Assignment # 1

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) ISBN numbers for books. (Look up the format on the Web.)

(i) Ability to pass light in terms of the following values: opaque, translucent, transparent.

(j) Military rank.

(k) Distance from the center of campus.

(l) Density of a substance in grams per cubic centimeter.

(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.)

2. What do you mean noise in Data Mining task? How you identify the presence of noise in given data?

3. Consider a document-term matrix, where tf_{ij} is the frequency of the i^{th} word (term) in the j^{th} document and m is the number of documents. Consider the variable transformation that is defined by $tf_{ij}{}' = tf_{ij} * log\frac{m}{df_i}$, where df_i is the number of documents in which the i^{th} 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?

(b) What might be the purpose of this transformation?

4. This exercise compares and contrasts some similarity and distance measures. (a) For binary data, the L1 distance corresponds to the Hamming distance; that is, the number of bits that are different between two binary vectors. The Jaccard similarity is a measure of the similarity between two binary vectors. Compute the Hamming distance, cosine, Euclidean and the Jaccard similarity between the following two binary vectors.
x = 0101010001
y = 0100011000

(b) Which approach, Jaccard or Hamming distance, is more similar to the Simple Matching Coefficient, and which approach is more similar to the cosine measure? Explain. (Note: The Hamming measure is a distance, while the other three measures are similarities, but don't let this confuse you.)

(c) Suppose that you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure, Hamming, cosine, Euclidean or Jaccard, you think would be more appropriate for comparing the genetic makeup of two organisms. Explain. (Assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise.)

(d) If you wanted to compare the genetic makeup of two organisms of the same species, e.g., two human beings, would you use the Hamming distance, the Jaccard coefficient, or a different measure of similarity or distance? Explain. (Note that two human beings share > 99.9% of the same genes.)

5. Explain why computing the proximity between two attributes is often simpler than computing the similarity between two objects.