Homework 01

Due Friday, January 24, 2025, 11:59 PM

YOUR NAME

Today's Date: 2025-01-20

Problem 1

Classify the following variables 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.
   - Binary, qualitative, nominal

b) Brightness as measured by a light meter.
   - Continuous, quantitative, ratio

c) Brightness as measured by people's judgments.
   - Discrete, qualitative, ordinal

d) Angles as measured in degrees between 0 and 360.
   - Continuous, quantitative, ratio

e) Bronze, Silver, and Gold medals as awarded at the Olympics.
   - Discrete, qualitative, ordinal

f) Height above sea level.
   - Continuous, quantitative, ratio

g) Number of patients in a hospital.
   - Discrete, quantitative, ratio

h) ISBN numbers for books. (Look up the format on the Web.)
   - Discrete, qualitative, nominal

i) Ability to pass light in terms of the following values: opaque, translucent, transparent.
   - Discrete, qualitative, ordinal

j) Military rank.
   - Discrete, qualitative, ordinal

k) Distance from the center of campus.
   - Continuous, quantitative, ratio

l) Density of a substance in grams per cubic centimeter.
   - Continuous, quantitative, interval

m) Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.)
   - Discrete, qualitative, nominal

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 variables, and so, my measure of product satisfaction must be a ratio variable. 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 best-selling 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 measure of satisfactions?
   - Boss is right, because large number of complaint can't fully represent the quality of product. Instead, using the complaint rate could better show the real satisfactions. That is complaint rate = number of complaints / number of comments.

b) What can you say about the variable type of the original product satisfaction variable?
   - The value of count of complaint is discrete and ratio since 0 is meaningful

Problem 3

Consider a document-term matrix, where f_{ij} is the frequency of the jth word (term) in the ith document, and n is the number of documents. Consider the variable transformation that is defined by

$$f_{ij}^* = f_{ij} \cdot \log \frac{n}{g_j},$$

where g_j is the number of documents in which the jth term appears and is known as the document frequency of the term. This transformation is known as the inverse document frequency transformation.

a) What is the effect of this transformation if a term occurs in one document? In every document?
   - A term occurs in one document, the result is $f_{ij}^* = f_{ij} \cdot \log n$.
   - A term occurs in every document, the result is 0.

b) What might be the purpose of this transformation?
   - To add penalty on terms appear in many documents, and add more weight on distinct terms which appear in less documents. Finally, many common and useless words like 'the', 'and' would be ignored; the real valuable words would have higher frequency.