Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Thomas Seidl
Max Berrendorf, Julian Busch

WS 2018/19
25.02.2019

EXAM

KNOWLEDGE DISCOVERY AND DATA MINING I

Exam ID: [redacted]

Notes

- **You have 90 minutes for this exam.**
- **You have to ask for extra sheets. Immediately upon receipt write your exam ID on them. You are not allowed to use your own sheets.**
- **Only use indelible pens in black or blue. In particular you are not allowed to use red or green pens or pencils. Answers written by other means are not graded.**
- **If you desire to cancel the exam, briefly and clearly write down your wish and put your signature under it. If your examination regulation does not allow cancellation, this will lead to failure of the exam.**

Task	Maximum Points	Reached Points	
1. General Questions	14	5,5	F.S.
2. Data Aggregation	7	0	MB
3. Data Privacy	5	4,5	F.S.
4. Frequent Itemset Mining	12	9	MH
5. Clustering	18	16	JB
6. Outlier Detection	8	4	MB
7. Classification	17	17	JB
8. Evaluation	9	6,5	F.S.
Sum:	90	26.5	
Grade:			

Task 1 General Questions (2+2+3+2+3+2 Points)

Some of the following subtasks contains multiple choice questions. Each row of those has to be regarded as a closed subtask, and may have multiple correct statements. Each row yields either 0 or 0.5 points.

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

(a) For each of the following functions $d_i : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, which axioms of metric distance functions are fulfilled? (2P) 0,5P.

Sym: $dist(x,y) = dist(y,x)$

Function	Symmetry	Identity of Indiscernibles	Triangle Inequality	Neither
$d_1(x,y) = \sqrt{(x-y)^2}$	☒	☒	☒	☐
$d_2(x,y) = 1$	☒	☐	☐	☐
$d_3(x,y) = \lvert x-y \rvert + 1$	☒	☐	☐	☐
$d_4(x,y) = x - y$	☐	☒	☐	☐

(b) 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 crosses are possible. (2P) 1,5P.

binning	equi-width	equi-height	neither
11-22-33	☐	☒	☐
12223-4566-777789	☒	☐	☐
111-478-999	☐	☒	☐
11-2344-567	☐	☐	☒

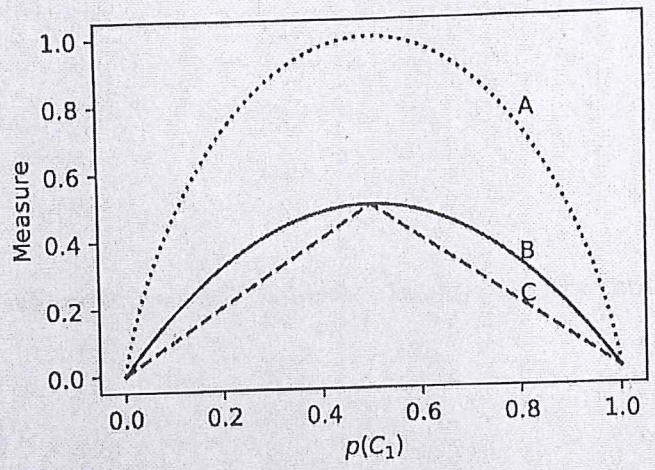
(c) Decide to which type(s) of clustering the following algorithms belong. (3P) 0,5 P

Algorithm	density-based	hierarchical	probabilistic model-based	neither
k-Means	☐	☐	☐	☐
OPTICS	☒	☐	☐	☐
Apriori	☐	☐	☐	☐
DBSCAN	☐	☐	☐	☐
Mean-Shift	☐	☐	☐	☐
Expectation Maximation	☒	☐	☒	☐

(d) Name the four ICES criteria for filter quality (2P) 1 P.

i. Indexable ✓
ii. Cee /
iii. E /
iv. Selectable ✓

(e) Assume a binary classification problem with classes C_1, and C_2. To build a decision tree, several different attribute selection criteria may be used. Given the following plot, for each line give the name of the criterion. (3P) 2 P.



A) Information Gain (wit Entropy) ✓
B) Gini Index ✓
C)

(f) Map the following algorithms to the three main tasks of process mining. (2P) 0 P.

