Exam DM868/DM870/DS804 spring 2022

The exam contains 18 questions summing up to 100 points. The point value of each question is stated at the beginning of its text in parentheses. All questions ask for evaluation of multiple statements with (yes/no) or (true/false) answers. You will gain the full point value of a question if you evaluate all of its statements correctly. Penalties apply to wrong evaluations. Hence, skipping to evaluate a statement is very likely to be more advantageous for you than making a random guess. The points earned from an X-point question are calculated by:

X*(C-W)/A

where

C : Number of correctly evaluated statements,
W: Number of wrongly evaluated statements,
A: Total number of statements in a question.

Note that C+W = A if you evaluate all statements and C+W < A if you skip evaluating at least one statement.

Page 1

(5 points)

Given the items $I = \{A, B, C, D, E, F, G, H, I\}$
and the set of transactions T:

TransID	Items
1	A B C F G H
2	A B D E G H
3	A B E F H I
4	A C F G
5	A C F G H I
6	A D E G H I
7	A I
8	B E F H I
9	B E I
10	D E H
11	G

For the minimum support of 3, we already determined the frequent 3-itemsets with the APRIORI algorithm:

$$L_3 = \{ABH, ACF, ACG, AEH, AFG, AFH, AGH, AHI,$$
$$BEH, BEI, BFH, CFG, DEH, EHI, FHI\}$$

Which of the following 4-itemsets are preliminary candidates in the next step of APRIORI (i.e., after the merging step but before pruning)?

	True	False
ACFG	☑	☐
AEFH	☐	☑
AFHI	☐	☑
AFGH	☑	☐
BEHI	☑	☐
BEFH	☐	☑
BFHI	☐	☑
CEGI	☐	☑

DEHI	○	✓
EFHI	○	✓

(5 points)

For some transaction database we found that the rule $\{A, B, C, D\} \Rightarrow \{E, F, G, H\}$ has a confidence below the confidence threshold.
Which of the following rules will therefore have a confidence below the confidence threshold as well?

	True	False
$\{A\} \Rightarrow \{B, C, D, E, F, G, H\}$	✓	○
$\{A, B, C, D\} \Rightarrow \{E, F\}$	○	✓
$\{A, C, D\} \Rightarrow \{B, E, F\}$	○	✓
$\{A, C\} \Rightarrow \{B, D, E, F, G, H\}$	✓	○
$\{A, D\} \Rightarrow \{B, E, F, G\}$	○	✓
$\{A, D\} \Rightarrow \{B, E, F, G, H\}$	○	✓
$\{A, D\} \Rightarrow \{B, C, E, F, G, H\}$	✓	○
$\{B, E, F\} \Rightarrow \{A, C, D\}$	○	✓
$\{C\} \Rightarrow \{A, B, D, E, F, G, H\}$	✓	○
$\{C, D\} \Rightarrow \{A, B, E, F, G, H\}$	✓	○

(5 points)

We have the following one-dimensional dataset:

ID	Value
A	2
B	4
C	6
D	10
E	14
F	16
G	18

In three attempts, k-means delivered the following three clustering solutions:

$$S_1 = \{A, B, C\}, \{D, E, F, G\}$$
$$S_2 = \{A, B\}, \{C, D\}, \{E, F, G\}$$
$$S_3 = \{A, B, C, D\}, \{E, F, G\}$$

We want to compare the solutions using TD^2. Which of the following statements are correct?

	True	False
S_1 is better than S_2 in terms of TD^2.	○	☑
S_2 is better than S_3 in terms of TD^2.	☑	○
S_1 and S_3 are equally good in terms of TD^2.	○	☑
S_3 is better than S_1 in terms of TD^2.	☑	○

(8 points)

We have the following one-dimensional dataset:

ID	Value
A	2
B	4
C	6
D	10
E	14
F	16
G	18

In three attempts, k-means delivered the following three clustering solutions:

$$S_1 = \{A, B, C\}, \{D, E, F, G\}$$
$$S_2 = \{A, B\}, \{C, D\}, \{E, F, G\}$$
$$S_3 = \{A, B, C, D\}, \{E, F, G\}$$

We want to compare the solutions using simplified Silhouette. Which of the following statements are correct?

