Cristina Blanco Garrido
A20431766

CS 584-04: Machine Learning

Fall 2018 Assignment 1

INSTRUCTIONS

1. Students should complete this assignment independently.
2. Students are encouraged to use Python to complete this assignment.
3. Students must submit their answers to Blackboard before 11:59 PM on September 5, 2018.

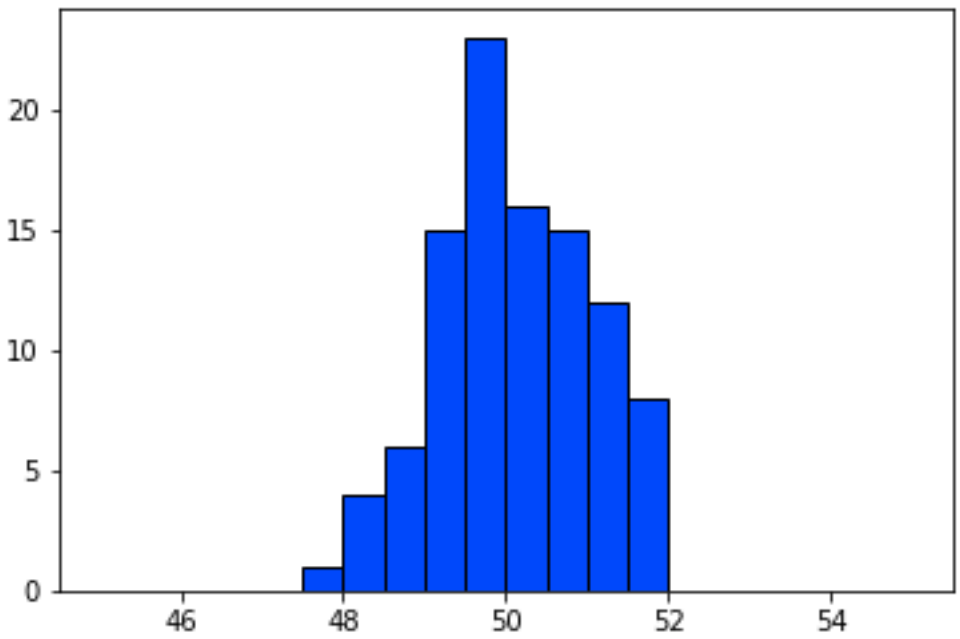
Question 1 (40 points)

Write a Python program to calculate the density estimator of a histogram. Use the field *x* in the NormalSample.csv file.

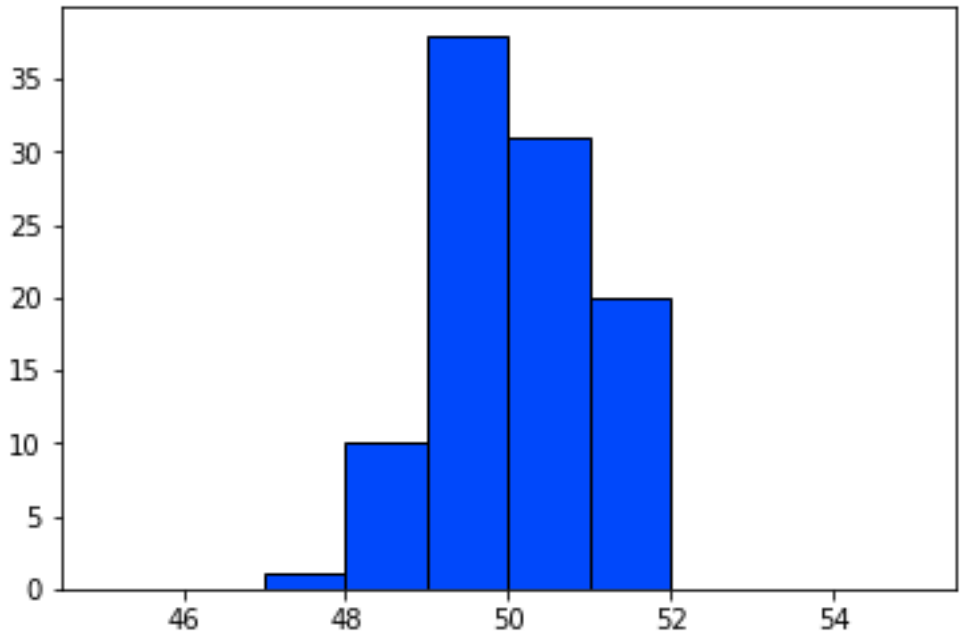
a) (4 points) According to Izenman (1991) method, what is the recommended bin-width for the histogram of x?
According to Izenman method h = 0.6237088427642296

b) (3 points) What is the bin-width after applying the beautification step?
After beautification h = 0.1

