Lonari, Nikit
1001108271

Q.2c

For the given dataset, Single classifier

1) tested on training dataset(100 examples)
   39 training examples are perfectly classified as ‘e’
   37 training examples are perfectly classified as ‘p’
   11 training examples are misclassified as ‘p’
   13 training examples are misclassified as ‘e’
   This gives us an accuracy 88% on training dataset examples.
   error_rate_on_train_data = (11+23)/100=0.34

2)tested on test dataset(50 examples)
   21 training examples are perfectly classified as ‘e’
   17 training examples are perfectly classified as ‘p’
   4 training examples are misclassified as ‘p’
   8 training examples are misclassified as ‘e’
   This gives us an accuracy 76% on training dataset examples.
   error_rate_on_test_data = (4+8)/50=0.24

In my code, 1-> e
2-> p
-1->can’t be classified.

However, we can also see that there is an overfitting problem because the decision tree learner continues to predict the class inspite of having an unknown attribute value(predictor).
Fortunately, we can use pre-pruning or post-pruning technique to stop growth of the tree wherever necessary or pos-prune it to rightly classify the given example.

Q.3b

1) For 10 different tree classifiers, using majority vote
   on train data :
      40 training examples are perfectly classified as ‘e’
      50 training examples are perfectly classified as ‘p’
      10 training examples are misclassified as ‘p’
      error_rate_on_train_data = (10+0)/100=0.1

   on test data :
      10 training examples are perfectly classified as ‘e’
      25 training examples are perfectly classified as ‘p’
      15 training examples are misclassified as ‘p’
      error_rate_on_test_data = (15+0)/50=0.3

2) For 50 different tree classifiers, using majority vote
   on train data :
      40 training examples are perfectly classified as ‘e’
      49 training examples are perfectly classified as ‘p’
      10 training examples are misclassified as ‘p’
      1 training examples are misclassified as ‘e’
      error_rate_on_train_data = (10+1)/100=0.11

on test data :

11 training examples are perfectly classified as ‘e’
25 training examples are perfectly classified as ‘p’
14 training examples are misclassified as ‘p’
error_rate_on_test_data = (14+0)/50=0.28

3) For 50 different tree classifiers, using majority vote

on train data :

39 training examples are perfectly classified as ‘e’
49 training examples are perfectly classified as ‘p’
11 training examples are misclassified as ‘p’
1 training examples are misclassified as ‘e’
error_rate_on_train_data = (11+1)/100=0.12

on test data :

11 training examples are perfectly classified as ‘e’
25 training examples are perfectly classified as ‘p’
14 training examples are misclassified as ‘p’
error_rate_on_test_data = (14+0)/50=0.28

Classifier	train	test
1	0.34	0.24
10	0.11	0.3
50	0.12	0.28
100	0.12	0.28

Since, I have randomly predicted the class for unknown attribute value like k,c, etc. to show the overfitting problem in this decision tree classifier. So the error rate on test data is low as compared to other multiple classifiers.
But, we can compare the error rates on train data. Due to bagging we have significantly decreased the error_rate on train data. Achieving approx. **90% accuracy.**