ML : Explainability

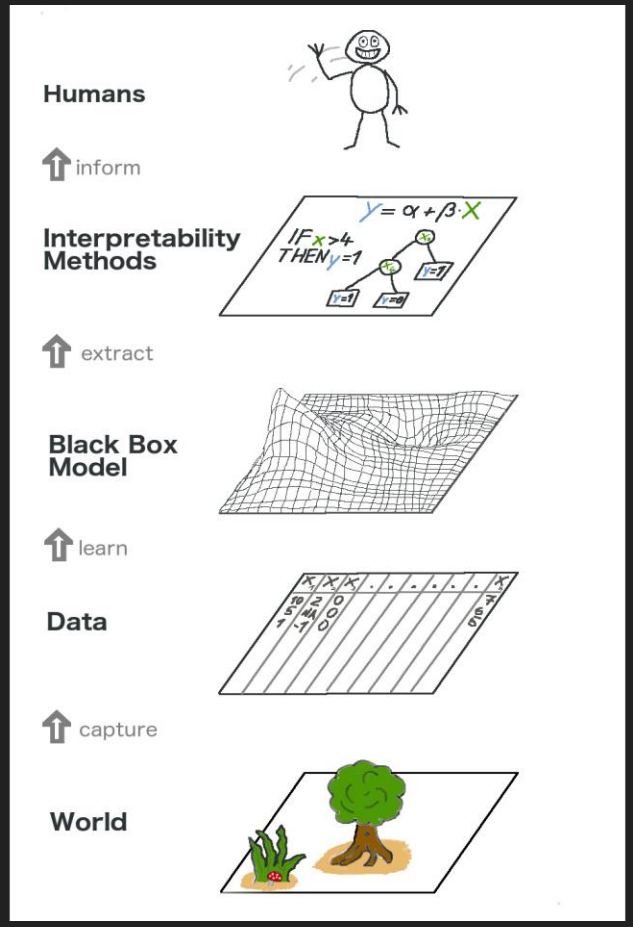
Sanjit Paliwal

Agenda

- Model Agnostic Explainability Landscape
- LIME
  - Intro
  - Process
- SHAP
  - Shapley Values
  - SHapley Additive Explanations
  - Kernel SHAP
- Demo

Model Agnostic Explainability Landscape

- World layer
  - It contains everything that can be observed and is of interest.
- Data layer
  - It contains anything from images, texts, tabular data and so on.
- Black Box Model layer
  - Machine learning algorithms learn with data from the real world to make predictions or find structures.
- Explainability Methods Layer
  - It helps us deal with the opacity of machine learning models. What were the most important features for a particular diagnosis? Why was a financial transaction classified as fraud?
- Human layer
  - Humans are ultimately the consumers of the explanations.



Source: Interpretable Machine Learning: *A Guide for Making Black Box Models Explainable by Christoph Molnar*.

LIME

Local Surrogate Model-Agnostic Explanations

- LIME is used to explain individual predictions of black box machine learning models[1].
- Characteristics of LIME[2]
  - **Interpretable**: Explanations use a representation that is understandable to humans, regardless of the actual features used by the model.
  - **Local Fidelity:** Provides a good approximation of the machine learning model predictions locally.
  - **Model Agnostic:** Able to explain any model.

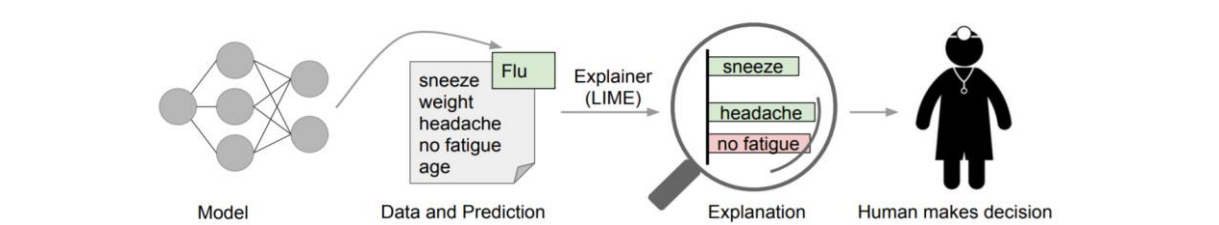


Figure 1: Explaining individual predictions. A model predicts that a patient has the flu, and LIME highlights the symptoms in the patient's history that led to the prediction. Sneeze and headache are portrayed as contributing to the "flu" prediction, while "no fatigue" is evidence against it. With these, a doctor can make an informed decision about whether to trust the model's prediction.

