Machine Learning Nano Degree Project 2

Kai Xiong

December 2015

1 Introduction

In this report, we analyse student intervention dataset. We extract 30 features to identify if a student needs early intervention. Thus this is a classification problem due to its binary outcome.

2 Statistical Analysis and Data Exploration

The general statistics of the Boston house price dataset is summarized as follows:

Statistics	Value
Total number of students	395
Number of students who passed	265
Number of students who failed	130
Graduation rate of the class (%)	67%
Number of features (excluding the label/target column)	30

Table 1: General statistics

We observe that the standard deviation is relatively high, which is about half of the mean.

3 Preparing the Data

We use all columns but the last column as features. The last column which denotes if a student pass or fail is the target column. There are a number of columns that have non-numeric values. For the columns only have "Yes" or "No", we convert them to 0/1 (binary) values. For other columns that have more than two values, we create as many columns as possible values and assign 1 to one of them and 0 to others. We then shuffle the data and split them into training and test sets. We use 300 data points, which is about 75% of the data, as the training set and others as the test set.

4 Training and Evaluating Models

We have used three supervised learning models, namely, SVM, Naive Bayesian, Boosting classifier to solve this problem.

SVM is a popular model that can be applied to classification or regression problems. It uses kernel to transform data into higher dimensions and then separate them using a margin as wide as possible in the higher dimensional feature space. New data points are also casted to the high dimensional space, where they are more easy to be separated. Therefore, SVM is good at classifying both linear and non-linear problems. Nonetheless, the disadvantages are that SVM only fits to a certain regularization parameter, kernel model and kernel parameters. The algorithm only optimizes over certain scenarios. It is hard for it to explore all kernel methods and regularization parameters.

We also apply Naive Bayesian classifier to the problem. NBC is part of the probabilistic classifiers based on Bayes theory with the assumption that all features are conditionally independent of each other. NBC is very efficient in linear classification problems. And it only requires small amount of data to train the model. The disadvantage of this algorithm is that its precision and recall are low when using a small dataset.

Boosting is an ensamble learning algorithm primarily for reducing bias and variance. It is known for converting weak classifiers to strong classifiers. Usually, it weights multiple realizations of classifiers and sum them up to get a more accurate results. In this project, Gradient boosting algorithm is applied as it allows optiomization of an arbitrary differentiable loss function. The disadvantage of Gradient boosting is that its training time is genarally longer as it builds tree models sequentially.

Performance of the supervised learning algorithms are presented below. From the tables, we see that Naive bayesian is the most time efficient algorithm. However, its F1 score on training set and test set is the lowest. SVM has intermediate speed and performance. It does well in both training set and test set. Boosting is the slowest algorithm. It has the highest score in the training set however, performance drops significantly in the test set. The

Training set sizes	Training time	prediction time	F1 score on training set	F1 score to test set
100	0.002	0.005	0.8481	0.8115
200	0.005	0.004	0.8868	0.8013
300	0.007	0.003	0.8761	0.7838

Table 2: Statistics for SVM

Training set sizes	Training time	prediction time	F1 score on training set	F1 score to test set
100	0.001	0.003	0.8308	0.6935
200	0.004	0.005	0.8102	0.7472
300	0.002	0.001	0.8037	0.7633

Table 3: Statistics for Naive Bayes

Training set sizes	Training time	prediction time	F1 score on training set	F1 score to test set
100	0.090	0.003	1	0.7335
200	0.112	0.001	0.9929	0.7735
300	0.150	0.002	0.9739	0.7794

Table 4: Statistics for Boosting

5 Best model

Boosting has high computational cost and may risk overfitting the data. Naive Bayesian has poor accuracy. Conclusively, SVM is the optimal classifier in terms of accuracy and speed as it looks for the max margin and deals with non linear data well.
The best parameter is with rbf kernel and with C equal to 1. The F1 score for the training set is 0.8760 and the F1 score for the test set is 0.7838. The prediction time is 0.003 for test set. Thus we see SVM works well both in sample and out of sample. And the performance in the test set is significantly improved.
In Layman terms, the Gradient boosting algorithm achieves its goal sequentially. Whenever new data point comes in, it includes it in the new model and optimise the problem over all data achieved so far. Thus, it solves the problem sequentially, in each step, it gets an optimal solution and then generalizes the local optimal solution to global optimal solution.

6 Conclusion

In this report, we study basic statistics of the boston student intervention dataset. We compare SVM, Naive Bayes and Boosting learning algorithm and find Gradient boosting is the optimal solution to this problem. We then use grid search to tune the model. Results show that Gradient boosting works well both in sample and out of sample.