

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

## *DRAFT*

*Elias Kassell, Fred Farrell*

September 2019

## 1 Abstract

LDR (Latent Dimensionality Reduction) is an algorithmic method for interpreting models by decomposing predictions from the original high dimensional feature space into a lower dimensional space. Both the output of the model and the certainty of its predictions are encoded in the condensed space. The method works out of the box on both classification and regression problems with input features relating to interpretable metrics. The output gives the ability to glimpse higher dimensional spaces, which are normally too complex for humans to understand, and is a powerful tool in understanding the underlying features generating the metrics.

## Acknowledgements

Thanks to Illumina® for supporting this research. The extremely high dimensional problems posed by genomic analysis provided the inspiration for this work.

## 2 Introduction

When training a model to make predictions in situations where interpretability is not required, the main goal is to maximise predictive accuracy [1]. Often black box models have a higher predictive accuracy than more interpretable solutions, particularly in higher dimensional or more complex problems [2]. Because of the uninterpretability of these models, they become impractical for situations where it is a necessity to understand why a model has chosen a particular output as its prediction, and what effect changes to values of the input features will have on the output of the model.

Some examples of uses of model interpretation in real world applications include

1. Understanding the quality of a systems current understanding [3].
2. Interpreting how the value of a feature, or subset of features, affects a model's prediction (henceforth referred to as "feature interpretation"). For

example, in a medical setting this interpretation can be used for suggesting the relation between features and a diagnosis. In another example, for manufacturing this interpretation can provide target values for the underlying processes involved in quality of produce.

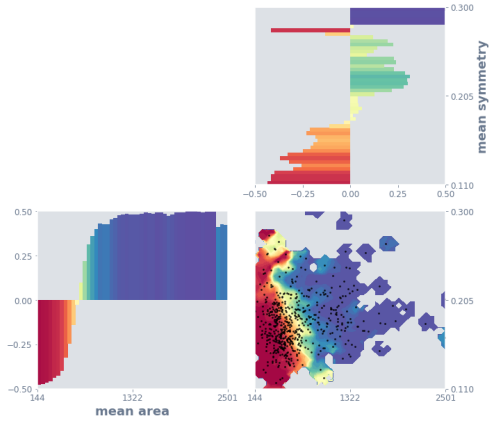
3. The ability to use a model when not all values for the input features are present. By integrating across features which are not present, the certainty of a models prediction can be estimated. Current methods, such as imputing values for features which are not present, ignores the variability of the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

## 3 The Algorithm

1. Normalize the original data so it falls into  $[0, 1]$  intervals.
2. Train a predictive model on the normalized data.
3. Train a OCSVM (One Class Support Vector Machine) [4], or an alternative model for detecting outliers, on the normalized data.
4. Create a KDE (kernel density estimate) [5] of the training points with a bandwidth equal to some resolution  $r$ . Optional: to improve the visualization, normalize the training to produce a constant density output.
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. Bin the samples according to regular intervals. For each dimension, group points with resolution  $r$ , reducing the value of the bin to the mean prediction across it.

## 4 Classification Example

The Wisconsin breast cancer dataset [6] (provided by scikit [7]) is used as it has 31 dimensions, a significant



**Figure 1:** RF prediction certainty, where red indicates certainty of malignancy, blue certainty of benality, and yellow a general lack of certainty. A lack of drawn contour indicates no samples were drawn from the sample space, as the kernel density was too low in those areas. The black dots are the locations of the samples. The top and left axis are the single dimensional decomposition, while the bottom right is the two dimensional decomposition. In the single dimension estimates, the resulting value is shifted down by 0.5 to create a flat midpoint for the bars.

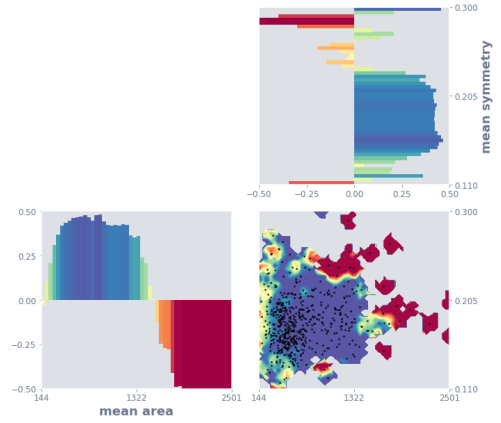
number. The aim with the dataset is to classify the tumor as either malignant or benign. There are 569 samples in total, with 357 benign and 212 malignant.

