Examination 2005-12-19

Data Mining (5 hours)

The last page of this exam is an "answer sheet" for some of its questions. 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.
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.

Don't Panic!

1. **Classification using KNN (4 pts total; 2 pts each).** Consider the one-dimensional data set shown in the table below.

value	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
class	−	−	+	+	+	−	−	+	−	−

   (a) Classify the data point x = 5.0 according to its 1-, 3-, 5-, and 9-nearest neighbors (using majority vote).

   (b) For the same sets of neighbors, classify the data point x = 5.0 using a distance-weighted voting approach (rather than using a simple majority vote) with weight $w_i = \frac{1}{d(x,x_i)^2}$ where x_i is the value of the i-th nearest neighbor and $d(\cdot)$ computes the distance.

2. **Classifier Accuracy (4 pts total; 1 pt each).** Consider the task of building a classifier from random data, where the attribute values are generated randomly irrespective of the class labels. Assume the data set contains records from two classes, "+" and "–". Half of the data set is used for training while the remaining half is used for testing.

   (a) Suppose there is an equal number of positive and negative records in the data and the decision tree classifier predicts every test record to be positive. What is the expected error rate of the classifier on the test data?

(b) Repeat the previous analysis assuming that the classifier predicts each test record to be positive class with probability 0.8 and negative class with probability 0.2.

(c) Suppose two-thirds of the data belong to the positive class and the remaining one-third belong to the negative class. What is the expected error of a classifier that predicts every test record to be positive?

(d) Repeat the previous analysis assuming that the classifier predicts each test record to be positive class with probability 2/3 and negative class with probability 1/3.

Recall that all error rates are percentages; a perfect classifier has 0% error rate.

3. **Clustering (6 pts total; 3 pts each).** The *leader algorithm* represents each cluster using a point, known as a leader, and assigns each point to the cluster corresponding to the closest leader, unless this distance is above a user-specified threshold.[1] In that case, the point becomes the leader of a new cluster.

   (a) What are the advantages and disadvantages of the leader algorithm as compared to K-means?

   (b) Suggest ways in which the leader algorithm might be improved.

4. **Clustering (5 pts total).** You are given two sets of 100 points that fall within the unit square (a square of size 1 × 1). One set of points is arranged so that the points are uniformly spaced. The other set of points is generated randomly from a uniform distribution over the unit square.

   (a) **(2pts)** Which set of points (if any) will typically have a smaller SSE for K=10 clusters?

   (b) **(3pts)** What will be the behavior of DBSCAN on the uniform data set? On the random data set?

   Briefly explain your answers. Note that for part (a) claiming that both sets of points have similar SSE (Sum of Squared Errors) is one of the possible answers.

5. **Association Rules (6 pts total).** Briefly explain all your answers to the questions below.

   (a) **(2pts)** Let c_1, c_2, and c_3 be the confidence values of the rules $\{p\} \rightarrow \{q\}$, $\{p\} \rightarrow \{q, r\}$, and $\{p, r\} \rightarrow \{q\}$, respectively. If we assume that c_1, c_2, and c_3 have different values, what are the possible relationships that may exist among c_1, c_2, and c_3? Which rule has the lowest confidence?

   (b) **(2pts)** Repeat the analysis in the previous sub-question (5a) assuming that the rules have identical support. Which rule has the highest confidence?

   (c) **(2pts)** Transitivity: Suppose the confidence of the rules $A \rightarrow B$ and $B \rightarrow C$ are larger than some threshold, *minconf*. Is it possible that $A \rightarrow C$ has a confidence less than *minconf*?

[1] The algorithm described here slightly differs from the original leader algorithm, which assigns a point to the first leader that is within the threshold distance. But do not let this detail distruct you.

Table 1: Market basket transactions for Question 6.

Transaction ID	Items Bought
1	{ Milk, Beer, Diapers }
2	{ Bread, Butter, Milk }
3	{ Milk, Diapers, Cookies }
4	{ Bread, Butter, Cookies }
5	{ Beer, Cookies, Diapers }
6	{ Milk, Diapers, Bread, Butter }
7	{ Bread, Butter, Diapers }
8	{ Beer, Diapers }
9	{ Milk, Diapers, Bread, Butter }
10	{ Beer, Cookies }

6. **Association Rules (4 pts total; 1 pt each).** Consider the market basket transactions shown in Table 1. Assume that *minsup* > 0.

   (a) What is the maximum size of frequent itemsets that can be extracted?

   (b) What is the maximum number of size-3 itemsets that can be derived from this data set. (An expression to calculate this number is also a valid answer to this sub-question.)

   (c) Find an itemset (of size 2 or larger) that has the largest support.

   (d) Find a pair of items, a and b, such that the rules $\{a\} \rightarrow \{b\}$ and $\{b\} \rightarrow \{a\}$ have the same confidence. Any such pair will do.

7. **Sequence Mining (5 pts total).** Consider the following frequent 3-sequences:

$$< \{1,2,3\} >, < \{1,2\}\{3\} >, < \{1\}\{2,3\} >, < \{1,2\}\{4\} >, < \{1,3\}\{4\} >,$$
$$< \{1,2,4\} >, < \{2,3\}\{3\} >, < \{2,3\}\{4\} >, < \{2\}\{3\}\{3\} >, \text{and} < \{2\}\{3\}\{4\} > .$$

   (a) **(3pts)** List all the candidate 4-sequences produced by the candidate generation step of the Generalized Sequential Patterns (GSP) algorithm.

   (b) **(2pts)** List all the candidate 4-sequences pruned during the candidate pruning step of the GSP algorithm (assuming no timing constraints).

8. **Web Mining (6 pts total; 2 pts for the ranking + 0.5 for each correct explanation)** Suppose that you have a web site with a number of pages that all link to some other page (i.e., you have no dangling links). You wish to improve your PageRank. Consider the following strategies:

   (a) Add links from your pages to high PageRank pages out on the web.

   (b) Remove links from your pages to high PageRank pages out on the web.

   (c) Add links from your pages to low PageRank pages out on the web.

   (d) Remove links from your pages to low PageRank pages out on the web.

(e) Force webmasters of low PageRank pages out on the web to link to your pages.

(f) Force webmasters of high PageRank pages out on the web to link to your pages.

(g) Force webmasters of low PageRank pages out on the web to remove links to your pages.

(h) Force webmasters of high PageRank pages out on the web to remove links to your pages.

For each of these strategies, explain in a line or two whether it will help your PageRank or hurt your PageRank. Also describe how markedly these strategies change your PageRank relative to each other (i.e., try to rank these strategies based on how effective they are). Write your answers in the space alloted in the "answer sheet" page.

Note: Your answer should be based on the details of the technique, and not on how likely you are to actually influence other people to change their web pages on your demand. Assume you have global control of the web.

Good luck !

Answer Sheet

Name:

Answer for Question 1

	1-NN	3-NN	5-NN	9-NN
(a) Class of x = 5.0 using majority voting				
(b) Class of x = 5.0 using distance-weighted voting				

Answer for Question 2

	(a)	(b)	(c)	(d)
Classifier error rate				

Answer for Question 6

(a) Max itemset size	
(b) Max number of 3-itemsets	
(c) k-itemset of largest support, $k \geq 2$	
(d) Pair of items whose rules have same confidence	and

Answer for Question 8

Overall Ranking	really bad strategy... no change ...very good strategy
(a)	
(b)	
(c)	
(d)	
(e)	
(f)	
(g)	
(h)	