CS 584-04: Machine Learning

Fall 2018 Assignment 1 Answer Key

# 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?

Using the describe() function, we found that there are 100 observations in the NormalSample.csv file. In addition, the 25% percentile (i.e., the Q1) of the field x is 49.4675 and the 75% percentile (i.e., the Q3) is 50.915. Thus, the interquartile range (i.e., the IQR) is 50.915 - 49.4675 = 1.4475. According to Izenman (1991) method, the recommended bin-width is  
h = 2 \* 1.4475 \* 100-1/3 = 0.62371 (rounded to the fifth decimal places).

1. (3 points) What is the bin-width after applying the beautification step?

The beautification steps are:

1. Calculate u = log10(h) = -0.20502.

2. Calculate v = sign(u) \* ceil(abs(u)) = -1

3. Calculate the beautified h = 10^v = 0.1

1. (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.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | Midpoint | Density |  | Midpoint | Density | | 45.25 | 0 |  | 50.25 | 0.32 | | 45.75 | 0 |  | 50.75 | 0.3 | | 46.25 | 0 |  | 51.25 | 0.24 | | 46.75 | 0 |  | 51.75 | 0.16 | | 47.25 | 0 |  | 52.25 | 0 | | 47.75 | 0.02 |  | 52.75 | 0 | | 48.25 | 0.08 |  | 53.25 | 0 | | 48.75 | 0.12 |  | 53.75 | 0 | | 49.25 | 0.3 |  | 54.25 | 0 | | 49.75 | 0.46 |  | 54.75 | 0 | |  |

1. (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.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Midpoint | Density | | 45.5 | 0 | | 46.5 | 0 | | 47.5 | 0.01 | | 48.5 | 0.1 | | 49.5 | 0.38 | | 50.5 | 0.31 | | 51.5 | 0.2 | | 52.5 | 0 | | 53.5 | 0 | | 54.5 | 0 | |  |

1. (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.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Midpoint | Density | | 46 | 0 | | 48 | 0.055 | | 50 | 0.345 | | 52 | 0.1 | | 54 | 0 | |  |

1. (3 points) Among the three histograms, which one, in your opinions, can best describe the distribution of the field x?

There is no right-or-wrong answer because it depends on your personal preference. However, in my opinion, I would go for the h = 0.5. My primary reason is this bin-width is closest to the Izenman’s recommended bin-width. Besides, the histogram shows an approximately symmetric density with a single mode.

# Question 2 (20 points)

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

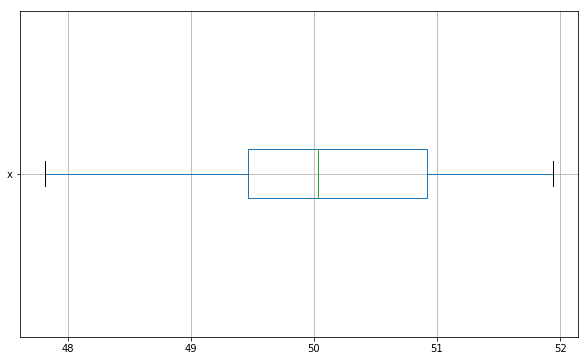
1. (2 points) What are the five-number summary of x?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Minimum | Q1 | Median | Q3 | Maximum |
| 47.82 | 49.4675 | 50.03 | 50.915 | 51.94 |

1. (3 points) What are the five-number summary of x for each category of Group?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Group | Minimum | Q1 | Median | Q3 | Maximum |
| 0 | 47.82 | 49.295 | 50.22 | 50.96 | 51.94 |
| 1 | 48.17 | 49.53 | 49.88 | 50.74 | 51.82 |

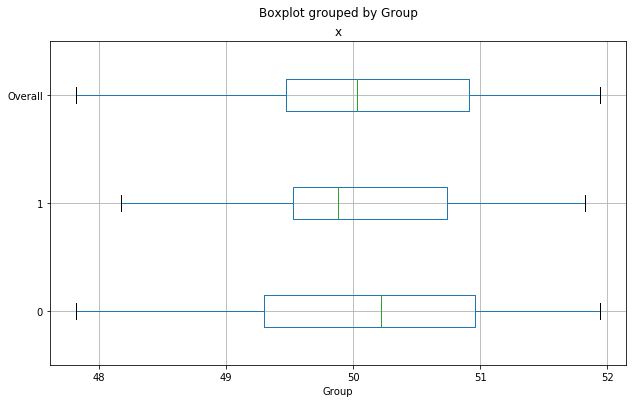
1. (5 points) Draw a boxplot of x (without Group) using the Python boxplot function. Can you tell if the Python’s boxplot has displayed the 1.5 IQR whiskers?

