Tania Soutonglang

CS 484: Introduction to Machine Learning

Autumn 2022 Assignment 1

Question 1 (25 points)

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

a) (5 points) Use the *Pandas describe()* function to find out the count, the mean, the standard deviation, the minimum, the 25th percentile, the median, the 75th percentile, and the maximum. The *describe()* function returns the following statistics, shown no more than 7 decimal places.

```
1.a
              group            x
count  1000.000000  1000.000000
mean      0.686000   314.154000
std       0.464349    13.981336
min       0.000000   263.000000
25%       0.000000   304.000000
50%       1.000000   315.000000
75%       1.000000   324.000000
max       1.000000   354.000000
```

b) (5 points) What is the bin width recommended by the Izenman (1991) method? Please round your answer to the nearest tenths (i.e., one decimal place).

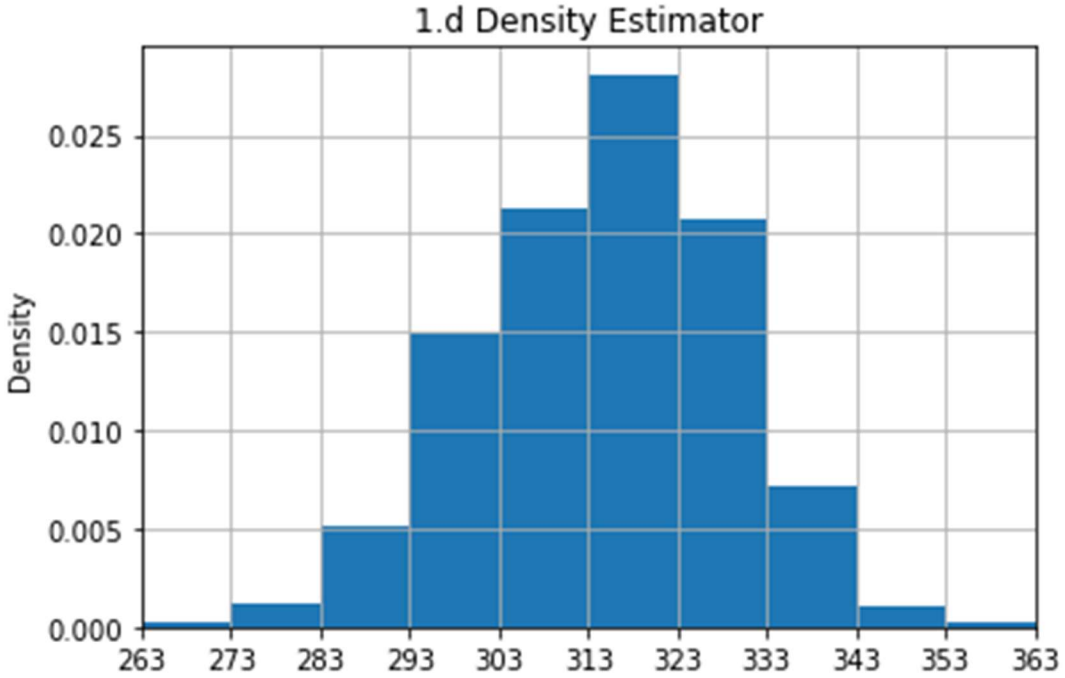
```
1.b
Bin Width =  4.0
```

c) (10 points) Use the Shimazaki and Shinomoto (2007) method and try d = 1, 2, 2.5, 5, 10, 20, 25, and 50. What is the recommended bin width if the number of bins must not fewer than 5 and not more than 20? You need to show your calculations to receive full credit.

```
1.c
      0          1      2      3      4     5
0   1.0 -67.929819  263.0  314.0  354.0  91.0
1   2.0 -76.463754  262.0  314.0  354.0  46.0
2   2.5 -86.595166  262.5  315.0  355.0  37.0
3   5.0 -89.112290  260.0  315.0  355.0  19.0
4  10.0 -95.172000  260.0  310.0  360.0  10.0
5  20.0 -79.357000  260.0  320.0  360.0   5.0
6  25.0 -76.479360  250.0  325.0  375.0   5.0
7  50.0 -49.338756  250.0  300.0  400.0   3.0
```

0 = Delta, 1 = C(Delta), 2 = Low X, 3 = Middle X, 4 = High X, 5 = N Bin
recommended bin width = 10

d) (5 points) Based on your recommended bin width answer in (c), list the mid-points and the estimated density function values. Draw the density estimator as a vertical bar chart using the matplotlib. You need to properly label the graph to receive full credit.



Question 2 (15 points)

The **NormalSample.csv** contains the *group* variable that has two values, namely, 0 and 1.