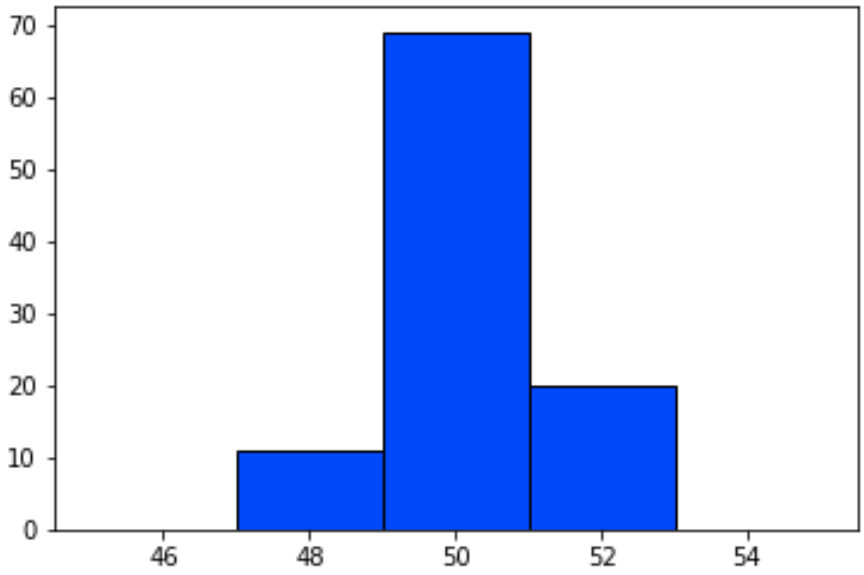
c) (10 points) Use h = 0.5, minimum = 45 and maximum = 55. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.
Coordinates of the density estimator:
[0.0, 0.0, 0.0, 0.0, 0.0, 0.02, 0.08, 0.12, 0.3, 0.46, 0.32, 0.3, 0.24, 0.16, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
Histogram:



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.
Coordinates of the density estimator:
[0.0, 0.0, 0.01, 0.1, 0.38, 0.31, 0.2, 0.0, 0.0, 0.0]
Histogram:



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.
Coordinates of the density estimator:
[0.0, 0.055, 0.345, 0.1, 0.0]
Histogram:



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

In my opinion, the histogram that best describe the distribution is c), which its bin-with is between the one calculated with the Izenman method and the value obtained after beautification.

Question 2 (20 points)

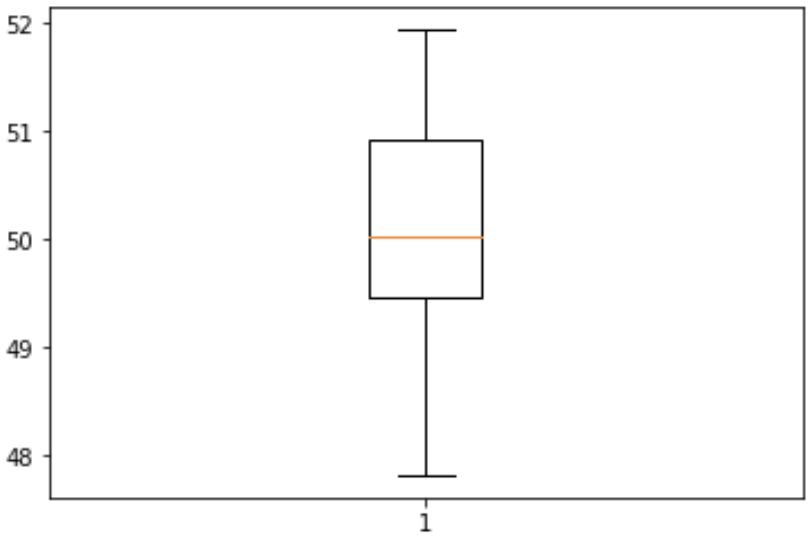
Use in the NormalSample.csv to generate box-plots for answering the following questions.

