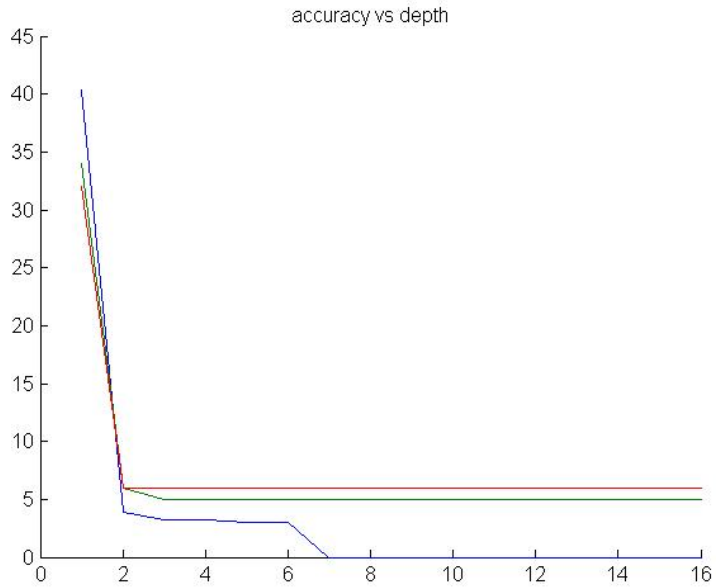
Machine Learning CSL341
Nipun Shrivastava – 2011cs50288
Assignment 3

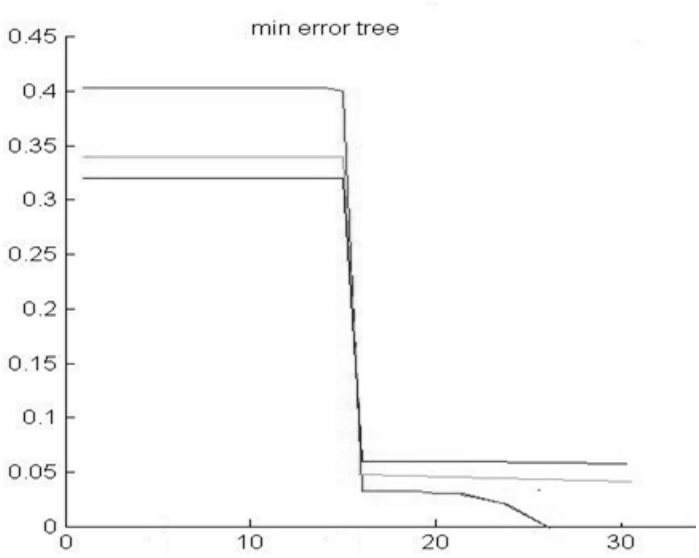
1 DECISION TREES FOR CLASSIFICATION

a. I have implemented decision trees in Matlab using the arrays. Each node is depicted by a sequence of 8 consecutive array elements, which is the proceeded by its child one after the other. Node details are:

- Partition Index
- Parent Distance
- Leaf
- Dominant Class
- 1st Child's distance from the start of the node
- 2nd Child's distance from the start of the node
- 3rd Child's distance from the start of the node
- Depth of Node

Net error was chosen as the criterion for choosing the best attribute to split on and in case of a tie, the attribute with the lowest index was chosen as the splitting attribute. For the first part, I am treating the missing values ("?") simply as another attribute value. Based on these information, each attribute has been split into 3 i.e. using the values y/n/"?".





