CS 584-04: Machine Learning

Fall 2019 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. (5 points) According to Izenman (1991) method, what is the recommended bin-width for the histogram of x?

There are 1,001 observations in the NormalSample.csv file. The 25% percentile (i.e., the Q1) of the field x is 30.4 and the 75% percentile (i.e., the Q3) is 32.4. Thus, the interquartile range (i.e., the IQR) is 32.4 – 30.4 = 2. According to Izenman (1991) method, the recommended bin-width is  
h = 2 \* 2 \* 1001-1/3 = 0.399867 (rounded to the sixth decimal places).

1. (5 points) What are the minimum and the maximum values of the field x?

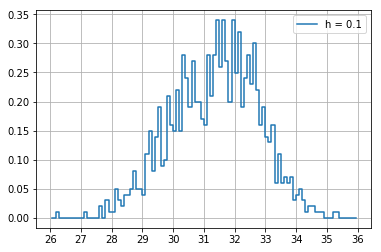
The minimum value of the field x is 26.3 and the maximum value of the field x is 35.4.

1. (5 points) Let a be the largest integer less than the minimum value of the field x, and b be the smallest integer greater than the maximum value of the field x. What are the values of a and b?

The value of a is 26 and the value of b is 36.

1. (5 points) Use h = 0.1, minimum = a and maximum = b. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.

| Midpoint | Density |  | Midpoint | Density |  | Midpoint | Density |  | Midpoint | Density |  | Midpoint | Density |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 26.05 | 0 |  | 28.05 | 0.0100 |  | 30.05 | 0.1499 |  | 32.05 | 0.2498 |  | 34.05 | 0.0400 |
| 26.15 | 0 |  | 28.15 | 0.0500 |  | 30.15 | 0.2198 |  | 32.15 | 0.3197 |  | 34.15 | 0.0500 |
| 26.25 | 0.0100 |  | 28.25 | 0.0300 |  | 30.25 | 0.1499 |  | 32.25 | 0.1898 |  | 34.25 | 0.0300 |
| 26.35 | 0 |  | 28.35 | 0.0200 |  | 30.35 | 0.2797 |  | 32.35 | 0.2398 |  | 34.35 | 0.0100 |
| 26.45 | 0 |  | 28.45 | 0.0400 |  | 30.45 | 0.2398 |  | 32.45 | 0.2797 |  | 34.45 | 0.0200 |
| 26.55 | 0 |  | 28.55 | 0.0400 |  | 30.55 | 0.1898 |  | 32.55 | 0.2298 |  | 34.55 | 0.0200 |
| 26.65 | 0 |  | 28.65 | 0.0500 |  | 30.65 | 0.2697 |  | 32.65 | 0.2997 |  | 34.65 | 0.0100 |
| 26.75 | 0 |  | 28.75 | 0.0799 |  | 30.75 | 0.1998 |  | 32.75 | 0.2198 |  | 34.75 | 0.0100 |
| 26.85 | 0 |  | 28.85 | 0.0500 |  | 30.85 | 0.1998 |  | 32.85 | 0.1598 |  | 34.85 | 0.0100 |
| 26.95 | 0 |  | 28.95 | 0.0500 |  | 30.95 | 0.1698 |  | 32.95 | 0.1898 |  | 34.95 | 0 |
| 27.05 | 0 |  | 29.05 | 0.0400 |  | 31.05 | 0.1598 |  | 33.05 | 0.1399 |  | 35.05 | 0 |
| 27.15 | 0.0100 |  | 29.15 | 0.1099 |  | 31.15 | 0.2797 |  | 33.15 | 0.1299 |  | 35.15 | 0 |
| 27.25 | 0 |  | 29.25 | 0.1499 |  | 31.25 | 0.2098 |  | 33.25 | 0.1598 |  | 35.25 | 0.0100 |
| 27.35 | 0 |  | 29.35 | 0.0799 |  | 31.35 | 0.2797 |  | 33.35 | 0.0599 |  | 35.35 | 0.0100 |
| 27.45 | 0 |  | 29.45 | 0.1399 |  | 31.45 | 0.3397 |  | 33.45 | 0.1099 |  | 35.45 | 0 |
| 27.55 | 0 |  | 29.55 | 0.1898 |  | 31.55 | 0.2597 |  | 33.55 | 0.0599 |  | 35.55 | 0 |
| 27.65 | 0.0200 |  | 29.65 | 0.0899 |  | 31.65 | 0.3397 |  | 33.65 | 0.0699 |  | 35.65 | 0 |
| 27.75 | 0 |  | 29.75 | 0.0999 |  | 31.75 | 0.2697 |  | 33.75 | 0.0599 |  | 35.75 | 0 |
| 27.85 | 0.0300 |  | 29.85 | 0.2098 |  | 31.85 | 0.1998 |  | 33.85 | 0.0699 |  | 35.85 | 0 |
| 27.95 | 0.0100 |  | 29.95 | 0.1598 |  | 31.95 | 0.3397 |  | 33.95 | 0.0300 |  | 35.95 | 0 |