a) (2 points) What are the five-number summary of x?
Q1 = 49.4675
Q2 = 50.03
Q3 = 50.915
Minimum = 47.82
Maximum = 51.94

b) (3 points) What are the five-number summary of x for each category of Group?
Group 0:
Q1 = 49.295
Q2 = 50.22
Q3 = 50.96
Minimum = 47.82
Maximum = 51.94

Group 1:
Q1 = 49.53
Q2 = 49.88
Q3 = 50.74
Minimum = 48.17
Maximum = 51.82

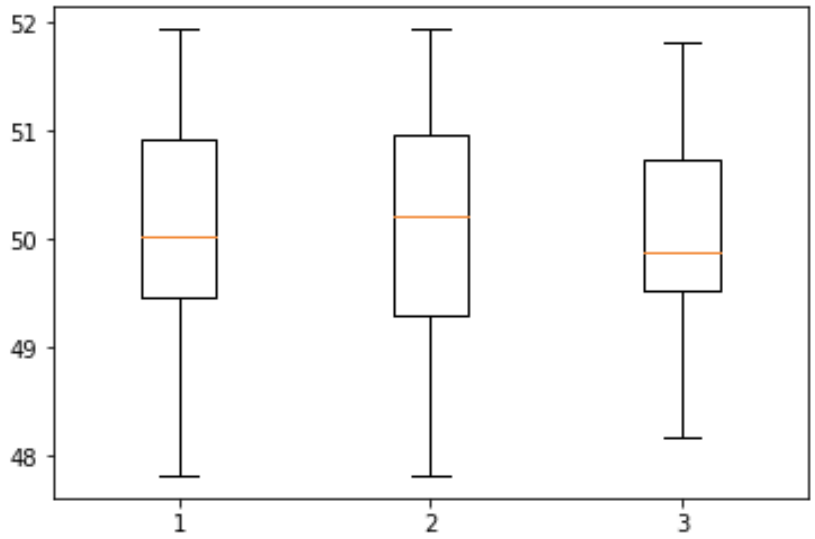
c) (5 points) Draw a boxplot of x (without Group) using the Python boxplot function. Can you tell if the Python's boxplot has displayed the 1.5 IQR whiskers?
Lower Whisker = 47.29625
Upper Whisker = 53.08624999999999
The Python boxplot function has the 1.5 IQR whiskers as default option, however the whiskers were not displayed at the boxplot.



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.

Group 0: (boxplot 2)
Lower Whisker = 46.7975
Upper Whisker = 53.457499999999996

Group 1: (boxplot 3)
Lower Whisker = 47.715
Upper Whisker = 52.55500000000001



Any of the three groups of data has outliers of x.

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.

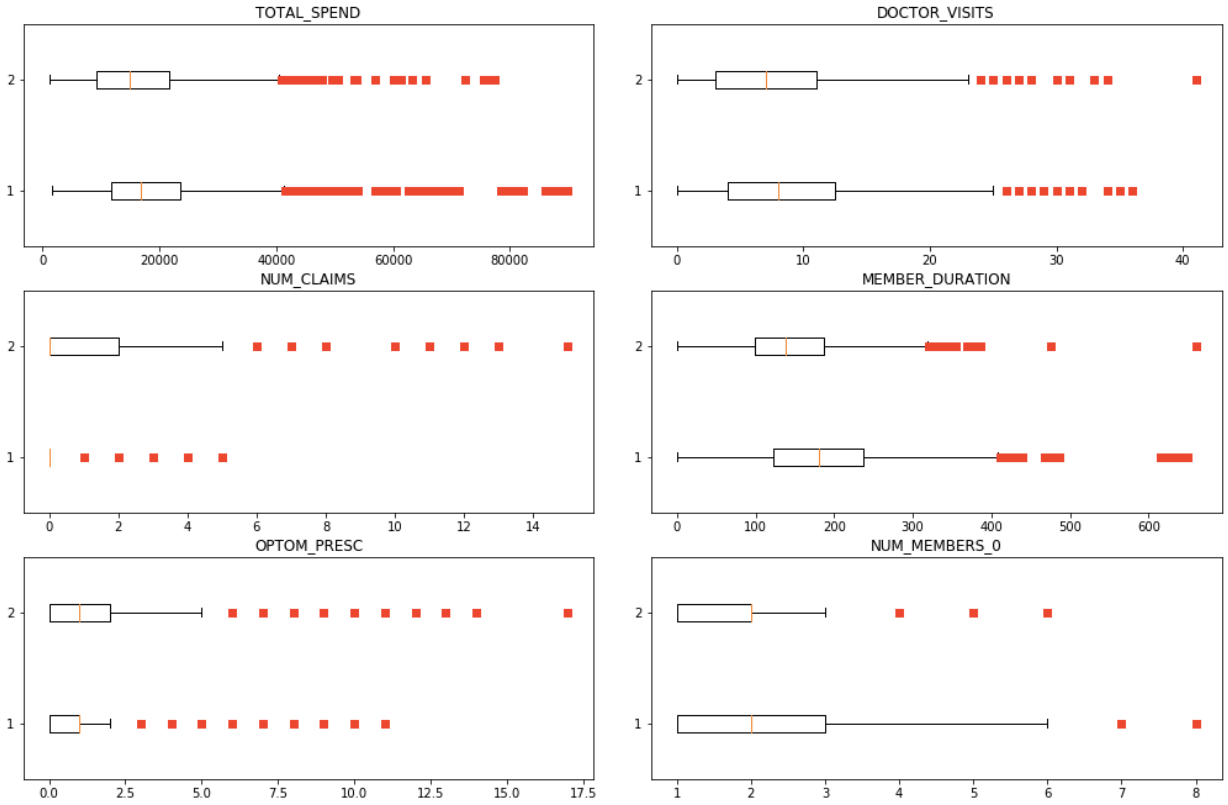
1. TOTAL_SPEND: Total amount of claims in dollars
2. DOCTOR_VISITS: Number of visits to a doctor
3. NUM_CLAIMS: Number of claims made recently
4. MEMBER_DURATION: Membership duration in number of months
5. OPTOM_PRESC: Number of optical examinations
6. NUM_MEMBERS: Number of members covered