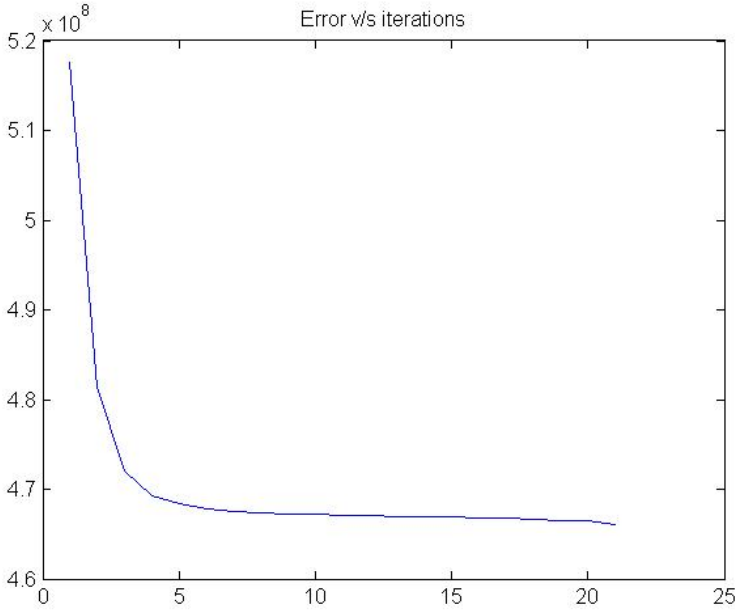
b. No, the tree are not very different as the tree trained on just the training data set perform quite well on the validation data too and there are considerable more number of examples in the training data, therefore, the best attribute is still defined predominantly by the training data, hence, the similar tree.
The error plot in both the cases come out to be almost identical.

c. The missing values can be treated as both y and n, when we construct the tree. Then based on the validation set, we can improve the accuracy by assigning "?" one of the two values.

2 K-MEANS FOR DIGIT RECOGNITION

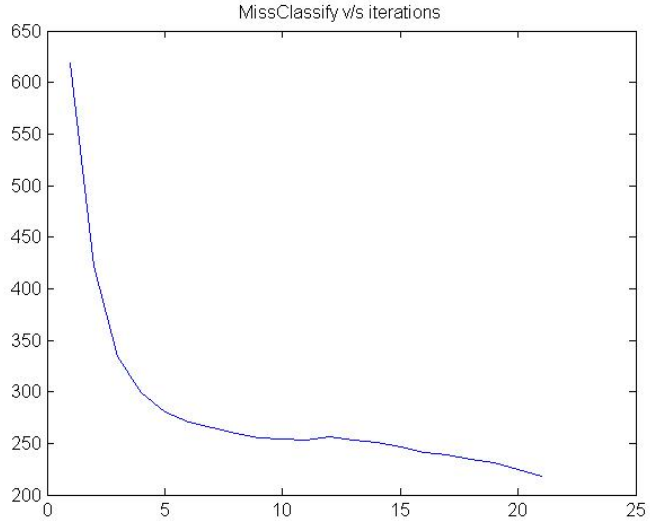
a. No explanation

b. The algorithm gave very good results for when we stopped the loop after around 30 iterations though there is a risk of k-means going haywire if more iterations are not included in the stopping criterion and the data is one of the exception cases.



c.

The distance from the mean goes on decreasing which is a given for the k-means algorithm. But one peculiar thing that I observed was that k-means failed to identify 5 after 20-30 iterations, mostly because of the fact that it is similar to 3 and the means of the two clusters can easily be confused in a general scenario.

