Homework 1 Solution

October 2, 2018

Question 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.

1. Speed of a vehicle measured in mph.
Answer:Continuous,quantitative, ratio

2. Altitude of a region.
Answer:Continuous, quantitative, ratio

3. Intensity of rain as indicated using the values: no rain, intermittent rain, incessant rain
Answer:Discrete, qualitative, ordinal

4. Brightness as measured by a light meter.
Answer:Continuous, quantitative, ratio

5. Barcode number printed on each item in a supermarket.
Answer:Discrete, qualitative, nominal

Question 2

The population for a clinical study has 500 Asian, 1000 Hispanic and 500 Native American people. What is good way of sampling this population to ensure that the distribution of various sub-populations is maintained if only 100 samples have to be chosen? Give the distribution of the various sub-populations in the final sample.

Answer: 25 Asian, 50 Hispanic, 25 Native American people.

Question 3

Justify your answers for the following:

1. Is the Jaccard coefficient for two binary strings (i.e., string of 0s and 1s) always greater than or equal to their cosine similarity?

Conclusion: For two binary strings, the Jaccard coefficient will always be less than their cosine similarity.

Proof: From the definition, we have equations below.

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} \qquad \cos(x, y) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

It's easy to see $f_{11} = \mathbf{x} \cdot \mathbf{y}$. Then we need to compare $f_{01} + f_{10} + f_{11}$ and $\|\mathbf{x}\|\|\mathbf{y}\|$. Let $N_{\mathbf{x}}$ denote the count of occurence of 1 in x, $N_{\mathbf{y}}$ denote the count of occurence of 1 in y. Then we can see the equations below.

$$\begin{cases} N_{\mathbf{x}} &= f_{10} + f_{11} \\ N_{\mathbf{y}} &= f_{01} + f_{11} \\ \|\mathbf{x}\| &= \sqrt{N_{\mathbf{x}}} \\ \|\mathbf{y}\| &= \sqrt{N_{\mathbf{y}}} \\ f_{11} &= \mathbf{x} \cdot \mathbf{y} \qquad \leq min(N_{\mathbf{x}}, N_{\mathbf{y}}) \end{cases}$$

Actually, we have

$$\because N_{\mathbf{x}} + N_{\mathbf{y}} - \sqrt{N_{\mathbf{x}} N_{\mathbf{y}}} \geq \sqrt{N_{\mathbf{x}} N_{\mathbf{y}}} \geq \min\left(N_{\mathbf{x}}, N_{\mathbf{y}}\right) \geq f_{11}$$

Extract the both ends of last inequality, then

$$\begin{array}{lll} \therefore N_{\mathbf{x}} + N_{\mathbf{y}} - f_{11} & \geq & \sqrt{N_{\mathbf{x}} N_{\mathbf{y}}} \\ \therefore (f_{10} + f_{11}) + (f_{01} + f_{11}) - f_{11} & \geq & \sqrt{N_{\mathbf{x}} N_{\mathbf{y}}} \\ \therefore f_{01} + f_{10} + f_{11} & \geq & \sqrt{N_{\mathbf{x}}}\sqrt{N_{\mathbf{y}}} \\ \therefore f_{01} + f_{10} + f_{11} & \geq & \|\mathbf{x}\|\|\mathbf{y}\| \end{array}$$

So For two binary strings, the Jaccard coefficient will always be less than or equal to their cosine similarity.

2. The cosine measure can range between [-1,1]. Give an example of a type of data for which the cosine measure will always be non-negative.

Answer: Since $\|\mathbf{x}\|\|\mathbf{y}\| \geq 0$, we only need to find data which satisfy $\mathbf{x} \cdot \mathbf{y} \geq 0$ to make the whole cosine measure non-negative. If we count the words in emails, build a data matrix $A(a_{ij})$ in which a_{ij} represents the count of the j-th word in the i-th email. It's obvious that $a_{ij} \geq 0$. If we compute the cosine measure of two records $A_{m,}A_{k,}$ in the data matrix, then we have,

$$\cos\left(A_{m,}A_{k,}\right) = \sum_{i=1}^{N} a_{mi} a_{ki} \geq 0 \qquad \text{where N is the total number of the words}$$

Question 4

The similarity between two undirected graphs G_1 and G_2 that have the same n vertices can be defined using:

$$S(G_1, G_2) = \frac{\sum_i \min\left(deg(v_i \in G_1), deg(v_i \in G_2)\right)}{2 \times \max\left(|G_1|, |G_2|\right)}$$

where $deg(v \in G)$ indicates the degree of a vertex v in graph G and $|G|$ indicates the number of edges in G. If $S(G_1, G_2) = 1$, are the two graphs equivalent? Provide an example to justify your answer.

