CSE497 Fall 2018 Assignment 3.

Assigned: Wednesday, September 12, 2018

Due: Wednesday, September 19, 2018

Maximum: 100 point

Note: This assignment is to be done by an individual student, no team work allowed.

---

1. (20%) 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) Military rank
   (b) Temperature as measured by a thermometer
   (c) Temperature as measured by people's judgments, e.g., cold, warm, hot, etc.
   (d) Distance from Nittany Lion Statue
   (e) Course number, e.g., CMPSC 497

2. (20%) Animal scientist measure elephants using following attributes: breed (e.g., Africa, Asian, etc), weight, height, trunk length, and ear area.
   (a) Based on breed only, what sort of similarity measure would you use to compare or group these elephants?
   (b) Based on weight, height, and trunk length, what sort of similarity measure would you use to compare or group these elephants?
   Justify your answer and explain any special circumstances.

3. (20%) Given a set of m objects that is divided into k groups, where the i-th group is of size m_i $(i = 1..k)$. You would like to obtain a sample of size n < m. You ask your friends Alice and Bob for help, and the following are their suggestions.

   Alice: You can randomly select $n \times \frac{m_i}{m}$ from each group with replacement.
   Bob: You can randomly select n elements from the data set with replacement, without regard for the group to which an object belongs.

   Please compare Alice and Bob's method, and explain which one is better?

4. (20%) For binary data, the L1 distance corresponds to the Hamming distance; that is, the number of bits that are different between two binary vectors. The Jaccard similarity is a measure of the similarity between two binary vectors. Compute the Hamming distance and the Jaccard similarity between the following two binary vectors.

$$x = 1010101110$$
$$y = 1011100111$$

5. (20%) Which approach, Jaccard or Hamming distance, is more similar to the Simple Matching Coefficient, and which approach is more similar to the cosine measure? Please explain. (Note: The Hamming measure is a distance, while the other three measures are similarities, but don't let this confuse you.)