

**Boston University**  
**Department of Computer Science**  
**CS 565 Data Mining**

Midterm Exam

Date: Oct 14, 2009

Time: 4:00 p.m. - 5:30 p.m.

Write Your University Number Here: \_\_\_\_\_

Answer all questions.

Good luck!

**Problem 1 [25 points]**

True or False:

1. Maximal frequent itemsets are sufficient to determine all frequent itemsets with their supports.
2. The maximal frequent itemsets (and only those) constitute the positive border of a frequent-set collection.
3. Let  $D$  be the Euclidean distance between multidimensional points. Assume a set of  $n$  points  $X = \{x_1, \dots, x_n\}$  in a  $d$ -dimensional space and project them into a lower-dimensional space  $k \geq O(\log n)$ . If  $Y = \{y_1, \dots, y_n\}$  is the new set of  $k$ -dimensional points, then, the Johnson Lindenstrauss lemma states that for all pairs  $(i, j)$  it holds that  $S(x_i, x_j) = D(y_i, y_j)$ . (All points  $x_i$  and  $y_i$  are normalized to have length 1.)
4. Computing the mean and a variance of a stream of numbers can be done using a single pass over the data and constant ( $O(1)$ ) space.
5. The *disagreement distance* between two clusterings is a metric.

**Problem 2 [10 points]**

Consider a dictionary of  $n$  terms (words)  $\mathcal{T} = \{t_1, \dots, t_n\}$ . Each term  $t_i$  is associated with its importance  $w(t_i)$  (a positive real value). Additionally, assume a collection of  $m$  documents  $\mathcal{D} = \{d_1, \dots, d_m\}$ , such that each document  $d_i$  uses a subset of terms in the dictionary (i.e.,  $d_i \subseteq \mathcal{T}$ ). You are asked to give a polynomial-time algorithm that finds a collection of  $k$  documents  $\mathcal{C} \subseteq \mathcal{D}$  that cover the most important terms in the dictionary. That is, the weighted sum of the terms covered by at least one document in  $\mathcal{C}$  is maximized. In other words, find  $\mathcal{C}$  such that  $F(\mathcal{C}) = \sum_{t_i \in (\cup_{d \in \mathcal{C}} d)}$   $w(t_i)$  is maximized. Discuss the optimality of the proposed algorithm.

**Problem 3 [10 points]**

Assume a set  $X$  of  $d$ -dimensional points  $X = \{x_1, \dots, x_n\}$ , and a metric distance function  $D$  between them. Let  $y$  be a  $d$ -dimensional point for which

$$\sum_{i=1}^n D(x_i, y)$$

TID	Items Bought
T1	I6, I1, I3
T2	I1, I2, I4, I5, I3
T3	I3, I2, I5
T4	I6, I7
T5	I1, I3, I2, I4, I5
T6	I1, I3, I6
T7	I1, I2, I5, I7
T8	I2, I8, I5, I1
T9	I4, I6
T10	I1, I2, I5

Table 1: Transaction database  $\mathcal{D}$ .

is minimized. Show that there exists a point  $x' \in X$  such that

$$\sum_{i=1}^n D(x_i, x') \leq 2 \sum_{i=1}^n D(x_i, y).$$

**Problem 4 [35 points]**

Consider the transaction database  $\mathcal{D}$  shown in Table 1, and let  $minsup = 0.4$ .

- Find all frequent itemsets and their supports in  $\mathcal{D}$  using the Apriori algorithm. At each level  $k$  of this algorithm show the candidate frequent itemsets  $C_k$  and the actual frequent itemsets at this level. (10 points)
- From the frequent itemsets you have mined generate all association rules with confidence 100%. (5 points)
- Report all maximal itemsets for the collection of frequent itemsets you have mined above. (5 points).
- Report all closed itemsets for the collection of frequent itemsets you have mined above. (5 points)
- For all maximal itemsets you have reported above, report their frequency and their expected frequency. (10 points)

**Problem 5 [30 points]**

Consider the dataset shown in Table 2 consisting of 6 records. Each record consists of 5 attributes. For two tuples  $x, y$  consider the *jaccard* similarity functions defined as follows:

$$Jaccard(x, y) = \frac{|x \cap y|}{|x \cup y|},$$

where  $x \cap y$  is a vector that has 1 in every dimension where both  $x$  and  $y$  have value 1. Similarly,  $x \cup y$  is a vector that has 1 in every dimension where at least one of the  $x$  or  $y$  vectors have value 1.

$x_1$	1	0	1	1	0
$x_2$	1	1	0	1	1
$x_3$	1	0	1	1	0
$x_4$	0	1	0	1	0
$x_5$	1	0	1	0	1
$x_6$	0	1	1	1	0

Table 2: Table  $X$  of binary points.

1. Apply Single-link clustering with Jaccard similarity function. (10 points)
2. Apply  $k$ -means clustering algorithm for  $k = 2$  and starting points  $(1, 0, 1, 0.6, 0.3)$  and  $(0.2, 1, 0.3, 1, 0.3)$  (10 points)
3. Compute the disagreement distance between the result of  $k$ -means clustering and the hierarchical clustering computed above. Consider the clustering that consists of 3 clusters in the hierarchical clustering. (10 points)

— **END OF PAPER** —