Report for Assignment 5

Group members:
2013A7PS152H – Simran Kapur
2013A7PS162H – Kinjal Jain
2013AAPS263H – Arth C Patel
2013A7PS387H – Aditya Chivukula

Submitted by (CMS): Aditya

Objective:
Comparison of multiple Machine Learning techniques on same dataset. The propsed methods to be used are Logistic Regression, Artificial Neural Networks, Decision Trees, Naive Bayes.

Dataset used : OnlineNewsPopularity.csv
The dataset contains several attributes on articles published by the media. The target variable is the number of times people have shared the article. We can binarize the target by setting a threshold value making it possible to apply classfication algorithms in place of continuous prediction algorithms.

Dimensions: 39644 * 58
Target: Popular (shares>1400) / Not popular (shares < 1400)

First 30,000 rows were used as train subset.
Next 9644 rows were used as test subset.

Method 1: Logistic Regression
Code written in Octave.
Functions for cost function using sigmoid activation function used to calculate total regularized cost and gradient function as well. Cost function optimised using fminunc.

Accuracy: Train – 59.641988%
Test – 62.782501%
Precision: Train – 50.751692%
Test – 39.508605%
Recall: Train – 60.206888%
Test – 52.576358%

Method 2: Neural Networks (15 hidden Units)
Code written in R.
Library used: nnet function in package “nnet”

A neural net with 15 hidden layers was trained with a decay factor of 1. This value was determined after multiple attempts by trial and error to produce highest train/test accuracy. Due to the random nature of the initial weights of the network. The program had to be seeded to replicate the best results.

Accuracy: Train – 59.536667%
Test – 61.758606%
Precision: Train – 67.48%
Test – 58.554541%
Recall: Train – 76.050805%
Test – 71.152052%

Method 3: Decision Tree
Code written in R.
Library used: rpart function in package “rpart”

A recursive partitioning method was used to learn a Decision Tree on the dataset. We also tried implementing Random Forests, but it didn't provide significant improvement without implementing complex tuning factors.

Accuracy:	Train – 62.220000%
	Test – 63.521360%
Precision:	Train – 48.520000%
	Test – 48.776441%
Recall:	Train – 60.547335%
	Test - 63.829291%

Method 4: Naive Bayes
Code written in R.
Library used: naiveBayes function in package “e1071”
naiveBayes classifier was trained using naiveBayes function used in earlier in first assignment. Laplace smoothin factor again did not result in changes in the conditional probability tables. Needs to investigated further.

Accuracy:	Train – 63.786324%
	Test – 62.874444%
Precision:	Train – 59.338719%
	Test – 61.787865%
Recall:	Train – 58.760000%
	Test – 62.891422%

Conclusion:
Decision Tree provided the best accuracy on the dataset. However it is a marginal benefit over the other algorithms. Therefore it can not be said that it is the best algorithm. Further fine-tuning of learning parameters may result in better accuracy by the algorithms.