CS4445 Data Mining and Knowledge Discovery in Databases. B Term 2010

Exam 2 - December 10, 2010

Prof. Carolina Ruiz
Department of Computer Science
Worcester Polytechnic Institute

NAME: _____

Problem I: _____ **(/30 points)** Rule-Based Classification

Problem II: _____ **(/27 points)** Association Analysis

Problem III: _____ **(/33 points)** Clustering Analysis

Problem IV: _____ **(/20 points)** Anomaly Detection

TOTAL SCORE: _____ **(/100 points)**

Instructions:

- Show your work and justify your answers
- Use the space provided to write your answers
- Ask in case of doubt

Problem I. Rule-Based Classification [30 Points]

1. **[8 Points]** Define validation set. What's the difference between validation set and training set? What's the difference between validation set and test set? Explain.

2. A validation set is used in some pruning approaches. Consider for example the following two rule pruning algorithms: IREP and RIPPER. Both algorithms apply the reduced-error pruning method to determine whether a rule needs to be pruned. The reduced-error pruning method uses a validation set to estimate the generalization error of a classifier. To determine whether a rule R should be pruned, these algorithms use the following measures, respectively:

   - IREP: $V_{IREP}(R) = \frac{p + (N - n)}{P + N}$
   - RIPPER: $V_{RIPPER}(R) = \frac{p - n}{p + n}$

   where,

   P is the total number of positive examples in the validation set
   N is the total number of negative examples in the validation set
   p is the number of positive examples in the validation set covered by the rule R
   n is the number of negative examples in the validation set covered by the rule R

   (Here, example = data instance.) Each method favors rules that have higher values for its metric.

   Consider the following pair of rules (note that R_1 is a pruned version of R_2):

   - R_1: If A then C
   - R_2: If A && B then C

   Assume that class C is the positive class. Consider a validation set that contains 500 positive examples and 500 negative examples.

   - For R1: assume that the number of positive examples covered by the rule is 200, and the number of negative examples covered by the rule is 50.
   - For R2: assume that the number of positive examples covered by the rule is 100, and the number of negative examples covered by the rule is 5.

   1. **[5 Points]** Would IREP prefer R_1 over R_2? Show you work and explain your answer.

2. **[5 Points]** Would RIPPER prefer R_1 over R_2? Show you work and explain your answer

3. **[2 Points]** Consider now the following rule R_0 (note that R_0 is a pruned version of R_1):

   R_0: If *true* then C (here "*true*" means no condition)

   Using the same information about the validation set above, what are the values of p and n for rule R_0?

4. **[5 Points]** Would IREP prefer R_0 over R_1? Show you work and explain your answer.

5. **[5 Points]** Would RIPPER prefer R_0 over R_1? Show you work and explain your answer.

Problem II. Association Analysis [27 Points]

Assume that the Apriori algorithm is used to generate association rules from a dataset of transactions. Assume also that the **complete list of frequent 4-itemsets** (i.e., all itemsets of size 4 that have enough support) that Apriori has generated is given on the right:
(Note that the dataset transactions are not provided.)

Level 4: Complete list of frequent 4-itemsets
{a, b, c, d}
{a, b, c, e}
{a, b, d, e}
{a, c, d, e}
{b, c, d, e}
{c, d, e, f}
{c, d, e, g}

Consider the 5-itemsets {a, b, c, d, e}, {b, c, d, e, f}, and {c, d, e, f, g}. Answer the questions below assuming that the **join** (= merge) and the **candidate prune** conditions are used to generate candidate itemsets in Level 5:

	Itemset {a, b, c, d, e}	Itemset {b, c, d, e, f}	Itemset {c, d, e, f, g}
[3 points each] Will this itemset be generated as a candidate 5-itemset using the **join** condition? Yes? No? Maybe? Justify your answer.			
[3 points each] If your answer above for the **join** condition was yes, will this itemset be eliminated by the **candidate prune** condition? Yes? No? Maybe? Justify your answer.			
[3 points each] Is this itemset frequent (i.e., does it have enough support)? Yes? No? Maybe? Justify your answer. (Note that the dataset of transactions is **not** given here, so your answer must be general.)			

