CS 484: Introduction to Machine Learning

Autumn 2022 Assignment 1

# Question 1 (25 points)

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

1. (5 points) Use the *Pandas describe()* function to find out the count, the mean, the standard deviation, the minimum, the 25th percentile, the median, the 75th percentile, and the maximum.

The *describe()* function returns the following statistics, shown no more than 7 decimal places.

1. (5 points) What is the bin width recommended by the Izenman (1991) method? Please round your answer to the nearest tenths (i.e., one decimal place).
2. (10 points) Use the Shimazaki and Shinomoto (2007) method and try = 1, 2, 2.5, 5, 10, 20, 25, and 50. What is the recommended bin width if the number of bins must not fewer than 5 and not more than 20? You need to show your calculations to receive full credit.
3. (5 points) Based on your recommended bin width answer in (c), list the mid-points and the estimated density function values. Draw the density estimator as a vertical bar chart using the matplotlib. You need to properly label the graph to receive full credit.

# Question 2 (15 points)

The **NormalSample.csv** contains the *group* variable that has two values, namely, 0 and 1.

1. (5 points) What is the five-number summary of for each category of the *group*? What are the values of the 1.5 IQR whiskers for each category of the group?
2. (10 points) Draw a graph where it contains the overall boxplot of , the boxplot of for each category of *group* (i.e., three horizontal boxplots within the same graph frame). Use the 1.5 IQR whiskers, identify any outliers of for the entire data and each category of the group. You must properly label your boxplots to receive full credits.

# Question 3 (35 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 = Fraud, 0 = Not Fraud. 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 frauds? This is the empirical fraud rate. Please give your answer up to 4 decimal places.
2. (10 points) Orthonormalize interval variables and use the orthonormalized columns for the nearest neighbor analysis. Use only the dimensions whose corresponding eigenvalues are greater than one.
   1. (5 points) How many dimensions are needed?
   2. (5 points) Please provide the transformation matrix? Show evidence that the orthonormalized columns are actually orthonormal.
3. (10 points) Use the KNeighborsClassifier module to execute the Nearest Neighbors algorithm using exactly five neighbors and the orthonormalized columns you have chosen in (b). The KNeighborsClassifier module has a score function.
   1. (5 points) Find out from the documentation the purpose of the score function.
   2. (5 points) Run the score function, show and explain the function return value.
4. (5 points) For an observation that has the median input variable values, find its **five** neighbors. Please list their input variable values and the target values. *Reminder: transform the input observation using the results in (b) before finding the neighbors*.
5. (5 points) Follow-up with (d), what is the predicted probability of fraud (i.e., FRAUD = 1)? If your predicted probability is greater than or equal to the empirical fraud rate (i.e., your answer in a), then the observation will be classified as a fraud.

# Question 4 (25 points)

I found the following flights from Chicago O’Hare Airport (ORD) to Shanghai Pudong Airport (PVG).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Flight** | **Carrier 1** | **Carrier 2** | **Airport 1** | **Airport 2** | **Airport 3** | **Airport 4** |
| A | American | Cathay Pacific | ORD | LAX | HKG | PVG |
| B | American | Cathay Pacific | ORD | SFO | HKG | PVG |
| C | American | China Southern | ORD | LAX | CAN | PVG |
| D | American | Virgin Atlantic | ORD | LHR |  | PVG |
| E | British Airways | Virgin Atlantic | ORD | LHR |  | PVG |
| F | United | Virgin Atlantic | ORD | LHR |  | PVG |
| G | United |  | ORD | DCA | EWR | PVG |
| H | United |  | ORD | DEN | LAX | PVG |
| I | United |  | ORD | EWR |  | PVG |
| J | United |  | ORD | IAD | EWR | PVG |
| K | United |  | ORD | LAS | LAX | PVG |
| L | United |  | ORD | LAX |  | PVG |
| M | United |  | ORD | LGA | EWR | PVG |

To answer the following questions, please replace empty string values in **Airport 3** with three underscore characters (i.e., ‘\_\_\_’).

1. (5 points) Generate a scatterplot of **Airport 3** (y-axis) versus **Airport 2** (x-axis). Please properly label the axes to receive full credits.
2. (5 points) Generate a frequency table of the airport codes in **Airport 2** and **Airport 3** combined.

1. (10 points) Suppose a new airline creates a new flight from ORD to PVG that makes one stop at LAX. I want to know which flight(s) most resembles this new flight. Use the Cosine Distance to measure the differences between this flight and the existing flights.
   1. Create a vector of word counts for each flight. This vector has as many elements as the number of unique values found in (b).
   2. Initialize all elements in the vector to zeros.
   3. Count the number of times the airport codes appeared in **Airport 2** and **Airport 3**.
   4. Calculate the Cosine Distance between the new flight and the Flights A to M.

You will list the Cosine Distances in a table.

1. (5 points) Which flight(s) have the shortest Cosine Distance from the new flight?