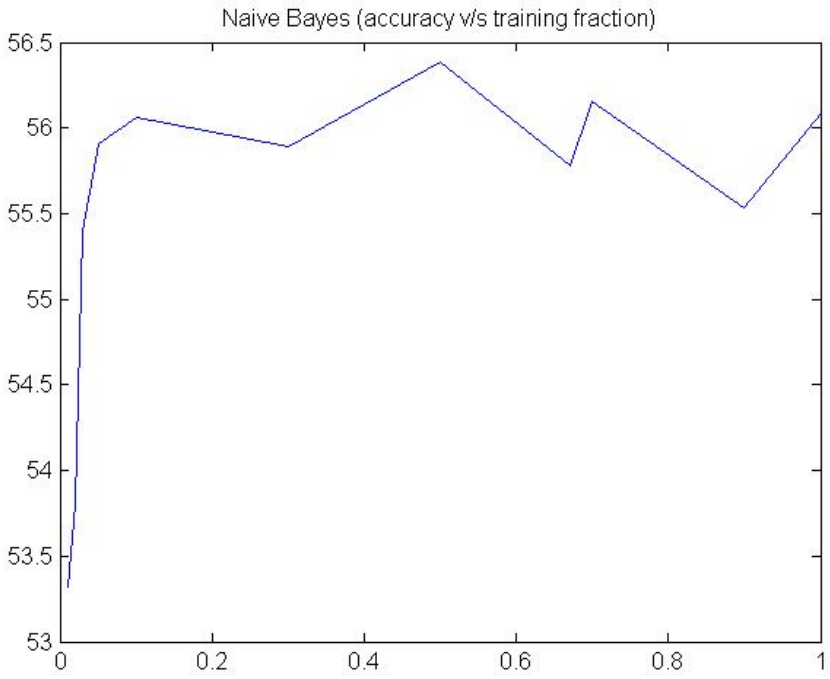


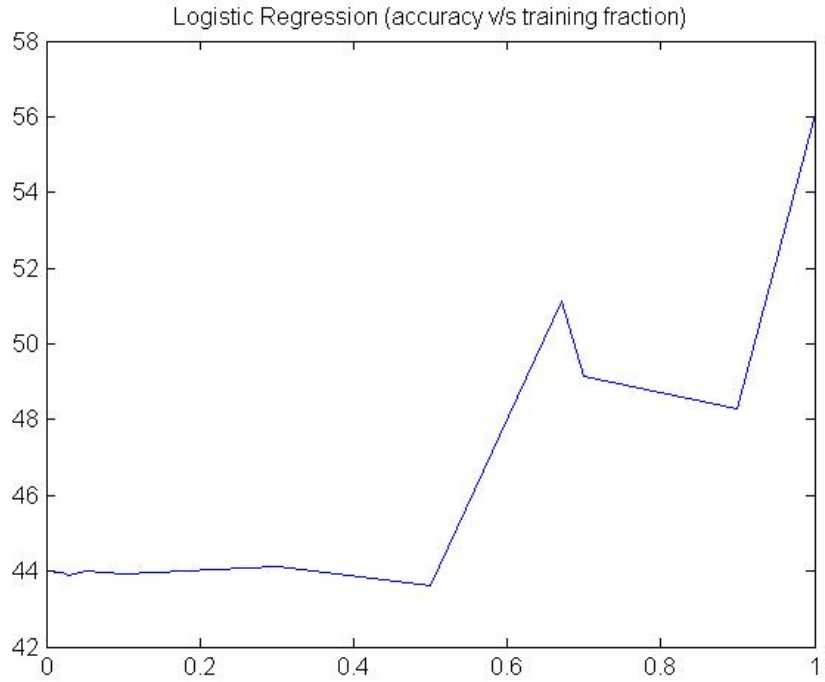
d.

We can see that the number of miss classification gradually reduces as we increase the number of iterations, but as we can see from the error plot that the error saturates around 15 iterations. Since we can only know the miss classifications if we know the label of the data, therefore, we cannot use miss classification as converging criteria. The two plots simultaneously show the pros and cons of k-means algorithm.

