HW 4 – 4803, Fall 2019

Instructor: Wenjing Liao

- HW 4 is due on Monday November 18 at the beginning of the class.
- You are strongly encouraged to type out your solutions using latex.
- Please write your solutions independently, and include your code at the end of your solutions.

Part I (Conceptual questions)

8.4 Exercises: 3,4
9.7 Exercises: 2,3
10.7 Exercises: 1,3,

Part II (Programming)

Programming Problem 1: This problem is about the Boston data set. It is an extension of Ex 7 in 8.4 Exercises. Split the data to two even subsets - one for training and the other for testing.

(a): Apply regression trees to predict the median value of owner-occupied homes in \$1000's from other variables. Describe your experiments and report the test mean squared error.

(b): Apply random forests to predict the median value of owner-occupied homes in \$1000's, using $m = 6$ that 6 random predictors are considered for each split of the tree. Try 25 and 100 trees respectively. Describe your experiments and report the test mean squared error.

(c): Apply classification trees to predict whether a given suburb has a crime rate above or below the median from other variables. Describe your experiments and report the test classification error.

(d): Apply random forests to predict whether a given suburb has a crime rate above or below the median from other variables, using $m = 6$. Try 25 and 100 trees respectively.

Describe your experiments and report the test classification error.

Programming Problem 2: Problem 7 in 9.7. Below are some explanations.

(b): The parameter "cost" is the misclassification cost: the C in Eq. (9.15).

(d): In (d), you are expected to plot the error versus "cost", "gamma", "degree".

Programming Problem 3: Download the 32×32 data file from the the Yale database: `http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html`. The file is in the "mat" format. If you use python or R, you can first load the file in matlab, and then save it in a "csv"or other format.

(a): Compute and display the mean of the faces. Then subtract the mean from all the faces. You are expected to display faces instead of vectors.

(b): Perform PCA on the whole data set. This can be down through the singular value decomposition of the data matrix.

(1) Display the singular values;

(2) Display the top 10 principal components. You are expected to display faces instead of vectors.

(c): Take an arbitrary face, display the reconstructed faces using k principal components while $k = 10, 50, 100, 500$. Plot the reconstruction error as k increases. You can try $k = 1, 50, 100, 150, 200, \ldots$. Denote the original and reconstructed images by x and x_k respectively. Then the reconstruction error can be defined as $\|x - x_k\|$.