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 investigation were found to be 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.

1: for the non-fraudulent observations 2: for the fraudulent observations



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.

Eigenvalues:

6847.28	8387.98	18064	315840	8.44539e+07	2.81233e+12

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

As all the eigenvalues are greater than one we will use all the dimensions. Therefore, six dimensions are used.

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

Transformation matrix:

-6.49862e-08	-2.41195e-07	2.69941e-07	-2.42526e-07	-7.90493e-07	5.96287e-07
7.31657e-05	-0.000294742	9.48856e-05	0.00177762	3.51604e-06	2.2056e-10
-0.0118697	0.00170828	-0.000768683	2.03673e-05	1.76401e-07	9.09939e-12
1.92524e-06	-5.37086e-05	2.32038e-05	-5.78328e-05	0.000108753	4.32672e-09
0.00083499	-0.00229965	-0.0072551	1.11508e-05	2.39239e-07	2.85769e-11
0.00210965	0.0105319	-0.00145669	4.85838e-05	6.76601e-07	4.66565e-11

Multiplying the transformation matrix of x with its transpose we obtain the identity matrix, which proof that the resulting variables are orthonormal.

1	-3.00432e-16	-4.6122e-16	5.45324e-15	1.20997e-15	-1.28912e-16
-3.00432e-16	1	-6.4445e-16	-2.76821e-14	-1.23512e-15	7.78891e-16
-4.6122e-16	-6.4445e-16	1	3.50891e-15	1.00614e-16	-2.25514e-16
5.45324e-15	-2.76821e-14	3.50891e-15	1	1.1486e-14	-3.47812e-15
1.20997e-15	-1.23512e-15	1.00614e-16	1.1486e-14	1	-6.31439e-16
-1.28912e-16	7.78891e-16	-2.25514e-16	-3.47812e-15	-6.31439e-16	1

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 return value of this function is: 0.8778523489932886

ii. (5 points) Explain the meaning of the function return value.
This means that the algorithm has an 87,79 % of accuracy. In other words, it will classify correctly in the 87,79% of the cases.

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.

Case ID	Fraud	Total Spend	Doctor Visits	Num. Claims	Member Duration	Optom. Presc.	Num. Members
589	1	7500	15	3	127	2	2
2898	1	16000	18	3	146	3	2
1200	1	10000	16	3	124	2	1
1247	1	10200	13	3	119	2	3
887	1	8900	22	3	166	1	2

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?

This input variable has a 100% of probability of fraudulent, then the observation is classified as fraudulent. In this case the misclassification is not possible, but in other cases this criterion could end up in misclassification, since a 20% of probability is too low to decide.