Sujad Patel

CS 422

Exercise 1

1. Discuss whether or not each of the following activities is a data mining task.

a. Sorting by gender in a database is not a data mining task.

b. This is also sorting in a database through query, so not a data mining task.

c. No, because it is a simple computation of summing each sale.

d. No, sorting in a database, therefore, not a database task.

e. Yes, because this activity does require some probability analysis, but could also

Argue that it is a simple computation.

f. Yes, because this activity requires predictive analysis.

g. Yes, this activity requires comparison of normal and abnormal heart rate associated variables and finding pattern for early detection.

h. Yes, because this activity also uses pattern analysis and data combing.

i. No, this is simple processing of sound wave using calculation.

3. For each of the following data sets, explain whether or not data privacy is an important issue.

a. According to the census.gov, “72-Year Rule” is applied to census data. Coincidently, 2022 is the year the data is publicly available. Therefore, it is not a data privacy issue.

b. IP-address can be used to identify someone, but websites might need some data to provide additional features. Still, it is a data privacy Issue.

c. Not a data privacy issue, because anyone see your house from outside.

d. No, because the information is already public.

e. No, you put your information for public to see.

2. 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. Binary (0 or 1), qualitative, nominal

b. Continuous, quantitative, ratio

c. if people are judging based on a scale. Discrete, qualitative, ordinal.

d. Discrete, quantitative, ratio.

e. Discrete, qualitative, Ordinal.

f. Continuous, quantitative, ratio.

g. Discrete, quantitative, ratio (count).

h. Discrete, qualitative, nominal

i. Discrete, qualitative, ordinal

j. Discrete, qualitative, ordinal.

k. Continuous, quantitative, ratio.

l. Continuous, quantitative, ratio.

m. Discrete, qualitative, nominal.

3. The boss is right, because the marketing director did not consider the number of sales of the books in his model. To fix the issue, I would include sales attribute and compare the ratio between each book sales and complaints. The attribute type is correct because it is counting and comparing complains. Discrete, quantitative, ratio.

7. Since daily temperature are more closely related in time due to their continuous nature, whereas rainfall is scattered and asymmetric.

12.

a. Noise is the random component of a measurement error, therefor not wanted. Outliers can be useful since they are data after all.

b. Yes, measurement error can occur at any component values.

c. No, not always.

d. No, can be part of the data set.

e. Yes noise can skew in any direction.