Interpretability of ML models

CE-477: Machine Learning - CS-828: Theory of Machine Learning
Sharif University of Technology
Fall 2024

Fatemeh Seyyedsalehi

References of the lecture are mentioned in the last slide

Interpretation

- A broad poorly defined concept.

- The set of methods falling under this umbrella spans everything from designing an initial experiment to visualizing final results.
  - Related to data science and applied statistics.

- We focus on the use of interpretations in the context of ML as part of the larger data science life cycle.

Interpretation

- Machine learning (ML) has recently received considerable attention for its ability to accurately predict a wide variety of complex phenomena.

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

- Discovering new knowledge from the problem domain
  - Medicine and science
- Trustworthy and Fairness
  - Judgment and medicine!

Interpretable machine learning

- Interpretable goal: extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model.

- Two approaches
  - Adopting models which are inherently interpretable
    - Short Decision trees or sparse linear models
    - Only simple models with low predictive ability
  - Post-hoc analysis: attempt to explain a model after the training process
    - An arbitrary ML model can be explained

Model-specific or model-agnostic?

- Model-specific
  - Interpretation tools limited to specific model classes.
  - Tools that only work for the interpretation of e.g. neural networks are model-specific.

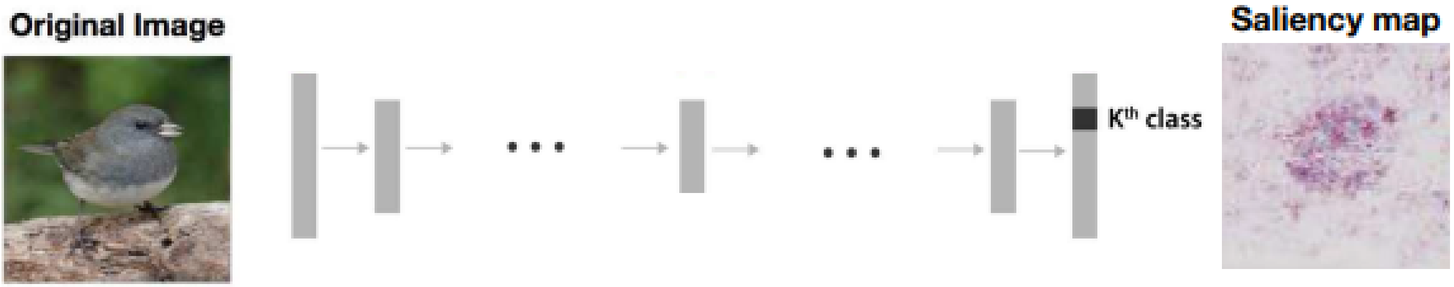
- Model-agnostic
  - Tools can be used on any machine learning model and are applied after the model has been trained (post hoc). These agnostic methods usually work by analyzing feature input and output pairs. By definition, these methods cannot have access to model internals such as weights or structural information ex. AI based APIs.
  - .

Evaluation of interpretability

- Application level evaluation (real task):
  - Put the explanation into the product and have it tested by the end user.
  - Needs an domain experts.
- Human level evaluation (simple task):
  - A simplified application level evaluation.
  - An example would be to show a user different explanations and the user would choose the best one.
- Function level evaluation:
  - Monitoring the change in evaluation measures when the interpretation results affect the model
    - Only giving selected features to a model

Feature importance

- One of common interpretation goal of ML models
- Feature importance vector
  - For each input feature calculates its importance for a decision made by a model
  - It may change in different locality of the input space
  - Therefore, an interpreter is a function over the input space
- In images, we call it saliency map



Interpretation method example: LIME

- Local surrogate models are interpretable models that are used to explain individual predictions of black-box machine learning models.

- Surrogate models are trained to approximate the predictions of the underlying black box model.

- LIME focuses on training local surrogate models to explain individual predictions.

Interpretation method example: LIME

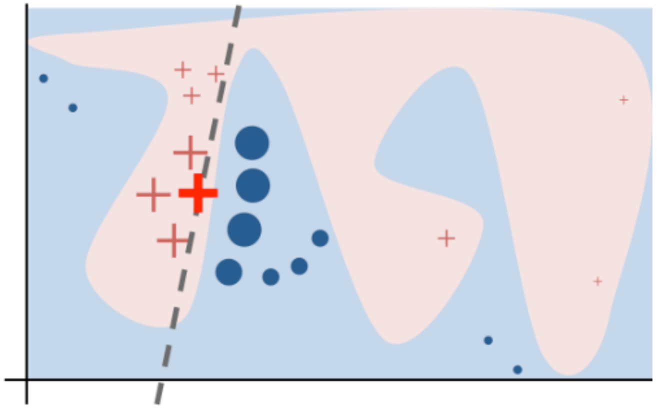
- LIME tests what happens to the predictions when you give variations of your data into the machine.

