(5+4+5+4+4 Points)

Task 1 General Questions

Some of the following subtasks contains multiple choice questions. Each row of those has to be regarded as a closed subtask and yields either 0 or 1 points.

Hint: Don't skip a block. There is no abstention.

(a) Which type of data are the following examples with the given domains? Mark the right answers with a cross. There is exactly one correct answer per line. (5P)

	Categorical	Ordinal	Numerical
Quality { poor, moderate, good, excellent}	☐	☒	☐
Shoe size { 32, 33, . . . , 49 }	☐	☐	☒
Gender { m, f, d }	☒	☐	☐
Frequency { daily, weekly, monthly, yearly }	☐	☒	☐
Weight in kg	☐	☐	☒

(b) For each of the following functions $d_i : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$, which axioms of metric distance functions are fulfilled? Multiple correct answers per line are possible. (4P)

	Symmetry	Identity of Indiscernibles	Triangle Inequality	Neither
$d_1(x, y) = \lvert x^2 - y \rvert$	☐	☐	☐	☒
$d_2(x, y) = 0$	☒	☐	☒	☐
$d_3(x, y) = \sqrt{(x - y)^2}$	☒	☒	☒	☐
$d_4(x, y) = (x - y)^2$	☒	☒	☐	☐

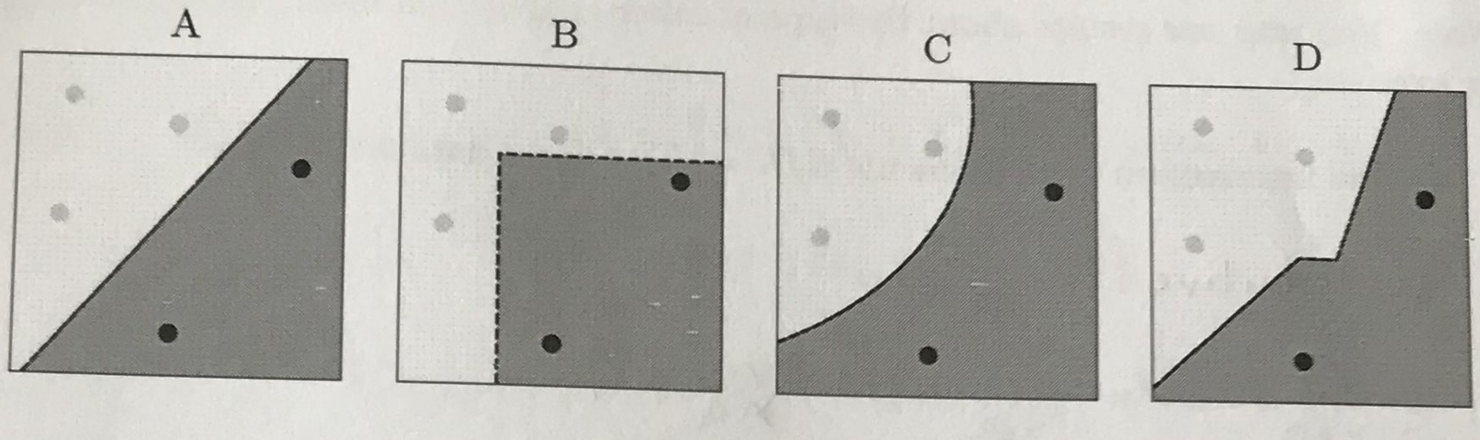
(c) Decide whether the following binnings are equi-width, equi-height or neither of both. "-" denotes the border between two bins, all elements are single-digit numbers. Multiple correct answers per line are possible. (5P)

	Equi-width	Equi-height	Neither
1112-344-556	☒	☐	☐
1123-4556-7778	☐	☒	☐
12-334-56	☒	☐	☐
11-234567-89	☐	☐	☒
1123-4556-7899	☒	☒	☐

(d) Name at least four basic OLAP operations (4P)

i. Roll-up
ii. Roll-down
iii. slice
iv. smaller cube

(e) The following four images show decision surfaces of different classifiers for two classes light grey and dark grey in a 2-dimensional continuous space. For each decision surface, decide which classification model was used. There is exactly one correct answer per line. (4P)



	Naive Bayes Classifier	Linear SVM	Nearest Neighbor Classifier	Decision Tree
A	☐	☒ ✓	☐	☐
B	☐	☐	☐	☒ ✓
C	☒ ✓	☐	☐	☐
D	☐	☐	☒ ✓	☐

4

Task 2 Data Aggregation (2+5+5 Points)

Let D be a database. For the following aggregation measures determine whether they are distributive, algebraic, or holistic. Proof your statement. For algebraic and holistic measures this includes proving exclusion from the former classes.
Note: You may use results about the type of other aggregation functions shown in the lecture or exercise.

