EE708: Fundamentals of Data Science and Machine Intelligence

Assignment 1

Based on Module 1: Foundational principles of data science and machine intelligence, Module 2: Statistical data analysis, visualization, and inference

1. Imagine we have two possibilities: We can scan and e-mail the image, or we can use an optical character reader (OCR) and send the text file. Discuss the advantages and disadvantages of the two approaches in a comparative manner. When would one be preferable over the other?
2. Assume we are tasked with building a system to distinguish junk e-mail.
   a. What is in a junk e-mail that lets us know it is junk?
   b. How can the computer detect junk through a syntactic analysis?
   c. What would we like the computer to do if it detects a junk e-mail: delete it automatically, move it to a different file, or just highlight it on the screen?
3. Let us say we are tasked with building an automated taxi.
   a. Define the constraints.
   b. What are the inputs?
   c. What is the output?
   d. How can we communicate with the passenger?
   e. Do we need to communicate with the other automated taxis; that is, do we need a "language"?
4. Let us say our hypothesis class is a circle.
   a. What are the parameters?
   b. How can the parameters of a circle hypothesis be calculated in such a case?
   c. What if it is an ellipse?
   d. Why does it make more sense to use an ellipse instead of a circle?
5. A manufacturer says the Z-Phone smartphone has a mean consumer life of 42 months with a standard deviation of 8 months. Assuming a normal distribution, what is the probability that a given random Z-Phone will last between 20 and 30 months?
6. An experiment to investigate the survival time in hours of an electronic component consists of placing the parts in a test cell and running them for 100 hours under elevated temperature conditions. (This is called an "accelerated" life test.) Eight components were tested with the following resulting failure times:
   $$75, 63, 100^{+}, 36, 51, 45, 80, 90$$
   The observation 100^{+} indicates that the unit still functioned at 100 hours. Is there any meaningful measure of location that can be calculated for these data? What is its numerical value?
7. A gasoline manufacturer is investigating the "cold start ignition time" of an automobile engine. The following times (in seconds) were obtained for a test vehicle: 1.75, 1.92, 2.62, 2.35, 3.09, 3.15, 2.53, 1.91. Calculate the sample mean, sample variance, and sample standard deviation. Construct a box plot of the data. A second formulation of the gasoline was tested in the same vehicle, with the following times (in seconds): 1.83, 1.99, 3.13, 3.29, 2.65, 2.87, 3.40, 2.46, 1.89, and 3.35. Use these new data, along with the cold start times reported in the previous exercise, to construct comparative box plots. Write an interpretation of the information that you see in these plots.
8. A set of 10 hypothetical patient records from a large database are presented in Table 1. Patients with a diabetes value of 1 have type-II diabetes, and patients with a diabetes value of 0 do not have type-II diabetes.
   a. Create a new column by normalizing the Weight (kg) variable into the range 0-1 using the min-max normalization.
   b. Create a new column by binning the Weight (kg) variable into three categories: low (less than 60 kg), medium (60-100 kg), and high (greater than 100 kg).
   c. Create an aggregated column, body mass index (BMI, which is defined by the formula:
   $$\text{BMI} = \frac{\text{Weight (kg)}}{\left(\text{Height (m)}\right)^2}$$

Table 1 Table of Patient Records

Name	Weight (kg)	Height (m)	Systolic Blood Pressure (mm Hg)	Diastolic Blood Pressure (mm Hg)	Diabetes
P. Lee	50	1.52	68	112	0
R. Jones	115	1.77	110	154	1
J. Smith	96	1.83	88	136	0
A. Patel	41	1.55	76	125	0
M. Owen	79	1.82	65	105	0
S. Green	109	1.89	114	159	1
N. Cook	73	1.76	108	136	0
W. Hands	104	1.71	107	145	1
P. Rice	64	1.74	101	132	0
F. Marsh	136	1.78	121	165	1

9. Table 2 shows a series of retail transactions monitored by the main office of a computer store.
   a. Create a histogram of Sale Price ($) using the following intervals: 0 to less than 250, 250 to less than 500, 500 to less than 750, and 750 to less than 1000.
   b. Generate a contingency table summarizing the variables Store and Product category.
   c. Generate the following summary tables:
      i. Grouping by Customer with a count of the number of observations and the sum of the Sale price ($) for each row.
      ii. Grouping by Store with a count of the number of observations and the mean Sale price ($) for each row.
      iii. Grouping by Product category with a count of the number of observations and the sum of the Profit ($) for each row.
   d. Create a scatterplot showing Sales price ($) against Profit ($).

Table 2 Retail Transaction Data Set

Customer	Store	Product Category	Product Description	Sale Price ($)	Profit ($)
B. March	New York. NY	Laptop	DR2984	950	190
B. March	New York, NY	Printer	FW288	350	105
B. March	New York, NY	Scanner	BW9338	400	100
J. Bain	New York, NY	Scanner	BW9443	500	125
T. Goss	Washington, DC	Printer	FW199	200	60
T. Goss	Washington, DC	Scanner	BW39339	550	140
L. Nye	New York, NY	Desktop	LR21	600	60
L. Nye	New York, NY	Printer	FW299	300	90
S. Cann	Washington, DC	Desktop	LR21	600	60
E. Sims	Washington, DC	Laptop	DR2983	700	140
P. Judd	New York, NY	Desktop	LR22	700	70
P. Judd	New York, NY	Scanner	FJ3999	200	50
G. Hinton	Washington, DC	Laptop	DR2983	700	140
G. Hinton	Washington, DC	Desktop	LR21	600	60
G. Hinton	Washington, DC	Printer	FW288	350	105
G. Hinton	Washington, DC	Scanner	BW9443	500	125
H. Fu	New York, NY	Desktop	ZX88	450	45
H. Taylor	New York, NY	Scanner	BW9338	400	100

10. Write a code to implement the following exploratory data analysis: (use dataset *A1.csv*)
    a. Find the frequency of samples for each class.
    b. Generate data description and calculate the interquartile range for all four features.
    c. Plot a histogram of feature 1 for class A.
    d. Make the box plot for feature 2 for each class separately.
    e. Violin plot for feature 3 for each class separately.
    f. Scatter plots between feature 1 and feature 3 showing classes separately.
    g. Contour plot between feature 1 and feature 4 showing classes separately.
    h. Hexagonal bin plot for class A between feature 2 and 4.
    i. Correlation matrix for the four features.
    j. Pair plot for the four features showing classes separately.

Note: Submit code outputs in the assignment PDF and code as a. ipynb file separately.