Assignment 1
Author: Yuan Qu **Date:** 01/31/2017

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

1. Speed of a vehicle measured in mph.
2. Altitude of a region.
3. Intensity of rain as indicated using the values: no rain, intermittent rain, incessant rain.
4. Brightness as measured by a light meter.
5. Barcode number printed on each item in a supermarket.

Answer:

1. Continuous, quantitative, ratio.
2. Continuous, quantitative, ratio.
   For some altitude graph, the altitude of a region is divided into colorful groups, it will be the qualitative ordinal data.
3. Discrete, qualitative, ordinal.
4. Continuous, quantitative, interval.
   Just consider it is same as Celsius, because light meter is not a precise and scientific measurement which has continuous range.
5. 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: In totality, the ratios of subpopulations are:

$$\text{Asian} : \text{Hispanic} : \text{Native American} = 500 : 1000 : 500 = 1 : 2 : 1$$

So, to ensure that the distribution of various subpopulations in sample:
$100 \times (1 + 2 + 1) = 25$
Asian: $25 \times 1 = 25$
Hispanic: $25 \times 2 = 50$
Native American: $25 \times 1 = 50$

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?

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 1: No, the cosine similarity is greater than or equal to the Jaccard coefficient.

Consider two binary string x and y. There are k "1"s in x and m "1"s in y, and r "1"s in $x \wedge y$.

Obviously, $k, m, r \geqslant 0$ and $k \geqslant r, m \geqslant r$.

According to the defination,

$$\cos(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\|} \tag{1}$$

$$EJ(x, y) = \frac{x \cdot y}{\|x\|^2 \cdot \|y\|^2 - x \cdot y} \tag{2}$$

So, we have,

$$x \cdot y \geqslant 0, \|x\| = \sqrt{k}, \text{ and } \|y\| = \sqrt{m}$$

Because the numerators of (1) and (2) are same and $x \cdot y = r \geqslant 0$, so we only need to prove (3),

$$\|x\|^2 \cdot \|y\|^2 - x \cdot y \geqslant \|x\| \cdot \|y\| \tag{3}$$

According to the defination, we can transform (3) to (4) and we only need to prove (4),

$$k + m - r \geqslant \sqrt{k \cdot m} \tag{4}$$

Because $k, m \geqslant 0$, $k + m \geqslant 2\sqrt{k \cdot m}$. So,

$$k + m - r \geqslant 2\sqrt{k \cdot m} - r \tag{5}$$

According to the defination, $k \geqslant r, m \geqslant r$, so

$$2\sqrt{k \cdot m} - r = \sqrt{k \cdot m} + (\sqrt{k \cdot m} - r) \geqslant \sqrt{k \cdot m} \tag{6}$$

According to (5) and (6), the (4) is proved. So,

$$cos(x, y) \geqslant EJ(x, y)$$

The "greater than or equal to" could be "equal to" only if $k = m = r$.

Answer 2: According to the defination,

$$cos(x, y) = \frac{x \cdot y}{||x|| \cdot ||y||}$$

So, to make cosine measure be non-negative is equal to make $x \cdot y$ non-negative.

There are two situations:

1. When vectors are all in non-negetive numbers(or all in negative) the cosine measure will always be non-negative.
   e.g. the ratings range from 1 to 10 or a whatever postive integer, the stars(can be regard as range 1-5) and the number of transactions.

2. The data is highly centralized and far from the origin. In geometry, the cosine measure can be explain to the angle between vectors, so small angle(less than $\pi/2$) means non-negative cosine. If the data is highly centralized, the angles will be very small and the cosine will be non-negative.
   e.g. all the types of data could be in this situation if an improper model is chosen.

