CS4445 Data Mining and Knowledge Discovery in Databases. B Term 2014

Solutions Exam 2 - December 15, 2014

Prof. Carolina Ruiz
Department of Computer Science
Worcester Polytechnic Institute

NAME: __Prof. Ruiz ____

Problem I:	**(/20 points)** Rule-Based Classification
Problem II:	**(/20 points)** Association Analysis
Problem III:	**(/30 points)** Clustering Analysis
Problem IV:	**(/30 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 [20 Points]

Consider a training set that contains 100 positive data instances (class = "+") and 400 negative data instances (class = "-"). Consider the following candidate rules:

R_1: A → class = "+" (covers 4 positive and 1 negative data instances)

R_2: B → class = "+" (covers 30 positive and 10 negative data instances)

Consider the following metrics to measure the goodness of a candidate rule.

1. [15 Points] **FOIL's information gain**. Calculate FOIL's information gain (as done by the RIPPER algorithm) for each candidate rule AND state which of the rules is selected by this metric. Show your work.

Solutions: BTW, note that this problem is an apart from Chapter 5's Exercise 4 (p. 317) of the textbook.

Remember that FOIL's information gain is: $p_1 \times \left(\log_2 \frac{p_1}{p_1+n_1} - \log_2 \frac{p_0}{p_0+n_0}\right)$ where:

p_0 (resp. p_1) is the number of positive data instances (i.e., data instances with class = "+") covered by the rule before (resp. after) adding the candidate condition.

n_0 (resp. n_1) is the number of negative data instances (i.e., data instances with class = "-") covered by the rule before (resp. after) adding the candidate condition.

Here, for both rules R_1 and R_2: p_0 = 100 and n_0 = 400 since the rule *empty* → class = "+" covers 100 positive instances, and 400 negative instances.

FOIL's information gain for R_1: Here p_1 = 4 and n_1 = 1.

$$= p_1 \times \left(\log_2 \frac{p_1}{p_1+n_1} - \log_2 \frac{p_0}{p_0+n_0}\right) = 4 \times \left(\log_2 \frac{4}{4+1} - \log_2 \frac{100}{100+400}\right) = 4 \times \left(\log_2 \frac{4}{5} - \log_2 \frac{1}{5}\right) = 8$$

FOIL's information gain for R_2: Here p_1 = 30 and n_1 = 10.

$$p_1 \times \left(\log_2 \frac{p_1}{p_1+n_1} - \log_2 \frac{p_0}{p_0+n_0}\right) = 30 \times \left(\log_2 \frac{30}{30+10} - \log_2 \frac{100}{100+400}\right) = 30 \times \left(\log_2 \frac{3}{4} - \log_2 \frac{1}{5}\right) = 57.2$$

Hence, R_2 is chosen over R_1 if FOIL's information gain is used to select among candidate conditions.

2. [5 Points] **Rule Accuracy**. Calculate the accuracy of the rule over the training set (as done by the PRISM algorithm using the p/t ratio, where p is the number of positive instances covered by the rule and t is the total number of data instances covered by the rule) for each candidate rule AND state which of the rules is selected by this metric. Show your work.

   Solution:

   Accuracy (= p/t ratio) for R_1 : Here $p = 4$ and $t = 4+1$. So accuracy of the R_1 is: 4/5 = 0.8

   Accuracy (= p/t ratio) for R_2 : Here $p = 30$ and $t = 30+10$. So accuracy of the R_1 is: 30/40 = 0.75

   Hence, R_1 is chosen over R_2 if rule accuracy over the training set is used to select among candidate conditions.

Problem II. Association Analysis [20 Points]

Consider the credit dataset below. The instances of this dataset may be interpreted as transactions. Each transaction is a list of items. Each item is an attribute-value pair of the form attribute=value.

ID	**Credit History (CH)**	**Debt (D)**	**Collateral (Co)**	**Risk (R)**
1	bad	small	none	high
2	bad	small	none	moderate
3	bad	small	adequate	moderate
4	bad	large	none	high
5	unknown	small	none	moderate
6	unknown	small	adequate	low
7	unknown	small	none	low
8	unknown	large	none	high
9	unknown	large	none	high
10	good	small	none	low
11	good	large	none	high
12	good	large	none	moderate
13	good	large	none	low
14	good	large	adequate	low

Assume that the minimum support threshold is 40%, or equivalently, the minimum support count is 6.

1. [12 Points] Use the Apriori algorithm to generate *all* frequent itemsets, level by level. Show your work. (An example is listed for Level 1 to get you going.)

