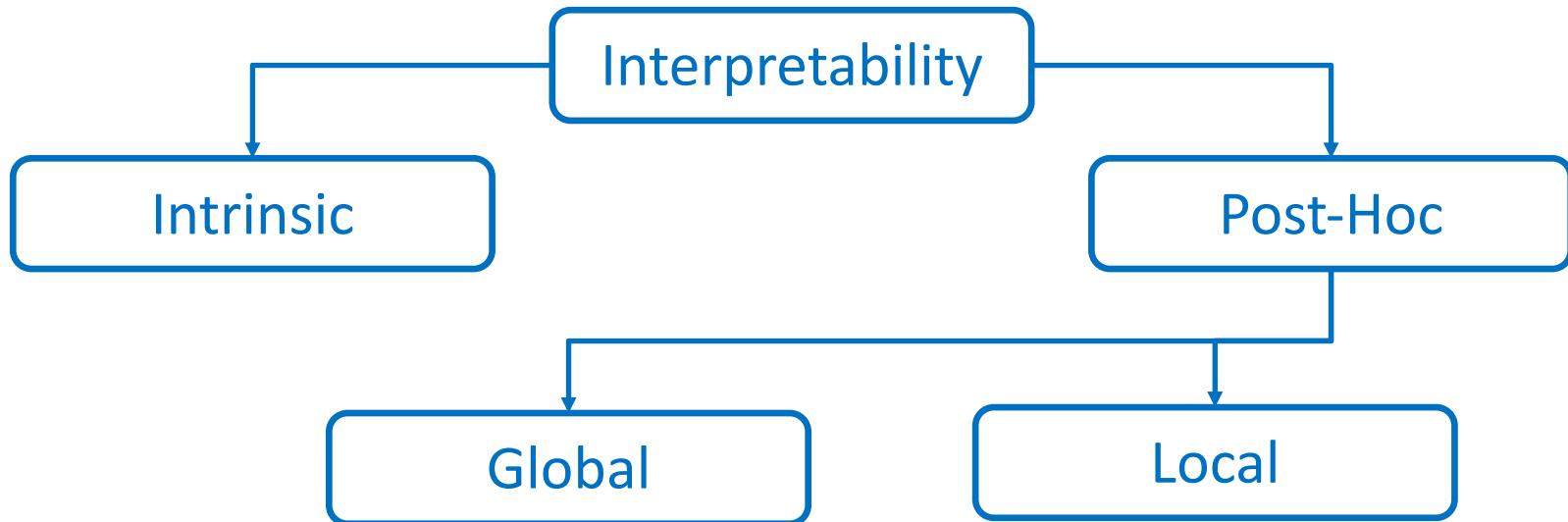


Taxonomy of Interpretability Methods

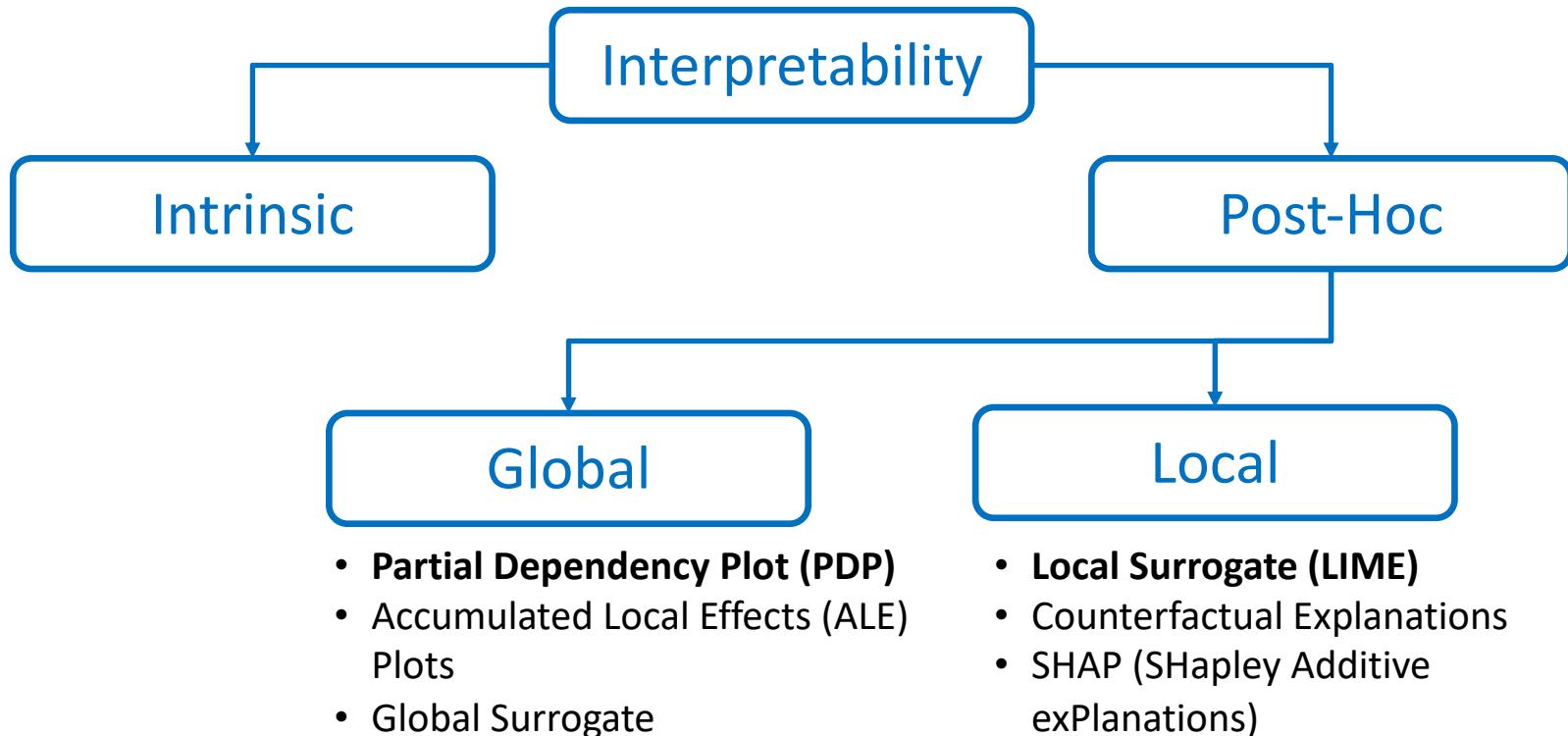


- Linear Regression
- Generalized Linear Models
(e.g., logistic regression)
- Decision Trees
- (k-Nearest Neighbors)