Answer: Before we give a direct response to the question. We try to see what we can get from the formula of $S(G_1, G_2)$.

For any non-negative real number $a, b \geq 0$, we have

$$\begin{array}{rcl} \max(a,b) & = & \frac{|a+b|}{2} + \frac{|a-b|}{2} \\ \min(a,b) & = & \frac{|a+b|}{2} - \frac{|a-b|}{2} \end{array} \tag{1}$$

If $S(G_1, G_2) = 1$, then

$$\sum_i \min\left(deg(v_i \in G_1), deg(v_i \in G_2)\right) = 2 \times \max\left(|G_1|, |G_2|\right)$$

From (1), we get

$$\sum_i \left(\frac{|deg(v_i \in G_1) + deg(v_i \in G_2)|}{2} - \frac{|deg(v_i \in G_1) - deg(v_i \in G_2)|}{2}\right) = 2 \times \left(\frac{|G_1| + |G_2|}{2} + \frac{\Big||G_1| - |G_2|\Big|}{2}\right)$$

Which equivalent to

$$\frac{1}{2}\left(\sum_i deg(v_i \in G_1) + \sum_i deg(v_i \in G_2)\right) - \frac{1}{2}\sum_i |deg(v_i \in G_1) - deg(v_i \in G_2)| = |G_1| + |G_2| + \Big||G_1| - |G_2|\Big|$$

Since for undirectional graph $\sum_i deg(v_i \in G) = 2 \times |G|$, the equations above changes to

$$\Big||G_1| - |G_2|\Big| + \frac{1}{2}\sum_i |deg(v_i \in G_1) - deg(v_i \in G_2)| = 0$$

Which is equivalent to

$$\left\{ \begin{array}{rcll} |G_1| & = & |G_2| & \\ deg(v_i \in G_1) & = & deg(v_i \in G_2) & \text{for each i} \end{array} \right. \tag{2}$$

So the original question transform to a new one, if undirectional graph G_1 and G_2 satisfy equations (2), are they equivalent? The answer is **NO**.

Consider graphs like below. For $i = 1, 2, \ldots, 6$, $deg(v_i \in G_1) = deg(v_i \in G_2)$, and also $|G_1| = |G_2|$, so for these two graph, $S(G_1, G_2) = 1$, but they are not equivalent.

Question 5

For every item i in a grocery store, a set s_i is used to represent the IDs of transactions in which i is purchased. Assume that the data set to be analyzed contains hundreds of thousands of such transactions.

1. In order to analyze the proximity between any two of these sets s_i and s_j , which measure, Jaccard or Hamming, would be more appropriate and why?

Answer: The data format for the s_i, s_j will be like,

Table 1: Data Matrix

	Transaction 1	Transaction 2	Trasaction 3	...
s_i	1	0	1	...
s_j	0	1	1	...

In this case, what we focus is the similarity of the two objects s_i and s_j, and for object i, compared to the rest items, it will only take a small propotion of all the transactions, which means the the matrix is asymmetric binary, so the **Jaccard Coefficient** is more suitable.

2. In order to analyze the proximity between any two of these sets s_i and s_j for items i and j that are often brought together (example: milk, bread), which measure, Jaccard or Hamming, would be more appropriate and why ?

Answer: Since i and j are often bought together, according to Table 1, s_i and s_j will share the same transactions in most cases. As a result, what determins how much they proximate each other is their dissimilarity, the less dissimilarity are there, the more proximation they have. In order to focus on the dissimilarity here, **Hamming Distance** is more suitable.

Extra Question

For the data set described below, give an example of the types of data mining questions that can be asked (one for each classification, clustering, association rule mining, and anomaly detection task) and the description of the data matrix (what are the rows and columns). If necessary, briefly explain the features that need to be constructed. Note that, depending on your data-mining question, the row and column definitions may be different.

a) A clinical dataset containing various measures like temperature, blood pressure, blood glucose and heart rate for each patient during every visit, along with the diagnosis information.

DM Task: Classification
Question: Does a patient has type 1 Diabetes, type 2 Diabetes or No Diabetes Row: A patient Column: blood presure, blood glucose, heart rate, info like prescriptions.
DM Task: Clustering
Question: What are the two patients have the similar disease? Row: A patient Column: temperature, blood presure, blood glucose, heart rate.
DM Task: Association rule mining
Question: What are the two measures often rise and decline together for a specific disease? Row: A patient Column: temperature, blood presure, blood glucose, heart rate and diagnostic result(root cause).
DM Task: Anomaly detection
Question: Indicate a rare disease with a quite low incidence. Row: A patient Column: temperature, blood presure, blood glucose, heart rate and info like prescription, doctor's suggestion.