Methods for understanding the variable importance of local explanations of black-box models

Abstract

Artificial Intelligence (AI) has seen a revitalization in recent years from the use of increasingly hard-to-interpret black-box models. In such models, increased predictive power comes at the cost of opaque factor analysis, which has led to the field of explainable AI (XAI). XAI attempts to shed light on these models, one such approach is the use of local explanations. A local explanation of a model give a point-estimate of linear variable importance in the vicinity of one observation. We extract explanations for each observation, and approximate data and this attribution space side-by-side with linked brushing. After identifying an observation of interest its local explanation is used as a 1D projection basis. We then manipulate the magnitude of the variable contributions with a technique called the tour. This tour animates many projections over small changes in the projection basis. Doing so allows a user to visually explore the data space through the lens of this local explanation and interrogate its variable importance. The implementation of our framework is available as an R package called cheem available at github.com/nspyrison/cheem.

1 Introduction

1.1 MODELING

1.2 XAI & interpretability crisis

1.3 Local explanations

1.4 Data visualization tours

2 SHAP local explanation

2.1 Variable importance, permuting over the X's included

2.2 Visualizing and break down plots

3 Interrogate variable imporances of the local explanations: Cheem

3.1 RF model

3.2 shap matrix

3.3 Linked global approximations of data and local attribution spaces

3.4 Applcation of the manual tour to intterogate local explanations

3.5 Classification task

3.6 Regression task

4 Application Design

4.1 1) Penguins speicies classification

4.2 2) FIFA wage regression

4.3 3)?

4.4 4)?

5 Software Infrastructure

5.1 Extend spinifex, consume DALEX & treeshap

5.2 Preprocess

5.3 Runtime rendering

6 Discsussion

7 Acknoledgements

8 References