Latent Dimensionality Reduction: An Algorithmic Method for Interpreting Black Box Models

Draft

Elias Kassell, Fred Farrell

TODO 2019

1 Abstract

LDR (Latent Dimensionality Reduction) allows users to interpret any model by decomposing predictions in a high dimensional space into a lower dimensional space, eliminating the requirement for feature reduction prior to model training. Both the certainty of the model of the output and the final value are encoded in the condensed space. This dimensionality reduction is done by approximating the models decision as function, drawing samples using the VEGAS algorithm, interpreting the space using Monte Carlo Integration, then using either binning or an ensuing model to estimate prediction certainty. This ability to glimpse higher dimensional spaces, which are normally too complex for humans to understand, represents a powerful tool in improving the underlying features generating the metrics.

Acknowledgements

TODO

2 Introduction

Black box models are notoriously difficult to interpret. This makes them impractical for situations where it is a necessity to understand why a model has made a decision, what changes to features can be changed to modify the result of the model in a particular way, The TODO method allows a high dimensional space to be decomposed into its

"Feature interpretation" here is used to describe understanding the effect a singular feature, or subset of features, has on the overall set of features used to train a model.

Some of the real world implications that interpreting black box models provides include

1. Interpreting how the value of a feature affects a models decision. In a medical setting, this can be used for suggesting the relation between features and a diagnosis in a medical setting. In manufacturing, this can providing target values for the underlying processes involved in quality of produce.

2. The ability to use a model when not all values for the input features are present. By integrating accross features which are not present, the certainty of a models prediction can be estimated. Current methods, such as using average values for features which are not present, ignores the variability of the absent features accross the feature space, which LDR provides.

3 The Algorithm

1. Normalize the original data so it falls into [0, 1] intervals.
2. Train the predicting model on the normalized data.
3. Train a OCSVM (One Class Support Vector Machine) [TODO] on the normalized data, or an alternative model for detecting outliers.
4. Create a KDE (kernel density estimate) [KDE] of the training points with a bandwidth equal to some resoltuion r. If the original model is a classification, create equal density accross all classes (can be done by duplciating points).
5. Sample n new points from the kernel density, using both classifiers to make predictions of each point. Weight the prediction according to the outlier classifier, where outliers have no model certainty in the prediction at that point.
6. Two options are available here, where single feature interpretation is easiest done with binning, while with multiple feature interpretation it is more sensible to use a regression SVM (support vector machine) [].
   (a) Binning: for each dimension, group points with resolution r, reducing the value of the bin to the mean prediction accross it.
   (b) SVM: for each dimension, train a regression SVM to estimate the prediction across the feature space. Use the SVM to make predictions across the space with resolution r.

4 Classification Example

5 Regression Example

6 Conclusion