Level 1

Itemset	**Support count**	**Frequent? (yes/no)**
CH=bad	4	no
…		

<u>Solution</u>: **(Taken from the solutions to CS4445 D term 2003 Exam 2)**

CH=bad	4	no
CH=unknown	5	no
CH=good	5	no
D=small	7	yes
D=large	7	yes
Co=none	11	yes
Co=adequate	3	no
R=low	5	no
R=moderate	4	no
R=high	5	no

Level 2

{D=small, Co=none}	5	no
{D=large, Co=none}	6	yes

Level 3 is empty since there is only one frequent 2-itemset, and no one else to join it with.

2. This part is independent from the frequent itemsets question above. Calculate the support and the confidence of the association rule

   CH=bad & Co=none => R=high

   relative to the given dataset above. You may leave your answers in the form of fractions.

   a. [4 Points] Support(CH=bad & Co=none => R=high) = ?

      Solution:

      Support of a rule is the percentage of instances in that dataset that contain all the items in the rule:

      Support = P(CH=bad & Co=none & R=high) = 2/14 = 1/7

   b. [4 Points] Confidence(CH=bad & Co=none => R=high) = ?

      Solution:

      Confidence of a rule is the percentage of data instances that contain all the items on the right-hand side of the rule (consequent) among those data instances that contain all the items on the left-hand side (antecedent) of the rule:

      Confidence = P(R=high | CH=bad & Co=none) = 2/3

Problem III. Clustering Analysis [30 Points]

For each of the following situations, describe what clustering method (among those covered in this course) you would use to solve the problem, why that method, and how you would solve the problem.

NOTE: The clustering methods chosen in the solutions provided below are not the only possible answers. Other clustering methods could be reasonable too, as long as they are well justified.

1. Determine groupings of documents that can help figure out topics, and subtopics within topics, that relate these documents.
   Your choice of clustering method [2 points], justification [2 points], and how you'd use it solve the problem [2 points]

   Solution: I'd use hierarchical clustering as this method produces nested clusters that can be used to identify topics and subtopics. I'd take the collection of documents, define a distance metric between pairs of documents based on similarity, and use (say single-link) hierarchical clustering to create a dendrogram. Then, I would analyze the hierarchical structure of the dendrogram to figure out topics relating documents in the same clusters, and subtopics between topics in the nested structure.

2. Assign students to a given number of shared offices based on similarity.
   Your choice of clustering method [2 points], justification [2 points], and how you'd use it solve the problem [2 points]

   Solution: I'd use k-means clustering as this method allows me to input the number of desired clusters, and it will partition the group of students into this number of clusters. I'd just run k-means with k = number of available office, and then assign students in the same cluster to an office.

3. Cluster tweets to discover current "hot topics" on Tweeter.
   Your choice of clustering method [2 points], justification [2 points], and how you'd use it solve the problem [2 points]

   Solution: I'd use a density-based clustering method like DBSCAN, as I expect hot topics on Tweeter to be more densely populated than other topics. I would experiment running DBSCAN with different input parameters until a reasonable number of core points is identified, and then I would look at the tweets in the neighborhoods defined by the core points to determine what their common topic is.

4. Find customers with shopping patterns that are very different from those of most other customers.
   Your choice of clustering method [2 points], justification [2 points], and how you'd use it solve the problem [2 points]

   Solution: I'd use k-means clustering because I can more easily define these outliers in terms to their distances to most other data instances. I'd define a distance metric that captures similarity in shopping patterns, and run k-means with small input values for k (k=2, 3, ..). For each of the resulting clusterings, I'd check if there are some very small clusters with few customers, or some customers that don't get in any cluster. If any, these would be my candidate customer outliers.

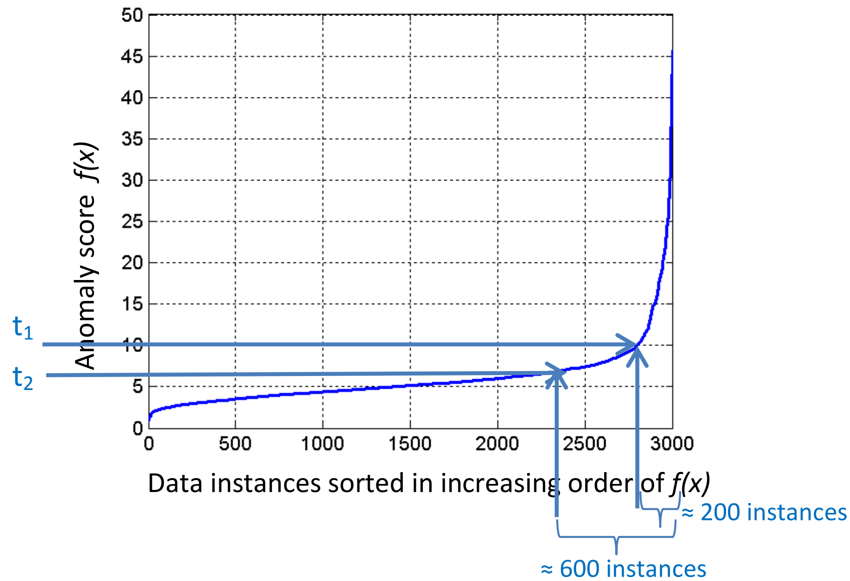
5. Identify a handful of shareholders to invite to a company's board meeting, in such a way that the chosen individuals would be good representatives of populations of shareholders.
   Your choice of clustering method [2 points], justification [2 points], and how you'd use it solve the problem [2 points]