	True	False
S_1 is better than S_2 in terms of simplified Silhouette.	☑	☐
S_2 is better than S_3 in terms of simplified Silhouette.	☐	☑
S_1 and S_3 are equally good in terms of simplified Silhouette.	☑	☐
S_3 is better than S_1 in terms of simplified Silhouette.	☐	☑

(4 points)

EM-clustering - which of the following statements are true:

	True	False
EM clustering makes use of Bayes' rule.	☑	☐
EM clustering assumes independence between attributes.	☑	☐
EM clustering is a generalization of *k*-means.	☑	☐
The principle of EM clustering is to find the MAP hypothesis.	☐	☑
Compared to *k*-means, EM has to fit more parameters with the same value of *k*.	☑	☐
Compared to *k*-means, EM works with stricter assumptions regarding the cluster distributions.	☐	☑

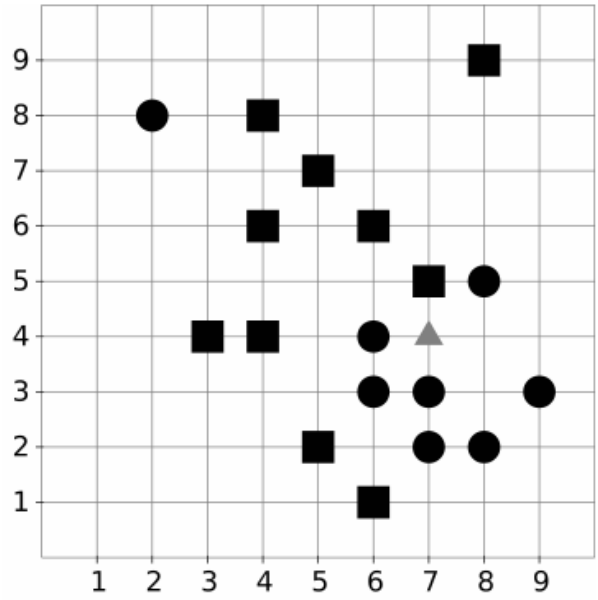
(4 points)

Which of the following statements are correct?

	True	False
The number of parameters to describe a *d*-dimensional normal distribution grows quadratically with the number of dimensions.	☑	☐
In 10-fold cross-validation, each object is used exactly ten times for training.	☐	☑
A non-parametric clustering algorithm is an algorithm that does not require any user-specified parameters.	☐	☑
The larger a decision tree grows, the more accurate will it be on the training data.	☐	☑
k-means, *k*-nearest neighbor classifiers, and EM clustering - all these three methods are parametric learning methods.	☑	☐

(7 points)

We have the following two-dimensional dataset:



Using Manhattan distance, for which choices of k would the kNN classifier classify the query point (triangle) as square?

	True	False
3	○	☑
6	○	☑
10	○	☑
13	☑	○
15	☑	○
16	☑	○
17	☑	○
18	☑	○

(5 points)

Given the true class of 10 test objects and the predictions of some classifier h, which statements are correct w.r.t. to the class specific evaluation measures recall and precision?

o	true class $(f(o))$	prediction $(h(o))$
o_1	A	A
o_2	A	A
o_3	A	C
o_4	A	B
o_5	B	B
o_6	A	A
o_7	B	B
o_8	B	B
o_9	B	C
o_10	C	B
o_11	C	A
o_12	C	C

	True	False
recall for class A > recall for class B	○	☑
precision for class A > precision for class C	☑	○
precision for class A > recall for class A	☑	○
precision for class B > recall for class B	○	☑
precision for class C > recall for class B	○	☑
recall for class A > precision for class C	☑	○

(8 points)