Taxonomy of Interpretability Methods



Taxonomy of Interpretability Methods



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)

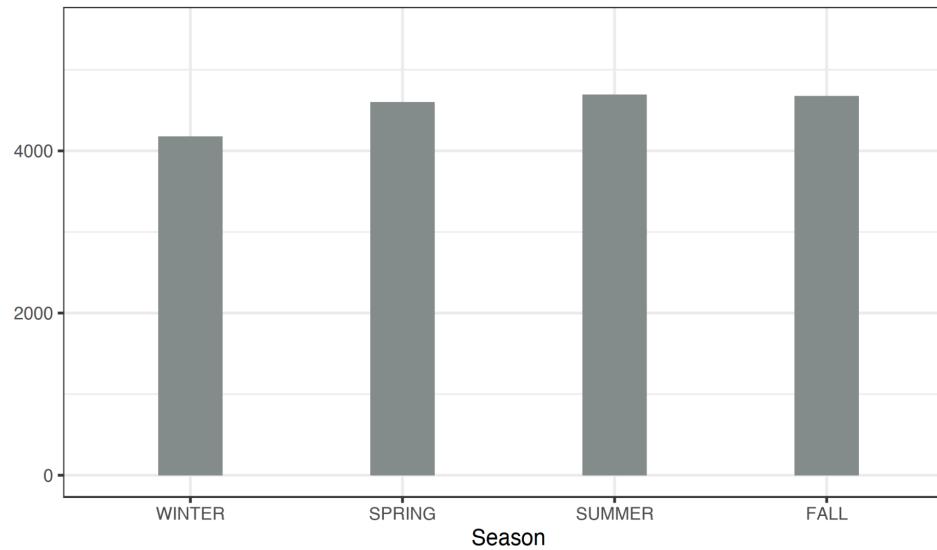
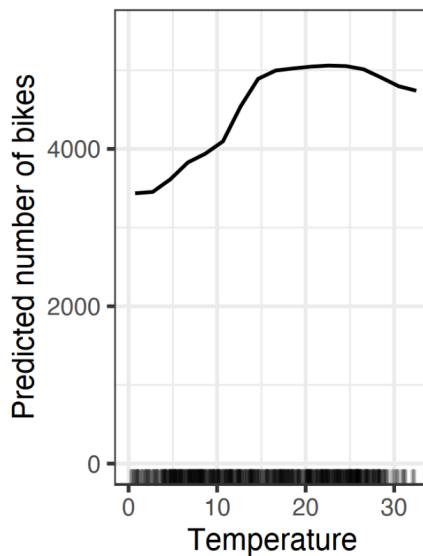


