**Solution Sketches December 3, 2018 Review for COSC 4355 Final Exam**

**1) Association Rule Mining**

a) What is the task of association rule mining?

Given a support threshold s, a confidence threshold c, and a transaction database T, find association rules r such that: conf(r)≥c and support(r)≥s

b) 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.

c) 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]

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

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) What skills are important to be hired as a Data Scientist?**

(see slides that discuss this topic)

* Should know R and/or Phyton
* Should have sound software development skills
* Should have some sound knowledge of Statistics
* Should have sound knowledge of the different data analysis tasks; e.g. clustering, classification, similarity assessment
* Should be knowledgeable in data visualization
* Data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others.”
* The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill in the next decades."
* But what’s even harder is finding people who have those skills *and* are good at communicating the story behind the data.”

Kind of an essay style question; use your own discretion to grade it!.

5) Classification

Problem 4 continued

a) 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%)

b) 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.

c) Assume you have a classification dataset that does not contain any inconsistent examples: examples that have identical values for all the attributes but belong to a different class. Assuming you use decision tree models what is the maximum training accuracy you can accomplish for such a datasets with no inconstant example? Give reasons for your answer! [3]

Accuracy is 1! [1.5]

Reason: as there are no inconsistent example, example always disagree in at least one attribute; therefore, it is always possible to further split nodes with purity less than 1 until a decision tree with training accuracy 1 is obtained.

d) Assume you have a decision tree that has a training accuracy of 100%. Is such a decision tree always a “good” model to classify unseen data? Give reasons for your answer! [3]

No [1] Should say something about overfitting [2]

e) Ensemble approaches have been quite successful for classification problems, achieving very high accuracies for many challenging dataset. Why do you believe this is the case? [3]

Although the base classifiers might have simple decision boundaries, the ensemble classifier is capable to learn quite complex decision boundaries. [3]

Combining a large number of independent, or at least somewhat dependent classifiers leads to very high accuracies, even if the base classifiers have quite low accuracies [3]

*Kind of an essay question; other answers might deserve credit*

**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]

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