Assignment 3

Jiarong Ye

September 19, 2018

Q1

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.

(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

- (a) Military rank
  - Discrete, Qualitative, Ordinal
- (b) Temperature as measured by a thermometer
  - Continuous, Quantitative, Interval
- (c) Temperature as measured by people's judgments, e.g., cold, warm, hot, etc.
  - Discrete, Qualitative, Nominal
- (d) Distance from Nittany Lion Statue
  - Continuous, Quantitative, Interval
- (e) Course number, e.g., CMPSC 497
  - Discrete, Qualitative, Nominal

Q2

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.

- (a) First quantify the categorical variable using one-hot encoding, with each label represented as a binary vector, then apply **the Jaccard Similarity** to compare the elephants based on their breed.

- (b) Since weight, height, and trunk length are all numerical, thus no encoding needed. However, depending on different scales applied, the varied values of these attributes might affect the classification result. Also it's possible that since similarity $= \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$, if based on the scale, $\|\mathbf{A}\|\|\mathbf{B}\| \gg \mathbf{A} \cdot \mathbf{B}$ or $\|\mathbf{A}\|\|\mathbf{B}\| \ll \mathbf{A} \cdot \mathbf{B}$, $\cos(\theta) \to 0$ or $\cos(\theta) \to \infty$. Thus probably cos similarity would not be a optimal choice under such circumstance. So first we vectorize record of each elephant with these 3 attributes (dimension=3), then normalize each row with $p = \frac{p-\mu_p}{\sigma_p}$, so that $p \sim N(0,1)$, then apply **the Euclidean Distance** $d(\mathbf{p}, \mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}$ between each pair to calculate their similarity and compare the elephants.

Q3

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?

Alice's method is better. (Alice: stratified random sampling; Bob: simple random sampling)

- By stratifying the entire dataset into k groups of similar distribution and do the random sampling proportionally, the sampled subset would provide a better representation of the distribution of the whole dataset, and it ensures each subgroup within the dataset receives proper representation within the sample. The statistical distribution of each one of the k group should be similar to ensure that each group is homogeneous, otherwise, might cause misrepresentation or inaccurate reflection of the dataset. Also, the stratified random sampling is more accurate, because the variance of stratified samples tends to be smaller.

Q4

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.

$$x = 1010101110$$

$$y = 1011100111$$

$$M_{01} = 2, \text{(the number of attributes where x was 0 and y was 1)}$$
$$M_{10} = 1, \text{(the number of attributes where x was 1 and y was 0)}$$
$$M_{00} = 2, \text{(the number of attributes where x was 0 and y was 0)}$$
$$M_{11} = 5, \text{(the number of attributes where x was 1 and y was 1)}$$

$$Hamming = 3, \text{there are 3 binary numbers different between the x and y}$$
$$Jaccard = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} = \frac{5}{2+1+5} = 0.625$$

Q5

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.)

- the Hamming distance is similar to the Simple Matching Coefficient

  *

$$\text{Hamming Distance} = \text{number of different bits in a binary vector}$$

$$SMC = \frac{M_{00} + M_{11}}{M_{00} + M_{01} + M_{10} + M_{11}} = \frac{\text{number of the same bits}}{\text{number of bits in the binary vector}}$$
$$SMC = 1 - \frac{\text{Hamming Distance}}{\text{number of bits in the binary vector}}$$

- the Jaccard Similarity is similar to the Cosine Similarity

  *

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

  in the algorithm the 0s are neglectable, which is similar to the Jaccard Similarity:

$$\text{Jaccard Similarity} = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

where M_{00} is excluded.