Example – Bike Rental Shop

- 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 $y = f(x)$



Partial Function - Regression

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

Partial Function - Regression

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



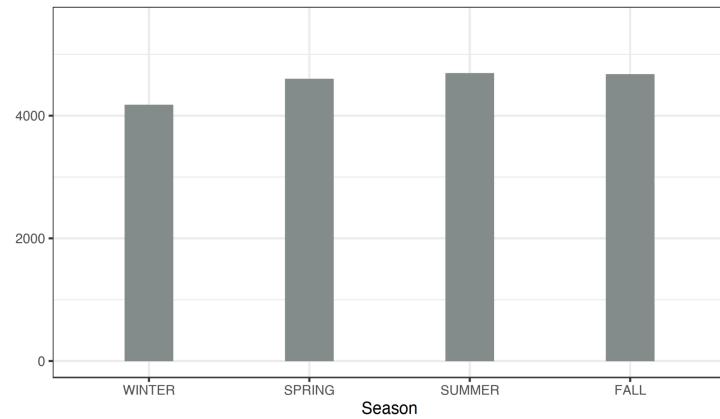
$$\widehat{f}_S(x_S) = \frac{1}{n} \sum_{i=1}^n \widehat{f}_S(x_S, x_c^{(i)})$$

Partial Function - Regression

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

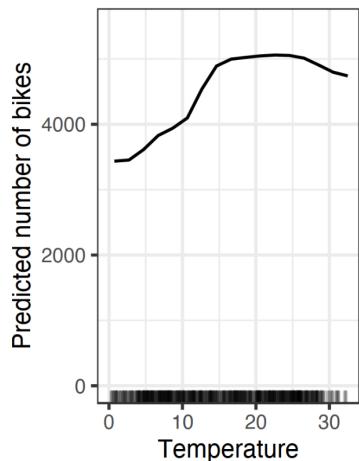
Partial Function - Regression

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



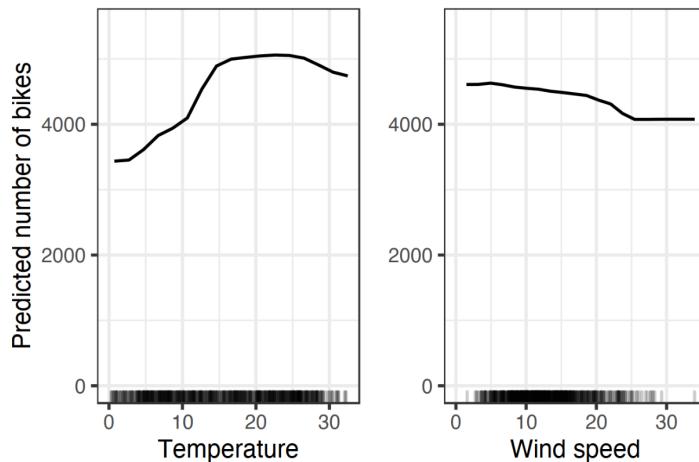
Partial Function - Regression

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



Partial Function - Regression

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



Partial Function - Classification

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

- 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

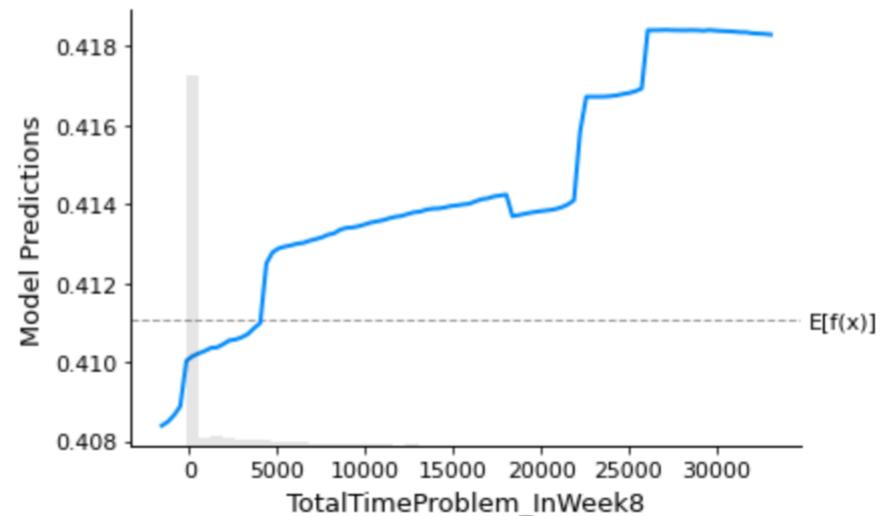
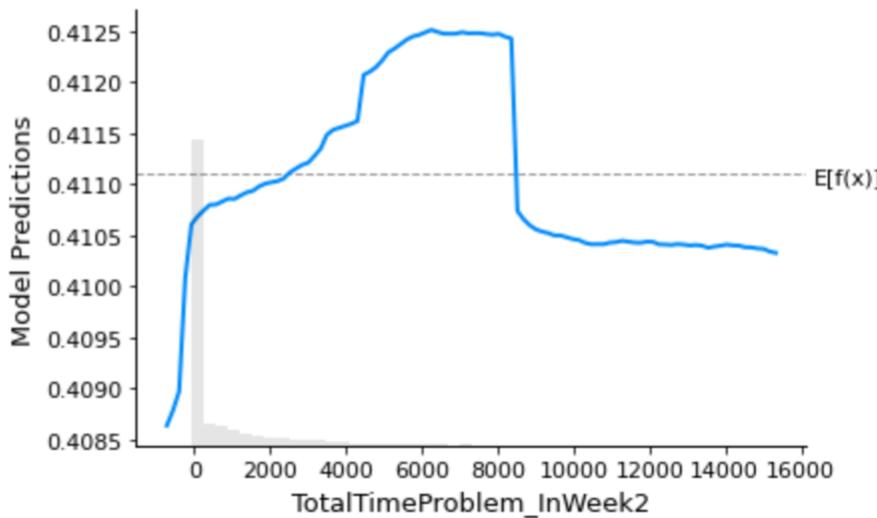
PDP – Strength & Weaknesses