- LIME generates a new dataset consisting of perturbed samples and the corresponding predictions of the black box model

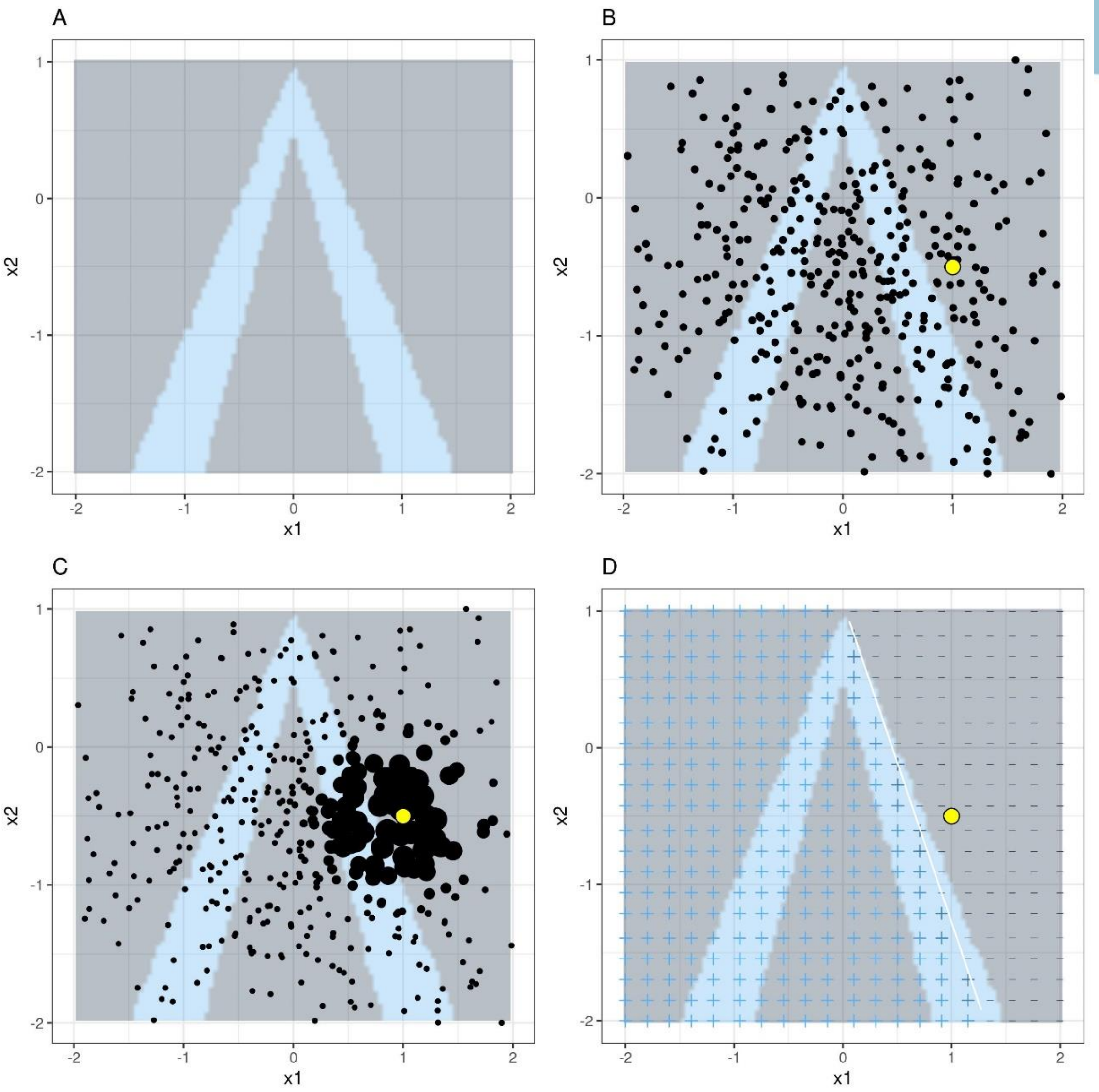
- On this new dataset LIME then trains an interpretable model, which is weighted by the proximity of the sampled instances to the instance of interest.
  - Linear regression
  - Remember locally weighted linear regression

Interpretation method example: LIME

- LIME steps:
  - Select your instance of interest for which you want to have an explanation of its black box prediction.
  - Perturb your dataset and get the black box predictions for these new points.
  - Weight the new samples according to their proximity to the instance of interest.
  - Train a weighted, interpretable model on the dataset with the variations.
  - Explain the prediction by interpreting the local model.

