Assignment 4 COL-774

Abhishek Agarwal - 2014MCS2114

May 2, 2016

1 Question 1

1.1 Part a

The script to visualize each image is in render.py. Takes index as input and shows the gray scale image.

1.2 Part b

Implemented K means in helper.m. The data is first processed using preprocess.py.
The iterations stopped after 15.
The distortion cost function as defined in assignment came out to be: 4.6483e+08.

1.3 Part c

The plot has been attached in partc.png.
The number of iterations have been kept fixed here: 20.
The plot contains quantity S as we vary the number of iterations from 1 to 20
The value of S decrease as we increase the iterations, this happens as on each iteration the points are assigned to a better cluster which in turn decreases the value of S as this captures how close the points within each cluster are.

1.4 Part d

The plot has been attached is partd.png.
The number of iterations have been kept fixed here: 20.
We observe that the ratio of missed classifications decrease as we increase the number of iterations.
This is because with each iteration we get a better cluster, better assignments, and thus lesser missed classifications.

2 Question 2

2.1 Part a

The ML estimate of the parameters in tabular format can be seen running the script. The log likelihood of the test data using the ML estimate of the parameters learned using train.data : -2515.27356509

2.2 Part b

The convergence criteria is : parameters do not change much (1e-7).

2.3 Part c

When working on train m1 data, the number of iterations taken are : 19.
The converged log likelihood over test data: -2514.68148479

When working on train m2 data, the number of iterations taken are : 31.
The converged log likelihood over test data: -2515.32456041

No, the change wasn't expected much. As the missing values were from the same data and EM learns from the same data about the missing values. The log likelihoods are not very apart from original case and are less than the original one.

The parameters can be seen by running the algorithm.

3 Question 3

3.1 Part a

SVM Linear: tolerance 1e-6
Train: 0.577331546631
Val1: 0.572885762859
Val2: 0.564018800627
Val3: 0.511183706124

SVM Gaussian: kernel rbf, gamma 0.012, C 1
Train: 0.92782
Val1: 0.8629
Val2: 0.796266666667
Val3: 0.7332

Decision Tree: entropy, max depth 20, min samples leaf 4, min samples split 10
Train: 0.942018840377
Val1: 0.752225074169
Val2: 0.681256041868
Val3: 0.637254575153

Naive Bayes:
Train: 0.542510850217
Val1: 0.538684622821
Val2: 0.539684656155
Val3: 0.519250641688

Random Forest: number of estimators 11, gini, max depth=8, min samples leaf=5, min samples split=10
Train: 0.77601552031
Val1: 0.749191639721
Val2: 0.718490616354
Val3: 0.699256641888

3.2 Part b

The report and various tests that were done for this part have been documented in a separate report. The criteria that served best for me was: QuadraticDiscriminantAnalysis with tol 1.0e-2 and reg param 0.08