Bootcamp VII: Explainable Machine Learning with LIME

Contents

- Motivation
- Feature Importances
- Feature Importances within a simple Linear Regression Framework
- Global Feature Importances (Random Forests)
- A Surrogate Approach
- Local-Interpretable Model Agnostic Explanations (LIME) – Explaining through local surrogate models

Motivation

- If a machine learning model performs well, **why do not we just trust the model** and ignore **why** it made a certain decision?
- “The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks.” (Doshi-Velez and Kim 2017)
- E.g. ML for Cancer Diagnosis

Explainable ML

- Various Approaches:
  - Feature Importances
  - Global Surrogate Model
  - Local Surrogate Model
  - Partial Dependence Plots
  - Individual Conditional Expectations
  - Shapley Value

- See the Interpretable ML book (available online for more)

Feature Importances

- A feature is “important” if shuffling its values increases the model error, because in this case the model relied on the feature for the prediction.
- A feature is “unimportant” if shuffling its values leaves the model error unchanged, because in this case the model ignored the feature for the prediction.
- Effects on model error <--> Loss functions

Feature Importance – Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

- We estimate the parameters usually with ordinary least squares

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j^{(i)} \right) \right)^2$$

Feature Importance – Linear Regression

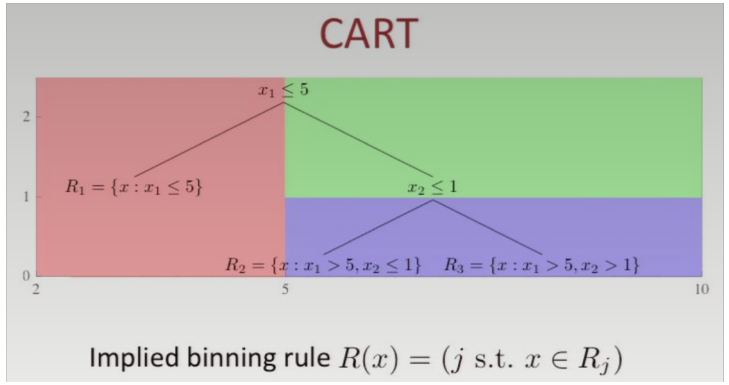
$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

- But how do we evaluate the importance of feature X_p ?
  - Maybe we can fit one regression including X_p , and one excluding X_p , and then use the differences in R^2 → Sequential R_2
  - But order of the regressors matter – what if X_1 is highly collinear with X_p ?
  - Maybe we can adjust for dependence on orderings by taking into account the orderings → Lindeman, Merenda & Gold (1980) implemented in the relaimpo package in R

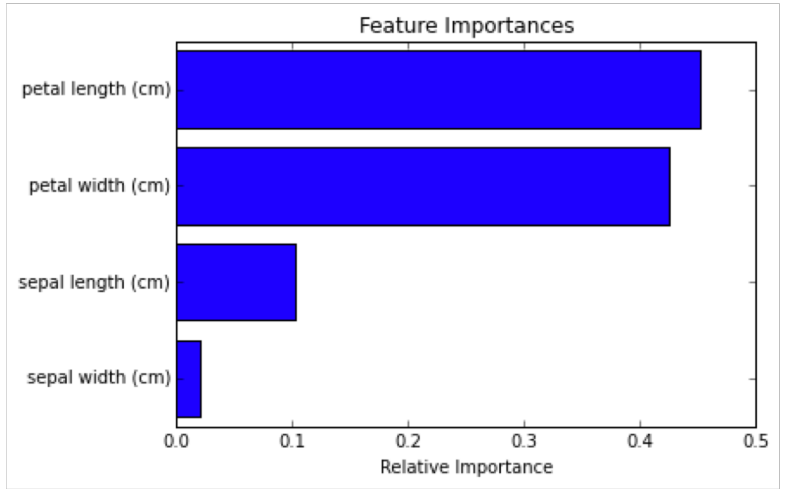
$$\mathrm{LMG}(x_k) = \frac{1}{p!} \sum_{r \text{ permutation}} seqR^2(\{x_k\}|r).$$

Global Feature Importances (Random Forests)

- For a classification tasks, Random Forests try to find ways to bin the data in order to minimize a loss function such as the Gini impurity / entropy



