Final Exam Solution Sketches

COSC 4335 *Data Mining*

December 10, 2018

Your Name:

Your student id:

Problem 1 --- Association Rule Mining [9]

Problem 2 --- Hierarchical Clustering [11]

Problem 3 --- Classification Techniques [16]

Problem 4 --- Anomaly/Outlier Detection [12]

Problem 5 --- Data Science [8]

Problem 6 --- Preprocessing [7]

:

**Grade:**



The exam is “open books and notes” but the use of computers is not allowed; you have 105 minutes to complete the exam. The exam will count approx. 20% towards the course grade.

**1) Association Rule and Sequence Mining [9]**

a) Assume we have the following Transaction Database

T1: {A,B,C,D}

T2: {A,C,D,E}

T3: {C,D,E,F}

T4: {B,C,D,E}

T5: {A,D,E}

What is the support and confidence the following association rule:

IF (C and D) THEN E? [3]

Support = 3/5 [1.5]

Confidence=3/4 [1.5]

b) Assume the APRIORI algorithm identified the following five 4-item sets that satisfy a user given support threshold: **abcd, acde, acdf, acdg adfg;** what initial candidate 5-itemsets are created by the APRIORI algorithm; which of those survive subset pruning? [4]

acdef, acdeg, acdfg [3] One error: at most one point!

None survives pruning [1]

c) Why are association rule mining systems interested in finding rules with high support? [2]

Rules with high support are more likely to predict the occurrence of an item based on the occurrences of other items in the transaction accurately; it is hard to learn accurate rules from just a few examples.

**2. Hierarchical Clustering [8]**

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| distance | A | B | C | D | E | F |
| A | 0 | 9 | 8 | 1 | 3 | 11 |
| B |  | 0 | 2 | 6 | 5 | 12 |
| C |  |  | 0 | 7 | 10 | 4 |
| D |  |  |  | 0 | 15 | 13 |
| E |  |  |  |  | 0 | 14 |
| F |  |  |  |  |  | 0 |

a) Assume single[[1]](#footnote-1) link hierarchical clustering is applied to the dataset! What dendrogram will be returned? [7]

A D E B C F

Partially correct (only 1 error): -4

Not drawn dendrogram: -3

b) How does hierarchical clustering differ from more classical clustering algorithms, such as K-Means and DBSCAN? [4]

Hierarchical clustering computes multiple clusterings that form a hierarchy whereas K-means and DBSCAN form a single set of clusters—a single clustering. [4]; it also organizes the objects in the dataset in a hierarchy with groups and subgroups [2]

**3) Classification [16]**

a) Assume you use an ensemble approach. What properties should base classifiers have to obtain a highly accurate ensemble classifier? [2]

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

b) Why does the AdaBoost increase the weight of an example that has been misclassified? How does AdaBoost change the weight of an example that has been misclassified by the previous classifier? Be precise! Give a verbal answer not a formula! [5]

To enhance the chance that they are classified correctly next time, leading to a classifier that classifies those examples correctly and which is therefore makes different errors in comparison to the previously learnt classifier, leading to a more diverse ensemble. [2]

The weight of this example is increased [1.5]; the weight increases in proportional to the accuracy of the previously learnt classifier---that is, if the accuracy of the previous classifier is low the weight increase will be less! [1.5]

c) 2-Layer Neural Networks do not use the Backpropagation Algorithm—why is the case? What does the back-propagated error depend on? [5]

Back propagation is used to calculate the error in the intermediate layers in multilayer neural network. Since, 2-layer neural network don’t have any intermediate layer, it does not use back-propagation. [2]

Back-propagated error depends on the derivative of the activation function for the input value that is fed into the node, the weight of the connection and the error that is back-propagated. [3]

d) Decision trees use tests <attribute> ≥ <value> (e.g. A≥33) for numerical attributes. How does the decision tree induction algorithm determine the value (e.g. 33) of the attribute A that is used in the decision tree node test? [4]

The algorithm exhaustively evaluates the merit of each possible split using purity gain, information gain, GINI…, and then selects the value with the maximum purity,…

among all the possible splits,

4. Outlier Detection

a. What is an outlier? [2]

The set of data points that are considerably different than the remainder of the data.

Other answers might get full or partial credit.

b. Why is it usually more challenging to detect object anomalies than attribute anomalies? [2]

Object anomalies are harder to identify as objects are usually described by multiple attributes. [2]

~~There can be noisy or irrelevant attributes [0.5].~~ The object can be anomalous with respect to only some attributes. [1]

Problem 4 continued

c. Propose either a Model-based Outlier Detection Approach or an outlier detection approach that uses clustering to determine object anomalies in a dataset[[2]](#footnote-2). Describe how you proposed approach detects outliers in some detail! [8]

5. Data Science [8+up to 2 extrapoints]

Data Science has gained a lot of importance in the last 5 years? Why do you believe this is the case? Write an essay answering this question, limiting yourself to 7-14 sentences! Use full sentences in your writing!

**6) Preprocessing [7]**

a) Dimensionality reduction is quite important in many data mining/data analysis projects. Why do you believe this is the case? [3]

* Alleviate the curse of dimensionality [0.5]
* Facilitates finding patterns, as finding patterns is easier in low dimensional spaces [1]
* Reduce the amount of time and memory required by data mining algorithms [1]
* Allow data to be more easily visualized [0.5]
* May reduce noise [0.5]
* May remove irrelevant [0.5] and redundant [0.5] features

At most 3 points!

b) What is the goal of Principal Component Analysis? What is a principal component? How are principal components used for dimensionality reduction? [4]

The goal is to reduce the dimensionality of a dataset by selecting a subset of the principal components that captures a large proportion of the variation in the dataset [2].

Principal component are linearly uncorrelated variables (a linear combinations of the original attributes in the dataset) [1].

Principle components are computed in the order of their contribution of the variation of the dataset and this process is stopped, if a certain percentage of the variation in the dataset is captured by the first k principal components [1]

1. When assessing the distance between clusters the minimum distance is used. [↑](#footnote-ref-1)
2. If you use an approach that is neither model-based nor cluster-based you will not receive any points for your answer! [↑](#footnote-ref-2)