CS 84: Introduction to Machine Learning

Spring Semester 2023 Assignment 1 Answer Key

# 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 25th percentile, the median, the 75th percentile, and the maximum of the feature ? Please round your answers to the seventh decimal place.

|  |  |  |  |
| --- | --- | --- | --- |
| Statistics | Count | Mean | Standard Deviation |
| x | 4804 | 75.0568006 | 27.4453554 |

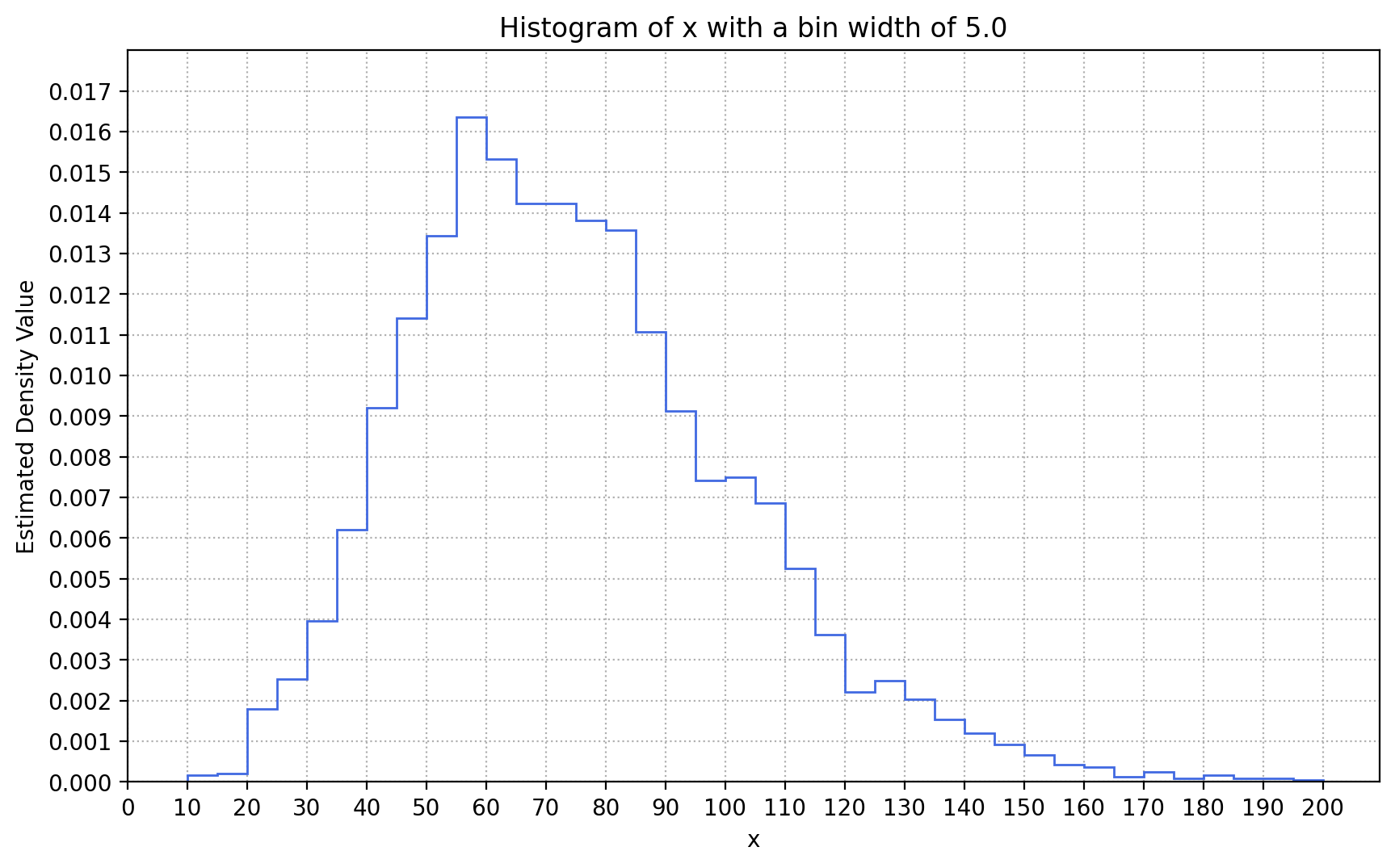
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Statistics | Minimum | 25th Percentile | Median | 75th Percentile | Maximum |
| x | 11.55 | 55.46 | 71.825 | 91.1725 | 195.6 |

1. (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 10 and 100 inclusively? You need to show your calculations to receive full credit.

|  |  |  |  |
| --- | --- | --- | --- |
| Bin Width | Number of Bins | Middle Cutpoint | Criterion |
| 0.1 | 1841 | 75.1 | -14.0324 |
| 0.2 | 921 | 75 | -373.8801 |
| 0.25 | 737 | 75 | -448.2729 |
| 0.5 | 369 | 75 | -576.4074 |
| 1 | 185 | 75 | -637.2623 |
| **2** | **93** | **76** | **-657.8490** |
| **2.5** | **75** | **75** | **-661.0811** |
| **5** | **38** | **75** | **-667.2308** |
| **10** | **19** | **80** | **-662.0940** |
| **20** | **10** | **80** | **-630.1331** |
| 25 | 8 | 75 | -606.6692 |
| 50 | 4 | 100 | -514.4502 |
| 100 | 2 | 100 | -236.3717 |

Only the number of bins of these bin widths 2, 2.5, 5, 10, and 20 are between 10 and 100 inclusively. Among these bin widths, = 5 yields the lowest criterion value. Therefore, we recommend a bin width of 5.

