Homework Problem Set 1

Due Jan. 19 at 11:59 PM
Upload a pdf to Canvas

Question 1:
For each data set given below, give specic examples of classication, clustering, association rule mining, and anomaly detection tasks that can be performed on the data. Use the same matrix for all four tasks. State how the data matrix should be constructed (i.e., specify the rows and columns of the matrix), including the target variable. If there is no appropriate target variable, state this. Use the same matrix for four tasks.

(a) ambulatory Medical Care data 1, which contains the demographic and medical visit information for each patient (e.g., gender, age, duration of visit, physicians diagnosis, symptoms, medication, etc).

(b) Stock market data, which include the prices and volumes of various stocks on dierent trading days.

(c) Database of Major League Baseball (MLB).

Question 2: Types of Attributes
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.

(a) Number of courses registered by a student in a given semester.

(b) Speed of a car (in miles per hour).

(c) Decibel as a measure of sound intensity.

(d) Hurricane intensity according to the Sar-Simpson Hurricane Scale.

(e) Social security number.

Question 3: Types of Attributes
Classify the following attributes as: 1) discrete or continuous, 2) qualitative or quantitative, 3) nominal, ordinal, interval, or ratio. Choose the most comprehensive attribute. If the attribute is both interval or ratio, choose

ratio.
Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

(a) Julian Date, which is the number of days elapsed since 12 noon Greenwich Mean Time of January 1, 4713 BC.

(b) Movie ratings provided by users (1-star, 2-star, 3-star, or 4-star).

(c) Mood level of a blogger (cheerful, calm, relaxed, bored, sad, angry or frustrated).

(d) Average number of hours a user spent on the Internet in a week.

(e) IP address of a machine.

(f) Richter scale (in terms of energy release during an earthquake).

(g) Salary above the median salary of all employees in an organization.

(h) Undergraduate level (freshman, sophomore, junior, and senior) for measuring years in college.

Question 4: State the type of each attribute given below before and after we have performed the following transformation.

(a) Hair color of a person is mapped to the following values: black = 0, brown = 1, red = 2, blonde = 3, grey = 4, white = 5.

(b) Grade of a student (from 0 to 100) is mapped to the following scale: A = 4.0, A- = 3.5, B = 3.0, B- = 2.5, C = 2.0, C- = 1.5, D = 1.0, D- = 0.5, E = 0.0

(c) Age of a person is discretized to the following scale: Age < 12, $12 \leq$ Age < 21, $21 \leq$ Age < 45, $45 \leq$ Age < 65, Age > 65.

(d) Annual income of a person is discretized to the following scale: Income $<$ \$20K, \$20K $\leq$ Income $<$ \$60K, \$60K $\leq$ Income $<$ \$120K, \$120K $\leq$ Age $<$ \$250K, Age $\geq$ \$250K.

(e) Height of a person is changed from meters to feet.

(f) Height of a person is changed from meters to (Short, Medium, Tall) .

(g) Height of a person is changed from feet to number of inches above 4 feet.

(h) Weight of a person is standardized by subtracting it with the mean of the weight for all people and dividing by its standard deviation.

Question 5: Data Preprocessing

Consider the following dataset that contains the age and gender information for 9 users who visited a given website.

UserID	1	2	3	4	5	6	7	8	9
Age	17	24	25	28	32	38	39	49	68
Gender	Female	Male	Male	Male	Female	Female	Female	Male	Male

(a) Suppose you apply an equal interval width approach to discretize the Age attribute into 3 bins. Show the userIDs assigned to each of the 3 bins.

(b) Repeat the previous question using the equal frequency approach.

(c) Repeat question (a) using a supervised discretization approach (with Gender as class attribute). Specically, choose the bins in such a way that their members are as pure as possible (i.e., belonging to the same class).

Question 6:

Consider an attribute X of a data set that takes the values $\{x_1, x_2, \cdots, x_9\}$ (sorted in increasing order of magnitude). We apply two methods (equal interval width and equal frequency) to discretize the attribute into three bins. The bins obtained are shown below:

Equal Width: $\{x_1, x_2, x_3\}$, $\{x_4, x_5, x_6, x_7, x_8\}$, $\{x_9\}$

Equal Frequency $\{x_1, x_2, x_3\}$, $\{x_4, x_5, x_6\}$, $\{x_7, x_8, x_9\}$

Explain what will be the effect of applying the following transformations on each discretization method, i.e., whether the elements assigned to each bin can change if you discretize the attribute after applying the transformation function below. Note that $\bar{X}$ denotes the average value and σ_x denotes standard deviation of attribute X.

The answer should be a dictionary with three keys: 'equal_width', 'equal_freq', and 'jusify'. The values of the first two keys are either 'Change' or 'No

change'. The value of 'justify' is an integer between 1 and 10, determined by the choices below, denoting the justification of you answer (an integer between 1 and 10, determined by the choices

1. The transformation leads to an inversion of the original order of values.
2. The distance between x_i and x_{i+1} does not change uniformly.
3. The average value $\bar{X}$ becomes the smallest value post-transformation.
4. The relative ordering of points changes
5. The transformation causes negative values to become positive and vice versa.
6. The transformation results in all values becoming equal.
7. The distance between x_i and x_{i+1} change uniformly.
8. The standard deviation σ_X becomes zero after the transformation.
9. No change in the relative ordering of points
10. The maximum and minimum values of X get swapped after the transformation.

