CS 584-04: Machine Learning

Fall 2018 Assignment 1

# INSTRUCTIONS

1. Students should complete this assignment independently.
2. Students are encouraged to use Python to complete this assignment.
3. Students must submit their answers to Blackboard before 11:59 PM on September 5, 2018.

# Question 1 (40 points)

Write a Python program to calculate the density estimator of a histogram. Use the field *x* in the NormalSample.csv file.

1. (4 points) According to Izenman (1991) method, what is the recommended bin-width for the histogram of x?
2. (3 points) What is the bin-width after applying the beautification step?
3. (10 points) Use h = 0.5, minimum = 45 and maximum = 55. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.
4. (10 points) Use h = 1, minimum = 45 and maximum = 55. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.
5. (10 points) Use h = 2, minimum = 45 and maximum = 55. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.
6. (3 points) Among the three histograms, which one, in your opinions, can best describe the distribution of the field x?

# Question 2 (20 points)

Use in the NormalSample.csv to generate box-plots for answering the following questions.

1. (2 points) What are the five-number summary of x?
2. (3 points) What are the five-number summary of x for each category of Group?
3. (5 points) Draw a boxplot of x (without Group) using the Python boxplot function. Can you tell if the Python’s boxplot has displayed the 1.5 IQR whiskers?
4. (10 points) Draw a graph where it contains the boxplot of x, the boxplot of x for each category of Group (i.e., three boxplots within the same graph frame). Use the 1.5 IQR whiskers, identify the outliers of x, if any, for the entire data and for each category of Group.

# Question 3 (40 points)

The data, FRAUD.csv, contains results of fraud investigations of 5,960 cases. The binary variable FRAUD indicates the result of a fraud investigation: 1 = Fraudulent, 0 = Otherwise. The other interval variables contain information about the cases.

1. TOTAL\_SPEND: Total amount of claims in dollars
2. DOCTOR\_VISITS: Number of visits to a doctor
3. NUM\_CLAIMS: Number of claims made recently
4. MEMBER\_DURATION: Membership duration in number of months
5. OPTOM\_PRESC: Number of optical examinations
6. NUM\_MEMBERS: Number of members covered

You are asked to use the Nearest Neighbors algorithm to predict the likelihood of fraud.

1. (5 points) What percent of investigations are found to be fraudulent? Please give your answer up to 4 decimal places.
2. (5 points) Use the BOXPLOT function to produce horizontal box-plots. For each interval variable, one box-plot for the fraudulent observations, and another box-plot for the non-fraudulent observations. These two box-plots must appear in the same graph for each interval variable.
3. (10 points) Orthonormalize interval variables and use the resulting variables for the nearest neighbor analysis. Use only the dimensions whose corresponding eigenvalues are greater than one.
   1. (5 points) How many dimensions are used?
   2. (5 points) Please provide the transformation matrix? You must provide proof that the resulting variables are actually orthonormal.
4. (10 points) Use the NearestNeighbors module to execute the Nearest Neighbors algorithm using exactly five neighbors and the resulting variables you have chosen in c). The KNeighborsClassifier module has the score function.
   1. (5 points) Run this function, provide the function return value
   2. (5 points) Explain the meaning of the function return value.
5. (5 points) For the observation which has these input variable values: TOTAL\_SPEND = 7500, DOCTOR\_VISITS = 15, NUM\_CLAIMS = 3, MEMBER\_DURATION = 127, OPTOM\_PRESC = 2, and NUM\_MEMBERS = 2, find its **five** neighbors. Please list their input variable values and the target values.
6. (5 points) Follow-up with e), what is the predicted probability of fraudulent (i.e., FRAUD = 1)? If your predicted probability is greater than or equal to your answer in a), then the observation will be classified as fraudulent. Otherwise, non-fraudulent. Based on this criterion, will this observation be misclassified?