1. (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.



A density estimator basically has the same shape as a histogram, but their vertical axes represent different statistical quantities. A histogram usually shows the numbers or the proportions of observations on the vertical axis. A density estimator shows the empirical density function value on the vertical axis. The function values are chosen such that the total area under the polygon curve is exactly one.

# 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. We initialize the random seed with the integer 20230101.

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?

There are 5,960 observations in the dataset. The frequency distribution of BAD is:

|  |  |
| --- | --- |
| BAD | Number of Observations |
| 0 | 4,771 (80.05%) |
| 1 | 1,189 (19.95%) |

The means and the standard deviations of DEBTINC, LOAN, MORTDUE, and VALUE are:

|  |  |  |
| --- | --- | --- |
| Feature | Mean | Standard Deviation |
| DEBTINC | 33.7799 | 8.6017 |
| LOAN | 18,607.9698 | 11,207.4804 |
| MORTDUE | 73,760.8172 | 44,457.6095 |
| VALUE | 101,776.0487 | 57,385.7753 |

1. (10 points). We first try the simple random sampling method. 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?

The Training partition contains 4,172 observations. The Testing partition contains 1,788 observations. The frequency distributions of BAD in the partitions are:

|  |  |  |
| --- | --- | --- |
| BAD | Number of Observation | |
| **Training** | **Testing** |
| 0 | 3,344 (80.15%) | 1,427 (79.81%) |
| 1 | 828 (19.85%) | 361 (20.19%) |

The means and the standard deviations of DEBTINC, LOAN, MORTDUE, and VALUE are:

| Feature | Training Partition | | Testing Partition | |
| --- | --- | --- | --- | --- |
| Mean | Standard Deviation | Mean | Standard Deviation |
| DEBTINC | 33.7680 | 8.4450 | 33.8079 | 8.9627 |
| LOAN | 18,609.4199 | 11,300.3408 | 18,604.5861 | 10,990.8810 |
| MORTDUE | 74,067.9985 | 44,640.1285 | 73,055.4724 | 44,040.9911 |
| VALUE | 101,716.9024 | 56,671.2579 | 101,912.3192 | 59,015.1227 |

1. (10 points). We next try the stratified random sampling method. 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?

The feature REASON contains missing values. After we replaced these missing values with the text ‘MISSING’, we observed six strata in the original data. The strata are:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| BAD | 0 | 0 | 0 | 1 | 1 | 1 |
| REASON | DebtCon | HomeImp | MISSING | DebtCon | HomeImp | MISSING |

The Training partition contains 4,173 observations. The Testing partition contains 1,787 observations. The frequency distributions of BAD in the partitions are:

|  |  |  |
| --- | --- | --- |
| BAD | Number of Observation | |
| **Training** | **Testing** |
| 0 | 3340 (80.04%) | 1431 (80.08%) |
| 1 | 833 (19.96%) | 356 (19.92%) |

The means and the standard deviations of DEBTINC, LOAN, MORTDUE, and VALUE are:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Training Partition | | Testing Partition | |
| **Mean** | **Standard Deviation** | **Mean** | **Standard Deviation** |
| DEBTINC | 33.6593 | 8.0192 | 34.0661 | 9.8448 |
| LOAN | 18,616.0316 | 11,355.9714 | 18,589.1438 | 10,855.9165 |
| MORTDUE | 73,896.1635 | 44,181.1012 | 73,445.3958 | 45,107.4886 |
| VALUE | 102,056.8430 | 59,351.5406 | 101,119.0448 | 52,509.8322 |

# 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.

Among the 5,960 observations, 1189 of them have FRAUD = 1. Therefore, the empirical fraud rate is 1189 / 5960 = 0.1995 rounded to the fourth decimal place.