- + 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

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

LIME - Recipe

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

1. Select your instance (sample) of interest



LIME - Recipe

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

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

LIME - Recipe

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

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

LIME - Recipe

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

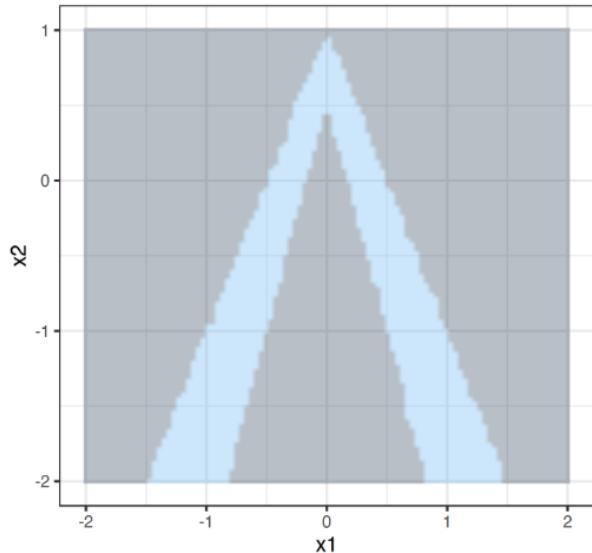
4. Train a weighted, interpretable model on the data set with variations

LIME - Recipe

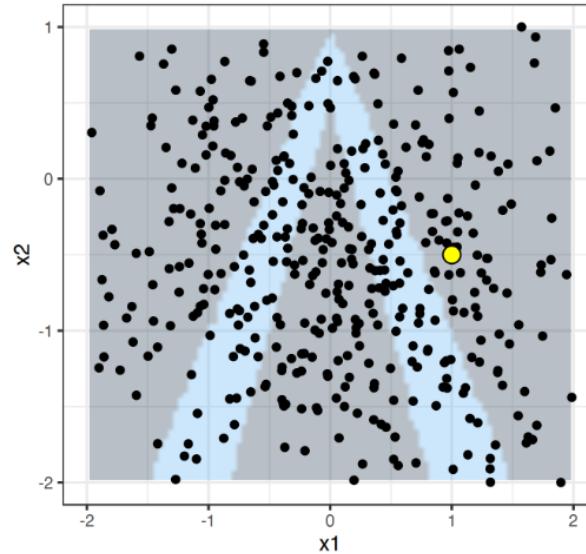
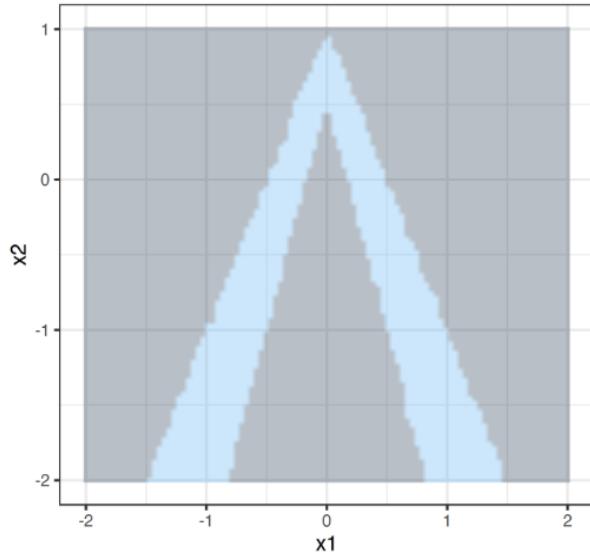
$$\text{explanation}(x) = \underset{g \in G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g)$$

