Examination 2004-12-22

Data Mining (5 hours)

Please write your name on each page you hand in. When you are finished, please staple these pages together in an order that corresponds to the order of the questions.
This examination contains **40** points in total and their 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.

1. **Classification (8 pts total, 4pts each).** Recall the CART (Classification and Regression Trees) splitting criteria in decision tree construction:

$$CART(D_0, D_1) = 2P(D_0)P(D_1)\sum_{i=1}^{k}|P(C_i|D_0) - P(C_i|D_1)|$$

After considering each binary split into two partitions D_0 and D_1, we can choose the one with the biggest CART value as the best split.

(a) Evaluate the CART value for all split points for Hair in the table below.

	Height	Hair	Eye	Class
1	Tall	Blond	Brown	C_1
2	Tall	Dark	Blue	C_1
3	Tall	Dark	Brown	C_1
4	Short	Dark	Blue	C_1
5	Short	Blond	Brown	C_1
6	Tall	Red	Blue	C_2
7	Tall	Blond	Blue	C_2
8	Short	Blond	Blue	C_2

Table 1: Table for Question 1a.

(b) In this sub-question, for simplicity, assume $k = 2$. What is the maximum and the minimum value CART can take? Under what circumstances do these values happen? [1]

[1] To avoid confusion, sub-question 1b does *not* refer to Table 1; it is a general one. Also, do not panic: this is the only question in this exam that requires some thinking, though not too much really...

2. **Clustering (9 pts total).**

(a) (**5pts**)

Suppose you want to cluster the eight points shown below using k-means.

	A_1	A_2
x_1	2	10
x_2	2	5
x_3	8	4
x_4	5	8
x_5	7	5
x_6	6	4
x_7	1	2
x_8	4	9

Assume that $k = 3$ and that initially the points are assigned to clusters as follows: $C_1 = \{x_1, x_2, x_3\}$, $C_2 = \{x_4, x_5, x_6\}$, $C_3 = \{x_7, x_8\}$. Apply the k-means algorithm until convergence (i.e., until the clusters do not change), using the Manhattan distance. (Hint: the Manhattan distance of point x_1 from the centroids of C_1, C_2, and C_3 in the initial clustering assignment is $5\frac{2}{3}$, $8\frac{1}{3}$, and 5, respectively.) Make sure you clearly identify the final clustering and show your steps.

(b) (**4pts**)

Consider the set of points given in Figure 1.

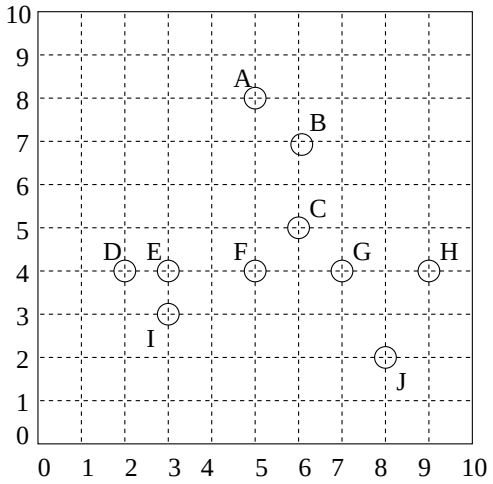

Figure 1: Points for question 2b.

Assume that $eps = \sqrt{2}$ and $minpts = 3$ (including the center point). Using Euclidian Distance find all the density-based clusters in the figure using the DBSCAN algorithm. List the final clusters (with the points in lexicographic order, i.e., from A to J) and outliers.

3. **Association Rules (8 pts total)**

   (a) (**4pts total; 2+2**) After i) finding all frequent itemsets with minimum support of 50% using the Apriori algorithm on the following database:

TID	Items
1	K, A, D, B
2	D, A, C, E, B
3	C, A, B, E
4	B, A, D

   ii) list all association rules with the same minimum support (50%) and with confidence of 75%. You do not have to show your steps, but if you chose to do so make sure that your final results are clearly identified.

   (b) (**4pts**) Consider the *partition algorithm* for association rule mining. It divides the database into p partitions, not necessarily of equal size, such that $D = \cup_{i=1}^{p} D_i$, where D_i is partition i, and for any $i \neq j$, we have $D_i \cap D_j = \emptyset$. Also, let $N_i = |D_i|$ denote the number of transactions in partition D_i. The algorithm first counts only locally frequent itemsets, i.e., itemsets that are frequent in each partition. In the second scan it takes a union of all locally frequent itemsets and computes the true frequent itemsets in the entire database. Prove that if a pattern is globally frequent in the database, then it must be locally frequent in at least one partition.

---

4. **Sequence Mining (7 pts)** In class we talked about mining (maximal) sequential patterns from market basket data and presented an efficient algorithm for it. In this question, we want to mine *maximal frequent sequences*, and the problem is the same as that of mining sequential patterns from market basket data where each transaction consists of only one item.

   Given the DNA sequence database (i.e., the only items that can be "purchased" are A, C, G, and T) below:

S_1: ACGTCACG
S_2: TCGA
S_3: GACTGCA
S_4: CAGTC
S_5: AGCT
S_6: TGCAGCTC
S_7: AGTCAG

   Find the maximal frequent sequences with minimum support = 4/7, i.e., appearing in four out of the seven sequences above. Make sure you clearly identify your final answer and for possible partial credit show your work.

---

5. **Web Mining (8 pts total)**

(a) (**4pts; max(0,number_correct-0.5*number_wrong)**)

Let $In(x)$ denote the set of pages which link to a page x, and let $Out(x)$ denote the set of pages to which page x links. Let $h(x), a(x)$, and $p(x)$ denote the "hubbiness", authority, and PageRank of page x, respectively. In a matrix similar to the one below, indicate whether each of the following statements is **always** true (T) or sometimes false (F). Note that no justification is required for these sub-questions; your answer should be just the matrix.

i. If $Out(i) \subseteq Out(j)$, then $h(i) \leq h(j)$.

ii. If $Out(i) \subseteq Out(j)$, then $p(i) \leq p(j)$.

iii. If $In(i) \subseteq In(j)$, then $p(i) \leq p(j)$.

iv. If $In(j) \subseteq In(i)$, then $a(i) \leq a(j)$.

	T	F
i		
ii		
iii		
iv		

(b) (**4pts; 2pts 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.

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

ii. 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 !