DAMI: Data Mining

Examination

October 26, 2018

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**. 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.

Note than 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	melon	apple
Actual	orange	1	1	3
	melon	2	1	2
	apple	3	1	1

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

1: The accuracy is (1+1+1) / (1+1+3+2+1+2+3+1+1).

2: The precision for class 'orange' is 1 / (1+1+3).

3: The recall for class 'melon' is 1 / (2+1+2).

4: The test data used in our experiment consists of 6 apples.

5: The classifier we have used misclassifies 3 'oranges' as 'apples'.

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

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

7: The Perceptron algorithm does not converge is the examples are not linearly separable.

8: As the flexibility of a classification model increases the variance decreases and the bias increases.

9: Consider a cancer diagnosis classification problem where almost all of the people being diagnosed do not have cancer. The probability of correct classification is the most important metric to optimize.

10: Increasing the dimensionality of the data will always result in lower classification error for any classifier.

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

11: K-medoids is not applicable when the feature space contains real values.

12: Hierarchical clustering methods are typically much slower than the K-means algorithm, especially as the number of data examples increases.

13: K-means is not guaranteed to converge if the distance metric used is not consistent with the "means".

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

14: All maximal itemsets are closed.

15: If an itemset is frequent then all its subsets have a relative support that is less than or equal to the relative support of that itemset.

16: Association rule {*orange*, *banana*}→{*apple*} with a lift of 3 means that if a customer buys and *orange* and a *banana*, then with 3% 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: In a neural network, all neurons should have the same activation function.

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

19: Having multiple Perceptrons can actually solve the XOR problem: this is because each individual Perceptron can partition off a linear part of the space itself, and the Perceptrons can then combine their results.

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

Part B (80 points)

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

(a) While the *Apriori principle* applies to the Apriori algorithm for pruning candidate itemsets given a required minimum support threshold, it does not apply for pruning candidate association rules for a required minimum confidence threshold. **Briefly explain why.**

Recall from the lecture the definition of confidence of rule X→Y:

confidence(X→Y) = support (X, Y) / support (X)

[5 points]

(b) Consider the following association rule found in a database of 100 transactions:

r: {*coffee*, *beer*} → {*cheese*}

1. If *coffee*, *beer*, and *cheese* appear together in 10 transactions, what is the support of this rule?

   [2 points]
2. If 50 transactions contain *cheese*, while 25 transactions contain *coffee* and *beer* together, what is the confidence of this rule?

   [2 points]
3. If we update the transactions so that 10 of them contain *coffee*, *beer*, and *cheese* together, while all other occurrences of items *coffee*, *beer*, and *cheese* (in any combination) are removed, what is the support and confidence of the rule?

   [4 points]
4. Consider the setup of (iii) but now we add 10 new transactions containing together *coffee*, *beer*, and *cheese*. What is the new support and confidence of the rule?

   [4 points]
5. Consider the setup of (iv). What is the lift of this rule?

   Recall from the lecture the definition of lift:

$$\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \times \text{supp}(Y)}$$

   [3 points]

Question 2 (Clustering) **[20 points]**

(a) In class, we discussed that *k-means* is sensitive to *initialization*. Provide an illustrative example of this claim. Briefly describe a solution to this initialization problem. [5 points]

(b) 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? [5 points]

(c) In class, we described the standard *k-means* algorithm that loads all data points together into the memory and then proceeds with findings the clusters given a value for *k*. In practice, however, we expect data to arrive in a stream, such that it is sequentially processed and deleted without storing anything in memory. The advantage of streaming algorithms is that their memory requirement is independent of the stream length. Thus, streaming algorithms are very useful in processing data that cannot fit into the memory.

Describe how one should update the *k-means* algorithm, so that it can handle streaming data input? That is, you should state how should one adapt *k-means* so that the input data examples are never stored in main memory. Discuss potential deficiencies of your algorithm compared to the standard version of *k-means*.

[10 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 C4.5 decision tree. Explain briefly why this holds for your drawn dataset. [5 points]

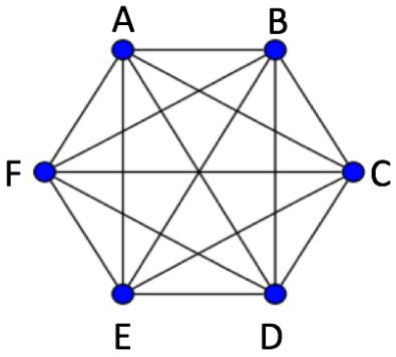
(b) A dataset of 10 examples was partitioned into a training set of 6 examples and a test set of 4 examples. A 1-NN model has an accuracy of 75% on the test set. It was subsequently found that the partitioning had been done incorrectly and that 1 example from the training set had been accidentally duplicated and had overwritten 1 example in the test set. What is the accuracy for the 3 cases that were truly part of the test set? [5 points]

(c) 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]

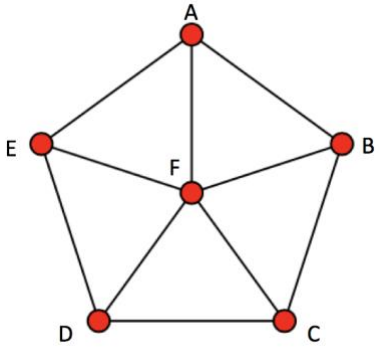
(d) Consider a split criterion for decision trees that favors splits resulting in groups with as evenly distributed classes as possible. What will be the effect on the resulting decision trees compared to using the information gain criterion? How would trees generated with the suggested criterion be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

Question 4 (Advanced Topics) **[20 points]**

(a) Consider the following **two graphs** containing 6 nodes. Note that both graphs are *undirected*, meaning that each edge leads to both directions, e.g., like in friendship graphs. In addition, graph A is *complete*, meaning that each node is directly linked to all other nodes in the graph.



Graph A



Graph B

(i) What is the InDegree score of <u>each node</u> in <u>each graph</u>? [2 points]

(ii) What is the Pagerank score of <u>each node</u> in <u>each graph</u>? [4 points]

(iii) If node F is removed from both graphs (along with all the edges that link to it), what would be the new Pagerank score of <u>each node</u> in <u>each graph</u>? [4 points]

(b) Consider a trained Artificial Neural Network with 4 input values and a single neuron. The weights of the input edges are given by the following vector: **w** = (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, 1, -1) 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 and third features are positive (non-zero), while the second and fourth features are negative (non-zero). If the true class of all examples is -1, 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 questions) [4 points]

(a) What is your favorite TV series?

(b) What is your favorite hangout in Stockholm?

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.

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