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

DRAFT - TENTATIVE

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 LDR 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 double resolution, $r \times 2$.

4 Classification Example

The Wisonsin breast cancer dataset is used as it has 30 dimensions. The aim with the dataset is to classify the tumor as either malignant or benign. There are 569 samples in total, with 357 benign and 212 malignant.

4.1 Random Forest and OCSVM Interpolation

The 100 estimator RF (Random Forest) achieved an F1 score of 0.975, while the OCSVM was trained to interpret 10% of training values as outliers. A value of 1.0 here indicates benign tumors, while 0.0 indicates malignant. All of the data is min-max scaled, a 0.7/0.3 training/testing data split is applied. LDR is then applied to the training data, integrating over 50,000 randomly selected points from the weighted kernel density of the sample space with a resolution of 1/50. Mean compactness, mean symmetry, and mean area are selected for visual inspection of their individual effect on the model's classification, which can be seen in figure 2.

4.2 Neural Network and OCSVM Interpolation

A 3 layer feed forward Neural Network (NN) was used; an input, a hidden layer where tanh is applied, and an output layer. Softmax is applied when making predictions, and the final certainty of the prediction is calculated by calculating the proportion of the largest weighted class out of all weightings. The NN was trained over 50,000 epochs and achieved an F1 score of 0.961, slightly worse than the RF. The NN was more binary with its decisions,

4.3 Analysis of Discrepancy

There is very little discrepancy between the two model, both of which have similar predictive success. However, random forests

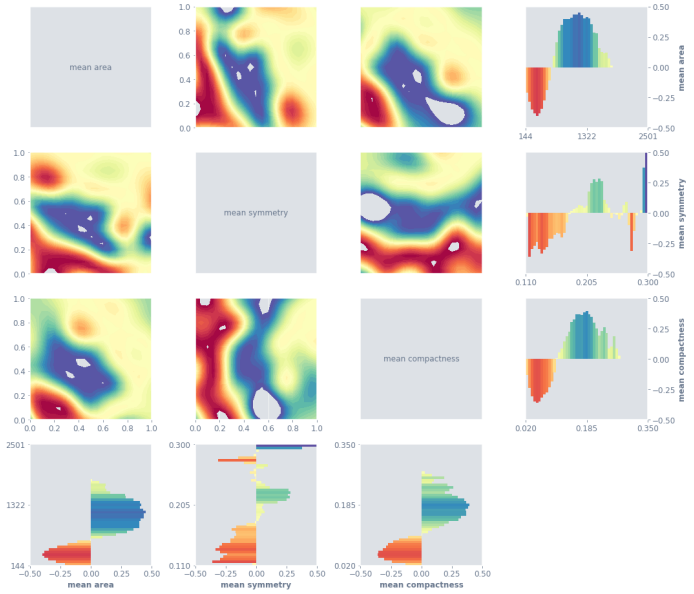

Figure 1: RF with OCSVM interpolation. Blue indicates model certainty of benality, yellow of uncertainty, and red of malignancy. Lack of contour indicates a prediction outside the sample space. The bottom and right axis are the single dimension decomposition, while the middle countours represent the density of the cross dimensional decomposition.

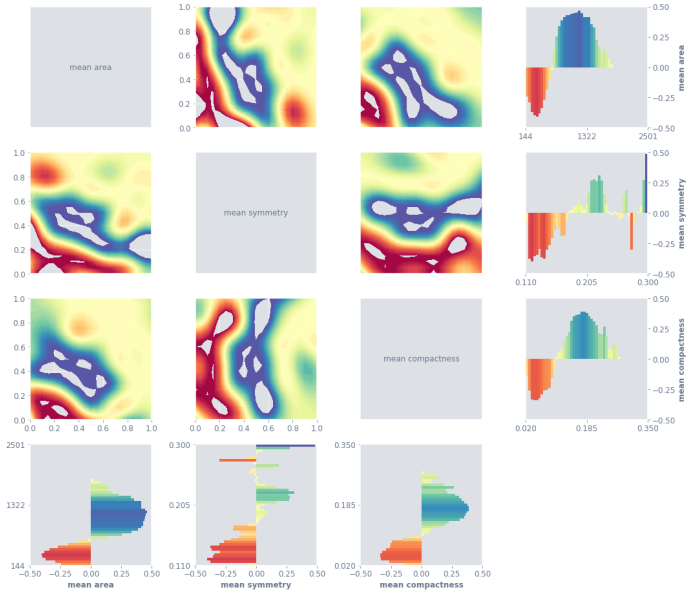

Figure 2: NN with OCSVM interpolation.

are widely regarded as more interpretable than neural networks []. The similarity of the two resultant visualizations is strong proof of the positive effect of LDR on model interpretability. The larger extreme value patches in the NN multi dimensional space are caused by the prediction density function interpreted by the final regression SVM being more steep to to the more binary classification of the SVM. This can be seen in the scatter plots of prediction certainties of the RF **??** compared to the NN 4.

