DAMI: Data Mining

Examination

October 21, 2019

Instructions:

You have **four (4)** hours to complete this exam. No textbooks, notes, or calculators are allowed during the exam. The total score that can be obtained in this exam is **100 points** plus a total of **7 extra bonus points**. In order to pass the exam, you should obtain at **least 60 points**. The Exam consists of two parts. Part A is worth **20 points** and Part B is worth **80 points**. For the Exam, you are allowed to use an English to Swedish (or vice versa) dictionary.

Part A

IMPORTANT: Your answers should be written on the provided answer sheet.

This part gives you **20 points** and **5 extra bonus points**. To obtain your bonus points, you can choose five questions for which you have more confidence about your answer to put one extra point. If your answer is correct then you also get the bonus point otherwise you do not get any points for that answer. Please indicate those five questions on the separate answer sheet provided to you on the last page of this exam paper.

Note: there is **no negative grading** for wrong answers!

Part B

You have **four questions** to answer and **one bonus question.** For each question, you are expected to answer all parts. You should provide concise answers. Note that short and precise answers are preferred to lengthy answers.

The mapping to letter grades will tentatively be as follows:

- A: 100 – 93
- B: 92 – 86
- C: 85 – 77
- D: 76 – 68
- E: 67 – 60
- Fx: 59 – 50
- F: 49 – 0

Dictionary: for solving the exam you are allowed to bring an English dictionary with you: English to your language or vice versa.

Part A (20 points)

All questions should be answered by indicating whether statements are correct or not.

(a) You are given the following confusion matrix that describes the results of classifying a test set (the columns represent the predicted values and the rows represent the actual values):

predicted

	classes	orange	banana	apple
	orange	1	0	10
actual	banana	10	1	0
	apple	9	0	0

For each of the following statements, indicate whether it is correct or not:

1: The accuracy is (1+1+0) / (1+0+10+10+1+0+9+0+0).

2: The precision for class 'orange' is 1/ (1+10+9).

3: The recall for class 'banana' is 1.

4: The test data used in our experiment consists of 20 oranges.

5: The classifier we have used misclassifies all 'apples' as 'oranges'.

(b) For each of the following statements about classification, indicate whether it is correct or not:

6: The training error of the 1-Nearest Neighbor classifier is always 0.

7: The VC dimension of a very flexible classifier (e.g., neural network) is higher than the VC dimension of a very rigid classifier (e.g., perceptron).

8: AUPRC summarizes the trade-off between recall and precision of a classifier for different thresholds related to the classifier's parameters.

9: For certain base learners and datasets, the test error of Adaboost may keep decreasing even if the training error becomes zero.

10: One difference between bagging and boosting is that bagging is always performed sequentially, while boosting can also be performed in a parallelized fashion.

(c) For each of the following statements about clustering, indicate whether it is correct or not:

11: k-medians is more suitable and reliable than k-means when the feature space contains discreate or binary values.

12: Purity is a metric that can be used for assessing the quality of clusters obtained for different values of parameter *k* of k-means.

13: k-means and Hierarchical clustering under the Ward distance are not applicable to real-valued feature spaces.

(d) For each of the following statements about frequent itemset mining, indicate whether it is correct or not:

14: Given the set of maximal frequent itemsets of a given transactional dataset along with their frequencies, we can infer all frequent itemsets in the dataset with their corresponding frequencies.

15: According to the Apriori Principle, if an itemset is found to be infrequent then all its subsets (combinations of items contained in the itemset) will also be infrequent.

16: Association rule {*orange*, *banana*}→{*apple*} with a *confidence of 0.3* means that if a customer buys an *orange* and a *banana*, then with 30% probability that customer may also buy an *apple*.

(e) For each of the following statements about deep learning and ranking indicate whether it is correct or not:

17: An auto-encoder refers to a neural network that tries to learn to reproduce its input using a learned encoding.

18: Back propagation refers to the transmission of error through the neural network to allow the weights to be adjusted so that the network can learn.

19: In a scale-free network the mean degree of the nodes is linear to the number of nodes in the network.

20: The ranking produced by Pagerank will be equal to that produced by InDegree, if the underlying graph is undirected.

Part B (80 points)

Question 1 (Frequent itemsets and association rules) **[20 points]**

(a) Consider a transactional database *D*, with A, B, and C being the possible items that can occur in *D*. Suppose that we have mined all *frequent closed* itemsets in *D* with a minimum support count threshold *min_sup = 3*.

These itemsets are:

{A, B, C}	with support count = 3
{C}	with support count = 5, and
{A, C}	with support count = 4.

Using only this information, infer the *remaining frequent itemsets* in *D* and their *support count* values. Provide a brief justification of your answer. [8 points]

(b) We are generally more interested in association rules with high confidence. However, often we are not interested in association rules that have a confidence of 100%. Briefly explain why would such rules be of no interest? Then specifically explain how could association rules with 99% confidence be used for anomaly detection. [4 points]

(c) Assume that the largest frequent itemset that can be obtained by Apriori (for a given support threshold) is of *size k*:

1. How many passes does the Apriori algorithm need in the worst case, before it terminates? [2 points]
2. What is the smallest number of frequent itemsets that will be generated? [2 points]

Note that the i_{th} *pass* of the Apriori algorithm is a complete loop starting from candidate itemset generation, to pruning, support counting, and finally reporting the frequent itemsets of *size i*. We also assume that the *first pass* is completed when the *itemsets of size 1* are reported.

(d) Consider the i_{th} *pass* of the Apriori algorithm, and suppose that a set of *n* frequent *i*-itemsets is found. If a transaction does not contain at least *x* frequent *i*-itemsets out of those found during this pass, then it is removed, since it will not contain a frequent itemset in any future pass.