#### 4.1 Random Forest and OCSVM Interpolation

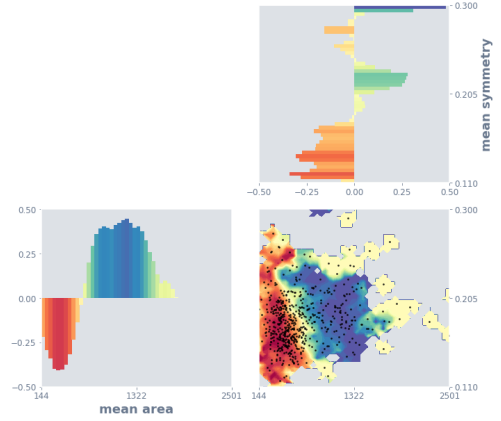
The 100 estimator RF (Random Forest) [8] achieved an F1 score of 0.975, while the OCSVM was trained to interpret 10% of training values as outliers. Benign was selected as the positive category, resulting in a value of 1.0 indicating benign tumors and 0.0 indicating malignant. All of the data is min-max scaled, then a 0.7/0.3 training/testing data split is applied. LDR is 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 symmetry and mean area are selected for visual inspection of their individual effect on the model's classification. The certainty of the RF is shown in figure 1, while the certainty of the OCSVM is shown in figure 2. The final combined certainty is shown in figure 3.

#### 4.2 Neural Network and OCSVM Interpolation

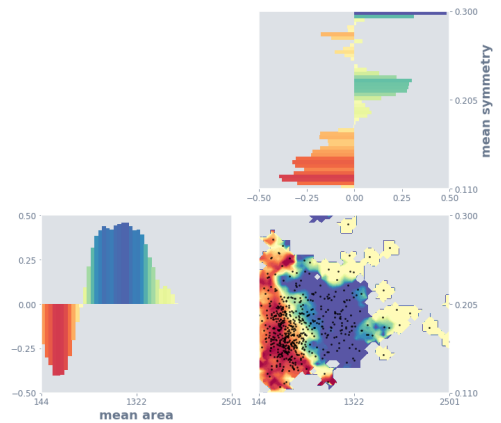
A 3 layer feed forward Neural Network (NN) was used (via pytorch [9]), consisting of an input layer, a hidden layer where tanh is applied, and an output layer. Softmax is applied when making predictions, and the final certainty of the prediction is calculated by calculating the proportion of the largest 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 resulting LDR of the same features can be seen in figure 4



**Figure 2:** OCSVM prediction certainty, where red indicates certainty of outlier, blue certainty of not outlier, and yellow uncertainty.



**Figure 3:** RF with OCSVM interpolation prediction certainty, where red indicates overall certainty of malignancy, blue indicates overall certainty of benality, and yellow general overall uncertainty.



**Figure 4:** NN with OCSVM interpolation prediction certainty.

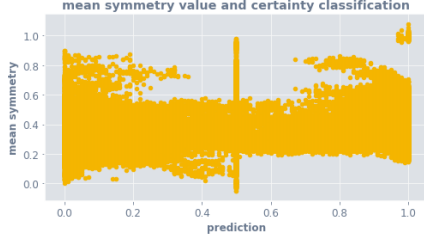


Figure 5: RF prediction certainty scatter plot of mean symmetry.

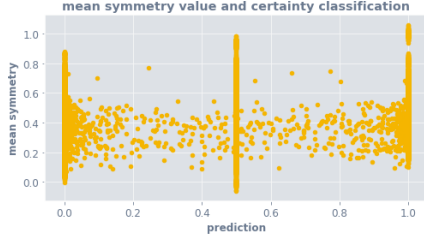


Figure 6: NN prediction certainty scatter plot of mean symmetry.

### 4.3 Analysis of Discrepancy Between RF and NN

There is very little discrepancy between the two LDR representations, both of which have similar predictive success. This is significant because random forests are widely regarded as more interpretable than neural networks [10], but LDR represents them both in an equally interpretable fashion. The only visual difference between the two LDR representation arises from the NN tending to be more certain of its classifications. This can be seen in the scatter plots of prediction certainties for mean symmetry of the RF in figure 5m compared to the NN in figure 6.

## 5 Regression Example

A similar method as with the classification was applied to the Boston House Price dataset [11] (also provided by scikit [7]). This dataset has 14 dimensions, where the target is to predict the house price based off of quantitative factors such as age of the house, the number of rooms, and the per capita crime rate of the time within which is situated. Some interesting relations are made immediately available from the LDR, shown in figure 7, are

1. An increasing crime rate tends to correlated with houses being worth less. Above the median level of crime however, the model becomes uncertain.
2. Any quantity of crime only has a negative effect on house price.
3. The lower the crime rate, the more rooms a house tends to have, and the more it is worth.
4. The older the house, the less it tends to be worth.