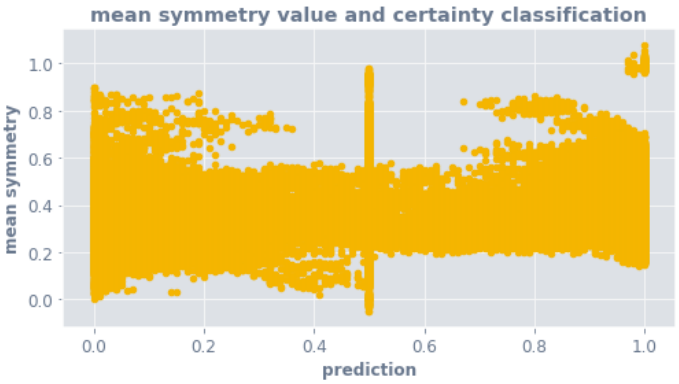

Figure 3: RF prediction certainty scatter plot of mean symmetry.

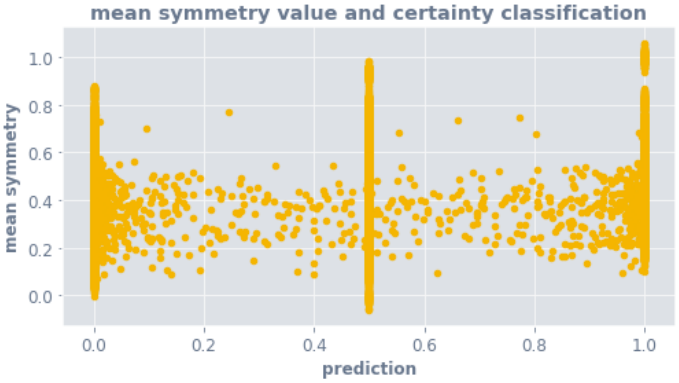

Figure 4: NN prediction certainty scatter plot of mean symmetry.

5 Regression Example

6 Mathematical Explanation

Let Ω be the total sample space of input data. Let $X \in \Omega$ describe an input vector of data of length n, where each $x_i \in [0,1]$ is the scaled value for feature i in the vector.

$$X = \{x_1, x_2, ..., x_n\}$$

Let Y be a single or multivariate output vector ($Y = y$ if single) where each y_i is a predicted variable.

$$Y = \{y_1, y_2, ..., y_n\}$$

Let $s \in [0,1]$ be a prediction of an outlier, s.t. 1 indicates an outlier and 0 not an outlier.

Let m describe a model as a function s.t. $m(X) = Y$, and o describe an outlier classifier function s.t. $o(X) = s$

$$m(X) = Y$$

$$o(X) = s$$

Let the outlier weighting and model prediction interpolation function f be defined as

$$f(X) = (m(X) - 0.5) \times o(X) + 0.5$$

Let the resolution of a KDE bandwidth be r. The KDE function g over a n samples can therefore be defined as

$$g(X) = \frac{1}{nr} \sum_{i=1}^{n} K \frac{x - x_i)}{h}$$

which is effectively an implementation of the VEGAS algorithm for the Monte Carlo Integration function I_Ω.

Let the set of points drawn from Ω be Q, where $q \in i * r_0^{1/r}$ while x_i are drawn from g. Q can therefore be described as

$$Q = x_1, x_2, ..., q, ..., x_n$$

One option for model certainty and outlier interpolation estimation is therefore given by

$$I_\Omega = \{ \frac{1}{n} \sum_{i=1}^{n} f(Q(X)) \}$$

While an alternative method is to use an SVM, or alternative predictive model, to model the function describing the dimensionality reduced space model certainty. Let this model be defined as h, resulting in

$$I_\Omega = h(Q(X))$$

7 Conclusion

This method will require either formal proof of efficacy or significant demonstration of reliable practical use before it should be used in the field. The ability to use this method on any dataset, removing the need for some visual exploratory work makes it a powerful tool.