Problem III. Clustering Analysis [33 Points]

1. **[10 Points]** How would you extend DBSCAN to appropriately cluster a dataset in which different regions have different densities? Explain your answer and illustrate with a picture.

2. **[8 Points]** Two metrics used in the evaluation of a clustering are *cohesion* and *separation*. Loosely speaking, cohesion refers to how similar (or close) data points in the same cluster are, and separation refers to how distinct or well-separated clusters in a clustering are. For each of the following notions, state whether or not the notion can be used to measure cohesion, or separation, or neither of the two: Sum of Squared Errors (SSE), min distance (single link), max distance (complete link), and group average distance. Explain your answers.

3. Consider a dataset that contains a target attribute. In this problem, you will be asked to evaluate a clustering of this dataset with respect to the target attribute (supervised evaluation). Assume that the dataset contains 5 data points: p1, p2, p3, p4, and p5.

   - Clusters: Assume that these points have been clustered into 2 clusters, without using the target attribute. p1, p2, and p3 belong to one cluster, and p4 and p5 belong to the other cluster.
   - Target Attribute: Assume that p1 and p2 share the same target value (or class); and that p3, p4, and p5 share the same target value, different from that of p1 and p2.

   The following matrices represent this same information:

cluster similarity matrix

Point	p1	p2	p3	p4	p5
p1	1	1	1	0	0
p2	1	1	1	0	0
p3	1	1	1	0	0
p4	0	0	0	1	1
p5	0	0	0	1	1

class similarity matrix

Point	p1	p2	p3	p4	p5
p1	1	1	0	0	0
p2	1	1	0	0	0
p3	0	0	1	1	1
p4	0	0	1	1	1
p5	0	0	1	1	1

   1. **[8 Points]** Discuss how the correlation value of the two matrices can be used to evaluate the clustering with respect to the target attribute. What is the range of values for this correlation? What would each of the extreme values of this range represent? Explain.

2. **[7 Points]** Based on the values of the two matrices above, one can create a two-way contingency table for determining whether pairs of data instances are in the same class and same cluster:

	same cluster	different cluster
same class	f_{11} the number of **pairs** of instances having the same class and the same cluster	f_{10} the number of **pairs** of instances having the same class and a different cluster
different class	f_{01} the number of **pairs** of instances having a different class and the same cluster	f_{00} the number of **pairs** of instances having a different class and a different cluster

The Rand statistic and the Jaccard coefficient are two commonly used metrics to evaluate cluster validity. Calculate the values f_{11}, f_{10}, f_{01}, and f_{00} in the table above. Use those values to calculate the Rand statistics and the Jaccard coeffient below:

$$\text{Rand statistics} = \frac{f00 + f11}{f00 + f01 + f10 + f11}$$

$$\text{Jaccard coefficient} = \frac{f11}{f01 + f10 + f11}$$

Problem IV. Anomaly Detection [20 Points]

Define the notion of outlier and provide an appropriate anomaly scoring function *f(x)* for **TWO** of the following anomaly detection approaches:

1. **Statistics-based**

   probabilistic-based definition of outlier:

   anomaly score function: Given a data instance *x,* the anomaly score of *x* is

   f(x) =

2. **Proximity-based**

   proximity-based definition of outlier:

   anomaly score function: Given a data instance *x,* the anomaly score of *x* is

   f(x) =

3. **Density-based**

   density-based definition of outlier:

   anomaly score function: Given a data instance *x,* the anomaly score of *x* is

   f(x) =

4. **Clustering-based**

   clustering-based definition of outlier:

   anomaly score function: Given a data instance *x,* the anomaly score of *x* is

   f(x) =