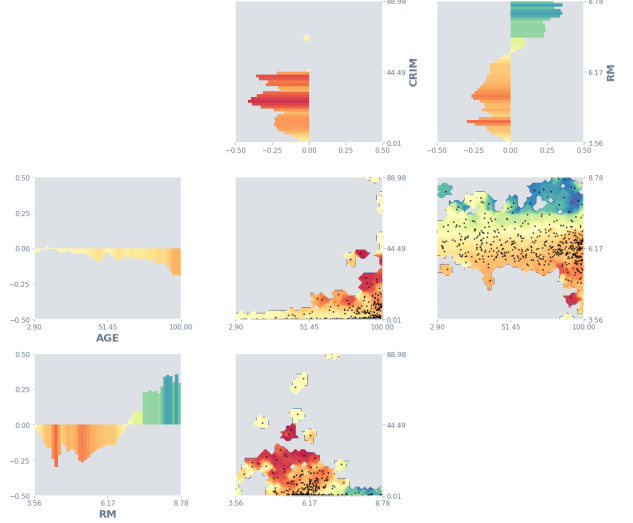


Figure 7: Regression RF with OCSVM interpolation. AGE is the age of the house, RM the number of rooms, and CRIM the per capita crime rate per town.

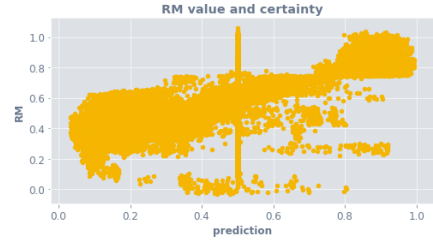


Figure 8: RF prediction certainty of number of rooms.

5. The age of the house and the number of rooms are for the most part independent.

The prediction certainty prior to be binning can be seen in figure 8.

## 6 Mathematical Explanation

Let  $\Omega$  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.

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

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

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

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  $I_\Omega$ .

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

$$Q = x_1, x_2, \dots, q, \dots, x_n$$

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

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

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

I have demonstrated that it is possible to interpret lower dimension subsets of high dimensional feature spaces by estimating the prediction of black box models trained on the feature spaces. However, before acceptance for clinical purposes, the model should be rigorously tested on more data sets. In addition to this, further research of extending LDR could be to interpret latent features that do not relate to uninterpretable input features, such as neural networks trained on images; a visualization of the prediction of the model based on a single pixel would not be of much use by itself.

## References

- [1] Leo Breiman. "Statistical Modeling: The Two Cultures". In: *Statistical Science* (Oct. 2001). DOI: <https://projecteuclid.org/euclid.ss/1009213726>.
- [2] Jurgen Schmidhuber. "Deep Learning in Neural Networks: An Overview". In: *Neural Networks* (Jan. 2014). DOI: 10.1016/j.neunet.2014.09.003.
- [3] Saleema Amershi et al. "Effective End-User Interaction with Machine Learning". In: *AAAI Conference on Artificial Intelligence* (Jan. 2001). DOI: [www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/viewPaper/3736](http://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/viewPaper/3736).
- [4] Bernhard Scholkopf et al. "Estimating the Support of a High-Dimensional Distribution". In: *Neural Comput.* 13.7 (July 2001), pp. 1443–1471. ISSN: 0899-7667. DOI: 10.1162/089976601750264965. URL: <https://doi.org/10.1162/089976601750264965>.
- [5] Emanuel Parzen. "On Estimation of a Probability Density Function and Mode". In: *Ann. Math. Statist.* (1962).
- [6] Olvi L. Mangasarian Dr. William H. Wolberg W. Nick Street. *Breast Cancer Wisconsin (Diagnostic) Data Set*. Nov. 1995. DOI: [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Diagnostic\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)).
- [7] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [8] Andy Liaw and Matthew Wiene. "Classification and Regression by randomForest". In: *R news* (Dec. 2002).
- [9] Adam Paszke et al. "Automatic Differentiation in PyTorch". In: *NIPS Autodiff Workshop*. 2017.
- [10] Peter Langfelder Lin Song and Steve Horvath. "Random generalized linear model: a highly accurate and interpretable ensemble predictor". In: *BMC Bioinformatics* volume (Jan. 2013).
- [11] D. Harrison and D.L. Rubinfeld. *The Boston house-price data*. Nov. 1978. DOI: <http://lib.stat.cmu.edu/datasets/boston>.