Solution Sketches

Final Exam

COSC 4335 *Data Mining*

May 10, 2016

Your Name:

Your student id:

Problem 1 --- Association Rule Mining [12]

Problem 2 --- PageRank [5]

Problem 3 --- Anomaly/Outlier Detection [12]

Problem 4 --- Classification Techniques [16]

Problem 5 --- Preprocessing [11]

:

**Grade:**



The exam is “open books and notes” but the use of computers is not allowed; you have 100 minutes to complete the exam. The exam will count approx. 20% towards the course grade.

**1) Association Rule and Sequence Mining [12]**

a) What is the anti-monotonicity property, also called APRIORI property, of frequent itemsets? How does the APRIORI algorithm take advantage of this property to create frequent itemsets efficiently? [4]

A⊆B 🡺 support(B)≤support(A) [1.5]

When creating k+1-itemsets from k itemsets [1.5]

For subset pruning of item-set candidates [1]

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

abcde

abcdf

abcef all three correct [3]; one error 1 points

none survives subset pruning; [1] other answers 0

c) Assume an association rule if drink\_wine then smoke has a confidence of 0.7 and a lift of 1. What does this tell you about the association between drinking wine and smoking? Do you believe this rule is useful in predicting the likelihood of a person being a smoker? [4]

The probability of a wine drinker to smoke is 0.7. [1]

The probability of smoking as well as the probability of a wine drinker smoking are both 0.7; P(Smoke|DrinkWine)/P(Smoke)=1 [1.5}

No[0.5]; drinking wine has not impact on smoking; it does not change the likelihood of somebody smoking [1]

2) PageRank [5]

How does the PageRank algorithm determine the importance of webpages—give a brief description how PageRank approaches this task.

.

PageRank assumes the Random Surfer Model: a person follows the weblinks at random, but jumps to a webpage at random with probability 1-d. [1] It defines an equation system that computes the importance of a webpage recursively as a function of the importance of webpages pointing to it [2.5]; the importance of pointing page is divided by the number of outgoing links of the pointing webpage[1]; moreover, the damping factor d plays a role [0.5]. [just giving the equation without much explanation:[2.5]) Next the equation system is initialized with initial webpage importance values and the equation system is updated until it converges, and the pagerank of each page is reported, as the result [1].

**3) Anomaly/Outlier Detection [12]**

a) What is the goal of outlier/anomaly detection? [2]

The goal is to identify objects in a dataset that significantly differ from the other objects in the dataset.

b) Propose and describe an approach *that uses k-nearest neighbor distances for outlier*/anomaly detection! [5]

No answer given!

c) How does the One-Class SVM approach identify outliers? [3]

It tries to match a sphere with center a and radius r to the dataset, minimizing the error of examples outside the sphere, that is measured by their distance to the sphere and maximizing its radius[2.5]. All points outside the sphere will be classified as outliers[1]. The distance of a dataset object o to the sphere center a can be used to order the points in the dataset—objects with the largest distances will the most likely outliers. [1]

d) What observations are considered as outliers by model-based statistical outlier detection approaches? [2]

Observations whose density is very low / observations that are very unlikely with respect to the statistical model that was fitted to the dataset.

**4) Classification and Other [16]**

a) SVMs have been successfully used in conjunction with kernel functions. How does this approach work; why do you believe it has been successful? [4]

Examples in the original attribute space are mapped into a higher dimensional attribute space and a hyperplane are learnt to separate classes in the mapped attribute space [2]. In a higher dimensional space, there are many more hyperplane to separate the two classes, making it more likely to find “better” hyperplanes in this space. Hyperplanes that separate examples in the mapped space correspond to decision boundaries in the original space that are no longer linear; that is, the kernel approach allows finding decision boundaries with quite complex shapes, which is not the case for the linear SVM. [2] 2 extrapoints to students who give both answers!

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



What does ||w|| in the above formula measure—why does this approach minimize ||w||? What is the relationship between ||w|| and ξi; how does changing one parameter affect the other parameter when selecting a hyperplanes? [4]

||w|| is inverse size of the margin [1.5; ”size of the margin” only 0.5 points] which is minimized by the above approach; in other words, it maximizes the size of the margin—the empty area that does not contain any examples to make the approach more tolerant with respect to noise[1]. As ||w|| decreases the margin becomes larger and more points will be on the “wrong side the respective class hyperplane”; that is, ξI increases in this case. [1.5]

c) 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! [3]

The weight of this example is increased [1.5]; the weight increase 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]

d)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? You do not need to justify your answer! [2]

Root test: A >= 0.29 no credit for other answers!

|  |  |
| --- | --- |
| A | Class |
| 0.22 | 0 |
| 0.29 | 0 |
| 0.31 | 1 |
| 0.31 | 0 |
| 0.31 | 1 |
| 0.41 | 1 |
| 0.43 | 1 |
| 0.43 | 1 |

e) Assume you fit a very large decision tree to a dataset that has a training accuracy of 100%. Do you believe this decision tree will be doing well in classifying unseen examples? Give a reason for your answer! [3]

No [1]; this will likely result in overfitting [1] as the decision tree induction algorithm will likely learn noise[1] in the very deep parts of the tree, as tests have to be generated based on just a few examples[1]. at most 3 points.

**5) Preprocessing [11]**

a) What is the purpose of aggregation when preprocessing raw data to obtain a dataset? Why is it, sometimes, desirable to aggregate sets of attributes/ objects into a single attribute/object? [3]

* + Data reduction
		- Reduce the number of attributes or objects
	+ Change of scale
		- Cities aggregated into regions, states, countries, to facilitate knowledge discovery [1]
	+ More “stable” data
		- Aggregated data tends to have less variability; consequently, it is easier to find meaningful patterns.

Other answers might deserve credit!

b) Dimensionality reduction is quite important in many data mining projects. Why do you believe this is the case? [4]

Alleviate the curse of dimensionality [0.5]: When dimensionality increases, data becomes increasingly sparse in the space that it occupies and definitions of density and distance between points, become less meaningful, and the respective algorithms do no longer work well and the only hope is to reduce the dimensionality of the dataset [2.5]

Reduce amount of time and memory required by data mining algorithms [0.5]

Allow data to be more easily visualized [1]

May reduce noise [1]

May reduce irrelevant [0.5] and redundant [0.5] features

At most 4 points!

c) Assume you have to discretize a numerical attribute age that gives the age of a person in a dataset; give a sketch of an approach that transforms the age attribute into an ordinal attribute. [4; up to 2 extra points]

 No answer given!

Final Exam Curving

CL-USER 1 > (setq x (+ 28.5 33.5 37 35.5 41.5 42.5 13 38 47.5 33.5 32.5 32.5 41))

456.5

CL-USER 2 > (/ x 13.0)

35.115384

Exam points to number grade translation (the obtained number is still rounded):

(defun final (x) (+ 77 (\* (- x 35) 1.43)))

Average Number Grade Final Exam: 77.23

CL-USER 35 : 1 > (setq y (+ 68 75 80 78 86 88 46 81 95 75 73 73 86))

1004

CL-USER 36 : 1 > (/ y 13.0)

77.23077