We have a classification problem with two classes "+" and "−", three trained classifiers h_1, h_2, and h_3, with the following probabilities of the classifiers, given the training data D:

$$\Pr(h_1|D) = 0.2$$
$$\Pr(h_2|D) = 0.1$$
$$\Pr(h_3|D) = 0.7$$

For the three test instances o_1, o_2, o_3, the classifiers give the following class probabilities:

$$\begin{aligned}
o_1 : \Pr(+|h_1) &= 0.4 & \Pr(-|h_1) &= 0.6 \\
\Pr(+|h_2) &= 0.4 & \Pr(-|h_2) &= 0.6 \\
\Pr(+|h_3) &= 0.6 & \Pr(-|h_3) &= 0.4 \\
o_2 : \Pr(+|h_1) &= 0.8 & \Pr(-|h_1) &= 0.2 \\
\Pr(+|h_2) &= 0.4 & \Pr(-|h_2) &= 0.6 \\
\Pr(+|h_3) &= 0.5 & \Pr(-|h_3) &= 0.5 \\
o_3 : \Pr(+|h_1) &= 0.6 & \Pr(-|h_1) &= 0.4 \\
\Pr(+|h_2) &= 0.8 & \Pr(-|h_2) &= 0.2 \\
\Pr(+|h_3) &= 0.5 & \Pr(-|h_3) &= 0.5
\end{aligned}$$

We combine the three classifiers to get a Bayes optimal classifier. Which of the following class probabilities will we get from this Bayes optimal classifier?

		True	False
$o_1 : \Pr(+\|\text{Bayes optimal})$	$= 0.54$	☑	☐
$o_1 : \Pr(-\|\text{Bayes optimal})$	$= 0.45$	☐	☑
$o_2 : \Pr(+\|\text{Bayes optimal})$	$= 0.45$	☐	☑
$o_2 : \Pr(-\|\text{Bayes optimal})$	$= 0.45$	☑	☐

o_3 : Pr(+|Bayes optimal)	$= 0.55$	☑	☐
o_3 : Pr(−|Bayes optimal)	$= 0.55$	☐	☑

(8 points)

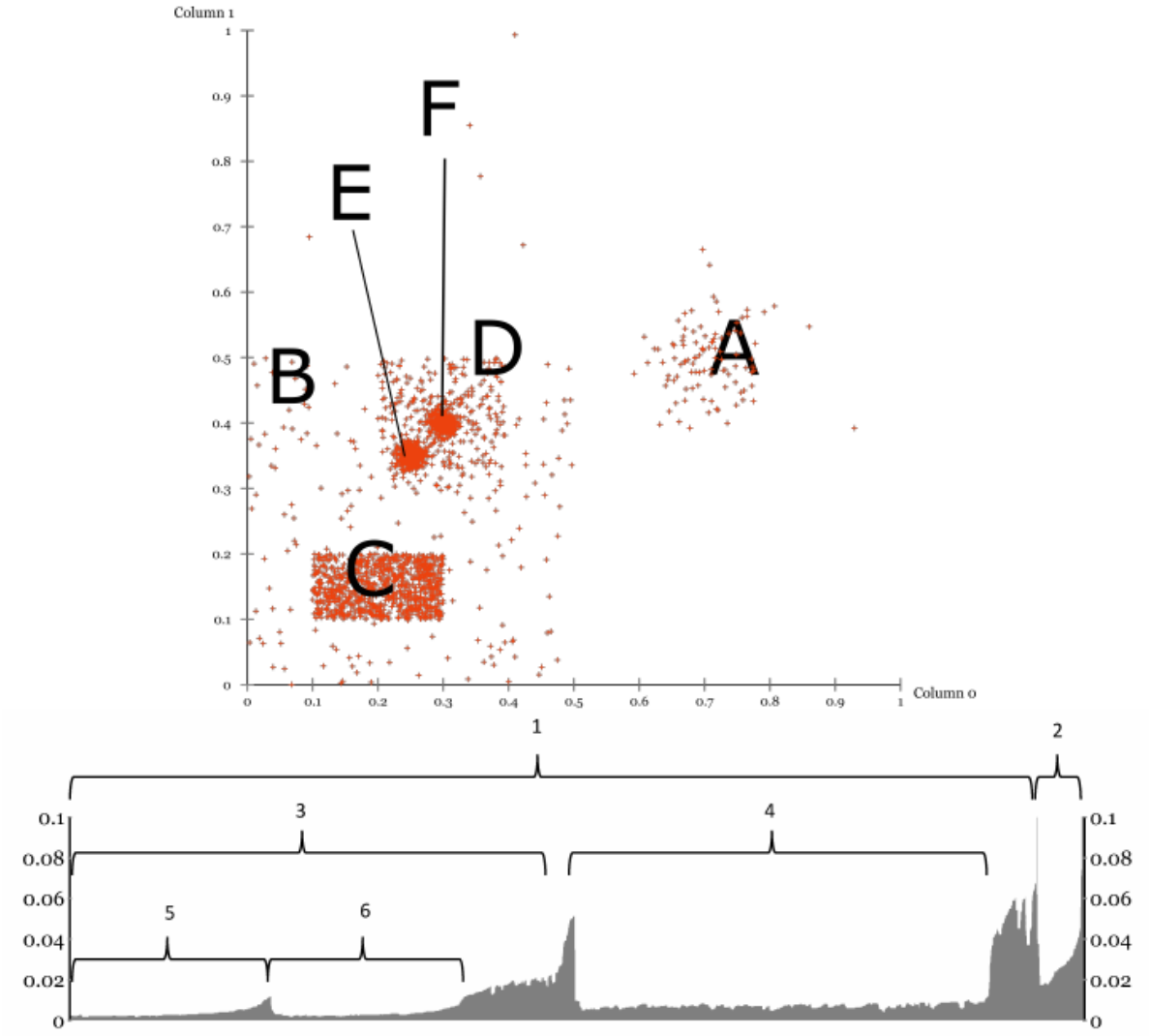
Navigating the society out of the lock-down, health inspector Mortensen shall give his opinion on allowing or forbidding several social activities. He takes orientation on the recommendations of medical, economical, and social science experts Olsen, Frandsen, and Jensen. As these three are not always in agreement, Mortensen wants to use a naïve Bayes classifier to combine their opinions. As training data, he uses the following observations from the past:

activity	**Olsen**	**Frandsen**	**Jensen**	**allow?**
$activity_1$	forbid	forbid	allow	yes
$activity_2$	forbid	forbid	forbid	yes
$activity_3$	forbid	forbid	allow	yes
$activity_4$	allow	forbid	forbid	yes
$activity_5$	allow	forbid	allow	yes
$activity_6$	allow	allow	forbid	yes
$activity_7$	allow	allow	allow	no
$activity_8$	forbid	allow	forbid	no
$activity_9$	forbid	allow	allow	no
$activity_{10}$	allow	forbid	forbid	no

Which of the following activities would the classifier recommend to allow (=yes)?

	Yes	No
activity **Olsen** **Frandsen** **Jensen** $activity_A$ allow allow forbid	☐	☑
activity **Olsen** **Frandsen** **Jensen** $activity_B$ allow forbid forbid	☐	☑
activity **Olsen** **Frandsen** **Jensen** $activity_C$ forbid forbid allow	☐	☑
activity **Olsen** **Frandsen** **Jensen** $activity_D$ forbid allow allow	☐	☑

(5 points)



Given the plotted dataset and the OPTICS plot – which statements are correct?

	True	False
The area 1 in the OPTICS plot relates to cluster B.	☑	☐
The area 2 in the OPTICS plot relates to cluster A.	☑	☐
The area 3 in the OPTICS plot relates to cluster D.	☑	☐
The area 4 in the OPTICS plot relates to cluster E or F.	☐	☑
The area 5 in the OPTICS plot relates to cluster C.	☐	☑

(8 points)

ID	forecast	humidity	wind	play tennis?
1	sunny	high	weak	no
2	sunny	high	strong	no
3	sunny	high	weak	yes
4	sunny	normal	weak	yes
5	sunny	normal	strong	no
6	rainy	high	weak	no
7	rainy	normal	weak	yes
8	rainy	normal	weak	yes
9	rainy	normal	strong	yes
10	rainy	high	strong	no

A decision tree is being trained on the above data set. As root of the tree, the attribute "forecast" was already selected.
Which attributes are selected as test nodes at the next level based on the Gini index?