(a) The intersection of all elements $i(D) = \bigcap_{x \in D} x$ for a database of sets. (2P)

distributive ✓

$\bigcap_{x \in D} x = (\bigcap_{x \in D_1} x) \cap (\bigcap_{x \in D_2} x)$ with $D = (D_1 \cup D_2)$ ✓

2

(b) The harmonic mean $h(D) = \frac{|D|}{\sum_{x \in D} \frac{1}{x}}$ for a database of positive real numbers $D \subset \mathbb{R}_{>0}$. (5P)

algebraic ✓ why? -2 why not dist? -2

$h(D) = \frac{|D|}{\sum_{x \in D} \frac{1}{x}} = \frac{count(D)}{sum(\frac{1}{x})}$ -2

1

(c) The mode $m(D)$, i.e., the most frequent object, for a database $D \subset U$ over a *finite* universe U. For simplicity, assume that the mode assumes a value *None* if there is no unique most frequent object. (5P)

holistic f

not distributive : $U = \{1,1,3,3,3\} \Rightarrow$ mode $= 3$

$U_1 = \{1,1\}$, $U_2 = \{3,3,3\}$

mode (mode(U_1), mode(U_2)) = None $\neq 3$ ✓ +2

not algebraic since there is no function that maps such cases accordingly

2

Task 3 Privacy

(10 Points)

Given the following table

Key	Quasi-Identifier			Sensitive
Name	Car	Age	City	Income
Aragon	no	40	Berlin	50k
Bilbo	yes	30	Chemnitz	50k
Celegorm	yes	30	Berlin	30k
Denethor	yes	30	Chemnitz	50k
Eowyn	yes	20	Aachen	200k
Faramir	yes	20	Aachen	50k
Gandalf	no	40	Berlin	100k
Halmir	no	40	Berlin	50k
Isildur	yes	30	Berlin	30k
Kili	no	20	Berlin	100k
Legolas	no	30	Berlin	100k

(a) Determine the largest $k \geq 1$ such that the table fulfils k-anonymity. To this end, show the equivalence classes and their sizes. Which equivalence classes contradict the $(k+1)$-anonymity? (8P) 7P.

Equivalence Class			Count
Car	Age	City	
no	40	Berlin	3 ✓
yes	30	Chemnitz	2 ✓
yes	30	Berlin	2 ✓
yes	20	Aachen	2 ✓
no	30	Berlin	1 ✓

⇒ table is 1-anonym ✓

because of (no 30 Berlin) ✓

its not 2-anonym

(b) One shortcoming of k-anonymity is that is does not consider the distribution of sensitive values within the equivalence classes regarding the quasi-identifiers. Give the name of an attack exploiting this. Which privacy notion was proposed to surpass this weakness? (2P)

Background - Knowledge - Attack ✓

L-diversity was introduced ✓

2P.

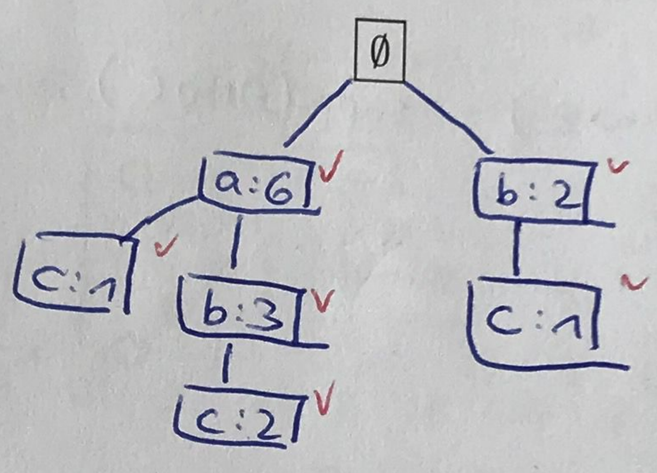
Task 4 Frequent Pattern Mining

(6+7+7 Points)

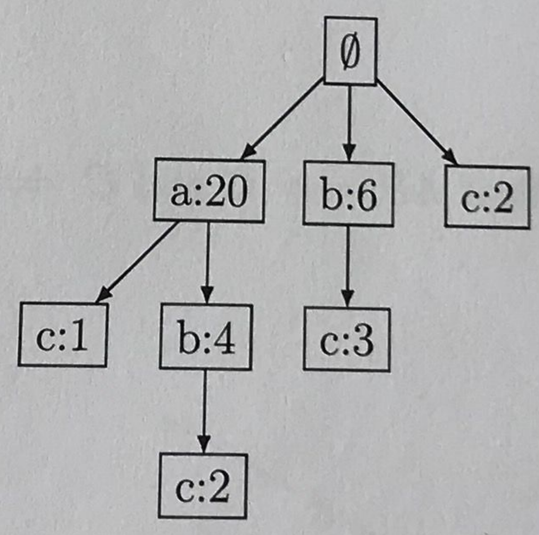
(a) Given is the following database of transactions that has already been preprocessed to be used with FP-Tree, i.e., the items are sorted in decreasing order by frequency.
Hint: One column is one itemset. For better readability the items have been aligned. The (absolute) frequency can be used to check your results.

TID	1	2	3	4	5	6	7	8	Frequency
Items	a	a	a	a	a	a			6
	b	b	b				b	b	5
	c	c		c			c		4

Without setting up the header table, construct the FP-tree starting at the given root node. You do not need to draw the links from the header table, too. (6P) 6?



(b) Given the following FP-Tree (obtained from a different database).



Fill out the conditional pattern base including the counts. (7P) 7P.

Item	Conditional Patterns
a	∅
b	a:4, ∅
c	a:1, ab:2, b:3, ∅

(c) Given is the following table of itemsets with their (relative) support.

Itemset	Support
A	7/10
B	8/10
C	5/10
AB	5/10
AC	3/10
BC	2/10
ABC	2/10

Compute as reduced fractions:

i. The support of $AB \Longrightarrow C$ (2P) 2 P.

