CSL603– Machine Learning
Lab 4

Due on 24/11/2017 11.55pm

Instructions: Upload to your moodle account one zip file containing the following. Please do not submit hardcopy of your solutions. In case moodle is not accessible email the zip file to the instructor at ckn@iitrpr.ac.in. Late submission is not allowed without prior approval of the instructor. You are expected to follow the honor code of the course while doing this homework.

1. **This lab can be completed in pairs. I and the TAs reserve the right to question you if required. If you are found unable to explain the code, you will receive no points for the lab.**

2. This lab must be implemented in Matlab/Python.

3. A neatly formatted PDF document with your answers for each of the questions in the homework. You can use latex, MS word or any other software to create the PDF.

4. Include a separate folder named as 'code' containing the scripts for the homework along with the necessary data files. Ensure the code is documented properly.

5. Include a README file explaining how to execute the scripts.

6. Name the ZIP file using the following convention rollnumberhwnumber.zip

---

In this lab you will be experimenting with clustering and dimensionality reduction techniques. The dataset is provided as part of the lab in the file data.txt Each line in the file represents one image consisting of 400 columns (20x20 image). The label information is present the file label.txt.

a. Using a k-means clustering implementation of your choice, perform k-means clustering on the MNIST hand written digits' dataset. Indicate in the report the source of your k-means clustering implementation. Perform clustering with k=10. Suppose we were to label each cluster with the most frequently occurring digit, what is the classification accuracy? What are the kinds of misclassifications (confusion matrix)? Suppose we were to increase k to 15, some of the digits which were previously represented as a single cluster will split into

multiple clusters. Which are the digits that get split further? Now if we were to reduce k to 5, some of the clusters will be combined. Do your new clusters make any sense? For example, do you observe that clusters with digits 7 and 1 get combined? Discuss your observations.

b. Use PCA to reduce the dimensionality of the digit images. Select the number of components such that the residual reconstruction error is under 0.1. Visualize at least the two 2 or 3 components and try to interpret what kind of variation in the data are they capturing?

   Note: Re-project the transformed data to the original space, by forcing the values in the higher components to be 0. Suppose the transformed space is 3 dimensional, then the remaining 397 dimensions are set to 0, and using the transformation matrix U, project it back to the original space. This would lead to some information loss (as some of the dimensions have been set to 0). The reconstruction error (Difference between the original data and the reconstructed data) will quantify the information loss. The re-construction can be achieved by just multiplying the transformed data with the transformation matrix U and adding the mean of the original data. If we were to use all the principal components for the projection, then the reconstruction error would be 0. The question being asked is how many principal components should we consider for achieving a reconstruction error of 0.1

c. Perform k-means clustering on the data projected onto lower dimensions. Repeat the experiments that were conducted in part (a) on the low dimensional dataset. How does clustering in the low dimensional space compare to the original space?

An important aspect of machine learning is reproducibility of the results presented in a paper/report. Therefore, we will run your code to see if the results are closely matching with what you have presented in the report. Any deviation beyond a reasonable threshold will be considered as fudging of results and will invite severe penalty.

[courtesy] Sumeet Agarwal