	True	False
For the branch of forecast=sunny, we test wind.	☑	☐
For the branch of forecast=sunny, we test humidity.	☐	☑
For the branch of forecast=rainy, we test wind.	☐	☑
For the branch of forecast=rainy, we test humidity.	☑	☐

(4 points)

Given the distance measure dist:

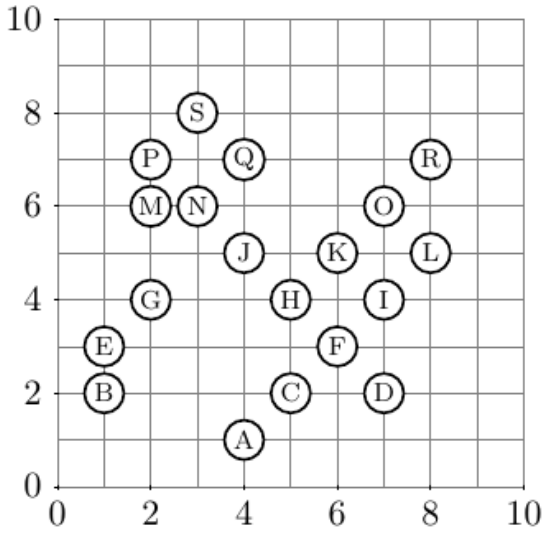
$$\text{dist}(x, y) = \sqrt{(x_1 - y_1, x_2 - y_2) \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix} (x_1 - y_1, x_2 - y_2)^\top}$$

for two-dimensional points and the point $p = (0, 2)$, which of the following points have the same distance as p from the origin $(0, 0)$?

	True	False
$(4, 0)$	☑	○
$(\sqrt{8}, 0)$	○	☑
$(8, 0)$	○	☑
$(\sqrt{2}, 0)$	○	☑
$(\sqrt{2}, \sqrt{3})$	○	☑
$(1, 0)$	○	☑
$(0, -2)$	☑	○
$(-\sqrt{8}, \sqrt{2})$	☑	○

(6 points)

Given the following dataset and Manhattan distance as distance function:

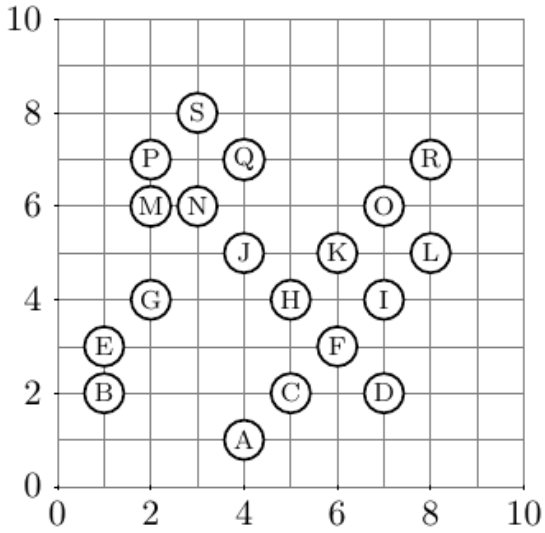


Given the definitions of DBSCAN and counting the query object as a member of its neighborhood: which of the following statements are correct?

	True	False
A is core point with $\varepsilon = 2$ and MinPts $= 4$.	○	☑
C is core point with $\varepsilon = 2$ and MinPts $= 4$.	☑	○
G is core point with $\varepsilon = 2$ and MinPts $= 4$.	○	☑
J is core point with $\varepsilon = 2$ and MinPts $= 4$.	☑	○
M is core point with $\varepsilon = 2$ and MinPts $= 4$.	☑	○
P is core point with $\varepsilon = 1$ and MinPts $= 2$.	☑	○
Q is core point with $\varepsilon = 1$ and MinPts $= 4$.	○	☑
R is core point with $\varepsilon = 1$ and MinPts $= 4$.	○	☑

(6 points)

Given the following dataset and Manhattan distance as distance function:

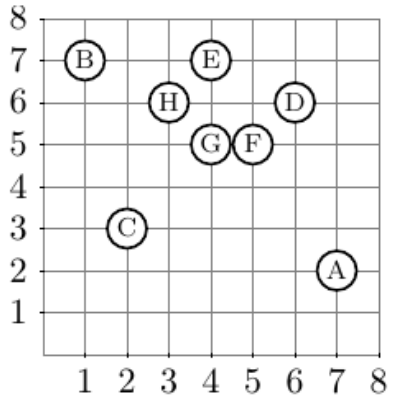


Given the definitions of DBSCAN and counting the query object as a member of its neighborhood: which of the following statements are correct?

