ECE 5403: Machine Learning **(Due: 05/18/2020)**

Submission Assignment #4

Instructor: Chun-Shu Wei *Name:* Student name, *Student Id:* Student Id

Course Policy: Read all the instructions below carefully before you start working on the assignment, and before you make a submission. For this assignment, please hand in the following two things: a pdf file and a ipynb file.

- PDF file: contains your results and explanations. Please name this file as **HW4_StudentID_Name.pdf** and **remember to type your Student ID and Name in pdf (e.g. HW4_9400000_chunshuwei)**.
- Ipynb file: write the comment to explain. Please name this ipynb file as **HW4_StudentID_Name.ipynb**
- Please name your assignment as **HW4_StudentID_Name.zip**. The archive file contains source code(ipynb file) and report (pdf file).
- Implementation will be graded by completeness, algorithm correctness, model description, and discussion.
- PLAGIARISM IS STRICTLY PROHIBITED.
- Please submit your assignment as ONE single zip file on the E3 system. Paper submission is not allowed. Inserting clear scanned image of handwritten derivations is accepted. Denote date and time on the first page.
- Submission deadline: **2020.05.18 11:55:00 PM**.

Problem 1: Clustering (40 points)

Notice: Please hand in with your code and results.
Note: You CAN use the ready-made functions, i.e. cluster function in scikit-learn.
Load the training set and testing set from the following website link http://yann.lecun.com/exdb/mnist/. It contains a handwritten digit collection of 28×28 images that is used for handwriting recognition benchmarks. Please use the testing set for cluster.
(a) Train an unsupervised clustering using k-means.
(b) Show the original test data distribution and the clustered data distribution, and mark the center point.

Problem 2: Dimension reduction (60 points)

Notice: Please hand in with your code and results.
Note: You CANNOT use the ready-made functions, i.e. PCA function in scikit-learn.
The attached file, "pca_dataset.zip", contains 982 images of in "4" variant appearance. Implement Principle Component Analysis to reduce the image dimension from 784 to 2, 16, 64, and 256, respectively.
(a) Compare the reconstruction result of them, visualize the image, and explain the reason why their performances are different. (Please use "four0.jpg" to visualize your result.)
(b) Perform PCA and drop the dimensionality down to two dimensions. Plot the projection as a 2D scatter plot like figure1.
(c) From the perspective of application, please explain the similarity and difference between LDA and PCA.
(d) From the perspective of mathematical derivation, please use objective function to explain the correlation and difference between LDA and PCA.

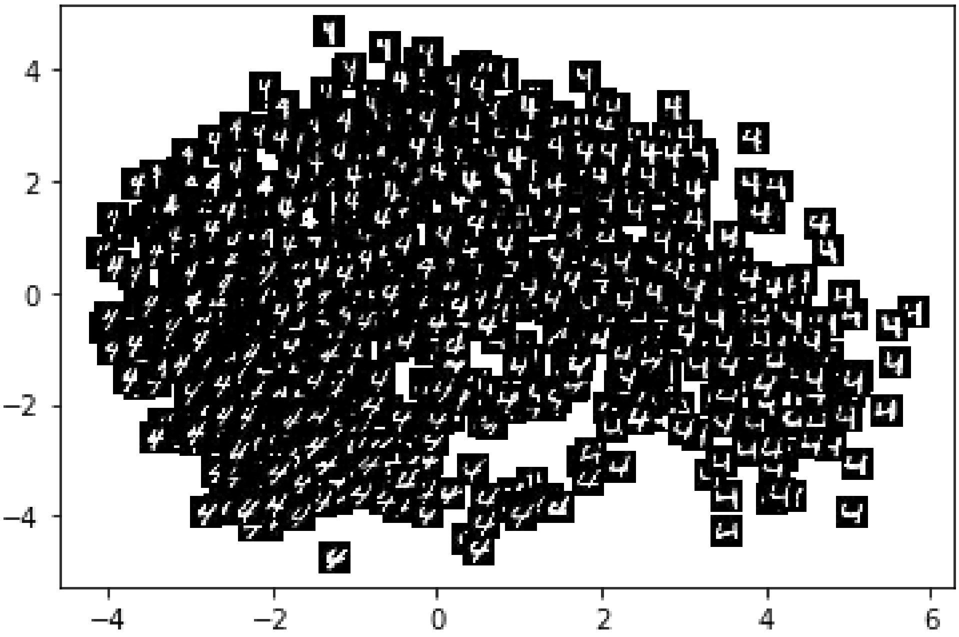

Figure 1: PCA result on digit 4.