Birla Institute of Technology & Science - Pilani, Hyderabad Campus
Second Semester 2014-2015
CS F415 / CS C415 : Data Mining
Test 1

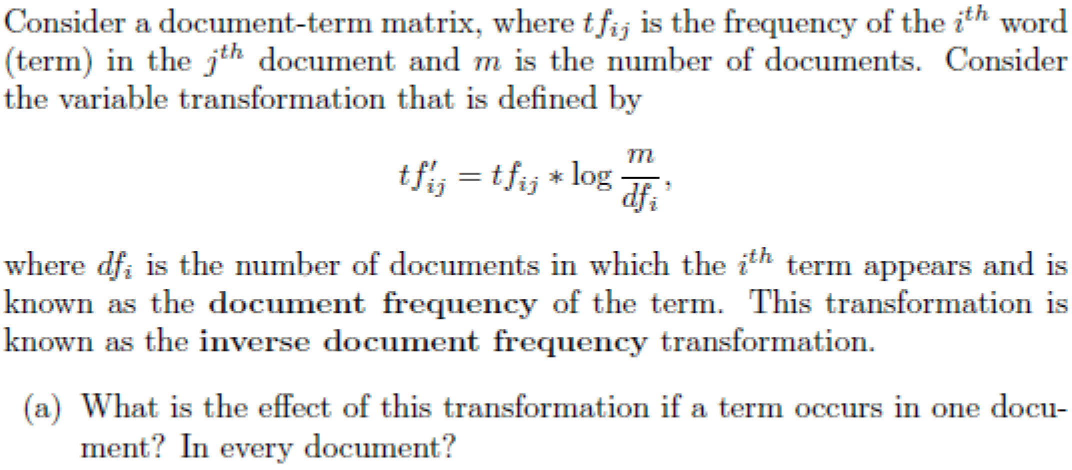
Type: Closed **Time: 60 mins** **Max Marks: 40** **Date: 23.02.2015**

All parts of the same question should be answered together.

1. Discuss whether or not each of the following activities is a data mining task. [4 Marks]
(a) Dividing the customers of a company according to their profitability.
(b) Computing the total sales of a company.
(c) Sorting a student database based on student identification numbers.
(d) Predicting the outcomes of tossing a (fair) pair of dice.
(e) Predicting the future stock price of a company using historical records.
(f) Monitoring the heart rate of a patient for abnormalities.
(g) Monitoring seismic waves for earthquake activities.
(h) Extracting the frequencies of a sound wave.

2. You are given a set of m objects that is divided into K groups, where the ith group is of size mi. If the goal is to obtain a sample of size n < m, what is the difference between the following two sampling schemes? (Assume sampling with replacement.) [3 Marks]
(a) We randomly select n ∗ mi/m elements from each group.
(b) We randomly select n elements from the data set, without regard for the group to which an object belongs.

3. [4 Marks]

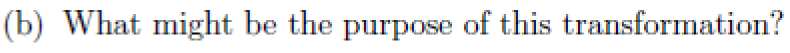


Consider a document-term matrix, where tf_{ij} is the frequency of the i^{th} word (term) in the j^{th} document and m is the number of documents. Consider the variable transformation that is defined by

$$tf'_{ij} = tf_{ij} * \log \frac{m}{df_i},$$

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. 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 y. [2 Marks]
(a) What is the corresponding interval (a, b) in terms of x?
(b) Give an equation that relates y to x.

5. Give at least two advantages to working with data stored in text files instead of in a binary format. [2 Marks]

6. Let I = {i_1, i_2, …., i_d} be the item set of a rule generation data mining problem. Prove that the total number of possible rules extracted from a data set that contain d items is $3^d - 2^{d+1} + 1$. [4 Marks]

7. Principal component analysis, or PCA, is a technique that is widely used for applications such as dimensionality reduction, lossy data compression, feature extraction, and data visualization. Derive 'k' principal components amongst 'n' features by presenting the problem as maximum variance formulation. Prove all steps that are required in this derivation. [7 Marks]

8.a. [4 Marks]

For each of the following measures, determine whether it is monotone, anti-monotone, or non-monotone (i.e., neither monotone nor anti-monotone).

> **Example:** Support, $s = \frac{\sigma(X)}{|T|}$ is anti-monotone because $s(X) \geq s(Y)$ whenever $X \subset Y$.

A discriminant rule is a rule of the form $\{p_1, p_2, \ldots, p_n\} \longrightarrow \{q\}$, where the rule consequent contains only a single item. An itemset of size k can produce up to k discriminant rules. Let η be the minimum confidence of all discriminant rules generated from a given itemset:

$$\begin{aligned}\eta(\{p_1, p_2, \ldots, p_k\}) \quad = \quad \min \big[\; & c\big(\{p_2, p_3, \ldots, p_k\} \longmapsto \{p_1\}\big), \ldots \\ & c\big(\{p_1, p_2, \ldots p_{k-1}\} \longrightarrow \{p_k\}\big) \;\big]\end{aligned}$$

Is η monotone, anti-monotone, or non-monotone?

8.b. Repeat the analysis in the above function by replacing the min function with a max function.

[4 Marks]

9. [6 Marks]

Consider the following set of frequent 3-itemsets:

$$\{1,2,3\}, \{1,2,4\}, \{1,2,5\}, \{1,3,4\}, \{1,3,5\}, \{2,3,4\}, \{2,3,5\}, \{3,4,5\}.$$

Assume that there are only five items in the data set.

(a) List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.

(b) List all candidate 4-itemsets obtained by the candidate generation procedure in *Apriori*.

(c) List all candidate 4-itemsets that survive the candidate pruning step of the *Apriori* algorithm.