What is the largest possible value of *x* that we can safely use to guarantee that we will never miss any frequent itemset? [4 points]

Question 2 (Dimensionality Reduction and Clustering) **[20 points]**

(a) Consider the following dataset

Example ID	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5
E1	1	10	100	1000	5
E2	2	20	200	2000	81
E3	3	30	300	3000	33
E4	4	40	400	4000	172
E5	5	50	500	5000	-84

We want to perform dimensionality reduction using *PCA*. What is the *smallest number of principal components* we would need to achieve a reconstruction error of 0? Motivate your answer carefully. [5 points]

(b) In class, we discussed that *k-means* is sensitive to *initialization*. Provide an illustrative example of this claim. Briefly describe how *k-means++* solves this problem. [4 points]

(c) Consider the two extreme cases of *k-means*: (c1) the number of clusters k is set to 1 and (c2) the number of clusters k is set to be equal to the number of data points (in the dataset to be clustered). For each case, explain after how many iterations will *k-means* converge and why? What will be the centroids for each case? [6 points]

(d) Consider a large dataset that contains points in the 2-dimensional space. After running *k-means* 10 times, we realize that the produced clusterings are not exactly the same. This is expected due to the way *k-means* is designed. We would, however, like to come up with a final clustering. Describe a simple approach that would decide on the final clustering of the points given the 10 different clusterings produced by *k-means*. [5 points]

Question 3 (Classification and model evaluation) **[20 points]**

(a) Draw a 2-dimensional real-valued dataset with two classes that can be classified with 100% accuracy by a 1-NN classifier, but not by a linear SVM. Explain briefly why this holds for your drawn dataset. [5 points]

(b) When generating forests of classification trees, it is often recommended that the individual trees should not be pruned to maximize the predictive performance of the forest. Explain why pruning may have a detrimental effect on the forest, even in cases when it improves the predictive performance of the individual trees in the forest. [5 points]

(c) Consider a split criterion for decision trees that favors splits resulting in groups with as evenly distributed classes as possible. What will be the effect on the resulting decision trees compared to using the information gain criterion? How would trees generated with the suggested criterion be expected to perform in terms of accuracy on independent test examples? [5 points]

(d) As we discussed in class, Adaboost uses "weak" learners to build a strong learner. One reason for this is to avoid overfitting. Explain briefly how would the choice of a "strong" base learner damage Adaboost. [5 points]

Question 4 (Advanced Topics) [20 points]

(a) Bob is doing his internship at Facebook and is highly interested in analyzing the friendship graph of Facebook. He takes the full dataset and creates a graph where each node corresponds to a Facebook user. Two nodes are connected by an edge if the two users are Facebook friends. Then, Bob implements and runs Pagerank and HITS on the graph he created. Using the results of the two algorithms, he identifies three sets of influential Facebook users:

a. Set A: the 100 users with the highest Pagerank scores.
b. Set B: the 100 users with the highest hubness score (given by HITS).
c. Set C: the 100 users with the highest authoritativeness score (given by HITS).

Happy about his findings, he runs to Zuckerberg's office, explains him his approach and then shows his results. His boss, Zuckerberg, notices that the three sets differ highly from each other. In addition, he quickly computes the in-degree (number of friendships) of each user and compares the scores to the Pagerank scores. He then turns to Bob and tells him that his approach and results are totally wrong and that he is fired. What did Zuckerberg observe that led him to that decision? [5 points]

(b) After being fired by Facebook, Bob claims to have hacked the company's data and has managed to obtain the complete social graph information of Facebook. Mary, who works for Google, becomes quite interested in this and offers Bob 1 billion SEK for his graph. However, before making the deal she demands to have a look at the data. Bob refuses to give her access to the graph itself, but instead he sends her a file containing a list that shows, for each node in the graph, the number of incoming and outgoing links. After making a simple plot of this list, Mary immediately becomes suspicious and rejects the offer. Based on the above information, why did Mary change her mind when she saw the file that Bob sent her? [5 points]

(c) Consider a trained Artificial Neural Network with **5 input values** and a **single** neuron. The weights of the input edges are given by the following vector: $\mathbf{w} = (1, 1, -1, -1, -1)$. The neuron uses the standard linear summation function, i.e., the function that computes the dot product between the input vector **v** and the weight vector **w**.

Recall that the dot product of two vectors **a** and **b** is the sum of the pairwise products of their corresponding coordinates:

---

according to Wikipedia:

The dot product of two vectors $\mathbf{a} = [a_1, a_2, ..., a_n]$ and $\mathbf{b} = [b_1, b_2, ..., b_n]$ is defined as:

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

The activation function of the neuron is the unit step-wise function:

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$$

A new example **v** = (1, 1, 0, 0, 0) is given to the network for classification.

(i) What is the predicted class label for the example? [2 points]

(ii) Consider a test set with 100 examples, for which the first two features are negative (non-zero), while the remaining features are positive (non-zero). If the true class of all examples is 0, what is the accuracy of the network on the test set? [4 points]

(iii) Consider the same test set as in (ii) but now with 50 examples of class 1 and 50 examples of class 0. What is the accuracy of the network on the test set? [4 points]

Question 5 (Bonus question) **[2 points]**

What is your favorite movie of all times?

Answer sheet for Part A

NOTE! This form (for Part A) must be handed in with the exam.

Room:..... Seat no:.....

You should indicate your answers **by checking (x) the appropriate boxes** below.

You can **choose five** of these questions to put your extra points by indicating them with a **x** in the fourth column (called **Extra**) below.

Question	**Answer**		
	True	**False**	**Extra**
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			