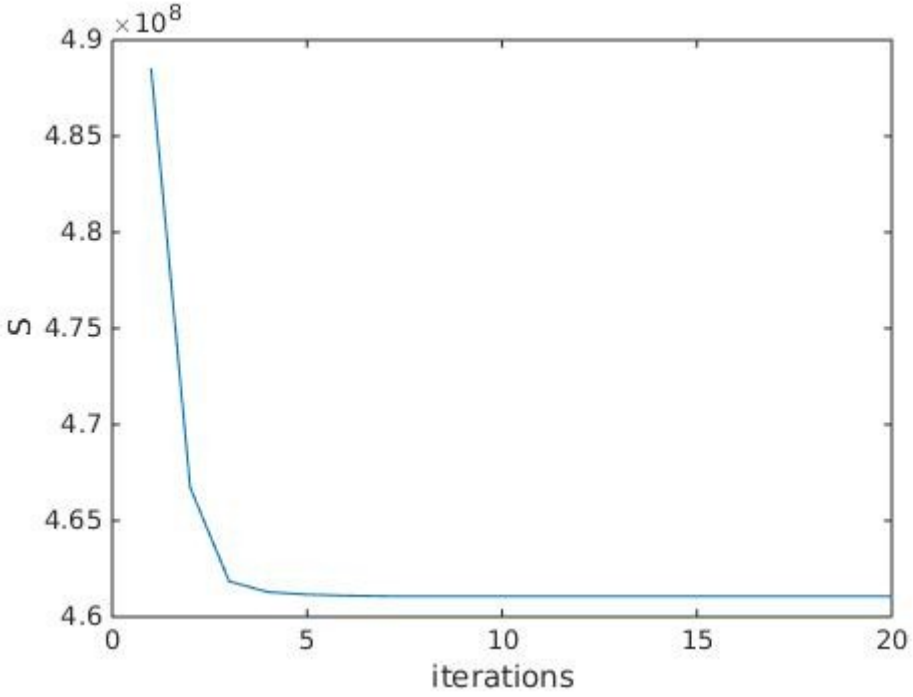
ASSIGNMENT 4

TANUMA PATRA
2015MCS2356

QUESTION1:
A)The script is written in the file Q1_a.m.The script takes a file **'input.csv'** as input from the user .**input.csv** contains an index of the position corresponding to which we need to output the image of the number.

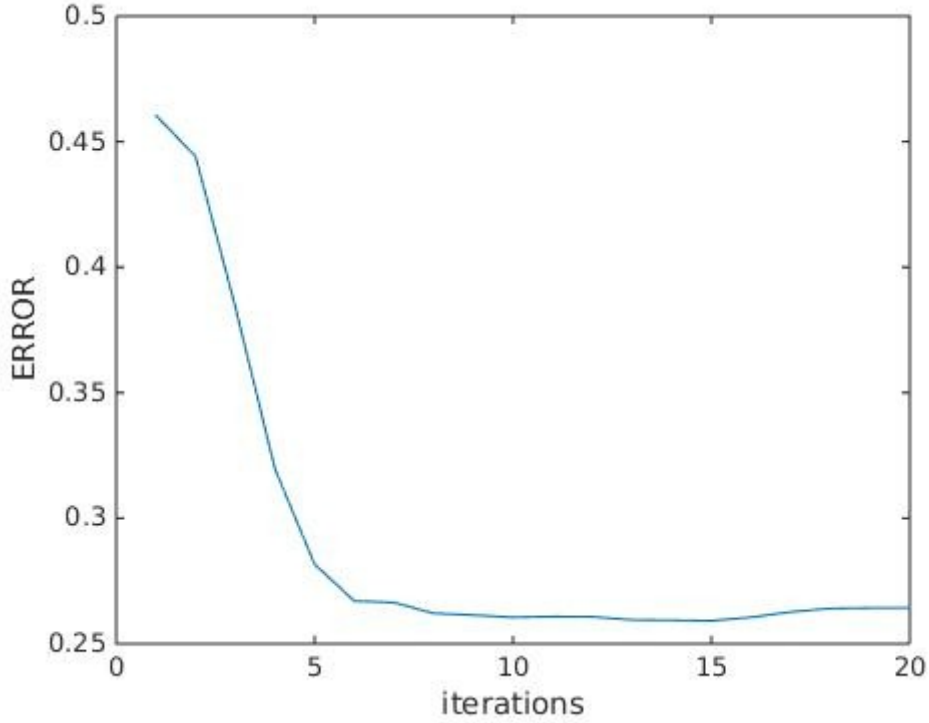
B)K-means was implemented as per the algorithm in class.The cluster means were randomly intialized by taking points from the data set .The iteration stops when the number of iterations reaches a maximum of 30 or when the cluster centroids do not change any more.
The accuracy obtained in this case was:
The accuracy fluctuates between 74%-82%
The number of points assigned to each cluster varies between 150-250.

C) The plot for this part of the question is given below:



As we can see from the graph,as the number of iterations increase,the value of the quantity S reduces exponentially.This signifies that with each iteration,the points in the cluster are getting closer to the actual centroids.that is the centroids are getting re-positioned correctly.