Subquestions to answer:

(a) $X \to X - \bar{x}$, (i.e., if the attribute values are centered).

(b) $X \to \frac{X-\bar{X}}{\sigma_x}$, (i.e., if the attribute values are standardized).

(c) $X \to \exp\left[\frac{X-\bar{X}}{\sigma_x}\right]$ (i.e., if the values are standardized and exponentiated).

Question 7:
An e-commerce company is interested in identifying the highest spending customers at its online store using association rule mining. One of the rules identified is:

$$21 \leq Age < 45$$

AND

$$\text{NumberOfVisits} > 50 \to \text{AmountSpent} > \$500,$$

where the Age attribute was discretized into 5 bins, NumberOfVisits was discretized into 8 bins, and AmountSpent was discretized into 8 bins. The confidence of an association rule $(A, B) \rightarrow C$ is defined as

$$\text{Confidence}((A, B) \rightarrow C) = P(C|A, B) = \frac{P(A, B, C)}{P(A, B)}$$

where $P(C|A, B)$ is the conditional probability of C given A and B, $P(A, B, C)$ is the joint probability of A, B, and C, and $P(A, B)$ is the joint probability of A and B. The probabilities are empirically estimated based on their relative frequencies in the data. For example, $P(\text{AmountSpent} > \$500)$ is given by the proportion of online users who visited the store and spent more than \$500.

(a) Suppose we increase the number of bins for the Age attribute from 5 to 6 so that the discretized Age in the rule becomes $21 \leq Age < 30$ instead of $21 \leq Age < 45$, will the confidence of the rule be non-increasing, non-decreasing, stays the same, or could go either way (increase/decrease)?

(b) Suppose we increase the number of bins for the AmountSpent attribute from 8 to 10, so that the right-hand side of the rule becomes $\$500 < \text{AmountSpent} < \$1000$, will the confidence of the rule be non-increasing, non-decreasing, stays the same, or could go either way (increase/decrease)?

(c) Suppose the values for NumberOfVisits attribute are distributed according to a Poisson distribution with a mean value equals to 4. If we discretize the attribute into 4 bins using the equal frequency approach, what are the bin values after discretization? Hint: you need to refer to the cumulative distribution table for Poisson distribution to answer the question.

Question 8: Measures of Similarity and Dissimilarity
Consider the following bindary vectors:

$$\begin{aligned} \mathbf{x_1} &= (1, 1, 1, 1, 1) \\ \mathbf{x_2} &= (1, 1, 1, 0, 0) \\ \mathbf{y_1} &= (0, 0, 0, 0, 0) \\ \mathbf{y_2} &= (0, 0, 0, 1, 1) \end{aligned} \tag{1}$$

(a) According to Jaccard coefficient, which pair of vectors--$(\mathbf{x}_1, \mathbf{x}_2)$ or $(\mathbf{y}_1, \mathbf{y}_2)$—are more similar to each other?

(b) According to simple matching coecient, which pair of vectors—$(\mathbf{x}_1, \mathbf{x}_2)$ or $(\mathbf{y}_1, \mathbf{y}_2)$–are more similar to each other?

(c) According to simple Euclidean distance, which pair of vectors—$(\mathbf{x}_1, \mathbf{x}_2)$ or $(\mathbf{y}_1, \mathbf{y}_2)$–are more similar to each other?

Question 9: Which similarity or distance measure is most eective for each of the domains given below:

(a) Which measure, Jaccard or Simple Matching Coecient, is most appropriate to compare how similar are the answers provided by students in an exam. Assume that the answers to all the questions in the exam are either True or False.

(b) Which measure, Jaccard or Simple Matching Coecient, is most appropriate to compare how similar are the locations visited by tourists at an amusement park. Assume the location information is stored as binary yes/no attributes (yes means a location was visited by the tourist and no means a location has not been visited).

(c) (c) Which measure, Euclidean distance or correlation coecient, is most appropriate to compare two ows in a network trac. For each ow, we record information about the number of packets transmitted, number of bytes transferred, number of acknowledgments sent, and duration of the session.

(d) Which measure, Euclidean distance or cosine similarity, is most appropriate to compare the coordinates of a moving object in a 2-dimensional space. For example, using GPS data, the object may be located at (31.4° West, 12.4° North) at time t_1 and (29.4° West, 12.5° North) at another time t_2 . Note: we may use +/- to indicate East/West or North/South directions when computing the similarity or distance measures.

(e) Which measure, Euclidean distance or cosine similarity, is most appropriate to compare the similarity of items bought by customers at a grocery store. Assume each customer is represented by a 0/1 binary vector of items (where a 1 means the customer had previously bought the item).

Question 10: Ten True/False questions. Please answer True or False, and provide a one sentence justification of your choice.

(a) Noise is not a problem with count data. Explain.

(b) For any two sets of real values, such as two vectors of size n ¿ 0, the correlation is a value between -1 and 1. Explain.

(c) For reducing the size of a daily time series, it would be better to sample than aggregate since sampling is a simpler process. Explain.

d) Noise and outliers are sometimes the same. Explain.

(e) If an object is an outlier, then it is noise.

(f) A binary attribute with values 0 or 1 is also an asymmetric binary attribute.

(g) If vectors of counts have a cosine measure of 1, the objects are identical. Explain.

(h) Discrete variables cannot be ratio.

(i) Quantitative variables are continuous.

(j) Converting ordinal variables to asymmetric binary variables does not lose any information.