$supp(AB \Rightarrow C) = supp(AB \cup C) = \frac{2}{10}$ ✓

ii. The confidence of $AB \Longrightarrow C$ (2P) 2 P.

$conf(AB \Rightarrow C) = \frac{supp(AB \cup C)}{supp(AB)} = \frac{\frac{2}{10}}{\frac{5}{10}} = \frac{2}{5}$ ✓

iii. The correlation/lift of $AB \Longrightarrow C$ and $C \Longrightarrow AB$. (3P) 0 P.

Exam-ID: [redacted]

Task 5 Clustering

15

(16+21+6+3+6+4 Points)

46

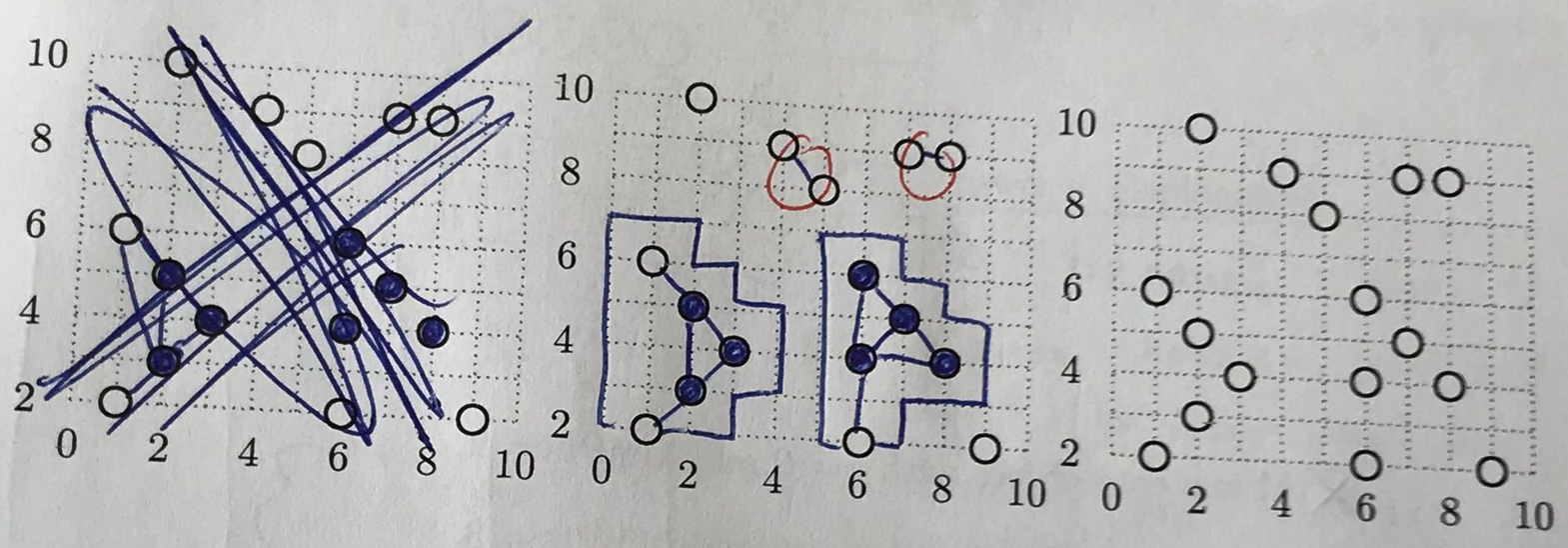
(a) Given the following 2-dimensional dataset, which contains 16 objects, apply DBSCAN with $\epsilon = 2$ and $MinPts = 3$. Use the Manhattan distance L_1 as the basic object distance. MinPts includes the object itself. Complete the following three tasks:

i. Fill out all core object circles. Border objects and outliers have to stay empty. (8P) +8

ii. Connect all pairs of directly density-reachable objects with lines. (6P) +5

iii. Outline all detected clusters with a significant contour. (2P) +2

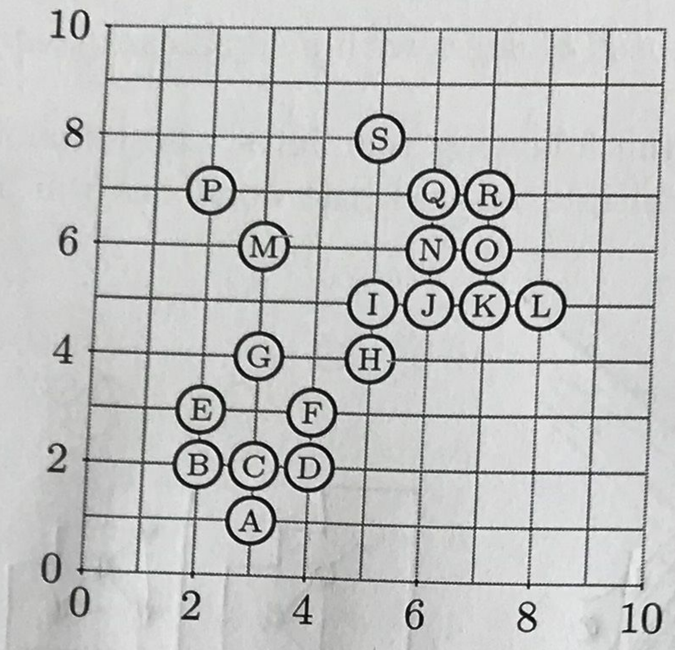
Use exactly one of the following templates, the remaining ones are intended as spare figures. In case of multiple answers, the worst one will be graded.



