**May 3, 2016 Review for COSC 4355 Final Exam**

**1) Association Rule Mining**

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

Frequent itemsets with their support are stored in a hash-table that has been created by the frequent item set mining phase; next rules are generated as follows:

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—those support values can be quickly retrieved from the hash-table; 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, bcdg, cdef.

What initial candidate 5-itemsets are created by the APRIORI algorithm; which of those survive subset pruning? [4]

acdef ,bcdef, bedeg, bcdfg [3]

None;[1]

c) Assume we have an association rule

 if Drink\_Tea and Drink\_Coffee then Smoke

that has a lift of 2. What does say about the relationship between smoking, and drinking coffee, and drinking tea? Moreover, the support of the above association rule is 1%. What does this mean? [3]

People who drink tea and coffee are twice as likely to smoke

1% of the transactions contain *Drink\_Tea* and *Drink\_Coffee* and *Smoke*.

d) Assume you run APRIORI with a given support threshold on a supermarket transaction database and you receive exactly 2 disjoint 8-item sets. What can be said about the total number of itemsets that are frequent in this case? [4]

Because all subsets of a frequent itemset are frequent, and due to the fact that the 2 sets are disjoint, implying their non-empty subsets are also disjoint—we do not need to worry about double counting—we obtain: 2x(2\*\*8 -1)= 2\*\*9 -2

2) Outlier Detection

a) Give a brief description of how model-based approaches for outlier detection work.

Fit a statistical model M to the data points of the dataset O; next, the density function dM of the model M is used to assess the likelihood of objects o belonging to O; objects with very values for dM(o) or log(dM(o)) are considered to be outliers in O

b) How do k-nearest neighbor-based outlier detection techniques determine the degree to which “*an object in a dataset is believed to be an outlier*”.

For each object the k-nearest neighbor distance—k is a parameter of the method;—to the other objects in the dataset is computed; objects with very high values for that distance are considered to be outliers

Remark: For example, boxplot based outlier detection approaches could be used to decide which low density/high k-NN distance objects are considered to be outliers for the two approaches, we just described.

**3) Classification**

a) The soft margin support vector machine solves the following optimization problem:



What does the first term minimize? Depict all non-zero ξi in the figure below! What is the advantage of the soft margin approach over the linear SVM approach? [5]

All other points

have **ξi** values

of 0!

width

width

The width of the margin with respect to the class1 and class2 hyperplane [1]. Depict [2; 2 errors=0 points]. Can deal with classification problems in which the examples are not linearly separable[2].

b) Referring to the figure above, explain how examples are classified by SVMs! What is the relationship between ξi and example i being classified correctly? [4]

Examples which are above the straight line hyperplane belong to the round class, and example below the line belong to the square class [1.5]. An example will be classified correctly if ξi is less equal to half of the width of the hyperplane; the width w is the distance between the class1 and class2 hyperplane. [2.5].

c) Assume you use an ensemble approach. What properties should base classifiers have to obtain a highly accurate ensemble classifier? [3]

The classifiers should be diverse; that is, they make different errors. [2]

The classifier should be “somewhat” accurate; their accuracy should be above 50%. [1]

d) 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.

e)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!

**f)** 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.

4) 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.

d) What is the goal of feature creation? Give an example of a problem that might benefit from feature creation.

To create new attributes that make it “easier” to find good classification models.

5) 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.