Question 4

The similarity between two undirected graphs G1 and G2 that have the same n vertices can be defined using:

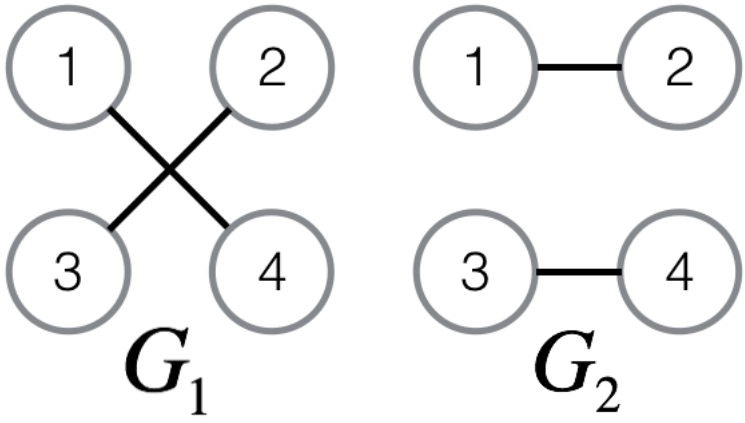
$$S(G_1, G_2) = \frac{\sum_i min(deg(v_i \in G_1), deg(v_i \in G_2))}{2 \times max(|G_1|, |G_2|)}$$

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: No.

If two graphs are quivalent, obviously, $S(G_1, G_2) = 1$. But for the converse situation, it is easy to built a counter-example like that:



In this case,

$$S(G_1, G_2) = \frac{\sum_{i=1}^{4} \min(\deg(v_i \in G_1), \deg(v_i \in G_2))}{2 \times \max(|G_1|, |G_2|)} = \frac{1+1+1+1}{2 \times 2} = 1$$

but G_1 and G_2 are different.

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 ?

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: Assumed that the number of total transactions is n and for the item r, the set $s_r = \mathtt{id}_1, \mathtt{id}_4, ..., \mathtt{id}_{n-2}, \mathtt{id}_n$. Obviously, each two sets may not have same length(same times be purchased).

To analyze the proximity between any of two sets, we have to standardize the sets s_i into constant length n. The standardization is to change IDs sets to binary set in n digit, in which the id_i are transformed to the "1" at i digit, for instance,

$$s_r = 10010...0101$$

Aussumed that $|s_i| = k, |s_j| = m$ and $|s_i \cap s_j| = l$, according to the defination, we have:

$$\texttt{Jaccard}: J = \frac{l}{k + m - l}, \text{ which measure similarity.}$$

and

$$\texttt{Hamming}: H = k + m - 2 \times l, \text{ which measure distinction.}$$

In the (1) situation, Jaccard measure would be more appropriate. In this case, the goal is to find the similar binary vectors, so we choose similarity measurement here.

In the (2) situation, Hamming measure would be more appropriate. In this case, the goal is to distinct the level or degree of similarity, so we choose distinction measurement here.

Question 6

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.

Example data: a collection of Web pages.

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.

Answer:

- DM Task: Classification of patients
  Question: What type of patient?
  Row: A record of a patient.
  Column: temperature, blood pressure, blood glucose, heart rate and labels.
  Label: Fever(high temperature, others normal), hypertension(high blood pressure), hypotension(low blood pressure), diabetes(high blood glucose), coronary disease(high blood pressure, high heart rate), healthy(all normal).
  Rules: Use these labels to classify all the patients.

- DM Task: Clustering of patients
  Question: What are the patients with similar diseases?
  Row: A record of a patient.
  Column: temperature, blood pressure, blood glucose and heart rate.
  Rules: Use clustering method to cluster data, find the group that the points are centralized.

- DM Task: Association rule mining
  Question: What are the symptoms that appear together frequently?
  Row: A record of a patient.
  Column: temperature, blood pressure, blood glucose and heart rate.
  Rules: For instance, most people with coronary disease have high blood pressure, high blood glucose and high heart rate as the symptoms, and most people with high blood glucose will also have high blood pressure.

- Anomaly detection
  Question: Is it a patient need first-aid?
  Row: A record of a patient.
  Column: temperature, is temperature critical(a boolean value), blood pressure, is blood pressure critical(a boolean value), blood glucose, is blood glucose critical(a boolean value), heart rate and is heart rate critical(a boolean value).
  Rules: According to the training set, we can get the distribution or clustering margin of each feature and set a critical value for each feature. When the new data come in, as the test set, judge the features are critical or not, if critical, set the record as anomaly.