1 : Interpretable Machine Learning: *A Guide for Making Black Box Models Explainable by Christoph Molnar*.
2: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016)

LIME : Process

- The recipe for training local surrogate models:
  - Select your instance of interest
  - 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.
- Mathematically, local surrogate models with interpretability constraint can be expressed as follows:

$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

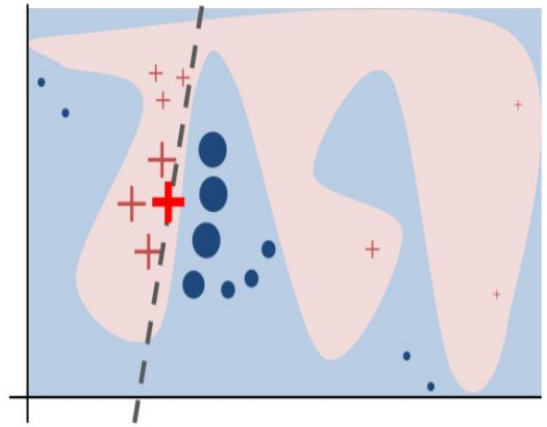
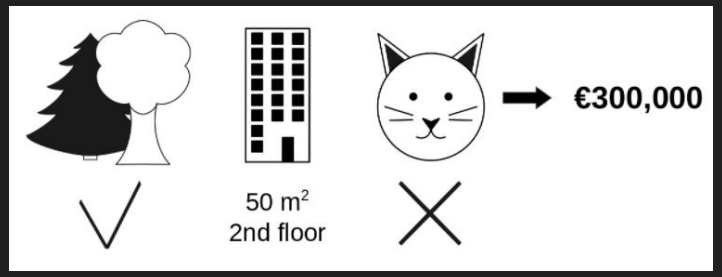


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

SHAP

Shapley Values

- Shapley value is the average marginal contribution of a feature value across all possible coalitions.
- It gives the contribution of each feature to the prediction as compared to the average prediction.



Park nearby (€30,000) +
area 50 square m (€10,000) +
floor 2nd (€0) +
Cat-banned (-€50,000) +
average prediction (€310,000) = actual prediction (€300,000)

Disadvantages

- Explanations created with the Shapley value method always use all the features.
- Shapley value cannot be used to make statements about changes in prediction for changes in the input, such as: "If I were to earn €300 more a year, my credit score would increase by 5 points.".

SHAP (SHapley Additive exPlanations)

- The goal of SHAP is to explain the prediction of an instance x by computing the contribution of each feature to the prediction.
- One innovation that SHAP brings to the table is that the Shapley value explanation is represented as an additive feature attribution method, a linear model.
- SHAP specifies the explanation as:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

g is the explanation model, z′ is the coalition vector and
φj is the feature attribution for a feature j, the Shapley values.

- **Missingness**: Missingness says that a missing feature gets an attribution of zero.

Kernel SHAP

KernelSHAP estimates for an instance x the contributions of each feature value to the prediction.

KernelSHAP consists of 5 steps:

- Sample coalitions $z'_k \in \{0,1\}^M, \quad k \in \{1,\ldots,K\}$ (1 = feature present in coalition, 0 = feature absent).
- Get prediction for each z'_k by first converting z'_k to the original feature space and then applying model f: $f(h_x(z'_k))$
- Compute the weight for each z'_k with the SHAP kernel.
- Fit weighted linear model.
- Return Shapley values ϕ_k, the coefficients from the linear model.

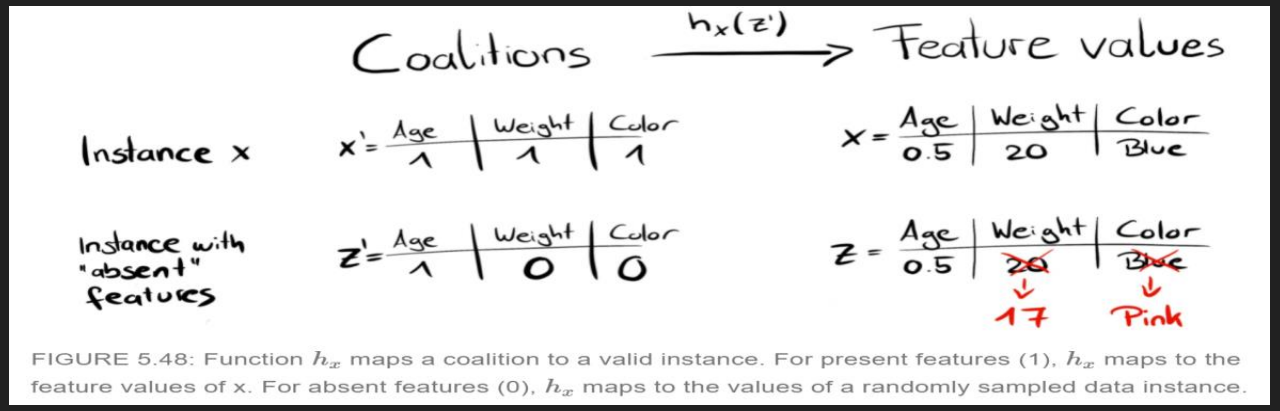


FIGURE 5.48: Function h_x maps a coalition to a valid instance. For present features (1), h_x maps to the feature values of x. For absent features (0), h_x maps to the values of a randomly sampled data instance.

References

1. Interpretable Machine Learning: *A Guide for Making Black Box Models Explainable by Christoph Molnar*
2. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016)
3. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems. 2017
4. https://github.com/interpretml/interpret
5. "Why Should I Trust you?" Explaining the Predictions of Any Classifier: https://www.youtube.com/watch?v=KP7-JtFMLo4


Q
&
A


THANK
YOU


That's all Folks!