Instruktioner för Quiz

Welcome to the exam of Data Mining 1!

You can work on the exam from 8:00 until 13:00 on March 18th.

The exam has 6 questions, each worth 10 points. You pass the exam if you achieve at least 30 (out of 60) points in total. For most students the exam is simply Pass/Fail. For the ones who started the course in a previous semester and who need to receive a grade, you obtain grade 3 if score ≥30 points, grade 4 if you obtain ≥42 points and grade 5 if you obtain ≥51 points.

During the exam you can ask us questions by joining the following zoom room:

https://uu-se.zoom.us/j/6922951831?pwd=VWxzM2t3a1p4ZGdUTTVzUTdWQ1BMZz09

Meeting ID: 692 295 1831
Passcode: 3489409239

Honor Code

You are allowed to use the course material from Studium while you are solving this exam. In particular, you may access the lecture slides, the lecture recordings and you can use the book "Introduction to Data Mining" by Tan, Steinbach, Karpatne and Kumar. You might also use a calculator for numerical computations.

We ask you to obey the following rules ("honor code"):

- While solving this exam, you do not discuss the exam questions or the exam solutions with anyone.
- You do not use any other materials than the ones mentioned above. In particular, you will not perform web searches to find solutions to the questions.

Good luck with the exam!
Florian and Stefan

Fråga 110 poäng

Suppose you obtain a data matrix which contains cars as objects. One of the attributes for each car is "maximum speed (in km/h)". Assume there are 6 different cars and their maximum speeds are given by the vector x = (120,150,200,165,270,180).

(a) Are the attribute values of "maximum speed (in km/h)" nominal, ordinal, ratio or interval? Explain your decision.

(b) Perform min-max rescaling for the vector x. Please provide x_min, x_max and write down the resulting rescaled vector. Round the numbers to two decimal places.

(c) Perform standardization for the vector x. Please provide the mean, the standard deviation and write down the resulting rescaled vector. Round the numbers to two decimal places.

Fråga 210 poäng

Your friend Sasha has a data matrix with 10000 objects and wants to perform a binary classification task. The dataset contains two classes: 9500 objects from class 1 and 500 objects from class 2. Your friend Sasha has randomly created a training dataset with 8000 objects and a test dataset with 2000 objects. Then Sasha built a k-Nearest Neighbor classifier with k=1000 on the training dataset. Sasha achieved an accuracy of 95% on the test dataset and is very happy.

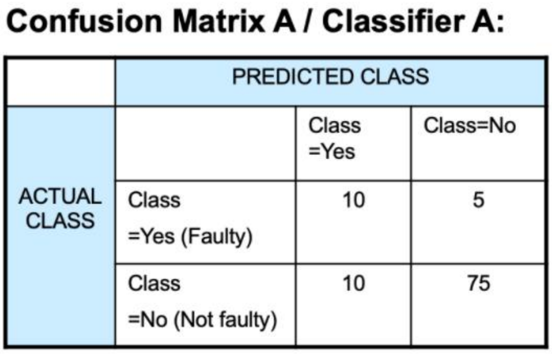
(a) Do you think that achieving an accuracy of 95% is a good result? Explain your answer.

(b) Suppose you have no access to the dataset and also no more information on the dataset than what is stated above. What value of k would you recommend to your friend Sasha?

(c) Now suppose you have access to the whole dataset. How would you proceed to find a suitable value for k?

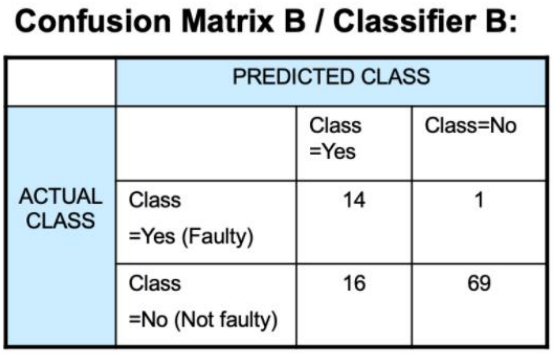
Fråga 310 poäng

Consider the following two confusion matrices.



