Statistical Machine Learning
Sharif University of Technology
Hamid R. Rabiee

May 27, 2020
CE 40-956

TH Quiz 8 (Interpretability)

Due June 11, 2020 (11:59 pm)

1. Below, we introduce two kind of interpretability technique title; first search about them. Indentify the model-specific or model-agnostic paramete of them and then use them to solve the problem of which should be used in following special cases in order to reduce complexity and making sutiatuion more interpretable.

   (a) Anchors

   (b) LOCO variable importance

   (c) LIME

   (d) Treeinterpreter

   (e) Shapley explanations

   Here's some situation. You should determine the usage of the disscused techniques for these cases (with explenations and ressons)

   i. Suppose a situation that we want to derive consistent local variable contributions to black-box model predictions. We know that our numbers are not large (or our trees are not deep) and there are some low-level codes.

   ii. Suppose a situation that we only have access to some most important local variables and we want to generate sparse or simplified, explanations using these variables. Totally we want to describe the average behavior of a complex machine-learned response function.

   iii. Suppose a situation that we want high precision and also we want to generates rules about the most important variables for a prediction.

2. A prediction can be explained by assuming that each feature value of the instance is a player in a game where the prediction is the payout. Shapley values tells us how to fairly distribute the payout among the features. Formally, the Shapley value is the average marginal contribution of a feature value across all possible coalitions.

   Suppose a linear model prediction for a single data point:

   $$\hat{f}(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

   Here x is the instance for which we want to compute the contributions. Each x_j is a feature value, with $j = 1, ..., p$. The $\beta_j$is the weight corresponding to feature j.
   The contribution is the difference between the feature effect minus the average effect. First, Calculate how much each feature contributed to the prediction.

   The Shapley value is defined via a value function val of players in S. The Shapley value of a feature value is its contribution to the payout, weighted and summed over all possible feature value combinations:

   $$\phi_j(val) = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j\}} \frac{|S|!(p-|S|-1)!}{p!} \left(val\left(S \cup \{x_j\}\right) - val(S)\right)$$

where S is a subset of the features used in the model, x is the vector of feature values of the instance to be explained and p the number of features. is the prediction for feature values in set S that are marginalized over features that are not included in set S:

$$val_x(S) = \int \hat{f}(x_1, \ldots, x_p) d\mathbb{P}_{x \notin S} - E_X(\hat{f}(X))$$

Now, suppose a concrete example: The machine learning model works with 4 features x_1, x_2, x_3 and x_4. Evaluate the prediction for the coalition S consisting of feature values x_1 and x_3 and compare it to feature contributions in the linear model. At the end, check if the Sharlpy value method satisfies the properties **Efficiency**, **Symmetry**, **Dummy** and **Additivity**. (Hint: Sharply value is a fair payout method)