Comp7103B: First Assignment

Discussing with other students, the professor and the teaching assistant is allowed and encouraged, however, everybody should submit his/her own solutions. Students submitting the same solutions (or very similar) for one or more exercises will be penalized or might fail the assignment. Students should write their solutions in a short report (5 pages max) and submit it through the moodle website (pdf or scanned copy) together with the two notebooks (one for each of the practical tasks).

Question 1 Frequent Itemsets, 20/100 points. a) Table 1 shows a list of baskets as well as the items they contain. For example, this could be the set of products bought by each customer during a single trip to a grocery store. Using the A-priori algorithm, find all frequent itemsets with support threshold 0.4 (i.e. in this example they occur at least 40% of 7 times, i.e. at least three times.) In particular you should specify for each pass of the algorithm, the frequent itemsets, as well as the counters (and their values) kept in main memory by the A-priori algorithm.

b) use the results from the previous step so as to compute the confidence and support of the rule $b \rightarrow d$. In order to answer this question, are the results from the previous step sufficient or some information is missing?

c) Table 2 shows a new list of baskets as well as the items they contain. Using the PCY algorithm, find all frequent itemsets with support threshold 0.33 (i.e. in this example they occur at least 33% of 6, i.e. at least two times.) To this end, we are going to use the hash function $f(i, j) = (i + j)\%3$ (% and *mod* are equivalent). In particular you should specify for each pass of the algorithm, the frequent itemsets, as well as all the counters (and their values) kept in main memory by the PCY algorithm.

ID	Baskets
1	a,b,c,e
2	a,d,b
3	c,b
4	a,b,d,e
5	b,d
6	a,b
7	a

Table 1: A list of baskets as well as the items they contain.

ID	Baskets
1	1,3,4
2	4,5
3	2,7
4	1,6
5	2,7
6	3

Table 2: A list of baskets as well as the items they contain.

Question 2 (k-means), 20/100 points.

1. We are given the following points in the 2-dimensional euclidean space. P1=$(0, 0)$, P2=$(0, 1/2)$, P3=$(1, 1/2)$, P4=$(1, 1)$, P5=$(4, 0)$, P6=$(4, 1)$, P7=$(5, 1)$. Suppose that P1 = (0, 0) and P4 = (1, 1) are chosen as initial centroids for the K-means algorithm, K=2. Show step by step the clustering you would obtain by running K-Means on the previous set of points, while specifying for each clustering the current set of centroids. Recall that the algorithm terminates when the current set of centroids does not change.

2. Provide an example for which the K-means algorithm produces at least an empty cluster, that is, the number of non-empty clusters is $< K$. Your example should contain at most 6 points in a one-dimension Euclidean space, while the number of points should not be smaller than K. We recall that the initial centroids are always chosen among the input points. Show all the steps of the algorithm until it terminates.

Question 3: Practical exercise on Clustering, 30/100 points. You are provided with a dataset ("clustering_data.csv") and a text document describing the task.

Question 4: Practical exercise on FI and AR, 30/100 points. You are provided with a dataset ("mammographic_masses.csv") and a text document describing the task.