CS4038 – Data Mining (Spring 2024)

Mam Ayesha Liaqat

Assignment 1

Submission Deadline: **Friday, February 23, 2024 (in the last lecture)**

Only hand-written solutions will be accepted.

Problem # 1: Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

Example: Age in years. Answer: Discrete, quantitative, ratio

a. Smoking status (Smoker/Non-smoker)
b. Weight in term of KG or Pound.
c. Brightness as measured by people's judgments.
d. Angles as measured in degrees between 0 and 360.
e. Bronze, Silver, and Gold medals as awarded at the Olympics.
f. Height above sea level.
g. Number of teacher in a school.
h. Blood types as A B A+ A- AB
i. Customer satisfaction rating on a scale of 1 to 5
j. Educational attainment (High school diploma/College degree)
k. Distance from your house of campus.

Problem # 2: Use the following 20 items price data and answer the questions below. Clearly write the formula for each and show all necessary calculations. No credit will be given for a direct answer.

5, 12, 52, 27, 28, 29, 33, 58, 56, 100, 104, 96, 87, 39, 190, 82, 88, 19, 49, 195

a. Calculate the mean, median, and mode.
b. Compute the standard deviation (SD) of the data. Interpret what this numbermeans.
c. Does the items price data follow a normal distribution, left skewed or right skewed? Support your answer with a histogram chart. Also, show your answers from Part (a) and (b) in the histogram chart.
d. Compute variance.
e. Draw a boxplot chart for the given data, compute and show the values of all quartiles. Also compute the IQR.
f. Compute the min – max normalized item price for all the values.
g. Compute the Z-score standardized item price for all the values.
h. Compute the decimal-scaling item price for all the values.
i. Based on your answers from Part e to g, discuss precisely that which normalization method would you prefer for this data.
j. Create four bins using equal width and equal frequency/depth methods. Which method in your opinion has done a better job for the input data?
k. Smooth the noisy data using bin boundaries and bin mean methods. Which method has performed well?

Problem # 3: Consider the following hypothetical data of ten persons with first three attributes (age, height, income) only for this exercise.

Age	Weight (kg)	Income (Rs)	Healthy?
29	65	45,000	Yes
21	45	25,000	Yes
24	45	32,000	Yes
26	55	30,000	No
31	50	75,000	Yes
28	63	50,000	No
34	56	100,000	No
29	42	55,000	Yes
26	52	98,000	No
27	56	75,000	Yes

a. Compute the correlation matrix. For each attribute, encircle the attribute with the highest correlation (i.e., identify which attribute has the highest correlation with Age and so on).
b. Draw a scatter chart for each pair of the attributes with highest correlation ONLY (that you encircled in Part a).

c. Compute the covariance matrix for the input data. For each attribute, encircle the attribute with the highest covariance.

Problem # 4: Use the data given in the previous problem for this exercise too.

a. Discretize the "*Age*" attribute by a method of your own choice.
b. By using the discretized "*Age*" attribute from Part (a), perform Chi-Square test of independence with class label "*Healthy?*" and significance level of 5%. State your hypotheses and show all the necessary steps.

Problem # 5: Distinguish between noise and outliers. Be sure to consider the following questions.

a. Is noise ever interesting or desirable? Outliers?
b. Can noise objects be outliers?
c. Are noise objects always outliers?
d. Are outliers always noise objects?
e. Can noise make a typical value into an unusual one, or vice versa?