Question 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.

Example: Age in years. **Answer:** Discrete, quantitative, ratio

1. Time in terms of AM or PM.
    - Binary, qualitative, ordinal
2. Brightness as measured by a light meter.
    - Continuous, quantitative, ratio
3. Brightness as measured by people's judgments.
    - Discrete, qualitative, ordinal
4. Angles as measured in degrees between 0 and 360.
    - Continuous, quantitative, ratio
5. Bronze, Silver, and Gold medals as awarded at the Olympics.
    - Discrete, qualitative, ordinal
6. Height above sea level.
    - Continuous, quantitative, ratio/interval
7. Number of patients in a hospital.
    - Discrete, quantitative, ratio
8. ISBN numbers for books. (Look up the format on the Web.)
    - Discrete, qualitative, nominal
9. Ability to pass light in terms of the following values: opaque, translucent, transparent.
    - Discrete, qualitative, ordinal
10. Military rank.
    - Discrete, qualitative, ordinal
11. Distance from the center of campus.
    - Continuous, quantitative, ratio
12. Density of a substance in grams per cubic centimeter.
    - Discrete, quantitative, ratio
13. 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

Question 5.

Can you think of a situation in which identification numbers would be useful for prediction?

Answer: Unique patient ID can be a good predictor of hospital admission/discharge rate.

Question 6.

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

1. **How would you convert this data into a form suitable for association analysis?**

Answer: For association analysis, binary attribute is important and suitable. We would have to convert the responses of all possible answers of 100 questions into binary form.

2. **In particular, what type of attributes would you have and how many of them are there?**

Answer: we would have 400 asymmetric binary attributes.

Question 7.

Which of the following quantities is likely to show more temporal autocorrelation: daily rainfall or daily temperature? Why?

Answer: Autocorrelation is a measure of correlation between nearby observation.

Temporal autocorrelation is a special case of correlation, and refers not to the relationship between two or more variables, but to the relationship between successive values of the same variables.

Daily temperature is going to show more temporal autocorrelation than Daily rainfall. Because, it is more common for physically close locations to have similar temperature than similar amounts of rainfall. For example, the amount of rainfall changes from one location to another in a city. However, temperature in the city is usually the same and not varies by large number.

Question 8.

Discuss why a document-term matrix is an example of a data set that has asymmetric discrete or asymmetric continuous features.

Answer: The document-term matrix is an example of a dataset that has asymmetric discrete features because in document-term matrix the rows represent the individual document and columns represent the individual words in a document and values represent either o or 1. The values in a document-term matrix depends on the number of times the corresponding term occurs in the document. Therefore, only non-zero entries are important.

The document-term matrix is an example of a dataset that has asymmetric continuous features because We can apply TF-IDF normalization to terms to reflect how important a word is to a document in a collection. This increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the collection that contain the word, which helps to adjust for the fact that some words appear more frequently in general. This creates a term-document matrix with continuous features. Still, the features are asymmetric because the normalization doesn't create non-zero entries for the entries that were previously zero.

Question 9.

Many sciences rely on observation instead of (or in addition to) designed experiments. Compare the data quality issues involved in observational science with those of experimental science and data mining.

Answer: In designed experiments, the data quality is prespecifies to be at the certain level. In real life, conducting the designed experiments to solve the problem is not always feasible. Collect the data for experiments and design the experiments are costly with respect to time or resources. Sometimes it is not ethical to conduct some experiments considering the harm of experiments to the participants.

Therefore, many sciences would have to reply on observations in addition to designed experiments. Relying on observations comes with its own cost. The data quality of observational data is not as good as designed experiments as the data quality is usually not prespecified before the data capture. Observational sciences have less control on the data quality in comparison with the experimental sciences.

Question 13.

Consider the problem of finding the K-nearest neighbors of a data object. A programmer designs Algorithm 2.3 for this task.

1. **Describe the potential problems with this algorithm if there are duplicate objects in the data set. Assume the distance function will return a distance of 0 only for objects that are the same.**

   Answer: There are many problems with the Algorithm:

   - The order of duplicate objects on a nearest neighbor list will depend on details of the algorithm and the order of objects in the data set.
   - If there are enough duplicates, the nearest neighbor list may consist only of duplicates.
   - An object may not be its own nearest neighbor.

2. **How would you fix this problem?**

   Answer: It depends upon the situation. We can keep only one object for each group of duplicate case. In this case, each neighbor can represent either a single object or a group of duplicate objects.

Question 16.

Consider a document-term matrix, where tfij is the frequency of the ith word (term) in the jth document and m is the number of documents. Consider the variable transformation that is defined by

tfij'=tfij × logmdfi, (2.31)

where dfi is the number of documents in which the ith term appears, which is known as the document frequency of the term. This transformation is known as the inverse document frequency transformation.

1. **What is the effect of this transformation if a term occurs in one document? In every document?**

   Answer: Terms that occur in every document have zero weight, while those that occurs in one document have maximum weight.

2. **What might be the purpose of this transformation?**

   Answer: This transformation helps to distinguish one document from another.

Question 17.

Assume that we apply a square root transformation to a ratio attribute x to obtain the new attribute x*. As part of your analysis, you identify an interval (a, b) in which x* has a linear relationship to another attribute.

1. **What is the corresponding interval (A, B) in terms of x?**

   Answer: a^2, b^2

2. **Give an equation that relates y to x.**

   Answer: $y = x^2$

Question 19.

For the following vectors, x and y, calculate the indicated similarity or distance measures.

1. x = (1, 1, 1, 1), y = (2, 2, 2, 2) cosine, correlation, Euclidean

2. x = (0, 1, 0, 1), y = (1, 0, 1, 0) cosine, correlation, Euclidean, Jaccard

3. x = (0, −1, 0, 1), y = (1, 0, −1, 0) cosine, correlation, Euclidean

4. x = (1, 1, 0, 1, 0, 1), y = (1, 1, 1, 0, 0, 1) cosine, correlation, Jaccard

5. x = (2, −1, 0, 2, 0, −3), y = (−1, 1, −1, 0, 0, −1) cosine, correlation

Vector	Cosine	Correlation	Euclidean	Jaccard
1) X = (1,1,1,1) Y = (2,2,2,2)	1.0	NAN	2.0	
2) X = (0,1,0,1) Y = (1,0,1,0)	0	-1	2	0
3) X = (0, -1,0,1) Y = (1,0,-1,0)	0	0	2	
4) X = (1,1,0,1,0,1) Y = (1,1,1,0,0,1)	0.75	0.25		0.6
5) X = (2, -1,0,2,0, -3) Y= (-1,1,-1,0,0, -1)	0	0		