5. Explain the prediction by interpreting the local model

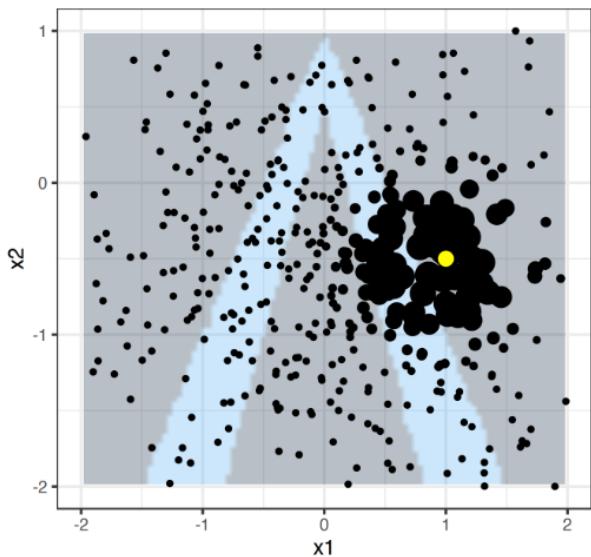
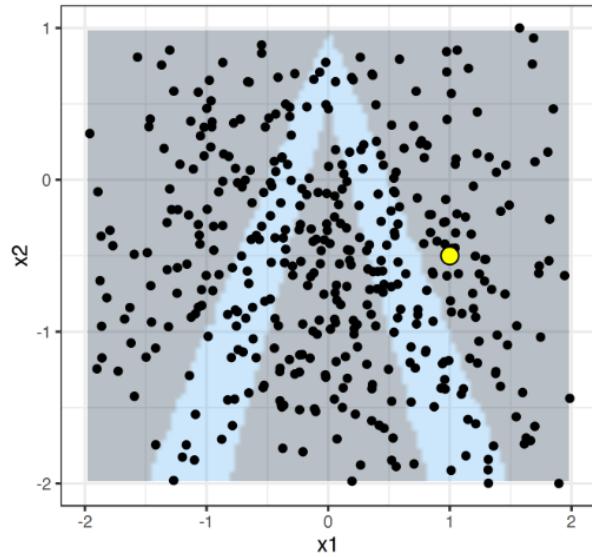
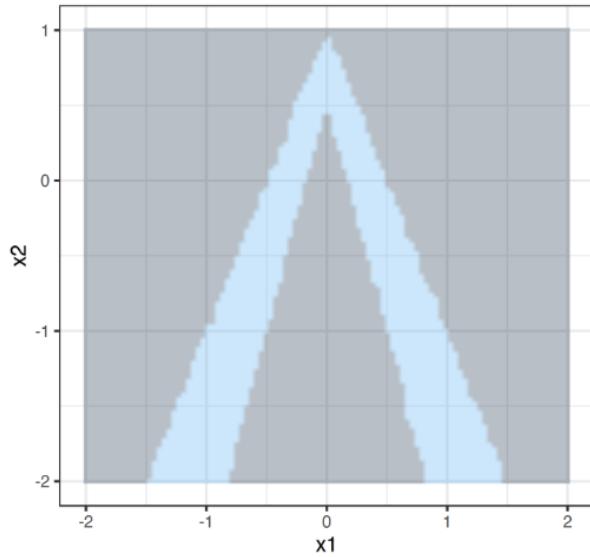
LIME – Perturbation of Sample



