CS4445 Data Mining and Knowledge Discovery in Databases. A Term 2008

Exam 2 - October 14, 2008

Prof. Carolina Ruiz
Department of Computer Science
Worcester Polytechnic Institute

NAME: ____Prof. Ruiz____

Problem I:	**(/25 points)** General Data Mining
Problem II:	**(/40 points)** Numeric Predictions
Problem III:	**(/10 points)** Instance Based Learning
Problem IV:	**(/35 points)** Clustering
TOTAL SCORE:	**(/100 points (+ 10 points of extra credit))**

Instructions:

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

Problem I. General Data Mining [25 points]

1. **[9 points]** What is the difference between supervised and unsupervised discretization? Explain decisively.

   Solution:
   In supervised discretization, the target attribute is used to help determine appropriate bins for the attribute being discretized. In unsupervised discretization, the attribute being discretized is split into bins that are determined without taking the target attribute (if any) under consideration.

2. **[8 points]** What is the difference between classification and regression?

Solution:

In classification, the target attribute is nominal. In regression, the target attribute is numeric.

3. **[8 points]** What is stratified sampling?

Solution:

A stratified sample is a sample that preserves the distribution of the target attribute.

Problem II. Numeric Predictions [40 points]

Consider the following dataset which is a small subset adapted from the Auto Miles-per-gallon (MPG) dataset, which is available at the University of California Irvine (UCI) Data Repository. Note that instances have been labeled with a number in parentheses so that you can refer to them in your solutions.

	car-name (disregard)	**cylinders**	**weight**	**model-year**	**mpg**
(1)	chevrolet	8	3504	70	18
(2)	chevrolet	4	2950	82	27
(3)	chevrolet	4	2395	82	34
(4)	toyota	4	2372	70	24
(5)	toyota	4	2155	76	28
(6)	toyota	4	2665	82	32
(7)	volkswagen	4	1835	70	26
(8)	volkswagen	4	1937	76	29
(9)	volkswagen	4	2130	82	44
(10)	ford	8	4615	70	10
(11)	ford	4	2665	82	28
(12)	ford	8	4335	77	16

The purpose of this problem is to construct a model/regression tree to predict the attribute **mpg** (miles-per-gallon) using the three predicting attributes (cylinders, weight, model-year).
The partial tree below is the result of applying the model/regression tree construction algorithm. The tree contains 5 leaves, marked with LM1, LM2, LM3, LM4, and LM5.

For the solutions of Problem II, see the solutions of Exam 2, CS4445 B term 2006 available at: http://www.cs.wpi.edu/~cs4445/b06/Exams/solutions_exam2_cs4445_b06.html

```
cylinders <= 6 :
|
|  model-year <= 79 :
|   |
|   |  model-year <= 73 :         LM1 This leaf contains instances: (4) and (7)
|   |
|   |  model-year >  73 :         LM2 This leaf contains instances: (5) and (8)
|   |
|  model-year >  79 :
|   |
|   |  weight <= ? :              LM3 This leaf contains instances: ____________________
|   |
|   |  weight >  ? :              LM4 This leaf contains instances: ____________________
|
cylinders >  6 :                  LM5 This leaf contains instances: (1), (10), and (12)
```

1. **Constructing the remaining internal node.** We need to determine the correct split point for the attribute **weight** to split the node marked with a "?" in the tree above. That is, the split point *x* of **weight** that will have the highest SDR.

    1. **(5 Points) Relevant data instances.** List all the data instances that need to be considered when splitting the node marked with "?" above. Suggestion: list these instances sorted in increasing order by **weight** value. (The table below is provided for your convenience, but you might need less or more rows than those provided.)

	car-name (disregard)	**cylinders**	**weight**	**model-year**	**mpg**
( )					
( )					
( )					
( )					
( )					
( )					

    2. **(5 Points) Candidate split points** List all the candidate split points for the attribute weight that need to be considered to find the correct value for "? in the nodes "weight <= ?" and "weight > ?" in the tree above.

3. **(12 Points) Evaluating candidate split points** Compute the SDR (Standard Deviation Reduction) of each of the candidate split points that you listed above. For your convenience, the following standard deviations (std) are provided:

   std({44, 34, 32, 28, 27}) = 6.8

std({44}) = 0	std({34, 32, 28, 27}) = 3.3
std({44, 34}) = 7.1	std({32, 28, 27}) = 2.6
std({44, 34, 32, 28}) = 6.8	std({27}) = 0

   SHOW YOUR WORK AND STATE WHAT FORMULA(S) YOU ARE USING EXPLICITLY.