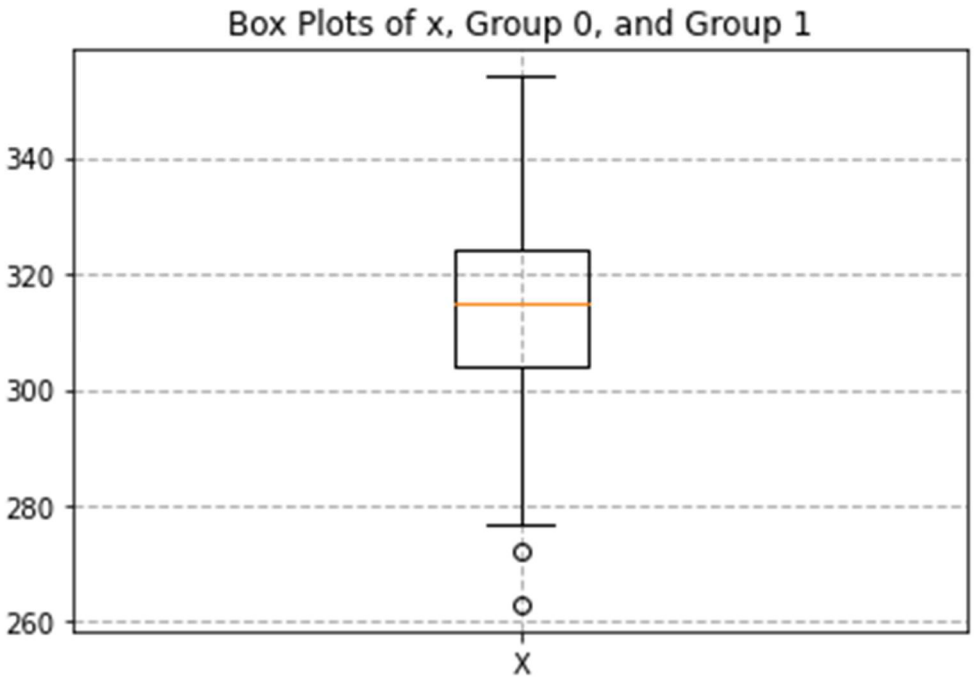
a) (5 points) What is the five-number summary of x for each category of the *group*? What are the values of the 1.5 IQR whiskers for each category of the group

```
2.a
Group 0:
        group           x
count  314.0  314.000000
mean     0.0  300.022293
std      0.0    9.749063
min      0.0  263.000000
25%      0.0  294.000000
50%      0.0  300.000000
75%      0.0  306.000000
max      0.0  322.000000
whisker 1:  288.0
whisker 2:  324.0

Group 1:
        group           x
count  686.0  686.000000
mean     1.0  320.622449
std      0.0   10.402359
min      1.0  291.000000
25%      1.0  314.000000
50%      1.0  321.000000
75%      1.0  327.000000
max      1.0  354.000000
whisker 1:  307.5
whisker 2:  346.5
```

b) (10 points) Draw a graph where it contains the overall boxplot of x, the boxplot of x for each category of *group* (i.e., three horizontal boxplots within the same graph frame). Use the 1.5 IQR whiskers, identify any outliers of x for the entire data and each category of the group. You must properly label your boxplots to receive full credits.

I couldn't get the Group 0 and Group 1 plots to show on the same graph, sorry.


Box Plots of x, Group 0, and Group 1
340
320
300
280
260
X

Question 3 (35 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 = Fraud, 0 = Not Fraud. 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 frauds? This is the empirical fraud rate. Please give your answer up to 4 decimal places.

```
3.a
number of frauds:  1189
number of rows:  5960
probability of fraud occuring:  19.9497
```

b) (10 points) Orthonormalize interval variables and use the orthonormalized columns for the nearest neighbor analysis. Use only the dimensions whose corresponding eigenvalues are greater than one.

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

6 dimensions

```
Eigenvalues of x =
 [6.84728061e+03 8.38798104e+03 1.80639631e+04 3.15839942e+05
 8.44539131e+07 2.81233324e+12]
Eigenvectors of x =
 [[-5.37750046e-06 -2.20900379e-05  3.62806809e-05  1.36298664e-04
   7.26453432e-03  9.99973603e-01]
 [ 6.05433402e-03 -2.69942162e-02  1.27528313e-02 -9.99013423e-01
  -3.23120126e-02  3.69879256e-04]
 [-9.82198935e-01  1.56454700e-01 -1.03312781e-01 -1.14463687e-02
  -1.62110700e-03  1.52596881e-05]
 [ 1.59310591e-04 -4.91894718e-03  3.11864824e-03  3.25018102e-02
  -9.99428355e-01  7.25592222e-03]
 [ 6.90939783e-02 -2.10615119e-01 -9.75101628e-01 -6.26672294e-03
  -2.19857585e-03  4.79234486e-05]
 [ 1.74569737e-01  9.64577791e-01 -1.95782843e-01 -2.73038995e-02
  -6.21788707e-03  7.82430481e-05]]
```