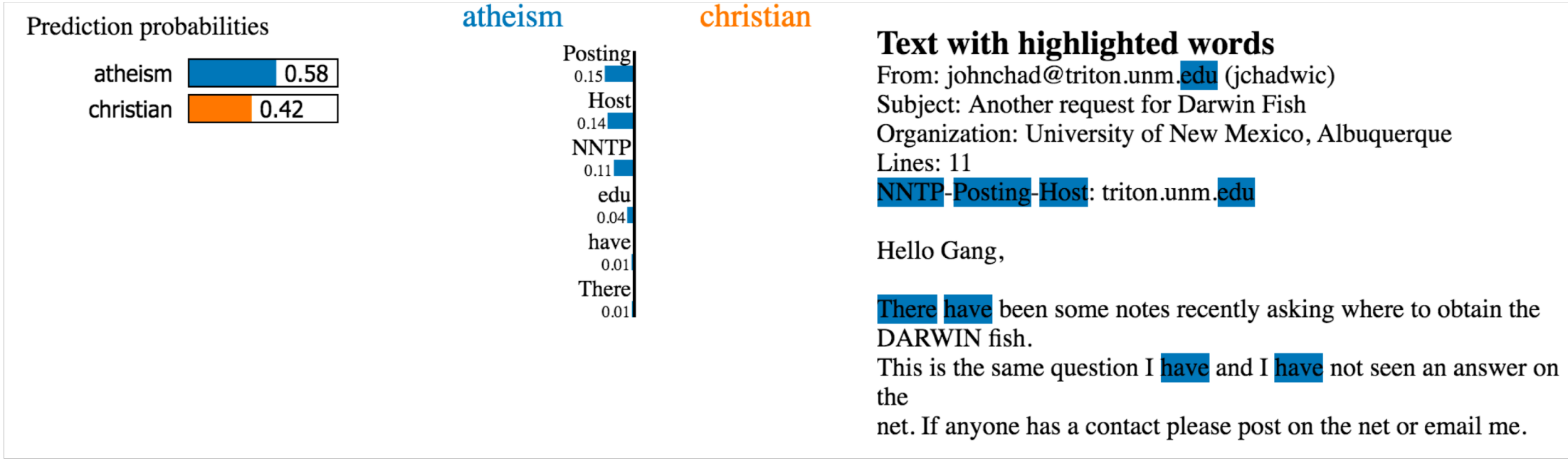
- We can calculate a feature importances by examining how much, on average, partitioning on that feature contributes to a decrease in the loss function



$$\mathrm{I}_G(p) = \sum_{i=1}^{J} p_i \sum_{k \neq i} p_k = \sum_{i=1}^{J} p_i(1-p_i) = \sum_{i=1}^{J} (p_i - {p_i}^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} {p_i}^2 = 1 - \sum_{i=1}^{J} {p_i}^2$$

But can we get an explanation for a particular example?

- We can't from the feature importance approach, but we can with from a local surrogate approach.
- Today we are going to try explaining how a trained RandomForest model classifies particular texts.