D)The accuracy obtained is:
It fluctuates between 72%-82%
The plot of the error vs the number of iterations is as follows:



QUESTION2:

A)The learned parameters are as follows:

H=0	**H=1**
0.8040	0.1960

	B=0	**B=1**
H=0	0.9502	0.0498
H=1	0.5837	0.4163

	L=0	**L=1**
H=0	0.9958	0.0042
H=1	0.7092	0.2908

	X=0	**X=1**
L=0	0.9779	0.0221
L=1	0.3924	0.6076

	F=0	F=1
B=0,L=0	0.9513	0.0487
B=0,L=1	0.4837	0.5163
B=1,L=0	0.9235	0.0765
B=0,L=1	0.3008	0.6992

B)EM algorithm was implemented for the above network.
The E-step included assigning probabilities for the positions in the data which had '?'.Probability is assigned for the value being 0 and 1.
The M-step included filling up the CPT by using the new data where the probabilities are used in place of '?'.
At the end of each E and M step,the log likelihood is calculated on the train data and the algorithm converges when the difference between two consicutive log likelihood values becomes <=0.00001

C)INITIAL SET OF PARAMETERS:

NO MISSING:

H=0	H=1
0.8040	0.1960

	B=0	B=1
H=0	0.9502	0.0498
H=1	0.5837	0.4163

	L=0	L=1
H=0	0.9958	0.0042
H=1	0.7092	0.2908

	X=0	X=1
L=0	0.9779	0.0221
L=1	0.3924	0.6076

	F=0	F=1
B=0,L=0	0.9513	0.0487
B=0,L=1	0.4837	0.5163
B=1,L=0	0.9235	0.0765
B=0,L=1	0.3008	0.6992

SINGLE MISSING:

H=0	H=1
0.8018	0.1982

	B=0	B=1
H=0	0.9535	0.0465
H=1	0.5902	0.4098

	L=0	L=1
H=0	0.9952	0.0048
H=1	0.7095	0.2905

	X=0	X=1
L=0	0.9794	0.0206
L=1	0.3949	0.6051

	F=0	F=1
B=0,L=0	0.9521	0.0479
B=0,L=1	0.4930	0.5070
B=1,L=0	0.9093	0.0907
B=0,L=1	0.2717	0.7283

DOUBLE MISSING:

H=0	H=1
0.8040	0.1960

	B=0	B=1
H=0	0.9502	0.0498
H=1	0.5837	0.4163

	L=0	L=1
H=0	0.9958	0.0042
H=1	0.7092	0.2908

	X=0	X=1
L=0	0.9779	0.0221
L=1	0.3924	0.6076

	F=0	F=1
B=0,L=0	0.9513	0.0487
B=0,L=1	0.4837	0.5163
B=1,L=0	0.9235	0.0765
B=0,L=1	0.3008	0.6992

FINAL SET OF PARAMETERS AFTER CONVERGENCE:

NO MISSING:

H=0	H=1
0.8040	0.1960

	B=0	B=1
H=0	0.9502	0.0498
H=1	0.5831	0.4169

	L=0	L=1
H=0	0.9958	0.0042
H=1	0.7092	0.2908

	X=0	X=1
L=0	0.9779	0.0221
L=1	0.3924	0.6076

	F=0	F=1
B=0,L=0	0.9513	0.0487
B=0,L=1	0.4837	0.5163
B=1,L=0	0.9235	0.0765
B=0,L=1	0.3008	0.6992

SINGLE MISSING:

H=0	H=1
0.8027	0.1973

	B=0	B=1
H=0	0.9529	0.0471
H=1	0.5831	0.4169

	L=0	L=1
H=0	0.9958	0.0042
H=1	0.7128	0.2872

	X=0	X=1
L=0	0.9794	0.0206
L=1	0.3778	0.6222

	F=0	F=1
B=0,L=0	0.9516	0.0484
B=0,L=1	0.4985	0.5015
B=1,L=0	0.9188	0.0812
B=0,L=1	0.2704	0.7296

DOUBLE MISSING:

H=0	H=1
0.8046	0.1954

	B=0	B=1
H=0	0.9429	0.0571
H=1	0.5825	0.4175

	L=0	L=1
H=0	0.9945	0.0055
H=1	0.7013	0.2987

	X=0	X=1
L=0	0.9769	0.0231
L=1	0.3797	0.6203

	F=0	F=1
B=0,L=0	0.9571	0.0429
B=0,L=1	0.4590	0.5410
B=1,L=0	0.9097	0.0903
B=0,L=1	0.2891	0.7109

VALUE OF LOG LIKELIHOOD:
No missing:-2.5153e+03
One missing:-2.5147e+03
Two missing:-2.5163e+03

QUESTION3:
(a) Various libraries are used as mentioned in part (a) on website:
The accuracy obtained was
1) Decsion Trees :
Train: 73.45 %
Average Validation accuracy : 70.23 %
Test : 68.73
2) Naive Bayes : Train: 68.56 %
Average Validation accuracy : 74.23 %
Test : 72.63 %
3) SVM Linear:
Train: 76.73 %
Average Validation accuracy : 74.23 %
Test : 75.38 %
4) SVM Gaussian:
Train: 74.73 %
Average Validation accuracy : 72.23 %
Test :73.58 %

(b)The following algorithms have been tried to get optimum accuracy:
Data Pre-processing: L. Wolf ’s paper on one-shot exemplar SVM was referred to pre-process the data and present in an efficient way. For this the absolute difference in the features for each image and element wise product of each feature is concatenated and this matrix is trained upon.
With this most of the time was spent on trying various parameters on gaussian SVM. Next, various parameters for Decision Tree Classifiers were tried to get improve accuracy. Other methods tried out were k-means, SVM with polynomial kernel,random forest and Ada Boost Regressor .The final atempt which gave the maximum accuracy was selecting a random subset(40%) of combination of the train and validation data and feeding it to gaussian SVM with C=5 and gamma=.09.