EEL709: Assignment 4

Weightage: 10%

April 11, 2015

1 Objective

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

2 Data

In this assignment you will work with two data sets that you have already used previously. Firstly, your personalised data set from Assignment 3, consisting of images of handwritten digits. Secondly, the real-world multivariable regression data set you used in the second part of Assignment 1.

3 Methodology

3.1 Part I

(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? [Viva: 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? [Viva: 2 marks]

For Extra Credit: 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). 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? [Maximum 5 extra marks; 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.]

3.2 Part II

Take your real-world multivariable regression data set from Assignment 1. You have previously looked at how to characterise the importance of different features in a supervised fashion. Now try using PCA to construct new features in an unsupervised fashion. You can experiment with different numbers of principal

components. Use these components as inputs to a linear regression model (rather than the original features). How does this affect the results which you obtained in Assignment 1? Also, are you able to interpret any of your principal components? Would you say the interpretability of your regression models is better, or worse, than in Assignment 1? Based on this, what are your conclusions about the relative usefulness of PCA and supervised feature selection, as two different approaches to reduce the dimensionality in the context of a supervised learning task? [Viva: 3 marks]

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 section. 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 portion, as mentioned above). The submission link is `http://web.iitd.ac.in/~sumeet/submit.html`; put everything into a single zip file or tarball, and name it as per the instructions given there. The submission deadline is **April 25th, 23:59**. Only results submitted by this deadline will be evaluated during the demo/viva.

- The schedule for demos/vivas will be announced by your respective TAs, in advance (planned dates are 26th to 28th April). 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.

- Anyone turning up for the demo without having submitted a report in advance will be evaluted only at the discretion of the TA; and if at all evaluated, their marks for this assignment will be downgraded by a factor of at least 10% $\times$ (no. of days between submission deadline and demo).