**April 30, 2015 Review for COSC 4355 Final Exam**

**1) Association Rule and Sequence Mining**

a) How are rules generated by APRIORI-style association rule mining algorithms? How are frequent itemsets used when creating rules? [3]

For each frequent itemset, create all possible rules that contain the all items in the item set. Then for each rule compute support(all attributes of the rule)/support(attributes that occur on the lefthand side) which is the confidence of that rule; keep those rules whose confidence is above the confidence threshold.

b) Assume the APRIORI algorithm identified the following seven 4-item sets that satisfy a user given support threshold:

acde, acdf, adfg, bcde, bcdf, bcef, cdef.

What initial candidate 5-itemsets are created by the APRIORI algorithm; which of those survive subset pruning? [3]

Acdef ,bcdef,[2]

None;[1]

c) APRIORI has been generalized for mining sequential patterns. How is the APRIORI property defined and used in the context of sequence mining? [5]

Property: see text book [2]

Use: Combine sequences that a frequent and which agree in all elements except the first element of the first sequence, and the last element of the second sequence.

Prune sequences if not all subsequences that can be obtained by removing a single element are frequent. [3

d) Assume the Apriori-style sequence mining algorithm described at pages 429-435 is used and the algorithm generated 3-sequences listed below:

Frequent 3-sequences Candidate Generation Candidates that survived pruning

Candidates that survived pruning:

<(1) (2) (3) (4)>

Candidate Generation:

<(1) (2) (3) (4)> 🡪 survived

<(1 2 3) (4)> 🡪 pruned, (1 3) (4) is infrequent

<(1) (3) (4 5)>🡪 pruned (1) (4 5) is infrequent

<(1 2) (3) (4)>🡪 pruned, (1 2) (4) is infrequent

<(2 3) (4 5)>🡪 pruned, (2) (4 5) is infrequent

<(2) (3) (4 5)>🡪pruned, (2) (4 5) is infrequent

What if the ans are correct, but this part of description isn’t giving?? Do I need to take any points off?? **Give an extra point if explanation is correct and present; otherwise subtract a point; more than 2 errors: 2 points or less!**

<(1) (2) (3)>

<(1 2 3)>

<(1) (2) (4)>

<(1) (3) (4)>

<(1 2) (3)>

<(2 3) (4)>

<(2) (3) (4)>

<(3) (4 5)>

What candidate 4-sequences are generated from this 3-sequence set? Which of the generated 4-sequences survive the pruning step? Use format of Figure 7.6 in the textbook on page 435 to describe your answer! [7]

e) What are the difficulties in using association rule mining for data sets that contain a lot of continuous attributes? [3]

**Association rule mining techniques have been designed for the discrete domain, and in order to apply this approach to datasets with continuous attributes the continuous attributes have to discretized. Discretization might loose valuable information and is quite difficult in practice. For example.** if the intervals are too long, we may lose some patterns because of their lack of confidence. If the intervals are too short, we may lose some patterns because of their lack of support. Finally, rules that have neighboring intervals in their left hand side and right hand side might need to be postprocessed to capture application semantics.

**2) Classification**

a) Compare Nearest Neighbor Classifiers and Support Vector Machines! What are the main differences between the two approaches? [6]

|  |  |
| --- | --- |
| Nearest Neighbor Classifiers | Support Vector Machines |
| Consider only the neighborhood to classify🡪local classifer | Maps objects into higher dimensional space to classify |
| Uses multiple decision boundaries that are edges pf convex polygons that can be computed using Vornoi tessalations | Uses a single, global decision boundary which is a usually high-dimensional |
| Lazy learner, so doesn’t build a model of the training data | Builds the model of the training data, and usually uses a kernel function to map the data in higher dimensions |
| No training cost  | Very expensive to learn |
| Distance function is critical; scale of attributes may impact predictions  | Kernel function is critical |
| Can handle multiple class output directly | Can handle only two class output. To handle multiple class outputs, several binary classifiers need to be learned to separate instance of class from rest of classes |
| Finds local maxima only | Finds global maxima |

b) Assume you use the support vector machine approach for the non-separable case, whose objective function is describes on page 268 of the textbook for a difficult classification problem. What approach would you use to find good values for the parameters *C* and *k* of the objective function? [3]

use n-fold cross-validation on a validation set!

