University of Coimbra

Pattern Recognition

Default of credit card clients

Author:
António Lima
2011166926

Author:
Pedro Janeiro
2012143629

April 22, 2016

Contents

1 Data Normalization

One of the first things we can assess about the provided dataset is that each of its feature dimensions uses different value distributions, making it harder to analyse potentially interesting variables of samples. As such, we first offer the possibility to normalize the data in of each dimension.

2 Feature Selection Reduction

Since the initial dataset boasts some 23 features for each of the 30 000 samples provided for the expected binary classification, it becomes apparent that some, if not most, of those 23 features might be useless.

2.1 Feature Selection

As far as selecting features goes, we provide the means to analyse the correlation and covariance between every pair of features to remove redundant variables, analyse the correlation and covariance of each feature with regard to the expected output and, last but not least, perform a Kruskal-Wallis test to assess the p-score of each feature. Any combination of these three feature selection methods can be chosen and, we've found that only 13 to 17 of the features yield any relevant or non-redundant information.

2.2 Feature Reduction

Regarding feature reduction, we offer the possibility of performing a Primary Component Analysis (PCA) or Linear Discriminant Analysis (LDA), combinations included. For PCA we also provide the possibility of selecting the most relevant primary components based on either a Skree test or the Kaiser criteria, both for varying thresholds.

3 Classification

3.1 Data Splitting

Since the provided dataset is not uniformly distributed (and probably couldn't be without increasing exponentially in size) we first shuffle the samples randomly, following with a stratifying split. What this means is that we guarantee that for a certain splitting threshold (70% for training and 30% for testing by default) the training set will indeed have X% of each class represented and vice-versa for the testing set. These Thresholds can be customized.

3.2 Classifiers

The following classifiers are available in our simulator:

- Matlab's Linear Discriminant Classifier
- Minimum Eucledian Distance Classifier
- Minimum Mahalanobis Distance Classifier

4 Classifier Performance Analysis

Looking at the accuracy, prevalence, sensitivity and specificity of our the predicted and expected results we can get a good feel of how each specific simulation performs compared ot the other ones.

5 To be implemented

- Run 30 of each simulation and record the average results.
- Experiment with various variables and combinations of available segments. For exemple, see how normalized data performs against unormalized data; how the Kaiser Criteria compares to the Skree test or neither, etc.
- Time each simulation.