CS 584-04: Machine Learning

Fall 2018 Assignment 1

Diego Martin Crespo A20432558

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.

a) (4 points) According to Izenman (1991) method, what is the recommended bin-width for the histogram of x?

Lzenman says that because of h* depends upon the unknown f though R(f´). An estimate f* of f can be plugged into h*. There is a robust rule that normally gives reasonable results.

Expression: $h^{*} = 2(IQR)n^{-1/3}$, being IQR the interquartile rage of the data.

Note1: as it is explained in the class notes in practice we will round h to a nicer value.

Expression: h = 10^ sign(log$_{10}$(h)) * ceil(abs(log$_{10}$(h)))

From the expression of h* above "n" is known, it is the number of samples (the NormalSample.csv has 100 samples of x). Otherwise the IQR needs to be calculated:

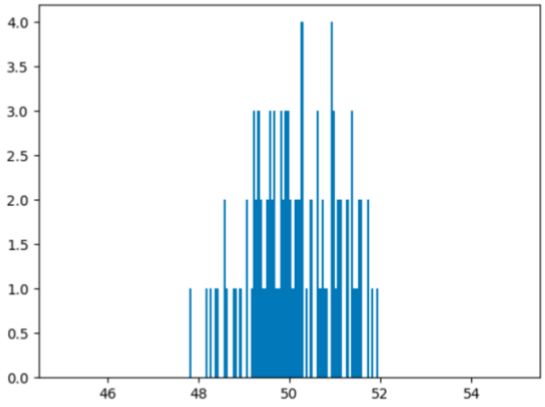
1. Find the median of the set x (needs to be ordered first)
2. Divide in two sets the array (the median or Q2, not included)
3. Find the median of each set, the first set median will correspond to Q1 and the second to Q3.
4. The IQR is defined as: IQR= Q3-Q1

With the given data and the proper calculations an h*= 0.6355582335594063 results from the python script.

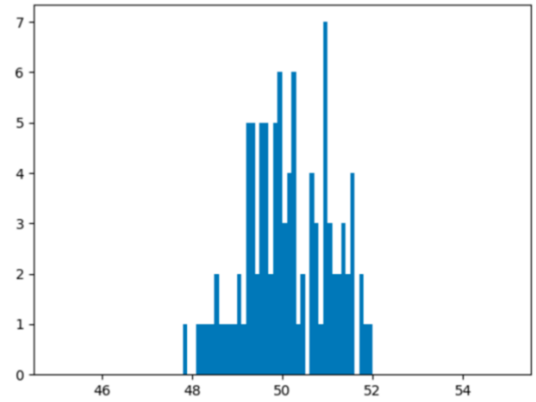
b) (3 points) What is the bin-width after applying the beautification step?

If we consider the note1 from a). After the process of beautification, the h* is rounded to hb=0.1

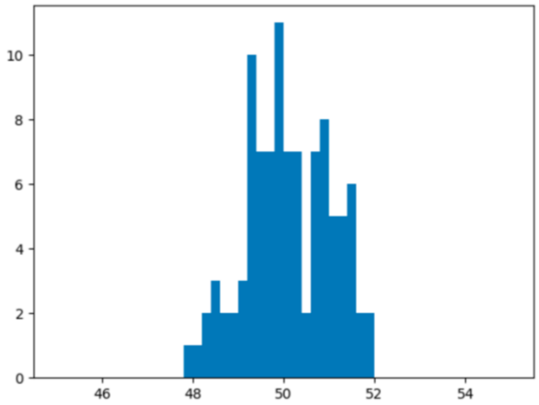
c) (10 points) Use h = 0.5, minimum = 45 and maximum = 55. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.



d) (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.



e) (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.



f) (3 points) Among the three histograms, which one, in your opinions, can best describe the distribution of the field x?
The 3 of them can describe the distribution of x. In c) is the bins are too small, you can appreciate many spaces in between the different values of x. However, in e) the bins are quite big, so you are “missing” some values of x (not missing, approximating them into the h range). Therefore, in my opinion the histogram that best represent the distribution is d).

Question 2 (20 points)

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

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

We already calculated Q1, Q2 and Q3. For the other two: min= Q1-1,5xIQR and max=Q3+1.5xIQR. Therefore, these are the 5 values:

- Min= 47.2325
- Q1= 49.445
- Q2= 50.03
- Q3= 50.92
- Max= 53.1325

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

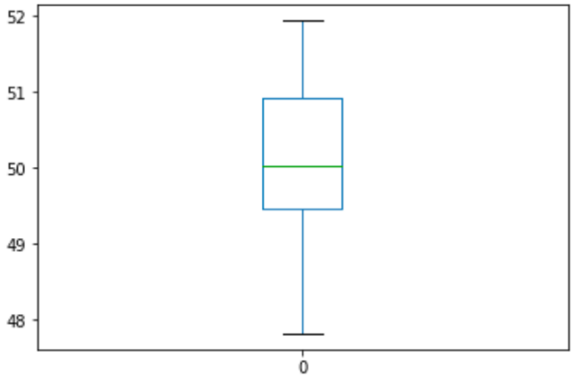
The five-number summary of x with group 0 are:

