CSC 5800 : Intelligent Systems

Homework 1

Total: 100 Points

Problem 1. Discuss whether or not each of the following activities is a data mining task. If the answer is yes, then also specify which one of the following categories it will belong to : (i) classification (ii) association analysis (iii) clustering (iv) regression or (v) anomaly detection) **(20 Points; 2.5 Points each)**

(a) Sorting a student database based on student identification numbers.

(b) By looking at a CT scan, a doctor wants to identify if a patient has cancer or not. There are a lot of labeled CT scans that the doctor will use for making the decision.

(c) An image analyst obtains some new images and wants to automatically detect the number of distinct objects in the image. He doesn't have any prior information about these objects

(d) Predicting the outcomes of tossing a (fair) pair of dice.

(e) Predicting the future stock price of a company using historical records.

(f) In an Internet search engine company, there is a need to find potential users who will click a particular advertisement on the webpage.

(g) Monitoring the heart rate of a patient for abnormalities.

(h) Extracting the frequencies of a sound wave.

Problem 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.
For example: Age in years. Answer: Discrete, quantitative, ratio

(10 Points; 2 Points each)

(a) Brightness as measured by a light meter

(b) Angles as measured in degrees between 0° and 360°.

(c) Bronze, Silver, and Gold medals as awarded at the Olympics.

(d) Time in terms of AM or PM.

(e) Military rank.

Problem 3. This problem is an R exercise **(30 Points; 10 Points each)**

(a) Load iris.dat file (available in R) – Give the basic description of the data matrix; no. of data points, no. of features, no. of classes

(b) Give some basic statistics (such as mean, median, standard deviation, min, max) for each of these features

(c) Plot the first two features of the data. Classes must be discriminated by using different symbols. Please label the figure.

Problem 4. Given a similarity measure s with values in the interval [0,1]. Plot and that compare $d1$ and $d2$ transform this similarity value into a dissimilarity value in the interval [0,∞].

$$d1 = (1\text{-}s)/s$$
$$d2 = -\log s.$$

(10 Points; 5 Points each)

Problem 5. 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. **(10 Points; 5 Points each)**

(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?

Problem 6. This problem compares and contrasts some similarity and distance measures.

(10 Points; 5 + 3 + 2)

(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 and the Jaccard similarity between the following two binary vectors.

x = 0101010101
y = 0100011000

(b) Write the Simple Matching Coefficient in terms of the number of bits and the Hamming distance.

(c) Do you see any similarity between the Jaccard measure and the cosine measure? If so, explain?

Problem7. For the following vectors, x and y, calculate the indicated similarity or distance measures. **(10 Points; 5 Points each)**

(a) x = (0, −1, 0, 1), y=(1, 0, −1, 0) cosine, correlation, Euclidean
(b) x = (0, 1, 0, 1), y = (1, 0, 1, 0) cosine, correlation, Euclidean, Jaccard