**c)** The following dataset is given (depicted below) with A being a continuous attribute and GINI is used as the evaluation function. What root test would be generated by the decision tree induction algorithm? What is the gain (equation 4.6 page 160 textbook) of the root test you chose? Please justify your answer![6]

Root test: A >=

|  |  |
| --- | --- |
| A | Class |
| 0.22 | 0 |
| 0.22 | 0 |
| 0.31 | 0 |
| 0.33 | 1 |
| 0.33 | 1 |
| 0.41 | 0 |
| 0.41 | 1 |

**Possible slits**

**A≤0.22: (0,2); (3,2)**

**A≤0.31: (0,3); (3,1)**

**A≤0.33: (2,3); (1,1)**

as A≤0.31has a purity of 100%/75% which is much higher than the purity of the other splits, this split will be selected.

d)Most decision tree tools use gain ratio and not GINI or information gain in their decision tree induction algorithm. Why? [3]

Information gain does not consider model complexity in terms of how many additional nodes added to a tree, whereas gain ratio does!

3) Preprocessing

a) What are the objectives of feature subset selection? [3]

* To reduce the dimensionality of data to facilitate creating models and finding patterns for a data mining algorithm🡪finding better solutions
* To remove redundant and irrelevant features from the dataset
* To reduce the time of execution (reduce computation) of a data mining algorithm🡪increase algorithm speed

b) Assume you have to mine association rules for a very large transaction database which contains 9,000,000 transactions. How could sampling be used to speed up association rule mining? Give a sketch of an approach which speeds up association rule mining which uses sampling! [5]

One Solution

1. Run the Association Rule Mining algorithm for a much smaller sample (e.g. 500,000 transactions) with a slightly lower support and confidence threshold[-1 if the same thresholds are use] obtaining a set of association rules R
2. For each rule go through the complete transaction database and compute its support and confidence, and prune rules which violate confidence or support thresholds.
3. Return the surviving rules

c) What does it mean if an attribute is irrelevant for a classification problem? [2]

The attribute is not helping in the classification; the distribution of the attribute does not provide any clues which class should be assigned to a particular example.

4) PageRank [7]

a) Give the equation system that PAGERANK would set up for the webpage structure given below: [4]

$$PR\left(P\_{1}\right)= \left(1-d\right)+d\left[\frac{PR\left(P\_{2}\right)}{2}+ \frac{PR\left(P\_{3}\right)}{2}+ \frac{PR\left(P\_{4}\right)}{1}\right]$$

$$PR\left(P\_{2}\right)= \left(1-d\right)+d\left[\frac{PR\left(P\_{1}\right)}{1}+ \frac{PR\left(P\_{3}\right)}{2}\right]$$

$$PR\left(P\_{3}\right)= \left(1-d\right)$$

$$PR\left(P\_{4}\right)= \left(1-d\right)+d\left[\frac{PR\left(P\_{2}\right)}{2}\right]$$

1 point for each correct equation

b) Which page of the 4 pages do you believe has the highest page rank and why? [2]

P1. Since all other pages have links to it/it has the maximum number of inbound links.

 c) The top 10 DM algorithm paper states “*The computation of page rank values can be done with the Power iteration method*”. Give a short verbal explanation how this method works! [3]

Given the set of linear equations with page ranks as unknowns (like you just found in part a), coefficient matrix as A (or AT) [1], the power iteration method tries to estimate the page ranks iteratively by computing the eigenvector corresponding to the largest eigenvalue of the coefficient matrix, which is 1 in this case [1]. At the end of iterations, the best estimate of the eigenvector is gives the page ranks. There must be an initial estimate for starting the iterative procedure (One way is to have all ones). [1]

5) Outlier Detection

*No review as this topic was discussed just 2 days ago!*