- Min= 46.76
- Q1= 49.28
- Q2= 50.22
- Q3= 50.96
- Max= 53.48

Max= 51.8 The five-number summary of x with group 1 are:

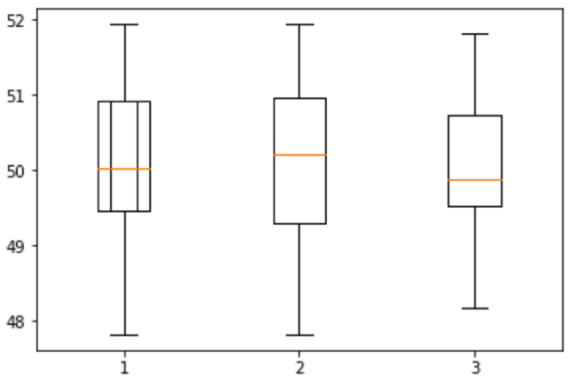
- Min= 47.665
- Q1= 49.51
- Q2= 49.88
- Q3= 50.74
- Max= 52.5845

c) (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?



Yes, it does the whiskers as it is shown in the previous image. This has been plot using the pyplot's python library.

d) (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 previous image shows the 3 boxplots for the hole group of x, only group 1 and only group 2. The first boxplot corresponds to x without group, the second is the boxplot of group 0 and the third is the boxplot of group 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.

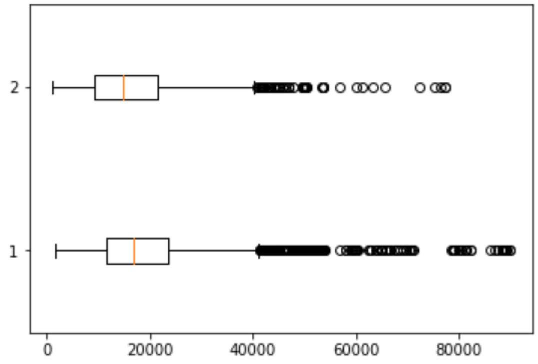
a) (5 points) What percent of investigations are found to be fraudulent? Please give your answer up to 4 decimal places.

   The 19.9497% of the investigations are fraudulent

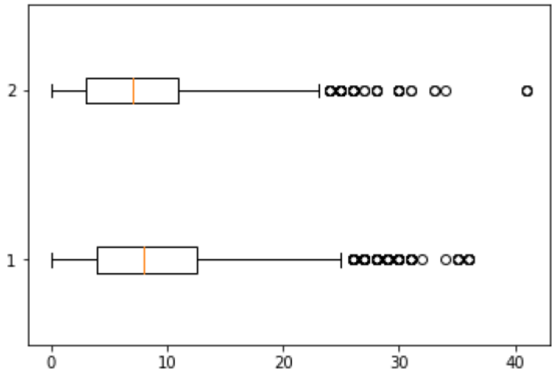
b) (5 points) Use the BOXPLOT function to produce horizontal box-plots. For each interval variable, one box-plot for the fraudulent observations, and another box-plot for the non-fraudulent observations. These two box-plots must appear in the same graph for each interval variable.

   These are the resulting box-plots:

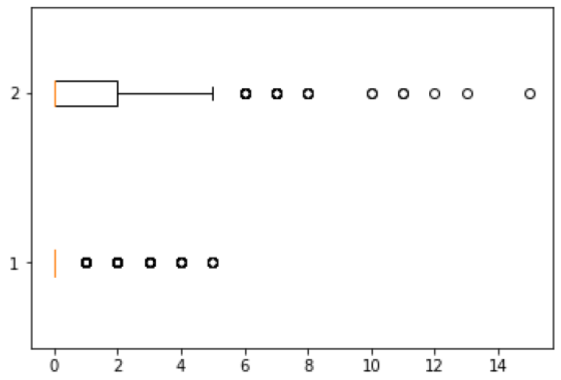
   Note: the row 1 is for non-fraudulent incidents and row 2 for fraudulent



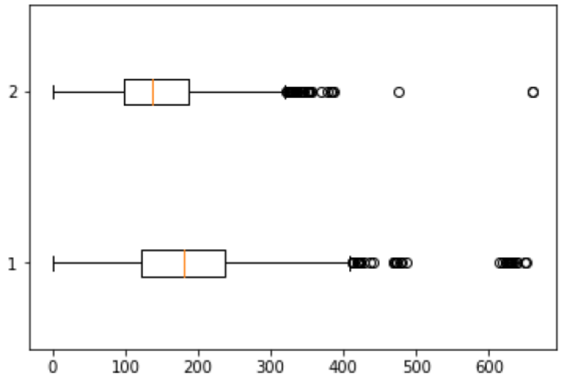
Box-plot of variable: 'TOTAL_SPEND'



