Examination 2006-08-17

Data Mining (5 hours)

This examination contains **5** questions and **40** points in total. The point distribution between sub-questions is clearly identifiable. Note that you will get **credit only for answers that are correct**. To pass, you must score at least **22**. To get VG, you must score at least **30**. The instructor reserves the right to lower these numbers. All answers should preferably be in English (however, if you are uncomfortable with English, you can of course write in Swedish).
You are allowed to use dictionaries to/from English, a simple (i.e., non-programmable) calculator, and the one A4 paper with notes that you have brought with you, but **no other material**. Whenever in *real doubt* for what a particular question might mean, state your assumptions clearly. Write readably and clearly. Solutions that cannot be read will of course not get any points, and unclear sentences run the risk of being misunderstood.
When you are finished, please place the exam cover page first, then the answer sheet, and then all other pages in an order that corresponds to the order of the remaining questions. Please staple all those pages together and remember to write your name on each page you hand in.

Don't Panic!

1. **Classification (11 pts total).**

   (a) **(4 pts)** In text of not more than two pages, present the main ideas (perhaps in the form of pseudocode) of the K-Nearest Neighbor (KNN) technique for classification.

   (b) **(2 pts)** What is the complexity of the KNN algorithm as a function of the number of elements in the training set (q), and the number of elements (n) to be classified?

   (c) **(3 pts)** Discuss issues that are important to consider when employing a Decision Tree based classification algorithm.

   (d) **(2 pts)** What are the main advantages and disadvantages of Decision Tree classification algorithms?

---

2. **Clustering (6 pts total).**

   (a) **(3pts)** Suppose that for a data set
   - there are m points and K clusters,
   - half the points and clusters are in "more dense" regions,
   - half the points and clusters are in "less dense" regions, and
   - the two regions are well-separated from each other.

For this data set, which of the following should occur in order to minimize the squared error when finding K clusters:

i. Centroids should be equally distributed between more dense and less dense regions.

ii. More centroids should be allocated to the less dense region.

iii. More centroids should be allocated to the more dense region.

Briefly justify your answer.

Note: Do not get distracted by special cases or bring in factors other than density in your reasoning. However, if you feel the true answer is different from any of the above, justify your response.

(b) **(3pts)** Describe the change in the time complexity of K-means as the number of clusters to be found increases and briefly justify why it is so.

---

3. **Association Rules (10 pts total)**

(a) (**6pts**) Consider the following transaction database:

TID	**Items**
01	A, B, C, D, F
02	A, B, C, D, E, G
03	A, C, G, H
04	B, C, D, E, H
05	D, E, F, H
06	A, B, C
07	A, D, F
08	B, E, I
09	C, D, F
10	A, B, D, E
11	C, D, H, I
12	C, E, F
13	B, C, D, F
14	A, B, C, D
15	C, H, I, J
16	A, D, E, F, H
17	F, G, H
18	A, D, H
19	D, E, F
20	B, C, D, E, H

Apply the Apriori algorithm with minimum support of 30% and minimum confidence of 75% to find all the association rules in the data set. Give details of your computation at each step, however, at each step give only the frequent itemsets that satisfy minimum support (i.e., itemsets which appear in at least 6 transactions). Also specify the confidence and support for each of the rules you discovered.

(b) (**4pts; 2pts each**) In text of not more than two pages present two methods to improve the efficiency of the 'basic' Apriori algorithm, discuss why these methods achieve the desired efficiency improvement, and mention situations in which their use is recommended.

---

4. **Sequence Mining (7 pts total).** Consider the sequence database shown below:

sequence id	time	itemset
1	1	abd
1	2	bcd
1	3	bcd
2	1	b
2	2	abc
3	1	ab
3	2	bcd

We can store the occurrence for each item using a bit vector representation, where we have one bit for each seqid and time pair, and we store whether or not an item appears at that instance. For example, the bit vector for a is 1000110. We obtain this bit vector by concatenating the fragments from each sequence. For example, in sequence 1, a appears at time 1, but not at time 2 and 3, so its bit vector fragment in this sequence is 100. In sequence 2, a has the bit vector fragment 01, and in sequence 3 the fragment is 10. Putting all three fragments together we get $100 \oplus 01 \oplus 10 = 1000110$.

(a) (**1pt**) Show the bit vectors for b, c and d.

(b) (**3pts**) Describe how we can count the frequency of all subsequences (with non-consecutive ones allowed) by using bit operations (like OR, AND, NOT) on these bit vectors. Show an example.

(c) (**3pts**) Show the bit vectors and the frequent sequences for minimum support 2. Briefly describe how you arrived in this result.

---

5. **Web Mining (6pts; 3pts each).** Suppose a Web graph is effectively undirected, i.e., page i points to page j if and only if page j points to page i. Are the following statements true or false? Briefly (i.e., in text of no more than a couple of lines if true, or with showing a counter-example if false) justify your answers.

(a) The hubbiness and authority vectors are identical, i.e., for each page, its hubbiness is equal to its authority.

(b) The matrix M that we use to compute PageRank is symmetric; i.e., $M[i, j] = M[j, i]$ for all i and j.

Good luck !