	True	False
P is directly density-reachable from N with $\varepsilon = 2$ and MinPts $= 4$.	☑	☐
S is directly density-reachable from M with $\varepsilon = 2$ and MinPts $= 4$.	☐	☑
A and L are density-connected with $\varepsilon = 2$ and MinPts $= 2$.	☑	☐
B and R are density-connected with $\varepsilon = 2$ and MinPts $= 2$.	☑	☐

(4 points)

Given the following dataset and Manhattan distance as distance function:



We have subsets of these points ordered by their outlier score (using the kNN-outlier score, the query point does not count to its own neighborhood here). For a correctly ordered subset of points, the top outlier among the selected points is listed first, later points follow with decreasing outlier score. Which of the following subsets are correctly ordered w.r.t. the given outlier method and parameter?

	True	False
A,C,B w.r.t. kNN ($k = 2$)	☑	☐
C,B,E w.r.t. kNN ($k = 2$)	☑	☐
C,E,B w.r.t. kNN ($k = 2$)	☐	☑
F,E,B w.r.t. kNN ($k = 2$)	☐	☑

(4 points)

In a dataset with ten points $\{A, B, C, D, E, F, G, H, I, J\}$, A and B are labeled outliers.
Five outlier detection methods, $m_1, \ldots, m_5$, deliver the following rankings (from left-to-right: top-rank to bottom-rank):

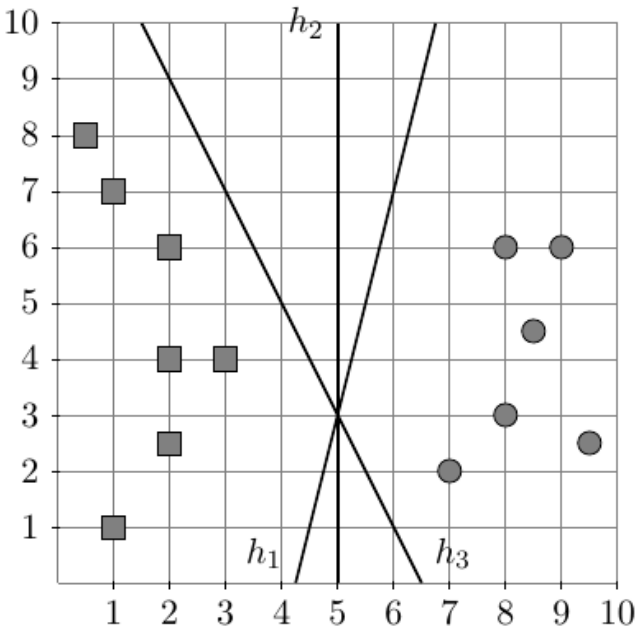
method	ranking
m_1	C,D,A,E,F,B,G,H,I,J
m_2	J,A,D,E,F,G,B,H,I,C
m_3	I,D,A,E,F,G,B,H,C,J
m_4	I,J,E,A,B,F,G,H,C,D
m_5	I,A,E,J,H,C,D,B,F,G

Based on ROC AUC as evaluation measure, which of the following statements are correct?

	True	False
m_1 and m_2 perform equally well.	☑	☐
m_1 is better than m_5.	☑	☐
m_2 is better than m_3.	☑	☐
m_2 is better than m_4.	☐	☑
m_2 is better than m_5.	☑	☐
m_3 is better than m_4.	☐	☑
m_5 is better than m_3.	☐	☑
m_5 is better than m_4.	☐	☑

(4 points)

The lines in this plot indicate the decision boundary derived by some classifier for the given training data:



In the following we conjecture, which classifier might have generated a given decision boundary. Which conjecture is possibly true?

	True	False
h_1: decision tree	○	☑
h_1: perceptron	☑	○
h_1: support vector machine (linear kernel)	☑	○
h_2: decision tree	☑	○
h_2: perceptron	○	☑
h_2: support vector machine (linear kernel)	☑	○
h_3: decision tree	○	☑
h_3: perceptron	☑	○
h_3: support vector machine (linear kernel)	☑	○