4. **(2 Points) Choosing the best candidate split point** According to the SDR's that you computed above select the best split point.

5. **(3 Points) Completing the tree** Replace the "?" marks in the following tree with the split point that you determined to be the correct one. ALSO, WRITE DOWN WHICH INSTANCES BELONG TO leaves LM3 and LM4.

```
cylinders <= 6 :
|
|  model-year <= 79 :
|  |
|  |  model-year <= 73 :          LM1 This leaf contains instances: (4) and (7)
|  |
|  |  model-year >  73 :          LM2 This leaf contains instances: (5) and (8)
|  |
|  model-year >  79 :
|  |
|  |  weight <= ?______ :         LM3 This leaf contains instances: ? _________
|  |
|  |  weight >  ?______ :         LM4 This leaf contains instances: ? _________
|
cylinders >  6 :                  LM5 This leaf contains instances: (1), (10), and (12)
```

2. **Constructing the leaves of the tree**
   1. **(5 Points) Regression Tree.** Assume that we will use the tree above as a regression tree. DESCRIBE how to calculate the leaf values (that is, the value that each of the leaf nodes will output as its prediction).

CALCULATE the precise value that the leaf marked as LM5 in the tree above will output. Show your work.

LM5 =

2. **(8 Points) Model Tree** Assume that we will use the tree above as a model tree. DESCRIBE how to calculate the leaf values (that is, the value that each of the leaf nodes will output as its prediction).

   ILLUSTRATE what the function/formula that the leaf marked as LM5 in the tree above will use to produce its output is like. (You don't have to produce the precise function just illustrate what the function will be like.) To simplify your answer, you can disregard the nominal attribute car-name.

   LM5 =

Problem III. Instance Based Learning [10 points]

Consider a dataset with a nominal target attribute (i.e., a nominal CLASS). Suppose that the dataset contains **n** instances. Consider the IBn: **n**-nearest neighbors classifier (without distance weighting) for this dataset. That is, the classifier that will use all the **n** nearest neighbors of a test instance to assign a classification (without any distance weighting) to the test instance.

1. **[5 Points]** What will be the classification assigned to a test instance? Explain.
2. **[5 Points]** What other classifier that you know always outputs the same result as this **n** nearest neighbors classifier (without distance weighting)? Explain your answer.

For the solutions of Problem III, see the solutions of Exam 2, CS4445 D term 2004 available at: http://web.cs.wpi.edu/~cs444x/d04/Exams/solutions_exam2_cs444X_d04.html

Problem IV. Clustering [35 points]

Consider the following variation of the k-means clustering algorithm called the **k-medoids clustering algorithm**. The k-medoids algorithm is identical to the k-means algorithm except that instead of using centroids (which might not be data instances in the dataset) as cluster representatives, it uses **medoids** (which are guaranteed to be data instances in the dataset) as cluster representatives. Remember that the centroid of a cluster is constructed as the average over all data instances in the cluster. In contrast, the medoid of a cluster is the data instance in the cluster that minimizes the average dissimilarity (i.e., average distance) to all the other data instances in the same cluster. That is, if a cluster C contains data instances $a_1, \ldots, a_n$, and $d(a_i,a_j)$ denotes the distance between instances a_i and a_j, (this is the same distance function used to cluster the dataset) then the medoid of the cluster C = $\{a_1, \ldots, a_n\}$ is the data instance a_i in C for which the average distance from a_i to each of the other data instances in the cluster

$$\left(\frac{1}{n}\right) * \sum_{j=1}^{n} d(ai, aj)$$

is the smallest, among all a*i, $1 \leq i \leq n$,* in the cluster C. This medoid is a most centrally located point in the cluster.

1. Assume that the following 3 instances belong to the same cluster and that the **Euclidean distance** is used to measure the distance between two data instances over all the four attributes: car-name, cylinders, model-year, and mpg, WITHOUT using normalization or attribute weights.

	car-name	**cylinders**	**model-year**	**mpg**
a_1	chevrolet	8	70	18
a_2	chevrolet	4	82	27
a_3	toyota	4	70	24

a. **[5 points]** Calculate the centroid of this cluster of 3 instances. Show your work.

