HW1_Data_Mining

Yifu Jason He

October 2019

Contents

1 Question 1

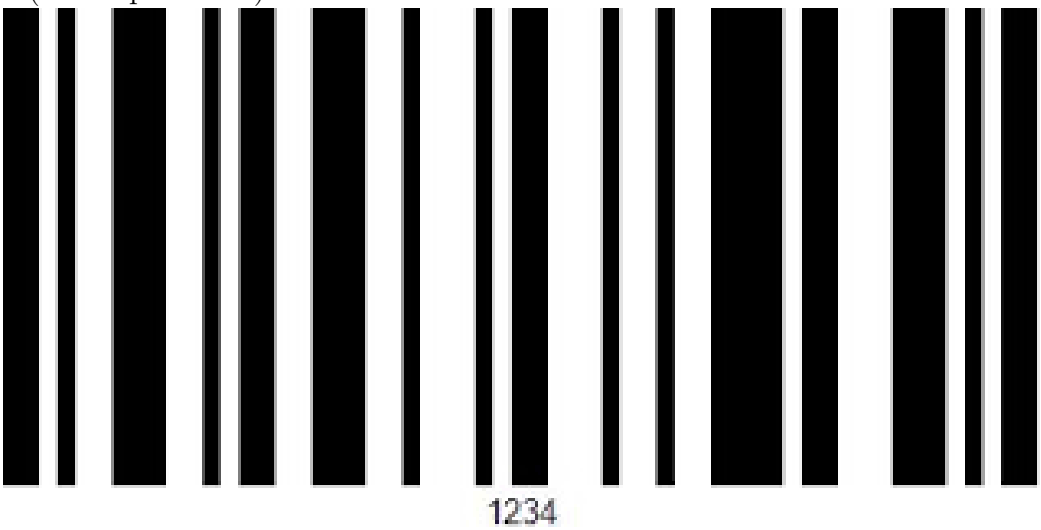
Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

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, interval
3.discrete, qualitative,ordinal
4.continuous, quantitative, interval
5.(2 interpretation)



a.discrete, qualitative, nominal. We can choose the shortest black or white individual bar and set it as a scale. The thickness can be expressed by number. And black and white can be enumerated as + and -. Then the Barcode can be convert into a certain sequence, such as { 3,-1,1,-3,5,-2,.....}

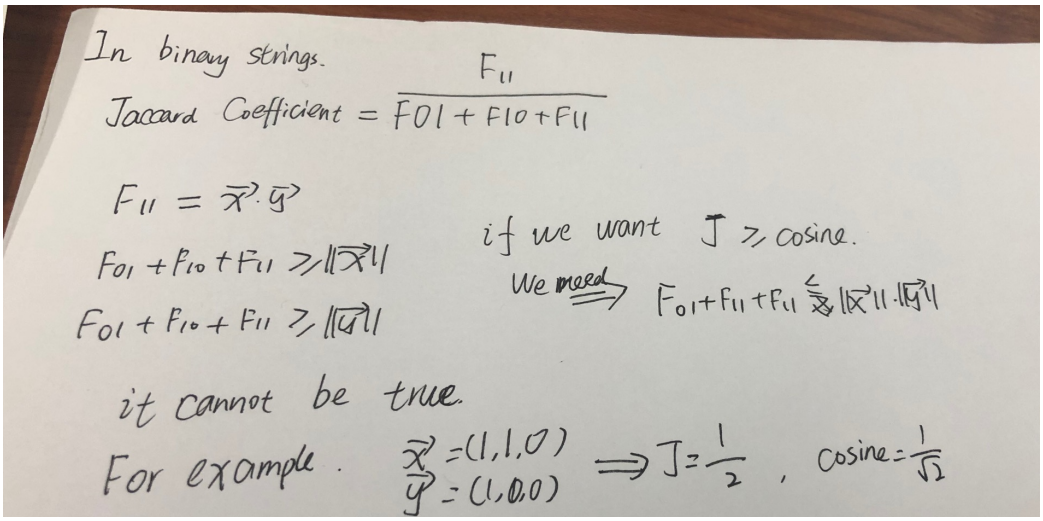
b.discrete, quantitative, interval. Also, choose the shortest one as a scale. However, split the total bar into every small unit and use 1,0 to express balck and white. This Barcode can be expressed as "111010001111100....".

2 Question 2

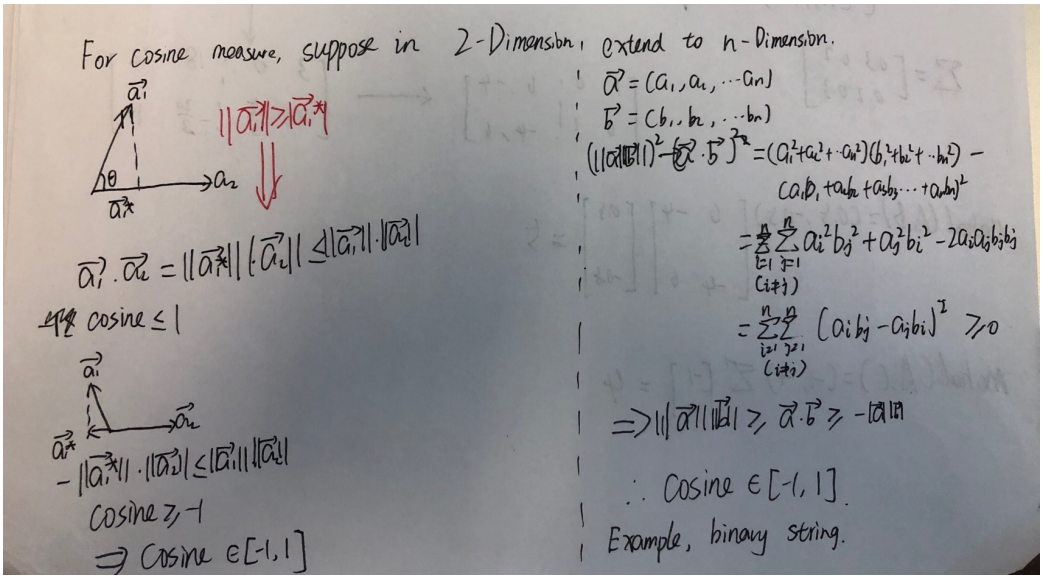
The principle of effective sampling is:
1. Using a sample will work almost as well as using the entire data sets.
2. A sample is representative if it has approximately the same property(of interest) as the original set of data.
So,we use stratified sampling and keep the ratio of each sub-population, 1:2:1.
The distribution is 25 Asian, 50 Hispanic and 25 Native American.

