**April 26, 2018 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]

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]

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]

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]

2) Outlier Detection

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

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*”.

3) Preprocessing

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

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]

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

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

4) PageRank [7]

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

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

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}

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

Problem 4 continued

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

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

**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[[1]](#footnote-1) link hierarchical clustering is applied to the dataset? What dendrogram will be returned? [4]

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]*

1. When assessing the distance between clusters the minimum distance is used. [↑](#footnote-ref-1)