12

(b) Determine the correct possibilities for the following propositions about DBSCAN. Assume the Manhattan-distance L_1 as the basic object distance. MinPts includes the object itself. Every statement is related to the figure below. Noise objects do not count as clusters.
Each small Latin number indicates a closed subtask. Each block of three statements can contain multiple correct statements and yields either 0, 1 or 3 points.

Hint: Don't skip a block. There is no abstention. (21P)



i. ☒ M is a core object with $\varepsilon = 2$ and MinPts $= 2$.

☒ M is a core object with $\varepsilon = 2$ and MinPts $= 3$.

☐ M is a core object with $\varepsilon = 2$ and MinPts $= 5$.

3

ii. ☐ L is border object with $\varepsilon = 1$ and MinPts $= 5$.

☒ L is border object with $\varepsilon = 2$ and MinPts $= 6$.

☐ L is border object with $\varepsilon = 3$ and MinPts $= 7$.

3

iii. ☒ P and S are both density-reachable from each other with $\varepsilon = 2$ and MinPts $= 3$.

2

☒ P and S belong to the same cluster for $\varepsilon = 4$ and MinPts $= 3$.

☒ P and S are density-connected with $\varepsilon = 2$ and MinPts $= 3$.

1

iv. ☐ Let MinPts $= 2$. DBSCAN finds only one cluster $\Longrightarrow \varepsilon \geq 1$.

☒ Let MinPts $= 2$. DBSCAN finds only one cluster $\Longrightarrow \varepsilon \geq 2$.

☒ Let MinPts $= 2$. DBSCAN finds only one cluster $\Longrightarrow \varepsilon \geq 3$.

0

v. ☐ minPts $= 3$ and $\varepsilon = 1 \Longrightarrow$ A and L are in the same cluster.

☒ minPts $= 3$ and $\varepsilon = 1 \Longrightarrow$ A and L are in different clusters.

☐ minPts $= 3$ and $\varepsilon = 1 \Longrightarrow$ L is a noise object.

3

vi. ☒ minPts = 22 $\Longrightarrow$ P is always a noise object.

☐ minPts = 12 $\Longrightarrow$ P is always a border object.

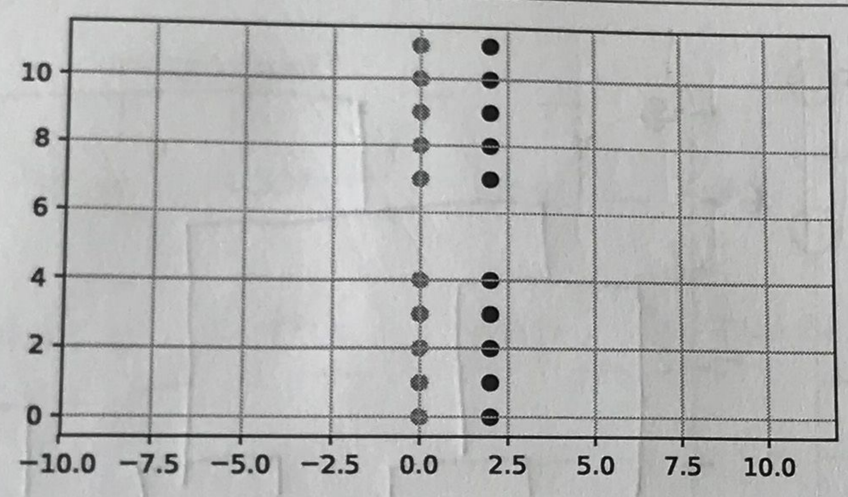
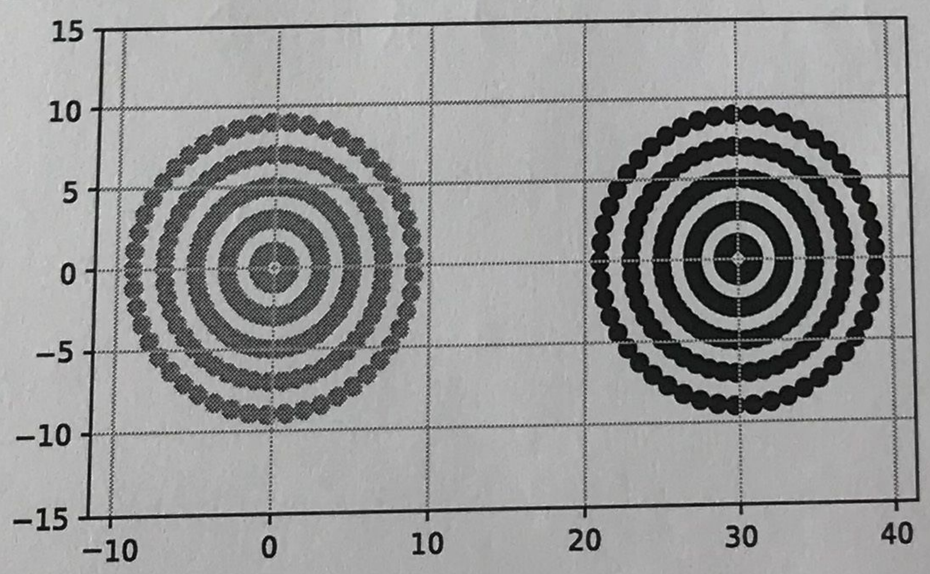
☒ minPts = 2 $\Longrightarrow$ P is always a core object.

vii. ☐ J is never a noise object for any choice of ε and MinPts.

