

Task 2: Modelling Report

Date: 17/11/2017

By K.M. Dinake

Model Motivation

I decided to utilize the Multi-Layer Perceptron Classifier (MLPC) because of its ability to learn non-linear models between the inputs and the outputs.

Model Performance

For this model's performance metric I have utilized the following:

1. During Training: I used the MLPC built in score metric.
2. During Testing: I used scikit learn's recall score performance metric for testing.

The MLPC model was trained using 190820 samples, after which it produced a score of: 0.999457371764.

During testing, the model used 93987 samples and was able to recall with an accuracy score of: 0.802469135802.

Model Improvements

Given that the dataset was imbalanced as it was skewed in favour of the more dominant class, several methods could be performed to improve the model's results.

1. The model itself could be placed through a hyper parameter tuning phase where arguments such as alpha can be optimized to avoid under- and overfitting. Also, playing around with the number of hidden layers, weight initializations can impact the performance of the model substantially.
2. As a second approach the data could be resampled such that the data has a 50/50 ratio between the classes. This can be done through either under-sampling or over-sampling the data, the usage of either depends on how much data we have. Over-sampling is chosen when the amount of data you have is limited, and under-sampling is preferred when you have a lot of data.
3. The most simple improvement is to collect more data to reduce the imbalance in the data.
4. Perhaps other classifiers or regressors can outperform the Multi-layer Perceptron classifier, so comparing their performance may work.