ELL784/EEL709: Assignment 3

Maximum Marks: 6+3

Submission deadline: **16 April, 23:59**

1 Objective

To explore the use of unsupervised learning methods for clustering data, and also for obtaining lower dimensional representations.

2 Part 1 (4 marks)

2.1 Data

For the previous assignment, you were provided a low dimensional representation of a data set of images. We now provide the corresponding original data: a personalised file for each of you, that contains 2000 images, each of size 28×28 pixels. You should download your file from `http://web.iitd.ac.in/~sumeet/A3/<EntryID>.mat` (for example, `http://web.iitd.ac.in/~sumeet/A3/2010EE50541.mat`). The file format is compatible with both MATLAB and Python; in case you have any difficulties with it and need the data in a different format, let us know. This file will contain two objects: `data_image`, which is a 2000×784 matrix, with each column corresponding to a pixel position (ordering is column-major) and each row corresponding to an image; and `data_labels`, which is a 2000×1 vector, giving the class label for each image (there are 10 classes, denoted by the labels 0 to 9).

2.2 Methodology

(a) Using an implementation of your choice, run K-means with $K = 10$ on your handwritten digits data set. Assess the clusters obtained: do they actually correspond to the digits 0–9? If you label each cluster with the digit that occurs most frequently within it, then what is your classification accuracy with this unsupervised method? What kinds of misclassifications are happening, and why? Now try re-running K-means with $K = 5$. Do your clusters make any sense in this case? Why or why not? (2 marks)

(b) Now we will try to reduce the dimensionality of our images, prior to clustering them. Use PCA for dimensionality reduction: you should select the minimum number of principal components such that the residual variance is under 10% (show a residual variance plot for this). Visualise at least the top few components, and try to interpret what kind of variation in the data they're capturing. Now run K-means with $K = 10$ on your PCA representations. Repeat the analysis as in (a): how does this clustering in PCA space compare to that in raw pixel space? Is it better or worse? Why? (2 marks)

2.3 Part 1EC (extra credit, 3 marks)

Write your own GMM implementation, using the EM algorithm for parameter learning. Learn a GMM with 10 components on your data in PCA space. You may initialise the parameters using your K-means results from (b) above. Does the GMM give you better clustering? Try to interpret some of the data points which have mixed membership, i.e., they have a significant posterior probability in more than one component. Also try a random initialisation of the parameters: how do the results change, compared to the K-means initialisation? [Code must be submitted for this. Code must be completely original and not make use of

any machine learning libraries. Any copying detected will result in all marks for this Assignment being nullified.]

3 Part 2 (2 marks)

For this part, you are provided with a mystery data set with 15000 data points and 127 features. It can be downloaded from `http://web.iitd.ac.in/~sumeet/A3/ass3_data.txt`. In this file, each row corresponds to a data point, and the feature values for each point are listed separated by single spaces. This is actually a labeled data set, but you are not given the labels or even the number of classes. Your job is to try and discover as much structure in this data set as you can, in an unsupervised fashion. Use K-means and/or GMMs, and first of all try to find out the number of classes/clusters. Then try to refine your clustering as much as possible: for instance, experiment with feature selection or reduction to see if that can give you tighter clusters. Once you feel you have a good clustering, submit your results to the leaderboard at `http://dubeya.com/ell784/`. Here it will be evaluated against the true labels, and a performance metric will be reported and used to rank your submission.

4 Evaluation

- You should prepare a report, compiling all your results and your interpretation of them, along with your overall conclusions. In particular, you should attempt to answer all of the questions posed in the previous sections. Any graphs or other visualisations should also be included therein. If you wish, you may also include code or other materials which are relevant, though this is not required (except for the extra credit part). Submission will be via Moodle: please carefully follow the instructions given there for organising and naming your submissions. The submission deadline is **April 16th, 23:59**. Any late submissions will be penalised.
- The schedule for demos/vivas will be announced by your respective TAs, in advance. If for any reason you cannot attend in your scheduled slot, you must arrange for an alternative slot with your TA well in advance. Last-minute requests for rescheduling will normally not be accepted.