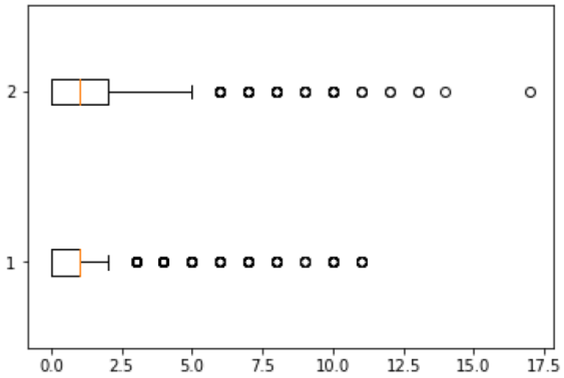
Box-plot of variable: 'DOCTOR_VISITS'



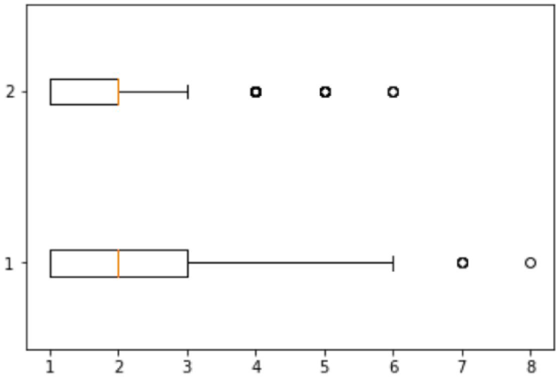
Box-plot of variable: 'NUM_CLAIMS'



Box-plot of variable: 'MEMBER_DURATION'



Box-plot of variable: 'OPTOM_PRESC'



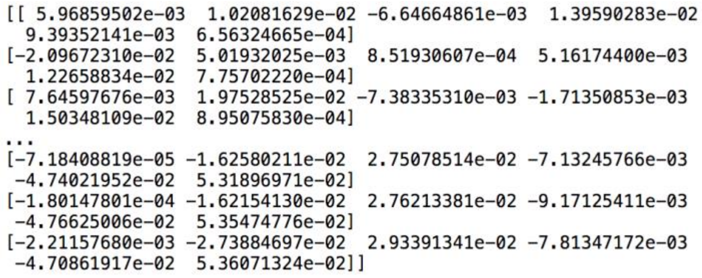
Box-plot of variable: 'NUM_MEMBERS'

c) (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.

i. (5 points) How many dimensions are used?

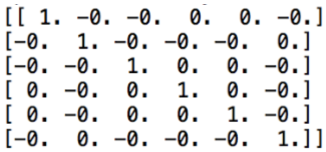
There are 6 eigenvalues greater that one. Therefore, 6 dimensions are used.
[6.84728061e+03 8.38798104e+03 1.80639631e+04 3.15839942e+05
8.44539131e+07 2.81233324e+12]

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



```
[[ 5.96859502e-03  1.02081629e-02 -6.64664861e-03  1.39590283e-02
   9.39352141e-03  6.56324665e-04]
 [-2.09672310e-02  5.01932025e-03  8.51930607e-04  5.16174400e-03
   1.22658834e-02  7.75702220e-04]
 [ 7.64597676e-03  1.97528525e-02 -7.38335310e-03 -1.71350853e-03
   1.50348109e-02  8.95075830e-04]
 ...
 [-7.18408819e-05 -1.62580211e-02  2.75078514e-02 -7.13245766e-03
  -4.74021952e-02  5.31896971e-02]
 [-1.80147801e-04 -1.62154130e-02  2.76213381e-02 -9.17125411e-03
  -4.76625006e-02  5.35474776e-02]
 [-2.21157680e-03 -2.73884697e-02  2.93391341e-02 -7.81347172e-03
  -4.70861917e-02  5.36071324e-02]]
```

The image above shows the transformation matrix of X= **XVD**$^{-1/2}$



```
[[ 1. -0. -0.  0.  0. -0.]
[-0.  1. -0. -0. -0.  0.]
[-0. -0.  1.  0.  0. -0.]
[ 0. -0.  0.  1.  0. -0.]
[ 0. -0.  0.  0.  1. -0.]
[-0.  0. -0. -0. -0.  1.]]
```

The matrix above proves that the transformation matrix is correct. It has been calculated with the formula: $(\mathbf{XVD}^{-1/2})\,{}^{t}(\mathbf{XVD}^{-1/2}) = \mathbf{I}$

d) (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.

i. (5 points) Run this function, provide the function return value

The resulting score is: 0.8416107382550335.

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

Returns the mean accuracy on the given test data and labels.

e) (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.

The 5 neighbors are: 588, 577, 582, 573 and 575.

- The input variable values are the training data set (columns) made by the variables: TOTAL_SPEND, DOCTOR_VISITS, NUM_CLAIMS, MEMBER_DURATIO, OPTOM_PRESC and NUM_MEMBERS.
- The target values are in the FRAUD column from the data set

f) (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?

The given observation is fraud and the predicted is also fraud. The predicted probability of the observation is 20% of being fraud. Therefore, as it is greater than the probability calculated in a) the observation is well classified.