ii. (5 points) Please provide the transformation matrix? Show evidence that the orthonormalized columns are actually orthonormal.

```
Transformation Matrix =
 [[-6.49862374e-08 -0.00000000e+00  0.00000000e+00  0.00000000e+00
   0.00000000e+00  0.00000000e+00]
 [ 0.00000000e+00 -2.94741983e-04  0.00000000e+00 -0.00000000e+00
  -0.00000000e+00  0.00000000e+00]
 [-0.00000000e+00  0.00000000e+00 -7.68683456e-04 -0.00000000e+00
  -0.00000000e+00  0.00000000e+00]
 [ 0.00000000e+00 -0.00000000e+00  0.00000000e+00  5.78327741e-05
  -0.00000000e+00  0.00000000e+00]
 [ 0.00000000e+00 -0.00000000e+00 -0.00000000e+00 -0.00000000e+00
  -2.39238772e-07  0.00000000e+00]
 [ 0.00000000e+00  0.00000000e+00 -0.00000000e+00 -0.00000000e+00
  -0.00000000e+00  4.66565230e-11]]
```

```
The Transformed x =
 [[-7.14848611e-05 -3.24216181e-03  0.00000000e+00  5.43628076e-03
  -2.39238772e-07  9.33130460e-11]
 [-8.44821086e-05 -2.06319388e-03 -1.53736691e-03  7.05559844e-03
   0.00000000e+00  4.66565230e-11]
 [-9.74793561e-05 -1.17896793e-03  0.00000000e+00  8.61708334e-03
  -2.39238772e-07  1.39969569e-10]
 ...
 [-5.79677238e-03 -4.42112974e-03  0.00000000e+00  1.22605481e-02
   0.00000000e+00  9.33130460e-11]
 [-5.83576412e-03 -4.12638776e-03  0.00000000e+00  1.23762137e-02
   0.00000000e+00  9.33130460e-11]
 [-5.84226274e-03 -4.42112974e-03  0.00000000e+00  1.27232103e-02
   0.00000000e+00  4.66565230e-11]]
```

```
Expect an Identity Matrix =
 [[ 1.18764498e-02  1.99237182e-02  2.14367936e-03 -7.66885102e-02
   2.09532785e-06 -6.67155864e-10]
 [ 1.99237182e-02  6.84696080e-02  5.39424528e-03 -1.74971946e-01
   4.06539743e-06 -1.46753335e-09]
 [ 2.14367936e-03  5.39424528e-03  4.68090585e-03 -1.99198842e-02
   6.36106245e-07 -1.70892424e-10]
 [-7.66885102e-02 -1.74971946e-01 -1.99198842e-02  7.77367494e-01
  -1.60964939e-05  5.72338641e-09]
 [ 2.09532785e-06  4.06539743e-06  6.36106245e-07 -1.60964939e-05
   1.39997274e-09 -1.51591791e-13]
 [-6.67155864e-10 -1.46753335e-09 -1.70892424e-10  5.72338641e-09
  -1.51591791e-13  6.40489026e-17]]
```

c) (10 points) Use the KNeighborsClassifier module to execute the Nearest Neighbors algorithm using exactly five neighbors and the orthonormalized columns you have chosen in (b). The KNeighborsClassifier module has a score function.

i. (5 points) Find out from the documentation the purpose of the score function.
From the documentation, the score function is meant to “return the mean accuracy on the given test data and labels.”

ii. (5 points) Run the score function, show and explain the function return value.
It’s the accuracy. I couldn’t get the score function to work properly, sorry.

d) (5 points) For an observation that has the median input variable values, find its **five** neighbors. Please list their input variable values and the target values. *Reminder: transform the input observation using the results in (b) before finding the neighbors*.
CASE_ID = 2980, FRAUD = 0, TOTAL_SPEND = 16300, DOCTOR_VISITS = 9, NUM_CLAIMS = 0, MEMBER_DURATION = 80, OPTOM_PRESC = 2, NUM_MEMBERS = 2
The fit() function couldn’t work for me.

e) (5 points) Follow-up with (d), what is the predicted probability of fraud (i.e., FRAUD = 1)? If your predicted probability is greater than or equal to the empirical fraud rate (i.e., your answer in a), then the observation will be classified as a fraud.
(d) was unsolved.

