Classification in health data

Kunal Sharma

October 30, 2019

Given: A data set with 1994 and 55 columns is given where each data point corresponds to a particular label. The task is to build a model to perform classification.

1 Assumptions & Preparations

- The original data set has a lot of missing data, addressing which becomes the first challenge.
- After some initial sanity checks, a short computation reveals that 23 features have less than 15% data in them.
- So they are removed, while the remaining features still have about 70% of data missing.
- To deal with this, imputation through MICE implemented with 65 iterations. It essentially performs linear regression on each of the data's columns.
- Alternatively, a maximum likelihood estimation could have been performed to complete the data set. Even more so, a statistical model could be build for each feature which could be then used to replace the missing values.
- Our approach makes sure to eliminate most of the bias and error for the missing values.
- A standing assumption made here is that the missing values are 'Missing at Random.'

2 Modeling

- Given its health data with lots of parameters on blood, fat etc we first study the possible correlation. To this, we find that most correlated features are 'OneKHz' and 'FourKHz' and the latter is removed to prevent issues.

- Following this, we begin to test the linearity of the class boundary with SVM. To measure the misclassification rate, we mainly rely on 'recall'. Moreover, as there are more than two classes, recall is calculated using *micro*.
- Here 'micro' is preferred than 'macro' as the former respects that class imbalances much better.
- The recall score obtained with linear SVM is 0.44 for the test data.
- Its interesting to point out that Linear Discriminant Analysis also produces linear boundaries and works well with multi classes. However, it assumes Gaussian densities and more importantly SVM optimized the *margin width* so that if the class labels represent a disease, SVM is preferred.
- For non-linear models, we try SVM with Polynomial kernel and Gaussian kernel. Here polynomial with 3rd degree is chosen and it provides a recall score of 0.525 for test data.
- To capture the possible complexity of the data, ensemble methods are also tried, namely Adaboost, Random Forest and Gradient Boosted Classifiers.
- Here the GBRTs, with 'staged_predict' perform much better giving recall score of 0.495.
- Finally, analysis of important features that contribute significantly to model reveals 'NeutralFat', 'Hmeatorcrit' and 'Hemoglobin' as the top three ones.
- In conclusion, the modeling here is heavily dependent on the data filling method and the metric for misclassification is recall which at best produces rate of approximately 53%. The measure of goodness here depends on the interpretation of class label. However, given health data it was chosen over precision.
- Other models that could be implemented here are 'OneVsOneClassifier' or 'OneVsRestClassifier'.