☐ If J is a border object, then $\varepsilon > 1$.

☐ If J is a noise object, then $\varepsilon < 2$

(c) Can the following clustering results be obtained using DBSCAN? If yes, give a value for ϵ and *minPts*. If not, explain why not. (6P)

Result	Answer + Explanation
	no because the distance between grey and black is lesser than the middle gap
	Yes ε = 0,25 minpts = 3
	Yes ε = 5 minpts = 5

3

(d) Give the definitions of single-link, complete-link and average-link distances. (3P)

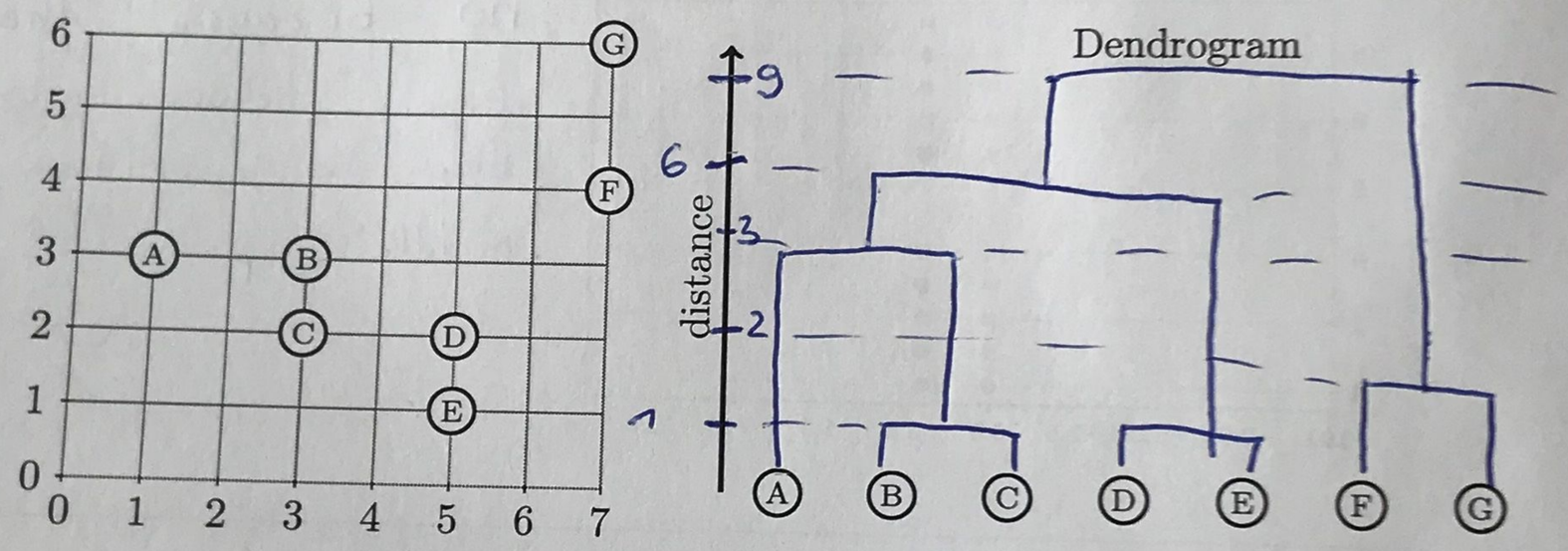
single link $\min_{x \in X, y \in Y} (dist(x,y))$ ✓

complete link $\max_{x \in X, y \in Y} (dist(x,y))$ ✓

average link $\frac{1}{|X| \cdot |Y|} \cdot \sum_{x \in X, y \in Y} dist(x,y)$ ✓

6

(e) Apply agglomerative hierarchical clustering using *complete-link* cluster distances on the 2-dimensional data set from the figure below using the Manhattan distance as the basic object distance. Draw the corresponding binary dendrogram on the right. Annotate each merging step with the corresponding distance value. (6P)



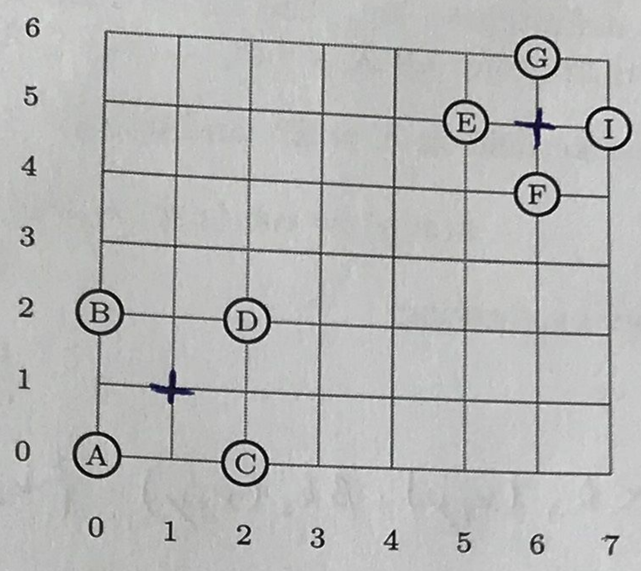
✓

(3P)

