Summary Homework 1-Machine Learning Saket Vishwasrao

Performance results

1. Information gain takes on an average 5 minutes to execute for the given news dataset on my machine (intel core i5-470m, 2GB RAM) which is quite slow.
2. Naïve Bayes Prediction method gives an accuracy of 60.89 % using the “trainData, trainLabels” for training and “testData,testLabels” for testing.
3. For a decision tree of depth 3, the decision tree algorithm is giving an accuracy of 15.25 percent. Running the algorithm for larger depths is infeasible as the approximate average running time is atleast

T= informationgain Time\*no. of nodes= 5\*(2depth-1) minutes

Which grows above and hour of execution time when my machine was hanging for depth greater than 5

1. Execution of decision tree by taking only first 5000 features gives an accuracy of 17% for depth=5 and accuracy of 24.8% for depth = 10.
2. Evaluating naïveBayes on first 5000 features with alpha=0 gives accuracy of 24.33%.
3. Evaluating Crossvalidation function is not feasible due to too much time as well as plotting variation with parameters.

Note: All the functions are verified for proper functioning on dataset of size 100, hence I am sure about the correctness of the code.

Observations:

1. From (4) and (5) the performance of both algorithms is nearly the same
2. In general I feel naïveBayes should work better than decision trees because the data set provided is sparse. Because of sparseness there isn’t significant information gain in the decision tree when we classify it on a particular feature and hence it would require decision trees of larger depth for achieving greater accuracy.