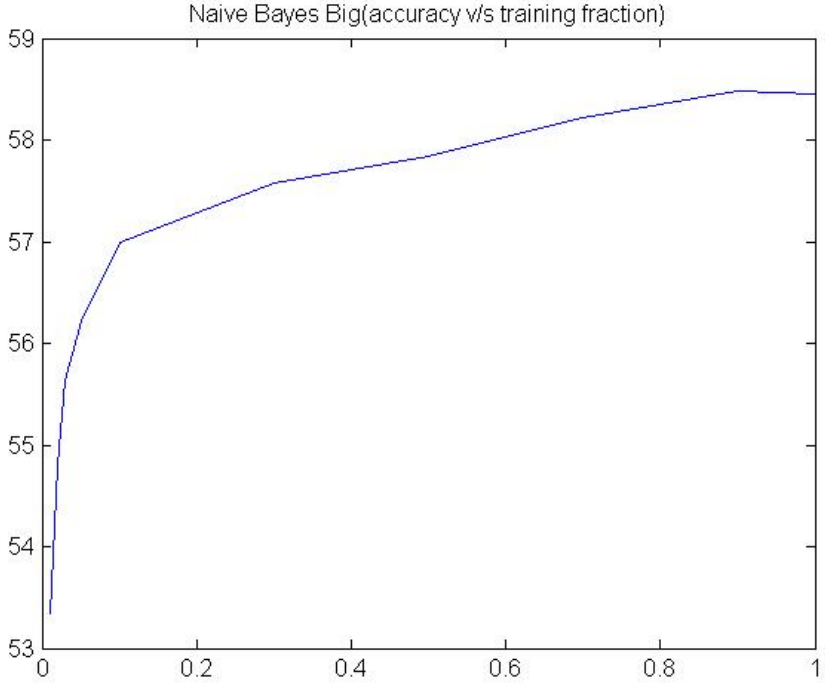
3 DISCRIMINATIVE VS. GENERATIVE CLASSIFICATION

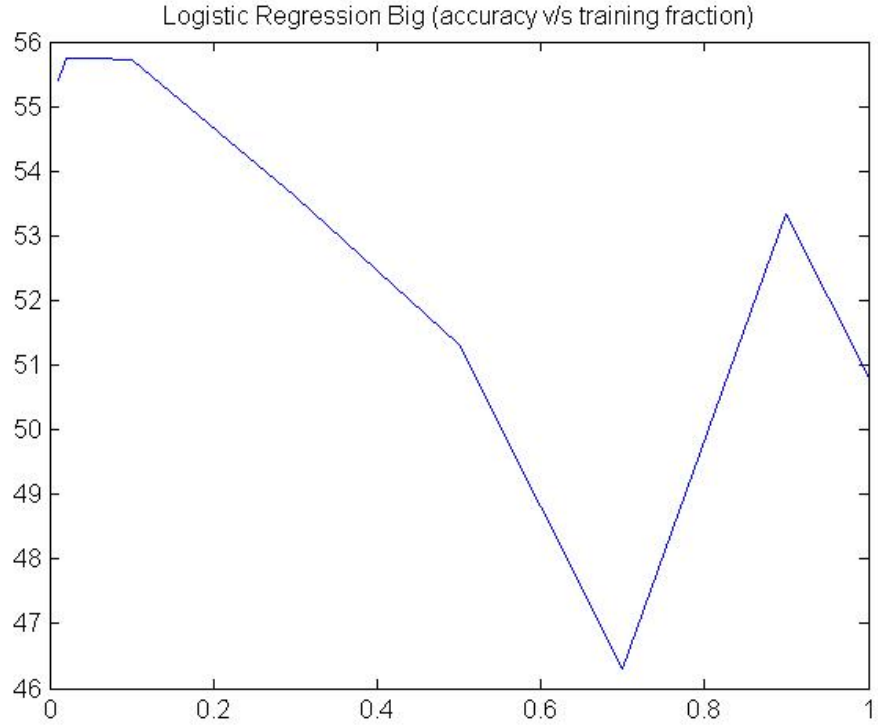
a. Learning rate for logistic regression = 0.00001
Stopping Criteria = Change in sum of square error < 0.00001





b. The sub data given to us shows a poor accuracy of 56 percent for both linear regression as well as naïve Bayes, suggesting the data is neither linearly separable nor the features are dependent on each other, i.e. the covariance of features is very small for any two pair.





c. The trend is repeats when all the features are used. In fact, linear regression performs worse for the bigger data suggesting the data has become less likely to be linearly separable. Naïve Bayes accuracy reached 59% which is still not good given label can take only 2 values and a random algorithm can reach the accuracy of 50% in the long run. Still suggesting that different features are mostly independent of each other.