   Solution: I'd use k-means clustering because this method would produce a partition of the shareholders into populations of similar individuals, and the centroid of each cluster would help me identify a shareholder that would be a good representative of the cluster. I'd run several experiments with k-means varying k and the initial seed to identify a small value for k (as only a handful of representatives will be invited) that produces a good clustering of the shareholders. Then for each cluster in this clustering, I would select the shareholder who is the closest to the cluster's centroid as its representative. [Note that the centroid of a cluster is constructed as the average among all the data instances in the cluster and hence the centroid may not be a data instance. In such case, a data instance close to this centroid needs to be identified.]

Problem IV. Anomaly Detection [30 Points]

Part IV.1: Assume that we are working with a dataset that contains 3,000 data instances. We want to identify data instances that may be anomalies. Let *f(x)* be the anomaly score function that we will use for that purpose. Given a threshold *t*, we say that a data instance *x* is an anomaly if and only if *f(x) > t.*

Assume that we plot below depicts the anomaly scores of the data instances, sorted in increasing order.



1. [5 Points] What would be a natural choice for the value of this threshold *t* based on the plot above? Explain your answer. Mark your chosen threshold value on the y-axis of the plot and label it "t_1". In this case, how many data instances (more or less) would be classified as anomalies?

   Solution:

   There is a clearly defined elbow in the plot corresponding to f(x) = 10. So a natural choice for the threshold would be t = 10. About 200 data instances would be classified as anomalies using this threshold. See plot above.

2. [5 Points] This question is unrelated to question 1 above. Assume that we want to classify 20% of the dataset instances as anomalies. In this case, what threshold value would you pick based on the plot above? Explain your answer. Mark your chosen threshold value on the y-axis of the plot and label it "t_2".

   Solution:

   There are 3,000 data instances in the dataset so 20% would be 600 instances. Looking at the plot, in order to classify 600 instances as anomalies, the threshold value should be around f(x) = 6 or 7.

Part IV.2: This part is unrelated to Part IV.1 above. The following are two different metrics that can be used to evaluate the effectiveness of an anomaly detection method. Below, the terms "detected" and "classified" are used interchangeably.

$$detection\ rate\ = \frac{number\ of\ anomalies\ correctly\ detected\ by\ the\ method}{total\ number\ of\ anomalies\ in\ the\ dataset}$$

$$false\ alarm\ rate\ = \frac{number\ of\ instances\ incorrectly\ classified\ as\ anomalies\ by\ the\ method}{total\ number\ of\ data\ instances\ classified\ as\ anomalies\ by\ the\ method}$$

These metrics can be calculated from the confusion matrix of the detection method. Let's denote by "TP" (True Positive), "TN" (True Negative), "FP" (False Positive), and "FN" (False Negative) the different quadrants of the confusion matrix as depicted below:

anomaly	not anomaly	← classified (= detected) as
TP	FN	anomaly
FP	TN	not anomaly

1. [5 Points] Rewrite the *detection rate* formula above in terms of just TP, TN, FP, and FN.

$$detection\ rate = \quad \frac{TP}{TP + FN}$$

2. [5 Points] Rewrite the *false alarm rate* formula above in terms of just TP, TN, FP, and FN.

$$false\ alarm\ rate = \quad \frac{FP}{TP + FP}$$

3. [5 Points] Write a formula for classification accuracy in terms of just TP, TN, FP, and FN.

$$accuracy = \quad \frac{TP + TN}{TP + FN + FP + TN}$$

4. [5 Points] Argue decisively that when the percentage of anomalies in a dataset is very small, *detection rate* and *false alarm rate* are better measures of the effectiveness of an anomaly detection method than accuracy is.

   Solution:

   If the number of anomalies is very small, a classifier can maximize its accuracy value by just classifying all data instances as non-anomalies (e.g., ZeroR). Since most instances are non-anomalies, the TN value of this classifier will be very high, and so will its accuracy. But this classifier is clearly a really bad anomaly detection method (it doesn't detect any anomalies!). Its detection rate would be 0, and its false alarm rate would be infinite, which more realistically represent the classifier's effectiveness.