CSCI-P556
Fall 2018
Assignment 3
Due 11:59PM, Nov. 2, 2018

Hasika Mahtta (hmahtta)

November 5, 2018

1 Introduction

We are being given the following four data files:

1. a3-train.data
2. a3-train.labels
3. a3-test.data
4. a3-test.labels

There are 2,000 training rows and 600 test rows, each row has 500 features, and there are only two labels, -1 and 1. The following tasks have been performed for the given data set. Firstly, exploratory data analysis has been done to get a better understanding of the data. Thereafter, Baseline models help us to know how the model will perform with the default parameters. And then we proceed on to feature engineering and tuning the models by changing the parameters to get better accuracy on the test set.

2 Exploratory Data Analysis

The first step in the machine learning pipeline is taking a look at the data to make sure that you understand what you are working with. This step includes tasks such as checking for unique identifier columns that have to be discarded, making sure there are no missing or weird values, are the labels balanced, etc.

1. Reading the train and test data - Firstly we read the train and test data sets using pandas.
2. Sampling the data - If we have a large data set ,we can take a sample of the data as an easy way to look at the data quickly. sample() function that is included in Pandas has been used to view both the train and test data samples.
3. Next step has been done to read the train and test labels.
4. Data Information - Info() function in the pandas data frame has been used which prints the information about the training data set including the index dtype and column dtypes, non-null values and memory usage.
5. Describing the data - The describe() function has been used to get the various summary statistics that excluded NaN values. This function returns the count, mean, standard deviation, minimum and maximum values and the quantiles of the test and train data.

6. Identifying missing values - isnull() function helps in identifying the missing values. In the given data set, three are no missing values.
7. Checking if the dataset is balanced or not and we see that the given dataset is balanced.

3 Baseline Models

The following baseline model with the default parameters have been used to get an idea of the minimum performance that can be achieved before performing feature engineering and optimizing the model's hyperparameters.

1. **Logistic Regression** - This Model performs regression analysis when the dependent variable is binary. In this case we are classifying into two labels -1 and +1. Using logistic regression with default parameters, train accuracy comes out to be 74% and test accuracy is 59% on the given data set.
2. **K- Nearest Neighbor** - KNN is a non-parametric data classification algorithm that estimates how likely is a data point is to be a member of one group or the other depending on what group the data points nearest to it are in. Using KNN with default parameter where K = 5, the train accuracy is 82% and test accuracy is 69% on the given data set.
3. **Random Forest Classifier** - Random forests or random decision trees are an ensemble learning method for classification and regression. It gives a great result even without hyper-parameter tuning. Random Forest Classifier with default parameters gives 100% train accuracy and 70% test accuracy on the given data set.

4 Feature Engineering

The feature engineering function returns a list of features that have been used in the models to achieve higher accuracy on the test data sets. First , we find the correlation among the features and set a threshold according to which the function will remove one of the pairs of the features with a correlation greater than this value. The next step is to find the correlation between each of the remaining features with the target labels and keep the important features which are highly correlated with the target labels. The correlation function returns 10 features which have been further used in the model building.

5 Model Building

1. **K-Nearest Neighbor** - A graph has been plotted to get the maximum k value (33) that gives the highest test accuracy by running a loop for k values from 5 to 40. And then set this parameter in the model on the transformed data sets to check the accuracy of the model. This gives 82% train accuracy and 88% test accuracy.
2. **Random Forest Classifier** - This model on the transformed data gives 100% train accuracy and 88% test accuracy.
3. **Decision Tree using Gradient Boosting and Adaboost Ensemble** Adaboost classifier gives 100% train accuracy and 91% test accuracy using decision trees. Whereas Gradient Boosting gives 100% train accuracy and 86% test accuracy.

6 Discussion

1. Logistic Regression - This model even after feature engineering does not reflect any changes in the test accuracy and is 59%. This does not gives a good as performance expected.

2. KNN - The test accuracy improves from 69% to 88% after model building using transformed data sets.
3. Random Forest Classifier - The test accuracy improves from 70% to 80% after model building using transformed data sets.
4. Decision Trees using Ensemble Classifier - Adaboost gives better test accuracy of 91% as compared to gradient boosting 86%.
5. Challenges - Feature enginneeing was a challenge. Finding the best possible features in a large data set without the domain knowledge is challenging. Accuracy for logistic regression and support vector machines does not improve even after feature engineering.
6. The accuracy can be improved if we further try to tune the hyper-parameters.
7. The best Model - Decision Trees with Adaboost Ensemble gives the highest test accuracy of 91%. Second highest test accuracy is obtained in KNN and random forest classifer with 88% accuracy.