Confusion Matrix A / Classifier A:

		PREDICTED CLASS	
		Class =Yes	Class=No
ACTUAL CLASS	Class =Yes (Faulty)	10	5
	Class =No (Not faulty)	10	75



Confusion Matrix B / Classifier B:

		PREDICTED CLASS	
		Class =Yes	Class=No
ACTUAL CLASS	Class =Yes (Faulty)	14	1
	Class =No (Not faulty)	16	69

(a) For both confusion matrices, compute the accuracy, the true positive rate and the false positive rate.

(b) Suppose you are a car manufacturer. After you finish assembling a car, you wish to check whether it has a malfunction or not. Now assume that the two confusion matrices above stem from two classifiers (A and B) which predict whether a car has a malfunction. Suppose that if a classifier predicts that a car is faulty, this incurs a reparation cost of $500 (regardless of whether the car has any faults or not). If a classifier predicts that a faulty car has *no* malfunction then this incurs a cost of $2000. If a classifier predicts that a car has no fault and this is correct, then this is without any cost. As an executive of the car company, which classifier would you prefer? Explain your answer.

Fråga 410 poäng

Let X and Y be two itemsets in a transactional database. An itemset measure f is called *monotone* if $f(X) \le f(Y)$ whenever $X \subset Y$, and f is called *anti-monotone* if $f(X) \ge f(Y)$ whenever $X \subset Y$. If a measure is neither monotone nor anti-monotone, it is called *non-monotone*.

(a) Give an example of an anti-monotone property that we have seen in class.

(b) Assume we are given an itemset $X = \{i1,...,ik\}$ with k items. Consider the k rules that can be formed by putting one item on the left-hand side of the rule, and all the remaining items on the right-hand side. Let $\alpha(X)$ be the minimum confidence of all these rules:
$\alpha(X)=\min[c(\{i1\}\rightarrow\{i2,...,ik\}),...,c(\{ik\}\rightarrow\{i1,...,ik-1\})]$,
where c denotes the confidence of a rule. Is the measure α monotone, anti-monotone or non-monotone? Explain why. *Hint*: Use the definition of confidence and simplify the expression for α.
(c) Similarly as in (b), but now we replace the minimum by the maximum. Let $\beta(X)$ be the maximum confidence of all these rules:
$\beta(X)=\max[c(\{i1\}\rightarrow\{i2,...,ik\}),...,c(\{ik\}\rightarrow\{i1,...,ik-1\})]$
Is the measure β monotone, anti-monotone or non-monotone? Explain why.

Fråga 5 10 poäng

Consider the following market transactional database:

Transaction ID	Purchased items
1	{Milk,Diaper}
2	{Beer,Milk,Diaper}
3	{Diaper,Beer,Chocolate}
4	{Beer,Milk,Chocolate}
5	{Diaper,Milk,Beer,Chocolate,Honey}

(a) What is the maximum number of association rules that can be extracted from the database (including rules that have zero support)?

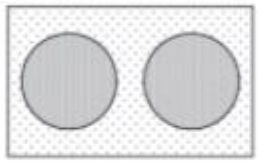
(b) Let minsup = 2. What is the size of the largest frequent itemset?

(c) Write an expression for the number of size-3 itemsets that can be extracted from the database (including itemsets that have zero support). Let minsup = 2. How many of these size-3 itemsets are frequent?

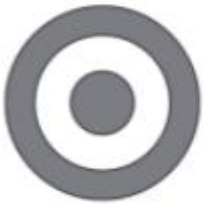
(d) Find an itemset of size 2 or larger that has the largest support.

Fråga 6 10 poäng

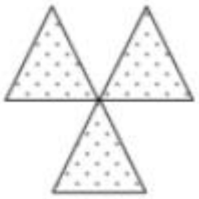
Consider the three (Euclidean) datasets given by the following figure. The darkness of the plots indicates that there are more data points in that area (higher density). For each dataset, we run three clustering algorithms: k-means, single link agglomerative clustering and DBScan. Briefly describe how the clusters (how many, shape, etc.) will look like for each algorithm. For DBScan, describe the resulting clustering in the final stages of the algorithm. For the other algorithms, you can assume the algorithms use ‘good’ parameters and initial conditions.



(a)



(b)



(c)

(a)

K-means (K=2):

Single Link:

DBScan:

(b)

K-means (K=2):

Single Link:

DBScan:

(c)

K-means (K=3):

Single Link:

DBScan: