**Solution Sketches April 26 Review for 2018 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) 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.

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.

5) Classification

a) k-NN employ a lazy approach to learning models from training examples. What does this mean? What disadvantages you see with k-NN’s lazy learning approach? [3+1 extra point}

No model is learnt and consequently no training is performed; the data themselves are the model. [1.5]; slow; major computations have to be performed at runtime when determining the class label of an example. [1.5]

b) What can be said about the number and shape of decision boundaries of a k-NN classifier? [3]

The shape consist of lines formed of edges of convex polygons of the Vonoroi tessellation[1]; in general, decision boundaries can be non-convex and can take many forms [1]; there can be many decision boundaries[[1]](#footnote-1).

Problem 4 continued

c) What are the characteristics of a “good” ensemble of base classifiers in ensemble learning? [2]

The classifiers are diverse and make different kind of errors, and have an “okay” but not necessarily high accuracy (above 50%)

d) Why does AdaBoost increase the weights of misclassified examples? [2]

To enhance the chance that they are classified correctly next time, leading to a classifier that classifies those examples correctly and which is therefore different from the previously learnt classifier, leading to a more diverse ensemble.

**6. Hierarchical Clustering [8]**

 A dataset consisting of object A, B, C, D, E with the following distance matrix is given:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| distance | A | B | C | D | E |
| A | 0 | 9 | 2 | 3 | 1 |
| B |  | 0 | 8 | 6 | 5 |
| C |  |  | 0 | 7 | 10 |
| D |  |  |  | 0 | 4 |
| E |  |  |  |  | 0 |

a) Assume single[[2]](#footnote-2) link hierarchical clustering is applied to the dataset? What dendrogram will be returned? [4]

A-E, A-E-C, A-E-C-D. A-E-C-D, A-E-C-D-B

One error: at most 1 point

b) Hierarchical clustering computes dendrograms; what is the dendogram? What is the value of creating dendrograms—what can they be used for? [4]

*A* ***dendrogram*** *is a* [*tree*](https://en.wikipedia.org/wiki/Tree_%28graph_theory%29) *diagram frequently used to illustrate the arrangement of the clusters produced by* [*hierarchical clustering*](https://en.wikipedia.org/wiki/Hierarchical_clustering)*. Edges of the dendrogram represent split/merge relationships between the nodes of the tree which represent clusters[2]*

Dendograms organize dataset hierarchically---identifying homogeneous groups at differerent levels of granularity---which is important in bio-informatics and social sciences as it allows for discovering meaningful categories/classes in data.[2]

1. Its number is only bounded by the number of examples in the training set. [↑](#footnote-ref-1)
2. When assessing the distance between clusters the minimum distance is used. [↑](#footnote-ref-2)