The left whisker should extend to the larger of the minimum value and Q1 – 1.5 \* IQR. Since the minimum value is 47.82 and Q1 – 1.5 \* IQR = 49.4675 – 1.5 \* (50.915 – 49.4675) = 47.2963, the left whisker extends to **47.82**. Similarly, the right whisker should extend to the smaller of the maximum value and Q3 + 1.5 \* IQR. Since the maximum value is 51.94 and Q3 + 1.5 \* IQR = 50.915 + 1.5 \* (50.915 – 49.4675) = 53.0863, the left whisker extends to **51.94**. Apparently, the boxplot below shows the 1.5 IQR whiskers.

1. (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.

The whiskers of the boxplots are:

|  |  |  |  |
| --- | --- | --- | --- |
| Group | IQR | Left Whisker | Right Whisker |
| 0 | 1.6650 | 46.7975 | 53.4575 |
| 1 | 1.2100 | 47.7150 | 52.5550 |
| Overall | 1.4475 | 47.2963 | 53.0863 |



Since the left (right) whiskers of the overall x and that of each category of Group have reached the minimum (maximum), there are no observations outside of the whiskers. Thus, there are no outliers in the entire data or in 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.

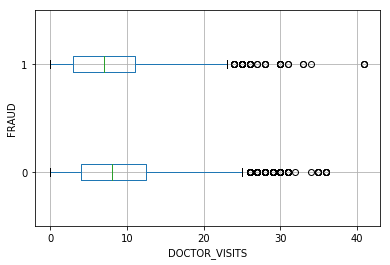
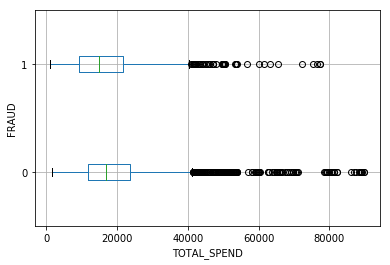
1. TOTAL\_SPEND: Total amount of claims in dollars
2. DOCTOR\_VISITS: Number of visits to a doctor
3. NUM\_CLAIMS: Number of claims made recently
4. MEMBER\_DURATION: Membership duration in number of months
5. OPTOM\_PRESC: Number of optical examinations
6. NUM\_MEMBERS: Number of members covered

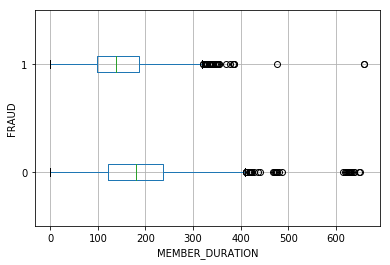
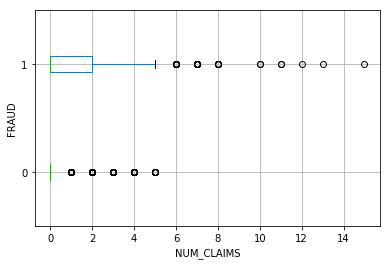
You are asked to use the Nearest Neighbors algorithm to predict the likelihood of fraud.

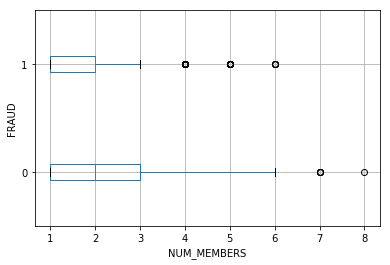
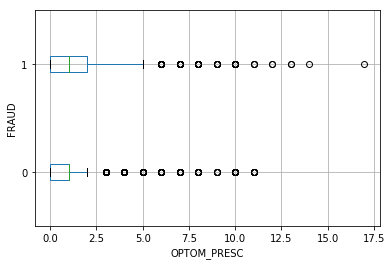
1. (5 points) What percent of investigations are found to be fraudulent? Please give your answer up to 4 decimal places.