Exam-ID: [ ]

4

(f) Using Euclidean distance and a window size of 4, perform the first iteration of the *Mean Shift* algorithm for the two points A and G only. To this end, draw the updated centers into the below-standing figure. Without performing the full algorithm until convergence, decide whether points A and G will end up in the same cluster. Briefly justify your answer. (4P)



+2

They won't end up in the same cluster since the gap between D and E, F is too big

+2

(6+6+8 Points)

Task 6 Kernel

(a) As known from the lecture, a Mercer kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ needs to fulfil

(1) Symmetry, i.e., $\kappa(x, y) = \kappa(y, x)$

(2) Positive semi-definiteness, i.e. the kernel matrix $\kappa(X) := (\kappa(x_i, x_j))_{ij} \in \mathbb{R}^n$ is positive semi-definite for all $X = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$.

Let κ_1, κ_2 be Mercer kernels on $\mathcal{X} = \mathbb{R}^d$, and let $\alpha, \beta \geq 0$ be real constants. Show that

$$\kappa(x, y) := \alpha\kappa_1(x, y) + \beta\kappa_2(x, y)$$

is also Mercer kernel by showing:

i. Symmetry (2P)

2

$k(x,y) = \alpha k_1(x,y) + \beta k_2(x,y) = \beta k_2(x,y) + \alpha k_1(x,y) = k(y,x)$

ii. Positive Semi-Definiteness (4P)

0

(b) Consider the $\kappa_S : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}, (x, y) \mapsto x^T Sy$ and $\phi_L : \mathbb{R}^2 \to \mathbb{R}^2, x \mapsto Lx$ with

$$S = \begin{pmatrix} 4 & 2 \\ 2 & 9 \end{pmatrix} \qquad L = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}$$

6

Show that $\kappa_S(x, y) = \langle \phi_L(x), \phi_L(y) \rangle$ for all $x, y \in \mathbb{R}^2$. (6P)

$k_S(x,y) = \langle \phi_L(x), \phi_L(y) \rangle$

$x^T \begin{pmatrix} 4 & 2 \\ 2 & 9 \end{pmatrix} y = \left\langle \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix} x, \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix} y \right\rangle$

$(4x_1 + 2x_2, 2x_1 + 9x_2) \cdot y = \langle (2x_1 + x_2, 3x_2), (2y_1 + y_2, 3y_2) \rangle$

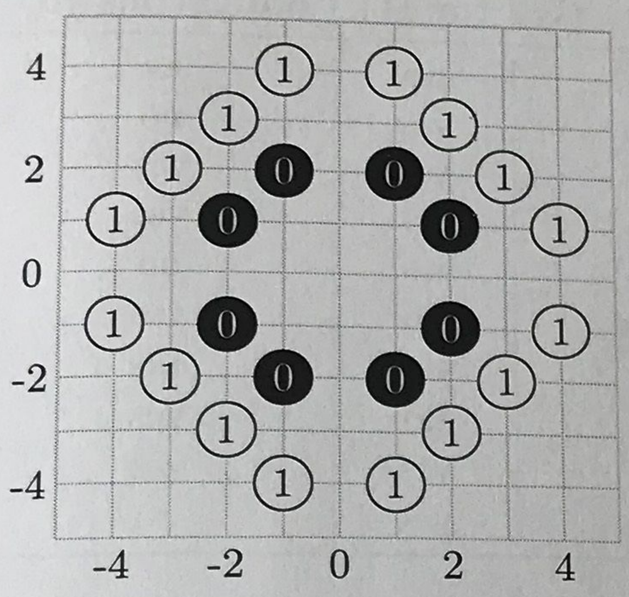
$= \langle (2x_1 + x_2, 3x_2), (2y_1 + y_2, 3y_2) \rangle$

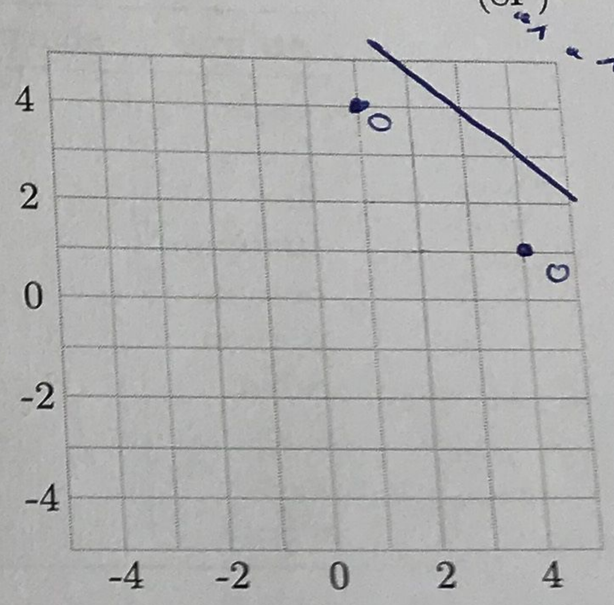
$4x_1y_1 + 2x_2y_1 + 2x_1y_2 + 9x_2y_2 = 4x_1y_1 + 2x_2y_1 + 2x_1y_2 + 9x_2y_2$

(c) For the following 2-dimensional data set with two classes black and white which is not linearly separable, give a (continuous) transformation $\phi : \mathbb{R}^2 \to \mathbb{R}^2$, such that the classes get linearly separable in the transformed space. Draw the transformed data points into the second empty plot, and insert a separating hyperplane.