1. (5 points) Use h = 0.5, minimum = a and maximum = b. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | Midpoint | Density |  | Midpoint | Density | | 26.25 | 0.0020 |  | 31.25 | 0.2537 | | 26.75 | 0.0000 |  | 31.75 | 0.2817 | | 27.25 | 0.0020 |  | 32.25 | 0.2557 | | 27.75 | 0.0120 |  | 32.75 | 0.2198 | | 28.25 | 0.0300 |  | 33.25 | 0.1199 | | 28.75 | 0.0539 |  | 33.75 | 0.0579 | | 29.25 | 0.1039 |  | 34.25 | 0.0300 | | 29.75 | 0.1499 |  | 34.75 | 0.0100 | | 30.25 | 0.2078 |  | 35.25 | 0.0040 | | 30.75 | 0.2058 |  | 35.75 | 0 | |  |

1. (5 points) Use h = 1, minimum = a and maximum = b. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Midpoint | Density | | 26.5 | 0.0010 | | 27.5 | 0.0070 | | 28.5 | 0.0420 | | 29.5 | 0.1269 | | 30.5 | 0.2068 | | 31.5 | 0.2677 | | 32.5 | 0.2378 | | 33.5 | 0.0889 | | 34.5 | 0.0200 | | 35.5 | 0.0020 | |  |

1. (5 points) Use h = 2, minimum = a and maximum = a. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Midpoint | Density | | 27.0 | 0.0040 | | 29.0 | 0.0844 | | 31.0 | 0.2373 | | 33.0 | 0.1633 | | 35.0 | 0.0110 | |  |

1. (5 points) Among the four histograms, which one, in your honest opinions, can best provide your insights into the shape and the spread of the distribution of the field x? Please state your arguments.

There is no definitive 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, that histogram is able to tell the mode with more precision.

# Question 2 (20 points)

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

1. (5 points) What is the five-number summary of x? What are the values of the 1.5 IQR whiskers?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Minimum | Q1 | Median | Q3 | Maximum |
| 26.3 | 30.4 | 31.5 | 32.4 | 35.4 |

The IQR = Q3 – Q1 = 32.4 – 30.4 = 2.

The lower whisker is Q1 – 1.5\*IQR = 30.4 – 1.5 \* 2 = 27.4.

The upper whisker is Q3 + 1.5\*IQR = 32.4 + 1.5 \* 2 = 35.4.

1. (5 points) What is the five-number summary of x for each category of the group? What are the values of the 1.5 IQR whiskers for each category of the group?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| group | Minimum | Q1 | Median | Q3 | Maximum |
| 0 | 26.3 | 29.4 | 30.0 | 30.6 | 32.2 |
| 1 | 29.1 | 31.4 | 32.1 | 32.7 | 35.4 |

For group = 0, the IQR = Q3 – Q1 = 30.6 – 29.4 = 1.2.

The lower whisker is Q1 – 1.5\*IQR = 29.4 – 1.5 \* 1.2 = 27.6.

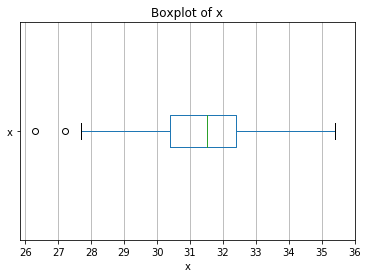
The upper whisker is Q3 + 1.5\*IQR = 30.6 + 1.5 \* 1.2 = 32.4.

For group = 1, the IQR = Q3 – Q1 = 32.7 – 31.4 = 1.3.

The lower whisker is Q1 – 1.5\*IQR = 31.4 – 1.5 \* 1.3 = 29.45.

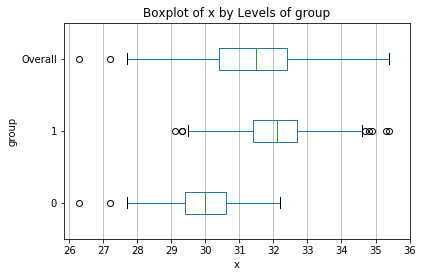
The upper whisker is Q3 + 1.5\*IQR = 32.7 + 1.5 \* 1.3 = 34.65.

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



Comparing the whiskers’ position with the answers in part a), we conclude that the Python’s boxplot has correctly displayed the 1.5 IQR whiskers.

1. (5 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 the group.  
   *Hint: Consider using the CONCAT function in the PANDA module to append observations*.