3 Question 3

1. No, it can be smaller. Proof:

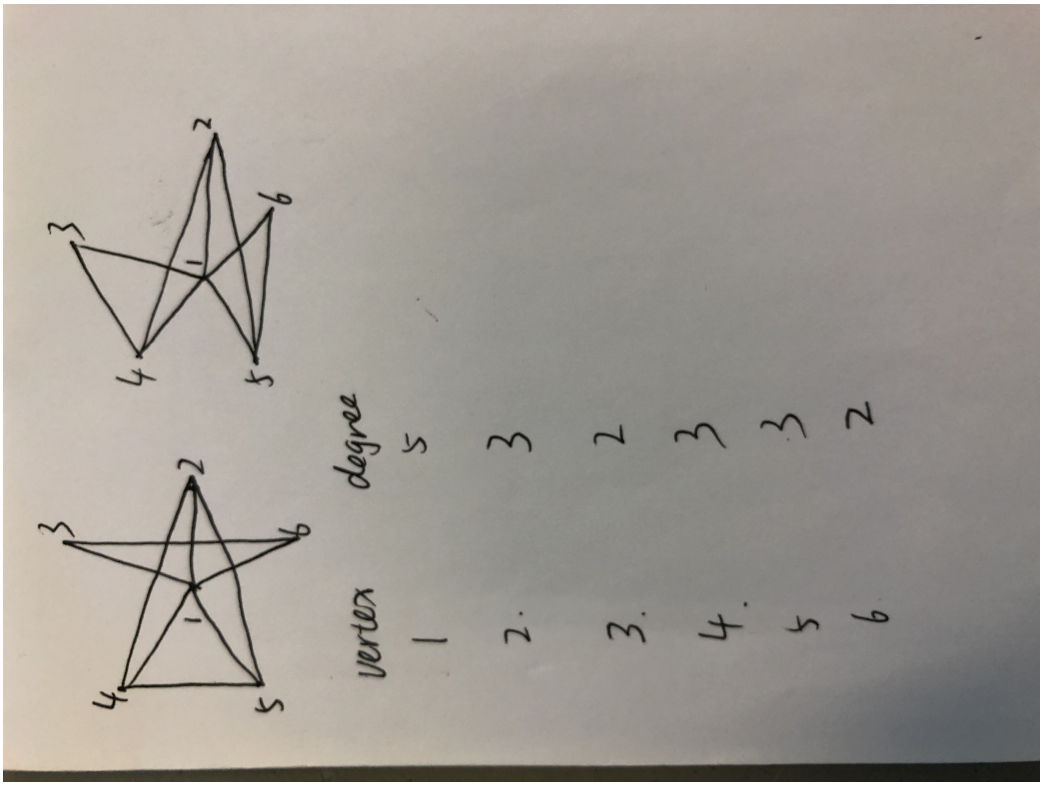


2. Proof is as following. The example is binary strings.



4 Question 4

No.Because in the undirected graph, the problem of equivalent is not just compare their vertexs and edges. The following example is a counter example.



Both are undirected graphs on the same six vertices, labeled "1" through "6". The vertex degrees on both are the same, vertex by vertex: one vertex ("1") has degree 5, two ("3" and "6") have degree 2, and three have degree 3. Therefore your similarity metric equals 1.

Note, however, that the three degree-3 vertices (labeled "2", "4", and "5") form a triangle at the left but not at the right. Therefore the two graphs are not isomorphic.

5 Question 5

5.1 subquestion 1

1.Hamming distance. According to the definition of Hamming distance, it is pretty much similar to the Simple Matching Coefficients(SMC), which will calculate the different rate of the whole dataset with the same importance.

5.2 subquestion 2

2.Jaccard. The setting is that we already have the two different item i and j. By using Jaccard coeffcient, we can compares members for two sets to see which members are shared and which are distinct.‘

6 Question 6

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.

6.1 classification of the clinical dataset

Question: What type of illness does the patient have? Row: The object of a patient's diagnosis information. Column: The health measures of the patient such as temperature, blood pressure, blood glucose and heart rate and the label of diagnosis result, such as cold, fever, cancer and so on.

6.2 Clustering of the clinical dataset

Question: What are the illness with similar health measures? Row: The object of a patient's diagnosis information. Column: Similar to the first one, but without the pre-defined label.

6.3 Association rule mining

Question: What are intervals or range of each health measure that appear together frequently? Such as high body temperature and the high percentage of white blood cell may appear together in the data of patient who has fever? Row: The object of a patient's diagnosis information. Column: Different health measures in the dataset.

6.4 Anomaly detection

Question: Is there any misdiagnosis that the doctor cannot recognize patient from healthy people? Row: The object of a patient's diagnosis information. Column: The health measures and the label of diagnosis information.