CSEN1083 – Data Mining
Problem Set #2

Problem 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

a) Time in terms of AM or PM.
b) Brightness as measured by a light meter.
c) Brightness as measured by people's judgments.
d) Angles as measured in degrees between 0 and 360.
e) Bronze, Silver, and Gold medals as awarded at the Olympics.
f) Height above sea level.
g) Number of patients in a hospital.
h) Ability to pass light in terms of the following values: opaque, translucent, transparent.
i) Military rank.
j) Distance from the center of campus.

Solution:

(a) Binary, qualitative, ordinal

(b) Continuous, quantitative, ratio

(c) Discrete, qualitative, ordinal

(d) Continuous, quantitative, ratio

(e) Discrete, qualitative, ordinal

(f) Continuous, quantitative, interval/ratio (depends on whether sea level is regarded as an arbitrary origin)

(g) Discrete, quantitative, ratio

(h) Discrete, qualitative, ordinal

(i) Discrete, qualitative, ordinal

(j) Continuous, quantitative, interval/ratio (depends)

CSEN1083 – Data Mining
Problem Set #2

Problem 2

You are approached by the marketing director of a local company, who believes that he has devised a foolproof way to measure customer satisfaction. He explains his scheme as follows: "It's so simple that I can't believe that no one has thought of it before. I just keep track of the number of customer complaints for each product. I read in a data mining book that counts are ratio attributes, and so, my measure of product satisfaction must be a ratio attribute. But when I rated the products based on my new customer satisfaction measure and showed them to my boss, he told me that I had overlooked the obvious, and that my measure was worthless. I think that he was just mad because our bestselling product had the worst satisfaction since it had the most complaints. Could you help me set him straight?"

(a) Who is right, the marketing director or his boss? If you answered his boss, what would you do to fix the meaaure of satisfaction?

(b) What can you say about the attribute type of the original product satisfaction attribute?

Solution:

(a) The boss is right. A better measure is given by

$$\text{Satisfaction(product)} = \frac{\text{number of complaints for the product}}{\text{total number of sales for the product}}$$

(b) Two products that have the same level of customer satisfaction may have different numbers of complaints and vice-versa.

CSEN1083 – Data Mining
Problem Set #2

Problem 3

An educational psychologist wants to use association analysis to analyze test results. The test consists of 100 questions with four possible answers each.

(a) How would you convert this data into a form suitable for association analysis?

(b) In particular, what type of attributes would you have and how many of them are there?

Solution:

(a) Association rule analysis works with binary attributes, so you have to convert original data into binary form as follows:

$Q_1 = A$	$Q_1 = B$	$Q_1 = C$	$Q_1 = D$	...	$Q_{100} = A$	$Q_{100} = B$	$Q_{100} = C$	$Q_{100} = D$
1	0	0	0	...	1	0	0	0
0	0	1	0	...	0	1	0	0

(b) 400 asymmetric binary attributes.