LIME – Perturbation of Sample



LIME – Perturbation of Sample



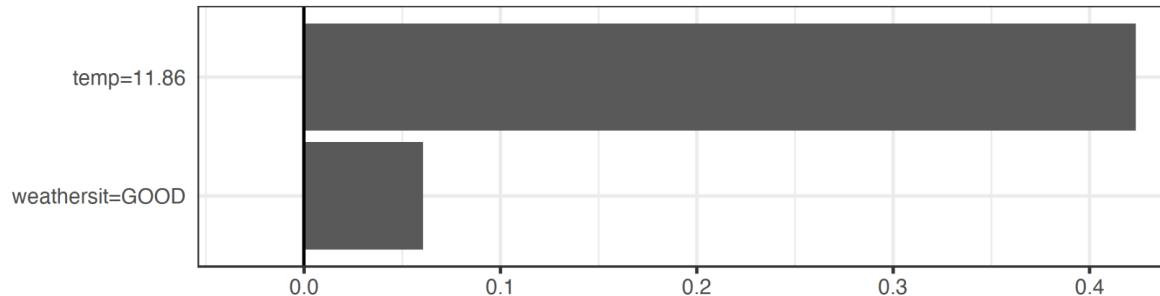
Example – Bike Rental Shop

- 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

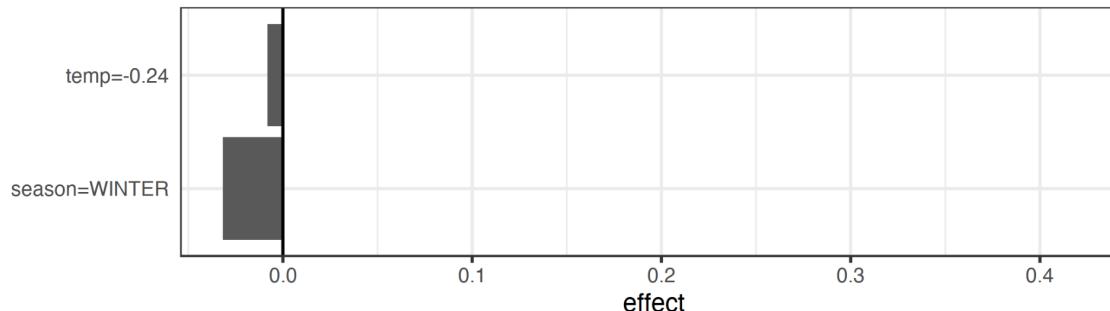
Actual prediction: 0.89

LocalModel prediction: 0.44



Actual prediction: 0.01

LocalModel prediction: -0.03



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

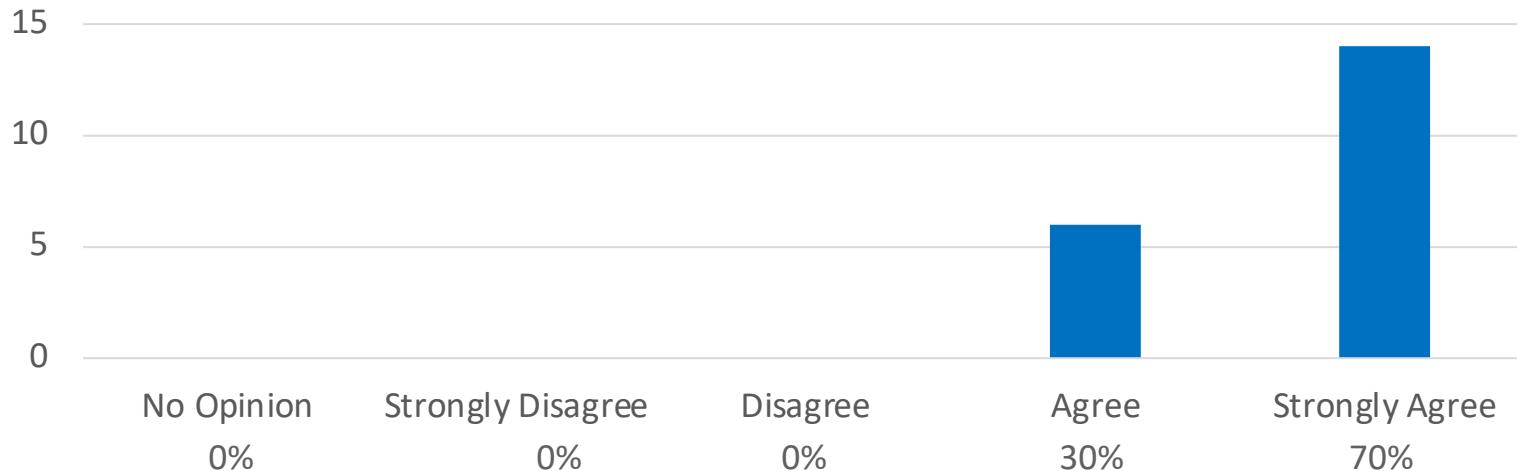
- Interpretability is important (not only for education)
 - We can use intrinsic interpretable models or post-hoc methods to get interpretable predictions
 - Methods can be categorized into global and local
 - PDP is easy to interpret, but has an independence assumption and is limited to a low number 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 (feedback, project, exam)
-

Indicative Feedback

- Overall, I think this course is good (based on 20 answers)



