CS 84: Introduction to Machine Learning

Fall Semester 2023 Assignment 1

# Question 1 (30 points)

Please use the field in the **Gamma4804.csv** file.

1. (10 points) What are the count, the mean, the standard deviation, the minimum, the 10th percentile, the 25th percentile, the median, the 75th percentile, the 90th percentile, and the maximum of the feature ? Please round your answers to the seventh decimal place.
2. (10 points) Use the Shimazaki and Shinomoto (2007) method to recommend a bin width. We will try = 0.1, 0.2, 0.25, 0.5, 1, 2, 2.5, 5, 10, 20, 25, 50, and 100. What bin width would you recommend if we want the number of bins to be between 5 and 50 inclusively? You need to show your calculations to receive full credit.
3. (10 points) Draw the density estimator using your recommended bin width answer in (b). You need to label the graph elements properly to receive full credit.

# Question 2 (30 points)

We need to create the Training and Testing partitions from the observations in the **hmeq.csv**. We will use all observations (including those with missing values in one or more variables) for this task. The Training partition will contain 70% of the observations. The Testing partition will contain the remaining 30% of the observations.

1. (10 points). Before we partition the observations, we need a baseline for reference. How many observations are in the dataset? What are the frequency distributions of BAD (including missing)? What are the means and the standard deviations of DEBTINC, LOAN, MORTDUE, and VALUE?
2. (10 points). We first try the simple random sampling method with the random seed of 202303484. How many observations (including those with missing values in at least one variable) are in each partition? What are the frequency distributions of BAD (including missing) in each partition? What are the means and the standard deviations of DEBTINC, LOAN, MORTDUE, and VALUE in each partition?
3. (10 points). We next try the stratified random sampling method with the random seed of 202303484. We use BAD and REASON to jointly define the strata. Since the strata variables may contain missing values, we will replace the missing values in BAD with the integer 99 and in REASON with the string ‘MISSING’. What are the frequency distributions of BAD (including missing) in each partition? What are the means and the standard deviations of DEBTINC, LOAN, MORTDUE, and VALUE in each partition?

# Question 3 (40 points)

The data **FRAUD.csv** contains the 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 quantitative variables contain information about the cases.

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

We will train 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 round your answers to the fourth decimal place.
2. (10 points). We will divide the complete observations into 80% Training and 20% Testing partitions. A complete observation does not contain missing values in any of the variables. The random seed is 202303484. The stratum variable is FRAUD. How many observations are in each partition?
3. (10 points). Use the KNeighborsClassifier module to train the Nearest Neighbors algorithm. We will try the number of neighbors from 2 to 7 inclusively. We will classify an observation as a fraud if the proportion of FRAUD = 1 among its neighbors is greater than or equal to the empirical fraud rate (rounded to the fourth decimal place). What are the misclassification rates of these numbers of neighbors in each partition?
4. (5 points). Which number of neighbors will yield the lowest misclassification rate in the Testing partition? In the case of ties, choose the smallest number of neighbors.
5. (10 points) Consider this focal observation where DOCTOR\_VISITS is 8, MEMBER\_DURATION is 178, NUM\_CLAIMS is 0, NUM\_MEMBERS is 2, OPTOM\_PRESC is 1, and TOTAL\_SPEND is 16300. Use your selected model from Part (d) and find its neighbors in the Training partition. What are the neighbors’ observation values? Also, calculate the predicted probability that this observation is fraudulent.