There are two outliers in the entire data (i.e., group = Overall). They are:

|  |  |
| --- | --- |
| Index | x |
| 70 | 27.2 |
| 295 | 26.3 |

There are two outliers in the group = 0. They are the same as those of the entire data.

|  |  |
| --- | --- |
| Index | x |
| 70 | 27.2 |
| 295 | 26.3 |

There are eight outliers in the group = 1. They are:

|  |  |
| --- | --- |
| Index | x |
| 30 | 35.3 |
| 107 | 29.3 |
| 297 | 35.4 |
| 812 | 34.9 |
| 846 | 34.7 |
| 907 | 34.8 |
| 938 | 29.3 |
| 975 | 29.1 |

# 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.

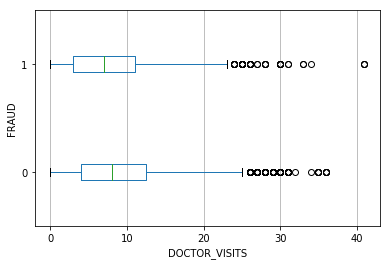
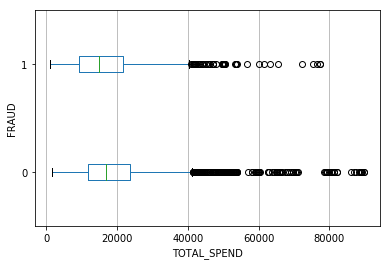
1. TOTAL\_SPEND: Total amount of claims in dollars
2. DOCTOR\_VISITS: Number of visits to a doctor
3. NUM\_CLAIMS: Number of claims made recently
4. MEMBER\_DURATION: Membership duration in number of months
5. OPTOM\_PRESC: Number of optical examinations
6. NUM\_MEMBERS: Number of members covered

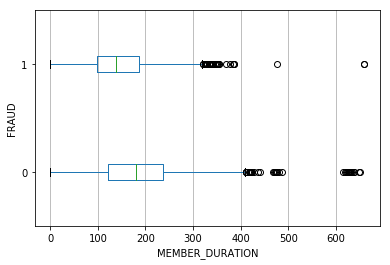
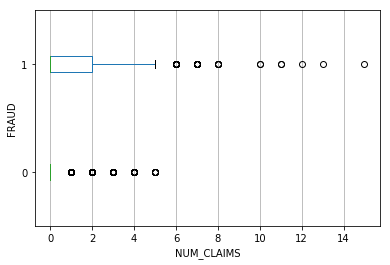
You are asked to use the Nearest Neighbors algorithm to predict the likelihood of fraud.

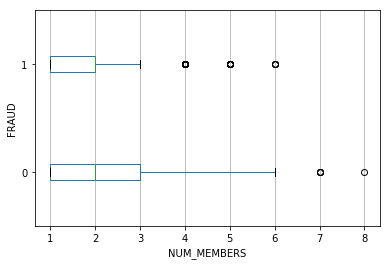
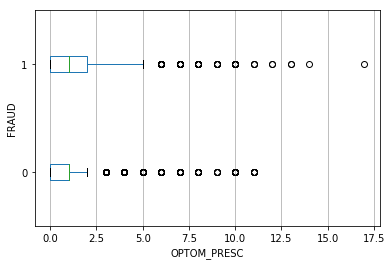
1. (5 points) What percent of investigations are found to be fraudulent? Please give your answer up to 4 decimal places.

We observed FRAUD = 0 in 4,771 observations and FRAUD = 1 in 1,189 observations. Thus, the percent of fraudulent investigation is 1189 / (4771 + 1189) = 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 6.8473e+03, 8.3880e+03, 1.8064e+04, 3.1584e+05, 8.4454e+07, and 2.8123e+12. Since all six eigenvalues are greater than one, 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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| -6.4986E-08 | -2.4119E-07 | 2.6994E-07 | -2.4253E-07 | -7.9049E-07 | 5.9629E-07 |
| 7.3166E-05 | -2.9474E-04 | 9.4886E-05 | 1.7776E-03 | 3.5160E-06 | 2.2056E-10 |
| -1.1870E-02 | 1.7083E-03 | -7.6868E-04 | 2.0367E-05 | 1.7640E-07 | 9.0994E-12 |
| 1.9252E-06 | -5.3709E-05 | 2.3204E-05 | -5.7833E-05 | 1.0875E-04 | 4.3267E-09 |
| 8.3499E-04 | -2.2996E-03 | -7.2551E-03 | 1.1151E-05 | 2.3924E-07 | 2.8577E-11 |
| 2.1096E-03 | 1.0532E-02 | -1.4567E-03 | 4.8584E-05 | 6.7660E-07 | 4.6657E-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 a score function.
   1. (5 points) Run the score function, provide the function return value

The score function returns a scalar number 0.8778523489932886.

* 1. (5 points) Explain the meaning of the score 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. *Reminder: transform the input observation using the results in c) before finding the neighbors*.

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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Index | TOTAL\_SPEND | DOCTOR\_VISITS | NUM\_CLAIMS | MEMBER\_DURATION | OPTOM\_PRESC | NUM\_MEMBERS | FRAUD |
| 588 | 7500 | 15 | 3 | 127 | 2 | 2 | 1 |
| 2897 | 16000 | 18 | 3 | 146 | 3 | 2 | 1 |
| 1199 | 10000 | 16 | 3 | 124 | 2 | 1 | 1 |
| 1246 | 10200 | 13 | 3 | 119 | 2 | 3 | 1 |
| 886 | 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 and, therefore, the predicted FRAUD is 1. The conclusion is since the original observation has FRAUD = 1, this observation is NOT misclassified.