Problem 2

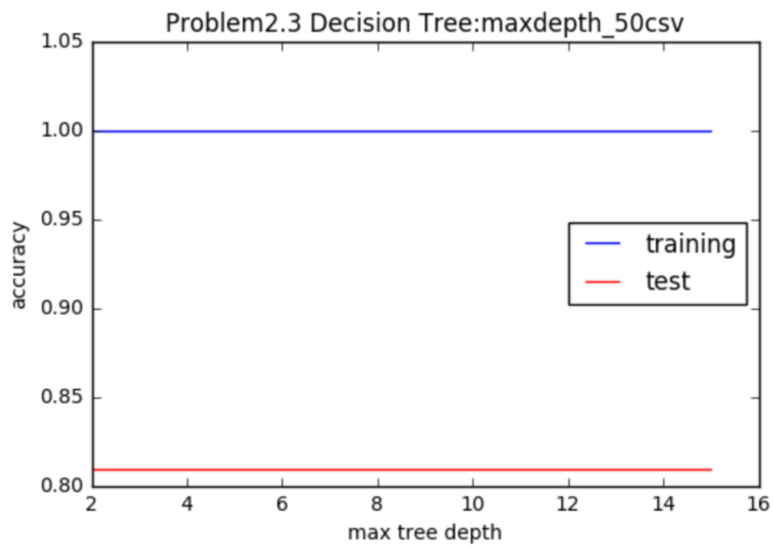
2. Accuracy of Gaussian Naïve Bayes over 3 datasets:

-50.csv: 0.890173410405

-200.csv: 0.964769647696

-400.csv: 0.96449704142

3.



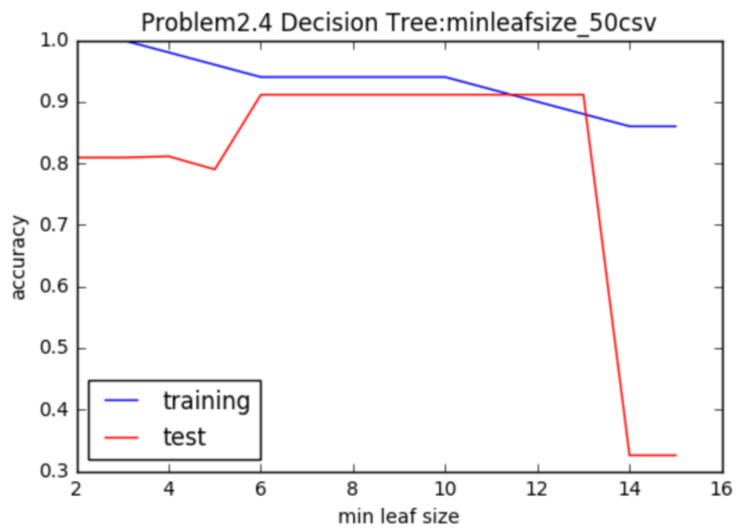
Accuracy of Decision Tree Classifier over -50.csv dataset:

0.809248554913

Gaussian Naïve Bayes has higher test accuracy over -50.csv dataset (0.890173410405) than Decision Tree (0.809248554913). Thus, Naïve Bayes has a better test error.

4.

-50.csv:

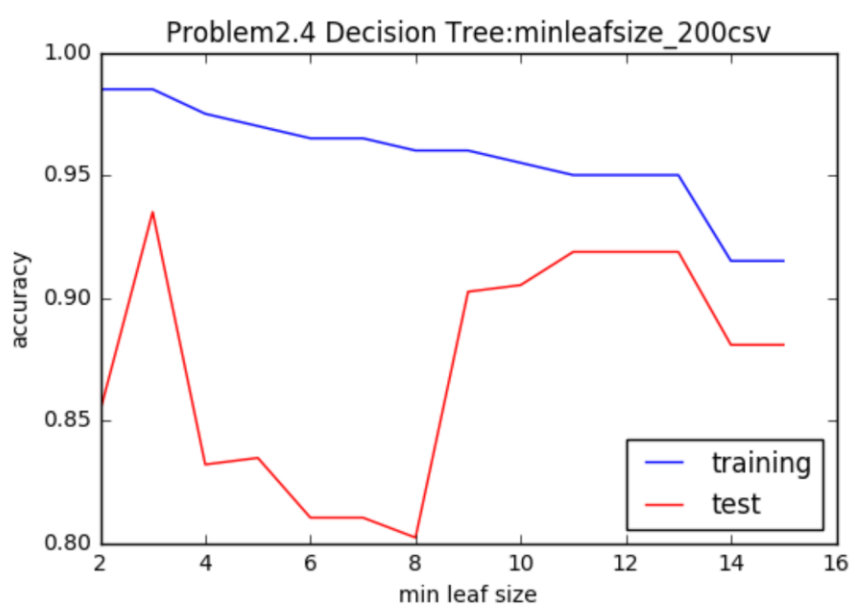


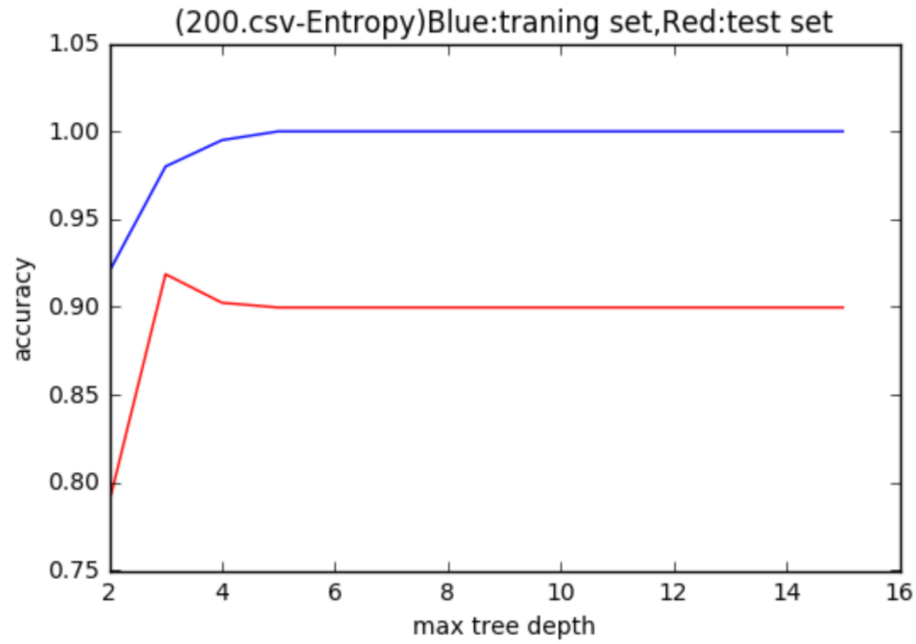
When $6 \leq$ minimum leaf size ≤ 13, the decision tree archives a higher accuracy compare to the maximum depth criterion and the Naïve Bayes.

-200.csv:

Compared to maximum depth criterion, when minimum leaf size is 3 or [9, 13], the decision tree archives a higher accuracy.

While obviously, the accuracy is lower than that of Naïve Bayes (0.964769647696). Thus, Decision tree with minimum leaf size doesn't perform better than Naïve Bayes.