FRAUD = 0 in 4771 observations and FRAUD = 1 in 1189 observations. Thus, the percent of fraudulent investigation is 1189 / 5960 = 0.199497 = 19.9497%.

1. (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.







1. (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?

The six eigenvalues are 6847.28, 8387.98, 18064, 315840, 8.44539e+07 and 2.81233e+12 which are all greater than one. Therefore, all six dimensions are used.

* 1. (5 points) Please provide the transformation matrix? You must provide proof that the resulting variables are actually orthonormal.

The transformation matrix is a 6 x 6 matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| -6.50E-08 | -2.41E-07 | 2.70E-07 | -2.43E-07 | -7.90E-07 | 5.96E-07 |
| 7.32E-05 | -0.000294742 | 9.49E-05 | 0.00177762 | 3.52E-06 | 2.21E-10 |
| -0.0118697 | 0.00170828 | -0.000768683 | 2.04E-05 | 1.76E-07 | 9.10E-12 |
| 1.93E-06 | -5.37E-05 | 2.32E-05 | -5.78E-05 | 0.000108753 | 4.33E-09 |
| 0.00083499 | -0.00229965 | -0.0072551 | 1.12E-05 | 2.39E-07 | 2.86E-11 |
| 0.00210965 | 0.0105319 | -0.00145669 | 4.86E-05 | 6.77E-07 | 4.67E-11 |

When we multiply the 5960 x 6 matrix of input fields with this 6 x 6 matrix, we got the 5960 x 6 matrix of transformed input fields. The product of the transpose of this matrix of transformed input fields and the matrix itself yield a 6 x 6 matrix which is shown below. Since this 6 x 6 matrix is practically an identity matrix, we can conclude that the columns of the matrix of transformed input fields are orthonormal.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1 | -1.55E-15 | 8.16E-16 | 6.27E-15 | 1.16E-15 | -6.68E-17 |
| -1.55E-15 | 1 | -5.88E-16 | -2.93E-14 | -1.50E-15 | 4.51E-16 |
| 8.16E-16 | -5.88E-16 | 1 | 2.88E-15 | -1.25E-16 | -1.25E-16 |
| 6.27E-15 | -2.93E-14 | 2.88E-15 | 1 | 1.13E-14 | -3.69E-15 |
| 1.16E-15 | -1.50E-15 | -1.25E-16 | 1.13E-14 | 1 | -7.22E-16 |
| -6.68E-17 | 4.51E-16 | -1.25E-16 | -3.69E-15 | -7.22E-16 | 1 |

1. (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

The score function returns a scalar number 0.8778523489932886.

* 1. (5 points) Explain the meaning of the function return value.

This is the fraction of observations which are correctly classified. The commonly used Misclassification rate is one minus this number. In our case, the misclassification rate is 0.1221476510067114 or approximately 12.21%.

1. (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.

Since the data is sorted in ascending order to TOTAL\_SPEND, we can find this observation from the input data. This observation has a CASE\_ID = 589 which corresponds to the index 588. The five neighbors have indices 588, 2897, 1199, 1246, and 886. Therefore, the five neighbors are

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| CASE\_ID | TOTAL\_SPEND | DOCTOR\_VISITS | NUM\_CLAIMS | MEMBER\_DURATION | OPTOM\_PRESC | NUM\_MEMBERS | FRAUD |
| 589 | 7500 | 15 | 3 | 127 | 2 | 2 | 1 |
| 2898 | 16000 | 18 | 3 | 146 | 3 | 2 | 1 |
| 1200 | 10000 | 16 | 3 | 124 | 2 | 1 | 1 |
| 1247 | 10200 | 13 | 3 | 119 | 2 | 3 | 1 |
| 887 | 8900 | 22 | 3 | 166 | 1 | 2 | 1 |

1. (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?

Since the FRAUD values of all five neighbors are 1, the predicted probability of fraudulent of the observation is also 1. The conclusion is the observation will be predicted fraudulent too and it will NOT be misclassified.