Hint: In case of overlapping points, it is sufficient to draw a single point at the respective position.

(8P)





8

$\phi : \mathbb{R}^2 \to \mathbb{R}^2 \quad (x, y) \mapsto (x^2, y^2)$

95

(15+6+6+13 Points)

Task 7 Classification

15

(a) Given is the following data set of animals. For each of them it is known whether the animal has (1) sharp teeth (yes/no), (2) at most two legs (yes/no), and (3) fur (yes/no). The target is to predict whether an animal is dangerous.

animal	sharp teeth (s)	≤2 legs (l)	fur (f)	dangerous (d)
eagle	no	yes	no	no
horse	no	no	yes	no
lion	yes	no	yes	yes
mosquito	no	no	no	yes
octopus	no	no	no	no
python	yes	yes	no	yes
rabbit	no	no	yes	no
shark	yes	yes	no	yes
tiger	yes	no	yes	yes
zebra	no	no	yes	no

Using the misclassification error criterion, determine which attribute shall be used for the *first* split. To this end, determine for each attribute the misclassification error after splitting by it. You do *not* have to construct the full tree!

i. Split by *sharp teeth* (5P)

$$Error_{SharpTeeth} = \frac{4}{10} \cdot \left(1 - \frac{4}{4}\right) + \frac{6}{10}\left(1 - \frac{5}{6}\right)$$

$$= \frac{6}{10} \cdot \frac{1}{6} = \underline{\underline{0,1}}$$

✓

⇒ split by sharp teeth

ii. Split by ≤ *2 legs* (5P)

$$Error_{2legs} = \frac{3}{10} \cdot \left(1 - \frac{2}{3}\right) + \frac{7}{10} \cdot \left(1 - \frac{4}{7}\right)$$

$$= \frac{3}{10} \cdot \frac{1}{3} + \frac{7}{10} \cdot \frac{3}{7} = \frac{1}{10} + \frac{3}{10} = 0,4$$

✓

iii. Split by *fur*

$$Error_{fur} = \frac{5}{10} \cdot \left(1 - \frac{3}{5}\right) + \frac{5}{10} \cdot \left(1 - \frac{3}{5}\right) \quad \text{(5P)}$$

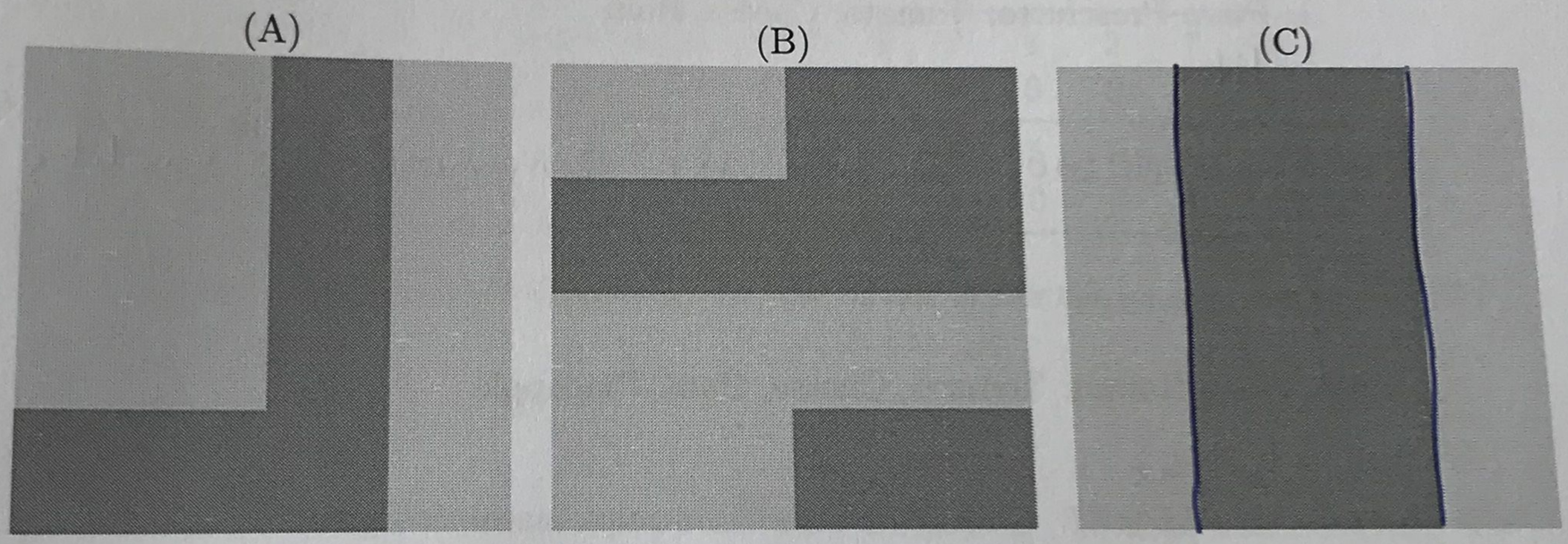
$$= \frac{5}{10} \cdot \frac{2}{5} + \frac{5}{10} \cdot \frac{2}{5}$$