4 EXPECTATION MAXIMIZATION

a. For train-0

g_prob <4x3 double>

	1	2	3
1	0.3008	0.3975	0.3017
2	0.0552	0.2459	0.6989
3	0.8999	0.0827	0.0174
4	0.4631	0.3187	0.2182

d_prob <2:

	1
1	0.5977
2	0.4023

i_prob <2x

	1
1	0.6989
2	0.3011

l_prob <3x2 double>

	1	2
1	0.0931	0.9069
2	0.3772	0.6228
3	0.9886	0.0114

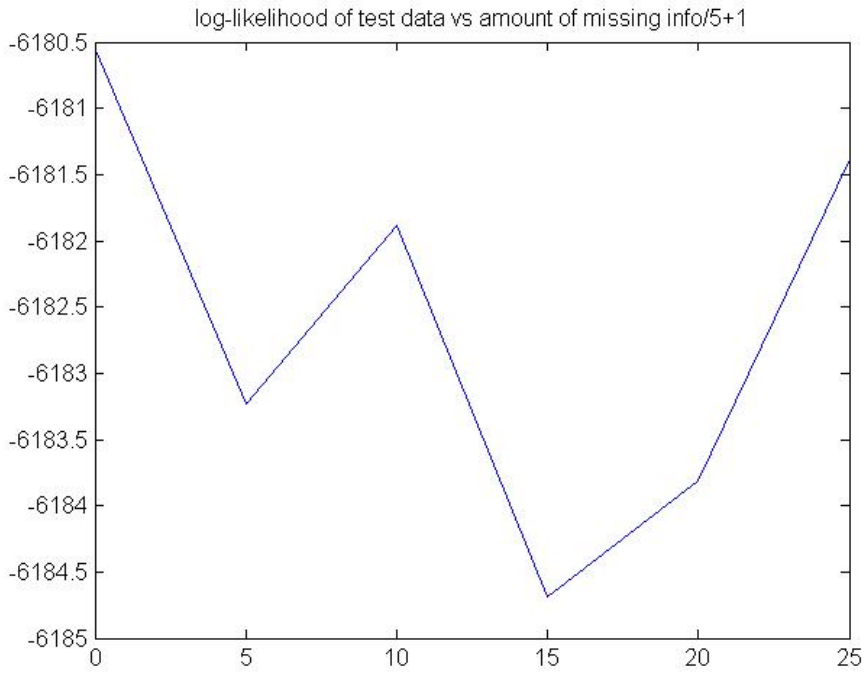
s_prob <2x2 double>

	1	2
1	0.6480	0.3520
2	0.6440	0.3560

b. First step generates the above probabilities using the examples which have all the values defined. Then in the loop, E step adds the probabilities of all the cases in missing examples to the counter matrix using the already learnt probabilities. The final counter matrix is then used to generate the new probabilities in M step. These new probabilities so generated are then again used to generate new counts related to probabilities for missing example which then adds to count of non-missing examples.
The stopping criteria is change in log likelihood to be less than 0.0001.

c.

The log likelihood doesn't really follow any specific pattern and there is minimal variation for different amount of the missing data, suggesting that data is highly dependent on each other and as long as only one of the features is missing from the set of 5, we can more or less predict the right value of the missing data.



The certain shape of the curve can be attributed to high variations in number of feature missing, i.e. one of the feature is missing relatively less or relatively less values.

But it's not making much difference as is clear from the values of the log likelihood.

5 PRINCIPAL COMPONENT ANALYSIS

a.

Average Face

b.

Eigen Faces

c. The projected image for index 55

The projected images looks similar to the original image suggesting that most of the information can be encoded in 50 dimensions of the faces instead of 19*19.