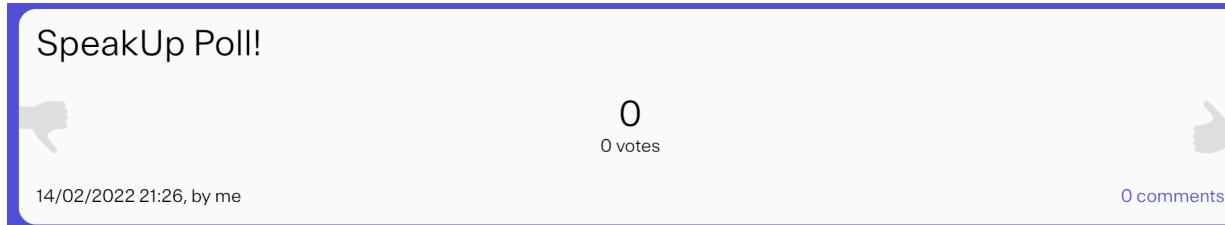
Indicative Feedback

- Learning Objectives
- More practice opportunities



Indicative Feedback

- Learning Objectives
- More practice opportunities



- a) Yes
- b) No

Would you regularly come to lab sessions to solve additional practice talks?

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 30 (on paper)
-

Project – Poster Presentations

- Poster Presentations on May 30 in the BC atrium, starting at 15:00
 - Send us your posters by May 24th 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
 - The slot from 16:35-16:45 is reserved for award nominees
-

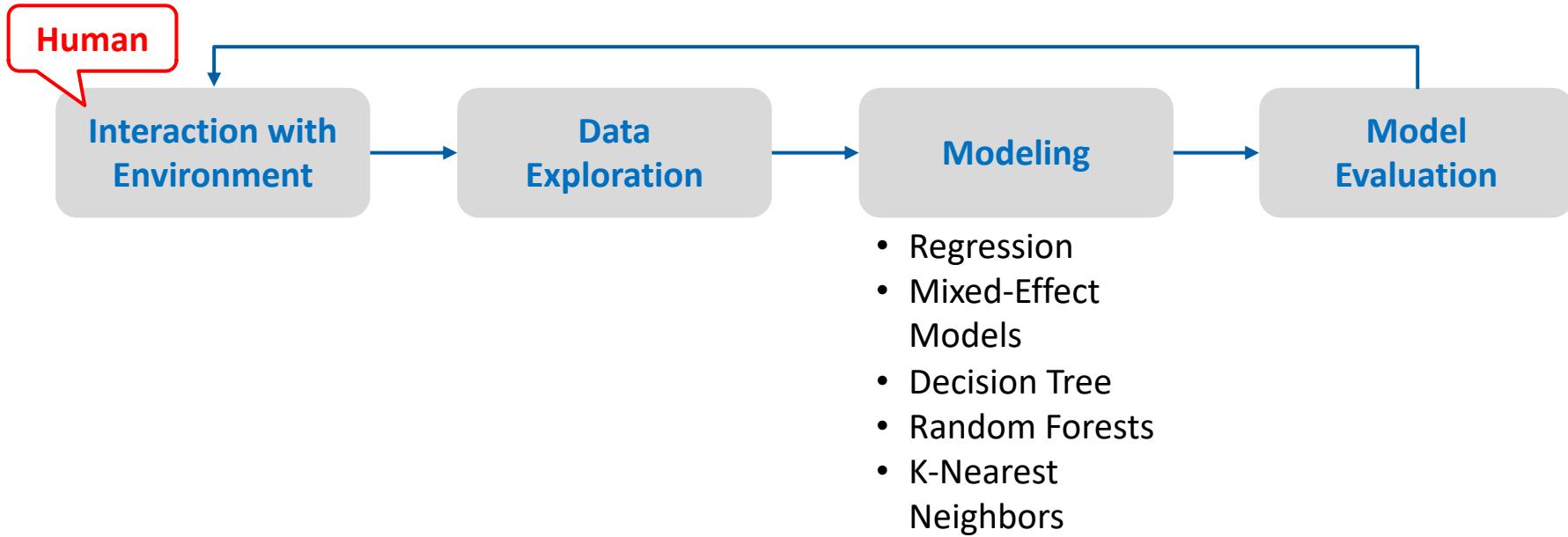
Project – Final Milestone

- Final project (Code + Report) to be delivered by **June 12, 2022 23:59 CET**
- Detailed guidelines (template and structure of report) will be posted on Moodle