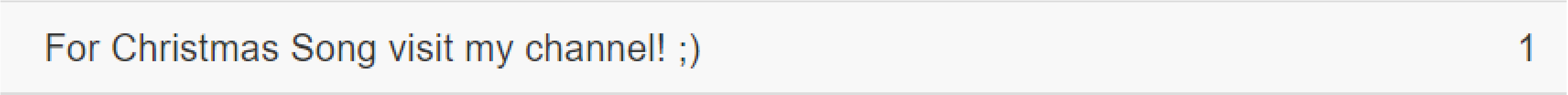


Interpretation method example: LIME



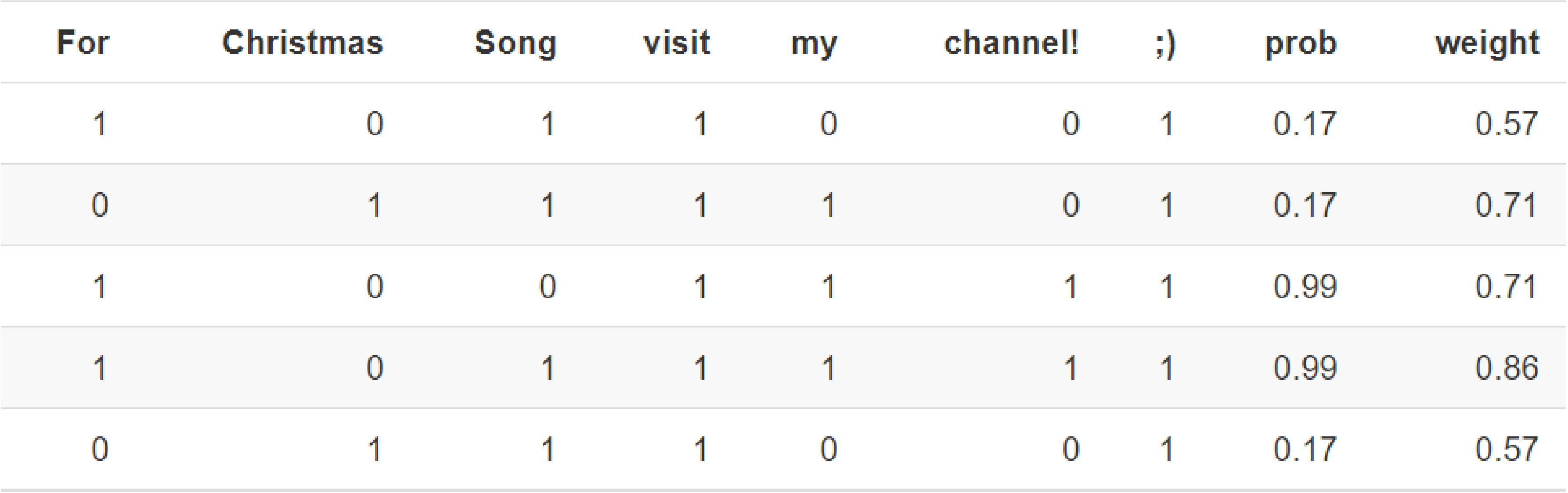
Interpretation method example: LIME

- Imagin a model trained to detect spams ☺
  - A spam example:



For Christmas Song visit my channel! ;)	1

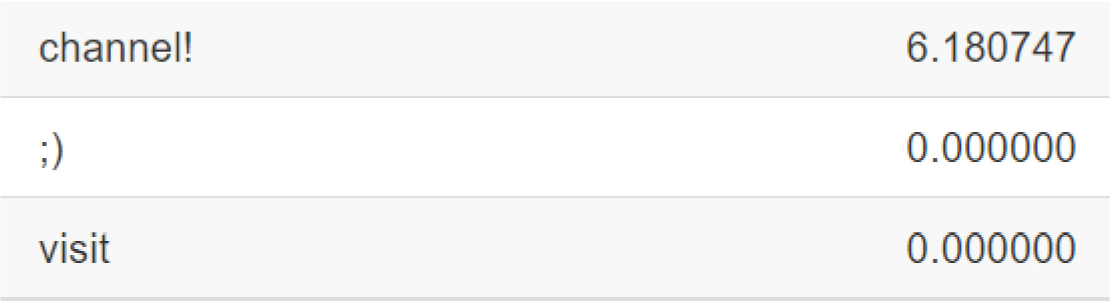
  - LIME generate different variations of this sentence



For	Christmas	Song	visit	my	channel!	;)	prob	weight
1	0	1	1	0	0	1	0.17	0.57
0	1	1	1	1	0	1	0.17	0.71
1	0	0	1	1	1	1	0.99	0.71
1	0	1	1	1	1	1	0.99	0.86
0	1	1	1	0	0	1	0.17	0.57

Interpretation method example: LIME

- Weights of a linear model for each feature
  - “channel” is the most important feature of this specific sentence which affect the model decision



channel!	6.180747
;)	0.000000
visit	0.000000

References

- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Interpretable machine learning: definitions, methods, and applications.
- https://christophm.github.io/interpretable-ml-book