CS 412 – MP4

pvijaya2

Adult dataset

NaiiveBayes

pauljv@paul-PC:~/Documents/coursework/fall14/cs412$ python NaiveBayes.py adult.train adult.test
The accuracy is 0.783800623053
356 308 39 902
The accuracy is 0.769802300039
6401 6081 1045 17429

The Naive Bayes accuracy is: 76.9802300039% on the testing set. The error rate is 100% - the accuracy rate.

TP: 6401
FN: 6081
FP: 1045
TN: 17429

Sensitivity: 0.51281845858
Specificity: 0.943434015373
Precision : 0.859656191244
Recall : 0.51281845858
FBeta_0.5 = 0.757228031988
FBeta_1 = 0.642412685668
FBeta_2 = 0.557831073308

Adaboost

The optimal size classifier would have 81.0591900312% accuracy on the training set 77.30792092% accuracy on the testing set and a k-value of 2. The error rate is 100% - the accuracy rate.

TP: 5635
FN: 4904
FP: 1811
TN: 18606

Sensitivity: 0.534680709745
Specificity: 0.911299407357
Precision : 0.756782164921
Recall : 0.534680709745
FBeta_0.5 = 0.69873273318
FBeta_1 = 0.626633305532
FBeta_2 = 0.568021450748

Poker dataset

NaiiveBayes

:~/Documents/coursework/fall14/cs412$ python NaiveBayes.py poker.train poker.test
The accuracy is 0.609404990403
449 108 298 186
The accuracy is 0.513274336283
242 113 217 106

The Naive Bayes accuracy is: 51.3274336283% on the testing set. The error rate is 100% - the accuracy rate.

TP: 242
FN: 113
FP: 217
TN: 106

Sensitivity: 0.681690140845
Specificity: 0.328173374613
Precision : 0.527233115468
Recall : 0.681690140845
FBeta_0.5 = 0.552259242355
FBeta_1 = 0.594594594595
FBeta_2 = 0.643959552954

Adaboost

The optimal size classifier would have 67.4664107486% accuracy on the training set 55.8997050147% accuracy on the testing set and a k-value of 70

TP: 310
FN: 150
FP: 149
TN: 69

Sensitivity: 0.673913043478
Specificity: 0.316513761468
Precision : 0.675381263617
Recall : 0.673913043478
FBeta_0.5 = 0.675087108014
FBeta_1 = 0.674646354733
FBeta_2 = 0.674206176599

Breast Cancer dataset

Naives Bayes

pauljv@paul-PC:~/Documents/coursework/fall14/cs412$ python NaiveBayes.py breast_cancer.train breast_cancer.test
The accuracy is 0.705555555556
37 34 19 90
The accuracy is 0.688679245283
18 22 11 55

The Naive Bayes accuracy is: 68.8679245283% on the testing set. The error rate is 100% - the accuracy rate.

TP: 18
FN: 22
FP: 11
TN: 55

Sensitivity: 0.45
Specificity: 0.833333333333
Precision : 0.620689655172
Recall : 0.45
FBeta_0.5 = 0.576923076923
FBeta_1 = 0.521739130435
FBeta_2 = 0.47619047619

Adaboost

The optimal size classifier would have 74.4444444444% accuracy on the training set 73.5849056604% accuracy on the testing set and a k-value of 21. The error rate is 100% - the accuracy rate.

TP: 19
FN: 18
FP: 10
TN: 59

Sensitivity: 0.513513513514
Specificity: 0.855072463768
Precision : 0.655172413793
Recall : 0.513513513514
FBeta_0.5 = 0.62091503268
FBeta_1 = 0.575757575758
FBeta_2 = 0.536723163842

LED dataset

Naives Bayes

:~/Documents/coursework/fall14/cs412$ python NaiveBayes.py led.train led.test

The accuracy is 0.765692381409
468 319 170 1130
The accuracy is 0.753968253968
253 181 98 602

The Naive Bayes accuracy is: 75.3968253968% on the testing set. The error rate is 100% - the accuracy rate.

TP: 253
FN: 181
FP: 98
TN: 602

Sensitivity: 0.582949308756
Specificity: 0.86
Precision : 0.720797720798
Recall : 0.582949308756
FBeta_0.5 = 0.688248095756
FBeta_1 = 0.644585987261
FBeta_2 = 0.606133205558

Adaboost

The optimal size classifier would have 76.9046478198% accuracy on the training set 77.0723104056% accuracy on the testing set and a k-value of 2. The error rate is 100% - the accuracy rate.

TP: 281
FN: 190
FP: 70
TN: 593

Sensitivity: 0.596602972399
Specificity: 0.894419306184
Precision : 0.80056980057
Recall : 0.596602972399
FBeta_0.5 = 0.749333333333
FBeta_1 = 0.683698296837
FBeta_2 = 0.628635346756

Observations and Analysis

For this MP, I noticed that the Adaboost algorithm would affect the accuracy differently based on the value of k (number of classifiers in ensemble) and therefore, I designed a test in testmain.py to select k based on the k that gives the best accuracy for each dataset. As you can see, in the results, the ks tend to vary for each dataset and I will analyze this more later in the report.

It is difficult to comprehensively evaluate the performance between the two algorithms, given that the parameters and performance metrics change quite a bit between each dataset. However, we can see that if we do parameter estimation for k like I have done, Adaboost actually performs better in every dataset and there tends to be at least a 2% increase in net accuracy.

Precision and the F1-score has improved in every dataset except the Adult dataset, even though the accuracy boost is shown to have increased for that dataset. The F2-score shows an increase because the recall of the dataset has improved through Adaboost. But as can be noted, while Adaboost is a kind of ensemble technique that improves the accuracy through an increased model complexity, there are no guarantees as to how it would perform on a test dataset.

Due to the increased complexity, there might also be a tendency for the model to overfit and therefore, perform slightly worse on the test set than on the training set. This can seen especially with the Adult dataset as the accuracy on the training data is 81.05% but drops steeply to 77.3% on the test set. For complex ensemble methods like Adaboost, whereas datasets that have a larger amount of data tend to be able to utilize the complexity of the ensemble methods to be able to produce better accuracy, so given that the model for the Adult dataset is already complex with significantly more attributes (113 more than the second-most complex dataset), it might require more data to be able to improve the accuracy further.

Another interesting note, was that I realised that Naive Bayes was not exactly a weak learner (a learner that will obtain at least 50% on a given dataset). This is because it works through an independence assumption of the random variables given the class, which may not always hold on some of the data. These data points are what the Naive Bayes algorithm will perform poorly on and they are the datapoints that will increase in weightage the more classifiers are trained in the Adaboost ensemble and therefore, future classifiers will be trained more regularly on those examples and will often obtain lesser than 50% accuracy on the dataset results in many poor learner which will be added to the ensemble. I, in fact, checked this by printing out the accuracy of each individual classifier and it turned out to be true. Therefore, the more complex datasets like the Adult dataset does horribly with Adaboost when trained on more than 3 classifiers with more than a 10% decrease in accuracy, whereas the other simpler datasets (which might have a better chance of meeting the independence assumptions) like the poker and breast cancer datasets, have a much higher k-value of 70 and 21 respectively. I think this was a very important learning from this assignment.

Therefore, we can determine that it may not always be the case, that the more complex or advanced the algorithm, the better the performance. There are numerous other factors such as dataset size and the actually complexity (number of attributes) of the data that better suits ensemble methods like Adaboost and in some simple scenarios Naiive Bayes alone may be the better model. But with the right parameter estimation technique, Adaboost could still be designed to perform better.