$$= \frac{2}{10} + \frac{2}{10}$$

$$= 0,4$$

3

(b) The following images show decision surfaces for two classes light grey and dark grey in a 2-dimensional continuous space. Can these regions be the result of a decision tree classifier of depth 2, where the root node is counted as depth 1? If yes, draw the splits and attach the depth. If not, briefly justify your answer. (6P)



no

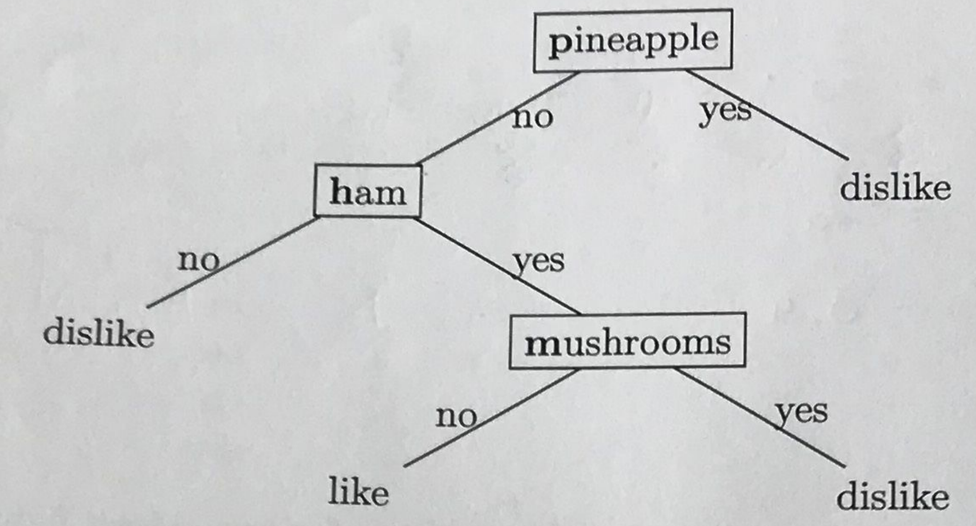
no +1

f

why?

6

(c) Norbert often eats pizza, but does not like every combination of toppings. Hence, from his experience he builds a decision tree to help him deciding whether he likes or dislikes the pizza. This tree is given below.



Now Norbert enters a restaurant and tries to predict whether he likes the following pizzas. Help him by evaluating the given decision tree. To this end, give the decision, as well as the path that led to it.

i. Pizza Prosciutto: Tomato, Cheese, Ham (2P)

like

pineapple : no → ham : yes → mushrooms : no → like

ii. Pizza Hawaii: Tomato, Cheese, Ham, Pineapple (2P)

dislike

pineapple : yes → dislike

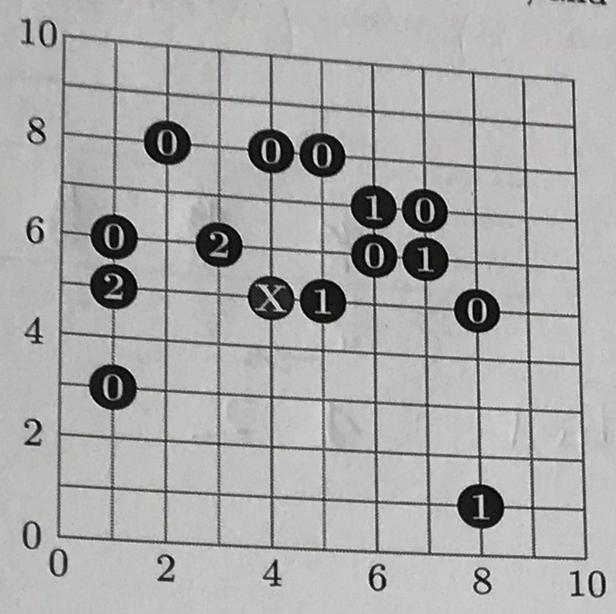
iii. Pizza Capricciosa: Tomato, Cheese, Ham, Mushrooms, Artichokes, Anchovies (2P)

dislike

pineapple : no → ham : yes → mushrooms : yes → dislike

11

(d) Consider the following dataset with class labels "0", "1", and "2".



Consider $X = (4, 5)$ and a k-Nearest Neighbor classifier with Manhattan distance and $k = 5$. You can use the following table sorted by distance to X.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
x_0	5	3	4	6	1	1	8	5	6	7	1	2	7	8
x_1	5	6	8	6	5	6	5	8	7	6	3	8	7	1
distance	1	2	3	3	3	4	4	4	4	4	5	5	5	8
label	1	2	0	0	2	0	0	0	1	1	0	0	0	1

i. Classify X using uniform weighting. In case of ties prefer the smaller number. (1P)

label : 0 +1

ii. Classify X using reciprocal square distance weighting. To this end, compute the weights for each of the points in the kNN set. In case of ties prefer the smaller number. (6P)

i	1	2	3	4	5
label	1	2	0	0	2
weight	1	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

+5

⇒ label : 1 +1

iii. Classify X using inverse a-priori probability weighting. To this end, compute the weights for each of the points in the kNN set. In case of ties prefer the smaller number. (6P)

i	1	2	3	4	5
label	1	2	0	0	2
weight					

$\frac{14}{4}$ $\frac{15}{4}$ $\frac{15}{2}$ $\frac{15}{8}$ $\frac{15}{8}$ $\frac{15}{2}$ f.f. +4

label : 1 2 f