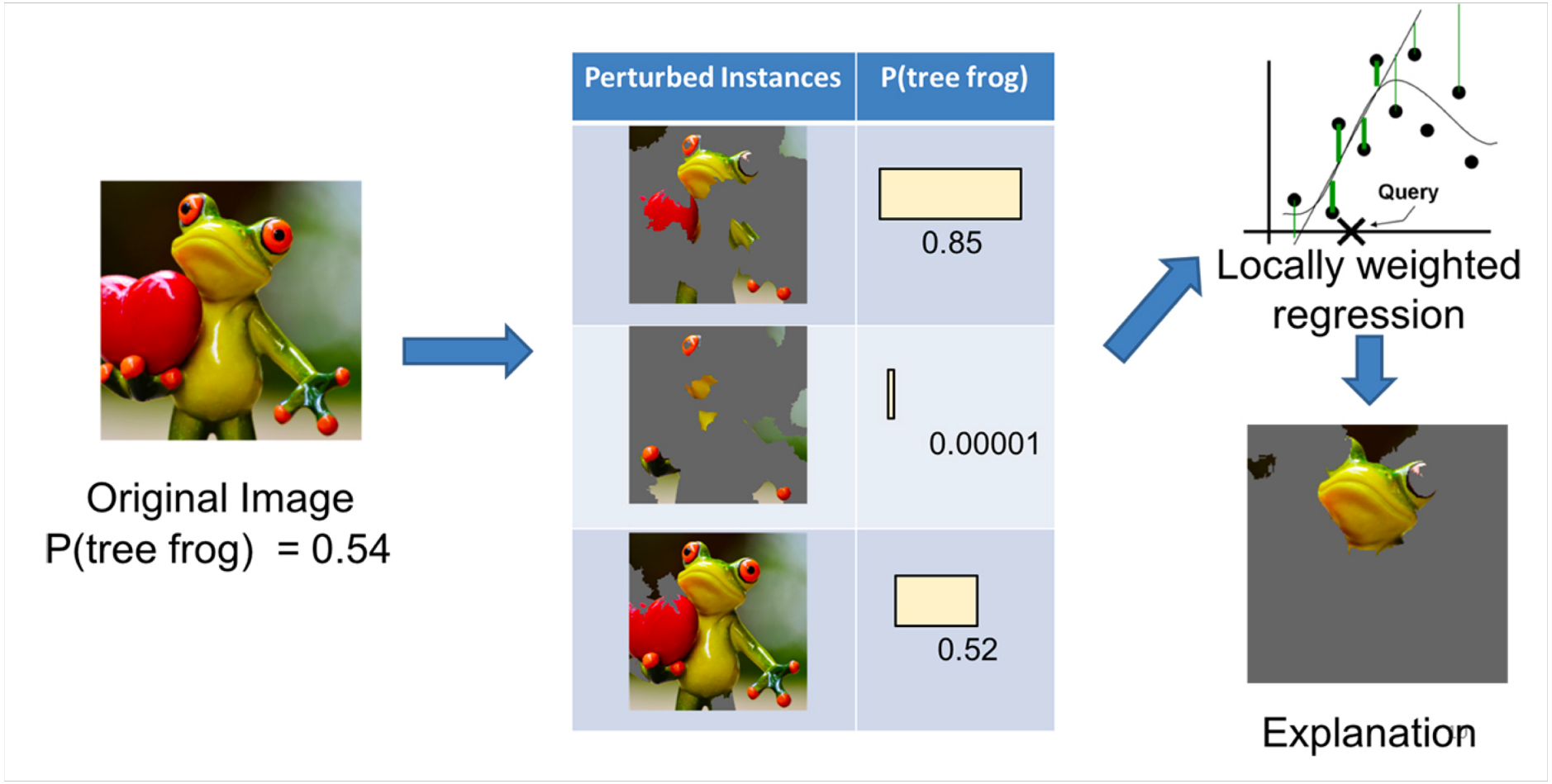


Global vs Local Surrogate Models

- Global local surrogate model is an interpretable model that is trained to approximate the predictions of a black box model.
  - Suppose you have a deep neural network that has been trained.
  - Let's approximate all of its predictions again with a simpler model such as linear regressions.

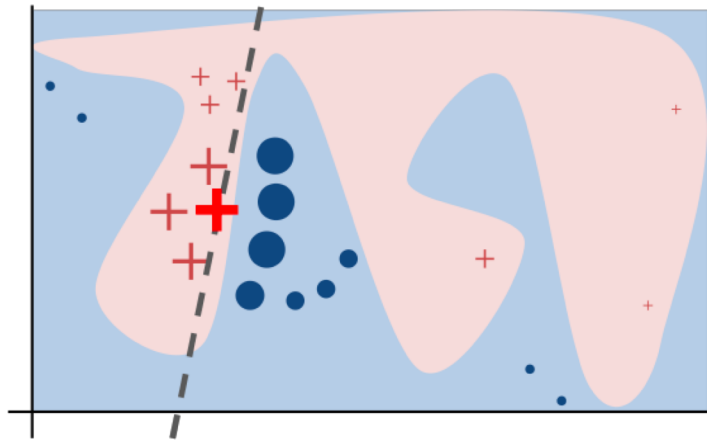
- Local surrogate models are interpretable models that are used to explain **individual** predictions of black box machine learning models.

LIME in Action



LIME: Intuition

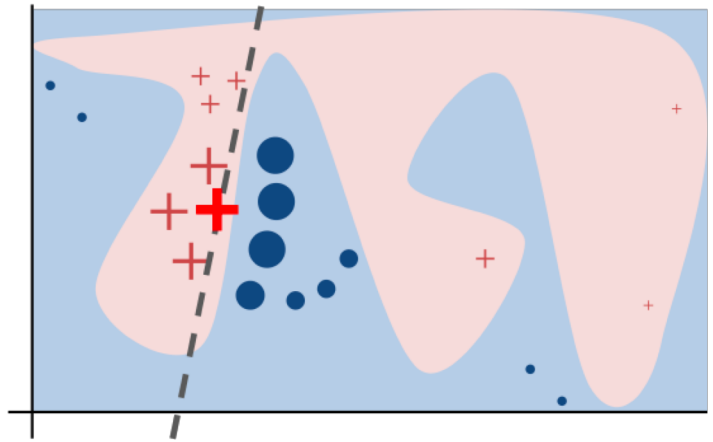
1. Select your instance of interest for which you want to have an explanation.
2. Sample instances in the neighbourhood of the instance of interest, get the black box predictions for these new points.
3. Weight the new samples according to their proximity to the instance of interest.
4. Train a weighted, interpretable model on the dataset with the variations.
5. Explain the prediction by interpreting the local model.



- Complex decision boundary (blue/pink)
- Instance of Interest: +
- Sample Points: + & ●

LIME: Notes

- Model-agnostic
- Robust to a wide variety of datasets
- Implemented in both Python & R
- Good LOCAL approximation to the decision boundary
- What is the correct definition of “the neighbourhood” of the instance of interest?



lime

build passing

This project is about explaining what machine learning classifiers (or models) are doing. At the moment, we support explaining individual predictions for text classifiers or classifiers that act on tables (numpy arrays of numerical or categorical data) or images, with a package called lime (short for local interpretable model-agnostic explanations). Lime is based on the work presented in this paper (bibtex here for citation). Here is a link to the promo video:

Our plan is to add more packages that help users understand and interact meaningfully with machine learning.

Lime is able to explain any black box classifier, with two or more classes. All we require is that the classifier implements a function that takes in raw text or a numpy array and outputs a probability for each class. Support for scikit-learn classifiers is built-in.

References

- Interpretable Machine Learning – A Guide for Making Black Box Models Explainable by Chrisoph Molnar
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). ACM.
- https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime
- https://github.com/marcotcr/lime