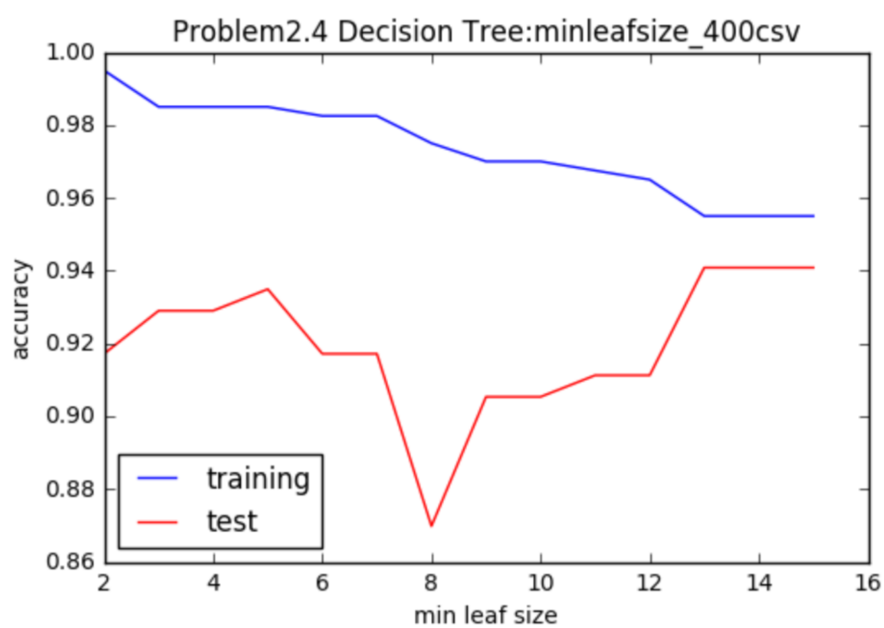


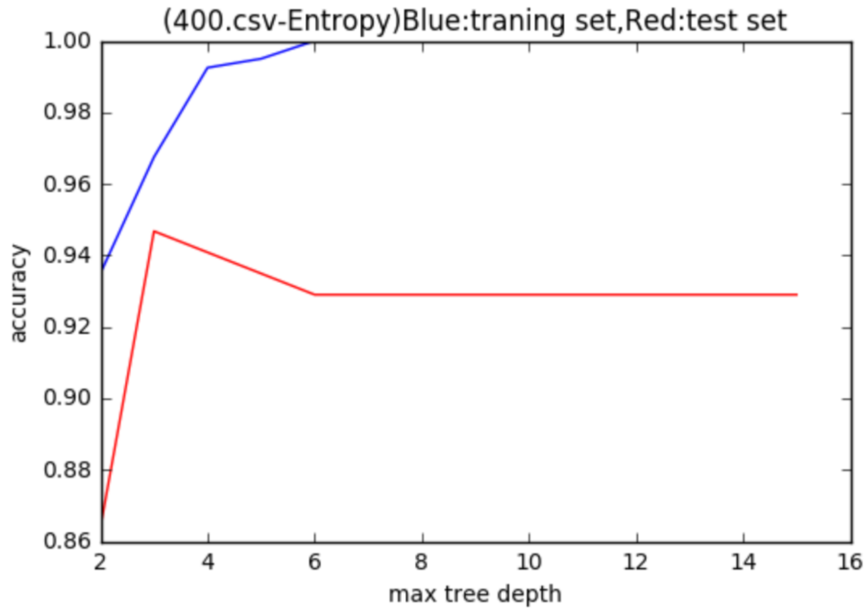


-400.csv:

Compared to maximum depth criterion, when minimum leaf size is among [13, 15], the decision tree archives a higher accuracy.

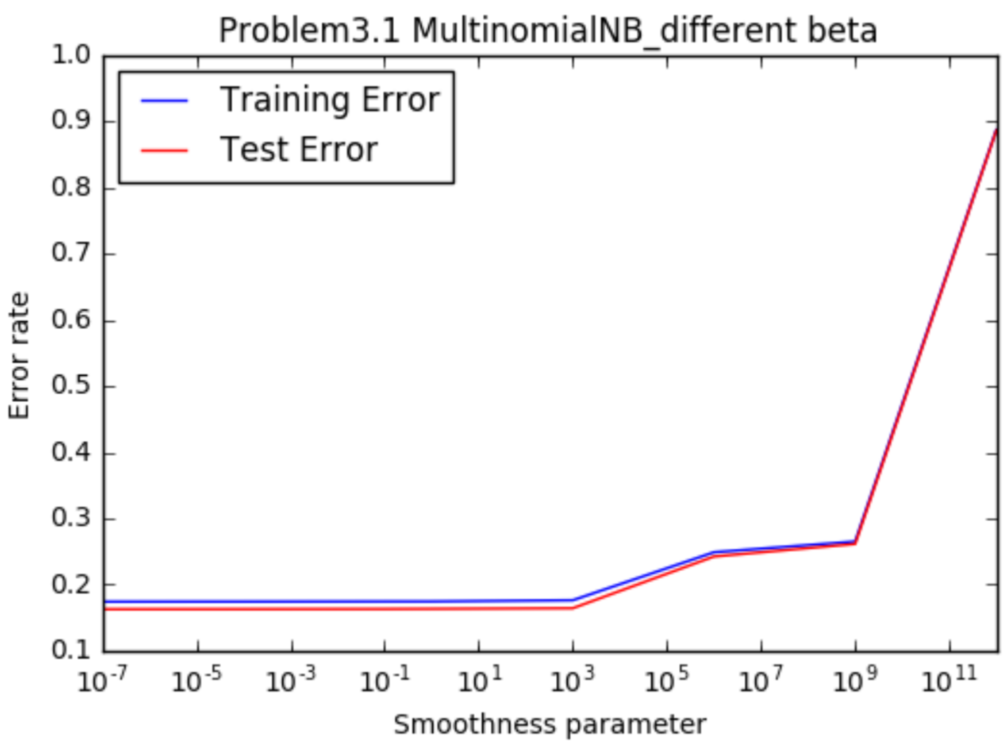
Obviously, the accuracy is lower than that of Naïve Bayes (0.96449704142). Thus, Decision tree with minimum leaf size doesn’t perform better than Naïve Bayes.




(400.csv-Entropy)Blue:traning set,Red:test set
accuracy
max tree depth
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
2
4
6
8
10
12
14
16

Problem 3

1.


Problem3.1 MultinomialNB_different beta
Training Error
Test Error
Error rate
Smoothness parameter
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
10^{-7}
10^{-5}
10^{-3}
10^{-1}
10^{1}
10^{3}
10^{5}
10^{7}
10^{9}
10^{11}

2. When $\beta \to \infty$, training error rate=0.901283333333, test error rate=0.902.

$\beta = \frac{n+\beta}{N+k\cdot\beta}$, k is the number of classes. When $\beta \to \infty$, n and N can be ignored, which means some features will not influence classifier accuracy anymore.