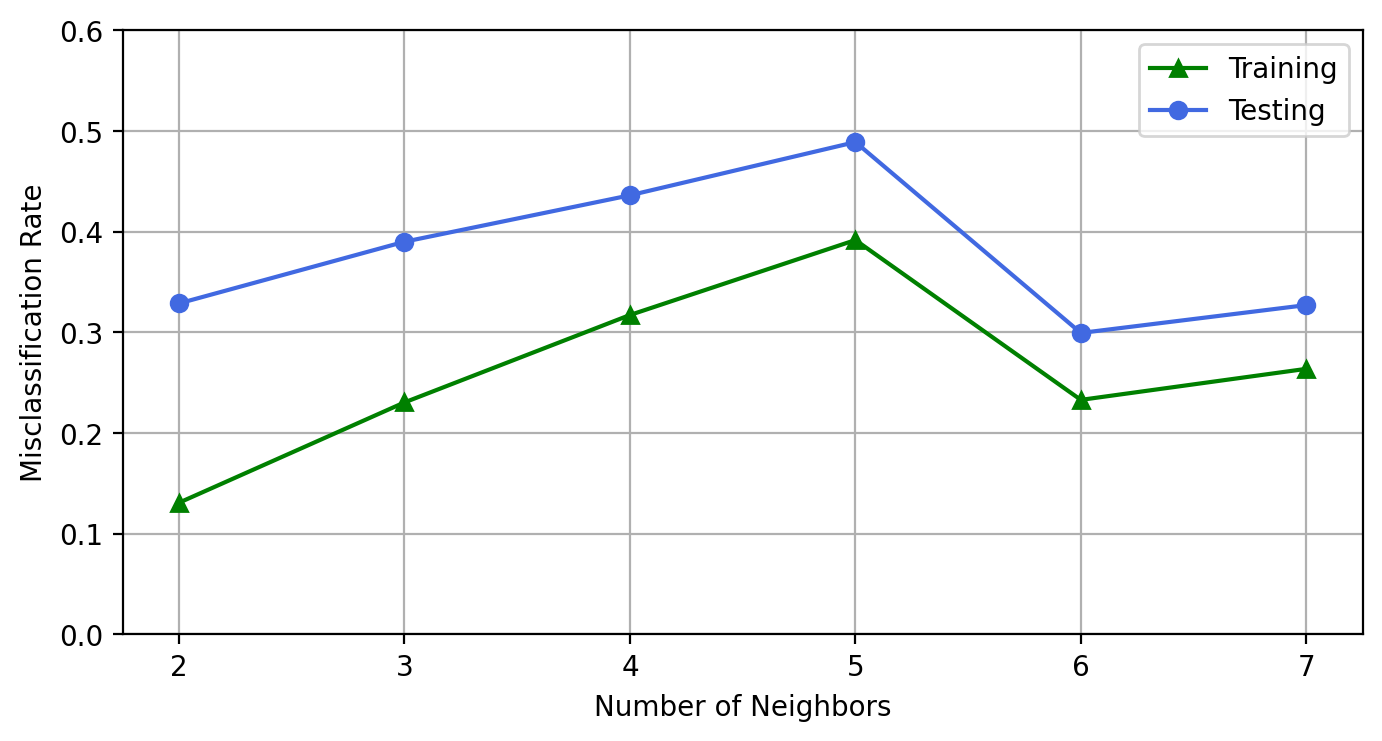
1. (10 points). We will divide the complete observations into 80% Training and 20% Testing partitions. A complete observation does not contain missing values in any of the variables. The random seed is 20230225. The stratum variable is FRAUD. How many observations are in each partition?

There are no missing values in the dataset, thus, all observations are complete. The Training partition has 4768 observations. The Testing partition has 1192 observations.

1. (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?

| Number of Neighbors | Misclassification Rate | |
| --- | --- | --- |
| Training | Testing |
| 2 | 0.1309 | 0.3289 |
| 3 | 0.2305 | 0.3901 |
| 4 | 0.3173 | 0.4362 |
| 5 | 0.3918 | 0.4891 |
| 6 | 0.2330 | 0.2995 |
| 7 | 0.2638 | 0.3272 |

1. (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.



When the number of neighbors is 6, the Testing Partition has its lowest misclassification rate.

1. (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. What are the neighbors’ observation values? Also, calculate the predicted probability that this observation is a fraud.

The six neighbors (in ascending distances) of this focal observation are:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Indices | DOCTOR\_VISITS | MEMBER\_DURATION | NUM\_CLAIMS | NUM\_MEMBERS | OPTOM\_PRESC | TOTAL\_SPEND | FRAUD |
| 2973 | 11 | 180 | 0 | 1 | 1 | 16300 | 0 |
| 2967 | 2 | 193 | 0 | 2 | 0 | 16300 | 0 |
| 2962 | 12 | 125 | 5 | 1 | 1 | 16300 | 1 |
| 2971 | 5 | 246 | 0 | 2 | 3 | 16300 | 0 |
| 2976 | 8 | 247 | 0 | 2 | 1 | 16300 | 0 |
| 2977 | 9 | 251 | 0 | 3 | 0 | 16300 | 0 |

Since FRAUD = 1 in one out of six neighbors, the predicted probability that the focal observation is a fraud if 1/6 = 0.1667. Finally, we will predict FRAUD = 0 for the focal observation because its predicted probability is less than the empirical fraud rate.