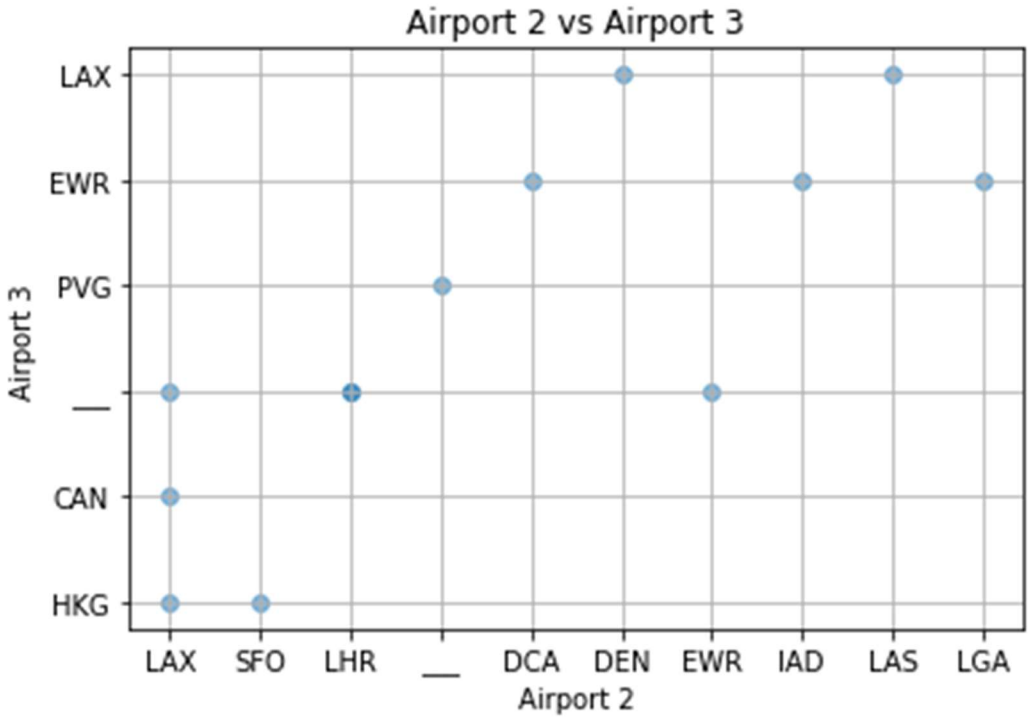
Question 4 (25 points)

I found the following flights from Chicago O'Hare Airport (ORD) to Shanghai Pudong Airport (PVG).

Flight	Carrier 1	Carrier 2	Airport 1	Airport 2	Airport 3	Airport 4
A	American	Cathay Pacific	ORD	LAX	HKG	PVG
B	American	Cathay Pacific	ORD	SFO	HKG	PVG
C	American	China Southern	ORD	LAX	CAN	PVG
D	American	Virgin Atlantic	ORD	LHR		PVG
E	British Airways	Virgin Atlantic	ORD	LHR		PVG
F	United	Virgin Atlantic	ORD	LHR		PVG
G	United		ORD	DCA	EWR	PVG
H	United		ORD	DEN	LAX	PVG
I	United		ORD	EWR		PVG
J	United		ORD	IAD	EWR	PVG
K	United		ORD	LAS	LAX	PVG
L	United		ORD	LAX		PVG
M	United		ORD	LGA	EWR	PVG

To answer the following questions, please replace empty string values in **Airport 3** with three underscore characters (i.e., '___').

a) (5 points) Generate a scatterplot of **Airport 3** (y-axis) versus **Airport 2** (x-axis). Please properly label the axes to receive full credits.



b) (5 points) Generate a frequency table of the airport codes in **Airport 2** and **Airport 3** combined.

```
4.b
Frequency Table:
 LAX    5
___     5
EWR     4
HKG     2
LHR     2
CAN     1
DCA     1
DEN     1
IAD     1
LAS     1
LGA     1
PVG     1
SFO     1
dtype: int64
```

c) (10 points) Suppose a new airline creates a new flight from ORD to PVG that makes one stop at LAX. I want to know which flight(s) most resembles this new flight. Use the Cosine Distance to measure the differences between this flight and the existing flights.
This would be most similar to Flight L.

i. Create a vector of word counts for each flight. This vector has as many elements as the number of unique values found in (b).
ii. Initialize all elements in the vector to zeros.
iii. Count the number of times the airport codes appeared in **Airport 2** and **Airport 3**.
iv. Calculate the Cosine Distance between the new flight and the Flights A to M.

You will list the Cosine Distances in a table.

d) (5 points) Which flight(s) have the shortest Cosine Distance from the new flight?