Solution:

For the centroid, we need to take the average of each attribute above

	car-name	**cylinders**	**model-year**	**mpg**
centroid	average(chevrolet, chevrolet, toyota) = **chevrolet**	average(8,4,4) = **5.33**	average(70,82,70) = **74**	average(18,27,24) = **23**

b. **[15 points]** Calculate the medoid of this cluster of 3 instances. Show your work. For this, you need to calculate the following Euclidean distances:

- d(a_1,a_2) = $\sqrt{\mathbf{0 + (8-4)^2 + (82-70)^2 + (27-18)^2}}$
  $= \sqrt{\mathbf{0 + 16 + 144 + 81}}$ = $\sqrt{\mathbf{241}}$
  $= \mathbf{15.52}$

- d(a_1,a_3) = $\sqrt{\mathbf{1^2 + (8-4)^2 + (70-70)^2 + (18-24)^2}}$
  $= \sqrt{\mathbf{1 + 16 + 0 + 36}}$ = $\sqrt{\mathbf{53}}$
  $= \mathbf{7.28}$

- d(a_2,a_3) = $\sqrt{\mathbf{1^2 + (4-4)^2 + (82-70)^2 + (27-24)^2}}$
  $= \sqrt{\mathbf{1 + 0 + 144 + 9}}$ = $\sqrt{\mathbf{154}}$
  $= \mathbf{12.4}$

Now calculate:

- The average distance from a_1 to the other two instances in the cluster

  <u>Solution:</u> ½ (15.52 + 7.28) = ½(22.8) = 11.4

  (If you divided by 3 instead of 2, as the formula or the average distance in the project statement above implied, that's fine too as it won't change the relative order of the averages.)

- The average distance from a_2 to the other two instances in the cluster

  <u>Solution:</u> ½(15.52 + 12.4) = ½(27.92) = 13.96

- The average distance from a_3 to the other two instances in the cluster

  <u>Solution:</u> ½(7.28 + 12.4) = ½(19.68) = 9.84

  Then, what data instance is the medoid of this cluster and why?

  <u>Solution:</u> a_3 is the medoid as it is the instance with the lowest average distance to the other instances in the cluster.

2. **[5 points]** In the k-means algorithm, is the centroid of a cluster always unique? Yes? No? Justify your answer.

   Solution:

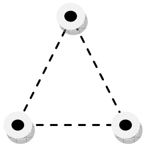
   No, the centroid might not be unique if there are nominal attributes in the dataset. The average/mode of a nominal attribute might not be unique, as there might be two or more nominal values that are the most common ones. For example, the average/mode(chevrolet, chevrolet, ford, toyota, toyota) is either chevrolet or toyota as both are equally common and the most common among the values of the attribute.

   In contrast, the average of a numeric attribute is unique.

3. **[5 points]** In the k-medoids algorithm, is the medoid of a cluster always unique? Yes? No? Justify your answer.

   Solution:

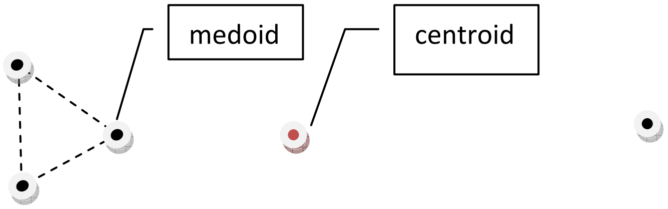
   No, the medoid might not be unique EVEN if there are NO nominal attributes in the dataset. Note that two or more data instances in the cluster might have the same average distance to the other instances in the dataset. For example, consider a cluster with 3 instances located at the vertices of an equilateral triangle in a two dimensional space. Then, the average distances of each data instance to the others is the same. Hence, any of the 3 instances might be the chosen medoid.



4. **[5 points]** The k-medoids algorithm is in general more robust to noise and outliers than the k-means algorithm. Explain why. Illustrate with an example.

   Solution:

   Outliers in a cluster skew the centroid towards the outliers, as the centroid is the average of all the instances in the cluster. In contrast, the medoid of a cluster tends to be the most centrally located point in the cluster. As an example, consider the following cluster of 4 points (in black), the one to the right being a clear outlier:


medoid
centroid