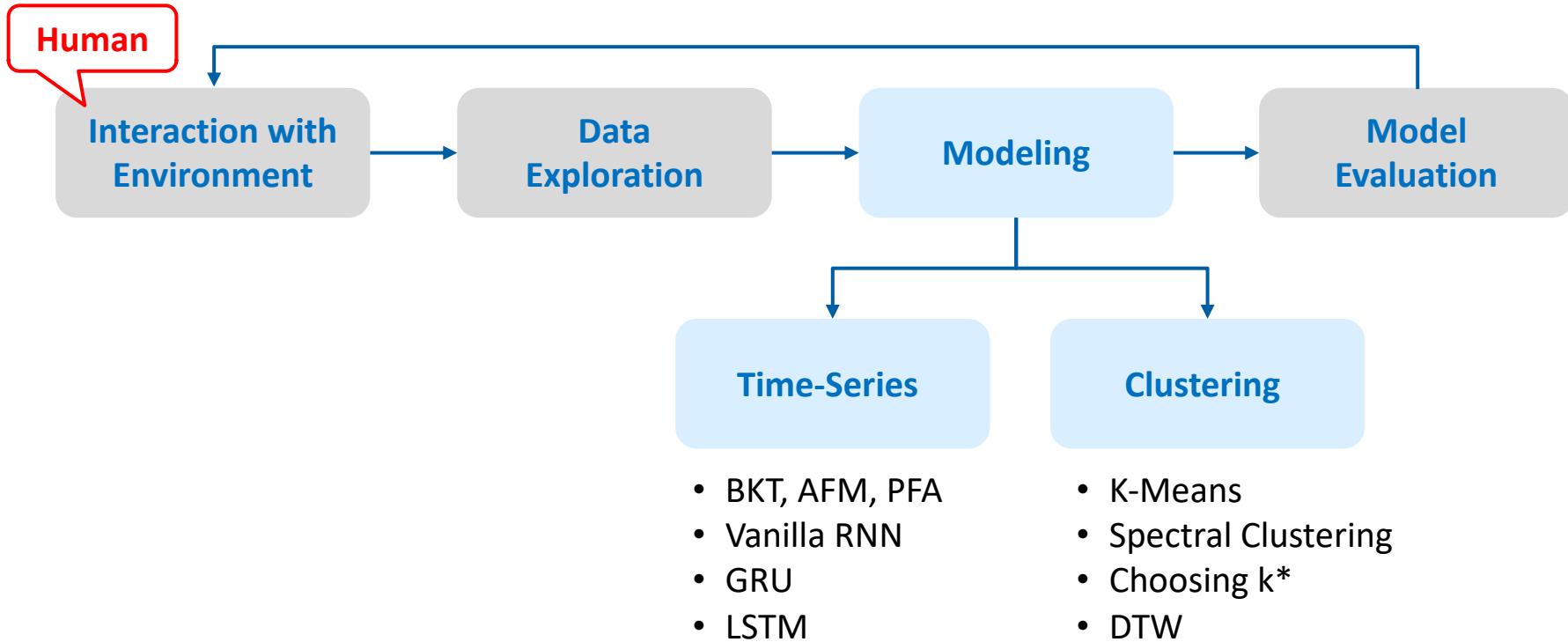
Final Exam - Content

- Mix of conceptual and coding questions
- In the exam: all topics covered in the lecture and tutorials until (including) May 30
- **Not** in the exam: topics covered by the guest lecturer (ML for neuroscience)

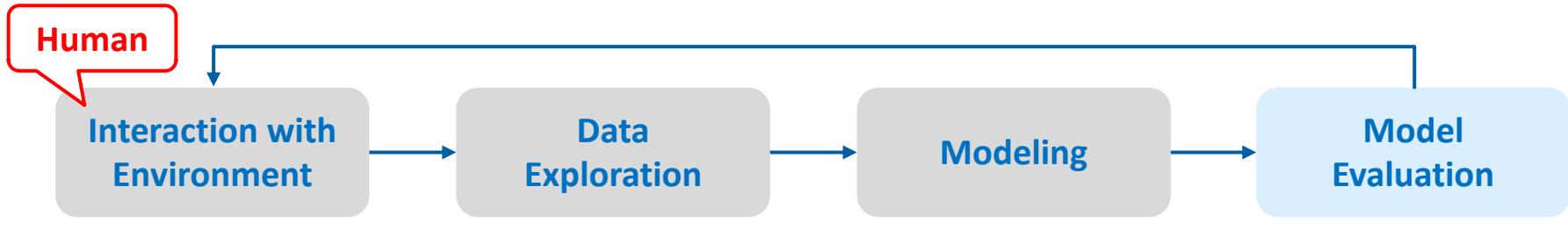
Final Exam - Content



Final Exam - Content



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
- Wednesday, July 6, 15:15-18:15 (CO020 and CO021)
- On campus, at the computer
- Environment:
 - Using EPFL NOTO
 - Packages will be pre-installed for you

Final Exam - Administrative

- Open-web: you are allowed to use any resources (lecture notebooks, lecture slides, the internet...)
- You are not allowed to communicate with other people



MOCK Exam

- We will offer a MOCK exam: 60 minutes + discussion of questions and solution
- **Tutorial session on June 1:**
 - 8:15-9:15: time to solve the MOCK exam
 - 9:15-10:00: questions and discussion of solution



Any Questions?