Algorithm	Process Discovery	Conformance Checking	Process Enhancement	Neither
α-Miner	☐	☐	☐	☐
k-Medoid	☐	☐	☐	☐
Token-Replay	☐	☐	☐	☐
AdaBoost	☐	☐	☐	☐

Task 2 Data Aggregation (1+3+3 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 union of all elements $u(D) = \bigcup_{x \in D} x$ for a database of sets. (1P)

(b) The mid-range $m(D) = (max(D) - min(D))/2$ for a database of real numbers $D \subset \mathbb{R}$. (3P)

(c) The geometric mean $g(D) = \left(\prod_{x \in D} x\right)^{1/|D|}$ for a database of real numbers $D \subset \mathbb{R}$. (3P)

(4+1 Points)

Task 3 Data Privacy

Given the following table

Key Name	Quasi-Identifier Semester	Age	Course	Sensitive Grade
Alice	1	20	Astronomy	3
Bob	1	20	Astronomy	1
Clara	2	20	Biology	2
Dave	2	21	Biology	2
Ellen	1	21	Chemistry	3
Felipe	1	21	Chemistry	3
Gwen	1	21	Biology	4
Henry	1	21	Biology	4
Irene	2	22	Biology	3
Jose	2	22	Biology	3
Kathleen	2	22	Biology	2

A = Astronomy
B = Biology
C = Chemistry

(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? (4P)

Equivalence Class (Semester, Age, Course)	Count
(1, 20, A)	2
(2, 20, B)	1
(2, 21, B)	1
(1, 21, C)	2
(1, 21, B)	2
(2, 22, B)	3

k Anonymity: ✓1-anonym

nicht (k+1)-anonym wegen Äquivalenzklasse (2, 20, B) oder (2, 21, B) beide count = 1 ✓

4 P.

(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? (1P)

Background-Knowledge-Attack ✓

↳ Suppression

privacy notion: l-diversity

0,5 P.

Task 4 Frequent Itemset Mining (8+2+2 Points)

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

TID	Items
1	A EFGH
2	A EFG
3	H
4	BC E
5	ABC EF H
6	A EF H
7	BC H
8	ABC EF
9	FG
10	ABC F
11	BC H

Note: The items are aligned to improve readability.

(a) For a minimal support of $minSup = 3$, the frequent itemsets of length 3 have already been computed:

$$L_3 = \{ABC, ABF, ACF, AEF, AEH, AFH, BCE, BCF, BCH, EFH\}$$

Construct all candidates of length 4 using the Apriori Algorithm. If a generated candidate is discarded, the reason has to be given. (8P)

L4 count threshold

ABCF ✓ 3 ABCF

AEFH ✓ 3 AEFH

BCEF ✓ 2 –

BCEH ✓ 1 –

BCFH ✓ 1 –

L3
ABC
ABF
ACF
AEF
AEH
AFH
BCE
BCF
BCH
EFH

discard before count because Subsets (CEF, CEH, CFH) not frequent

(b) Determine all frequent itemsets of length ≥ 4 for $minSup = 3$ using the Apriori Algorithm. Also provide the according frequencies. (2P)

2

(1) L4 Count # L5 L6

(2) ABCF 3 ✓ ABCEFH #1

(3) AEFH 3 ✓

(4) BCEF 2 BCEF prune wegen (4)(5)

(5) BCEH 1

(6) BCFH 1

(c) Calculate the confidence of the association rule $\{A, C\} \rightarrow \{B, F\}$. (2P)

2

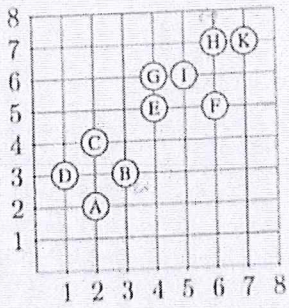
$$Conf(\{AC\} \rightarrow \{B,F\}) = \frac{supp(x \cup y)}{supp(x)}$$

$$= \frac{3 \checkmark}{3 \checkmark} = 1$$

16

Task 5 Clustering (3+2+2+3+3+3+2 Points)

(a) Given are the following points (3P)

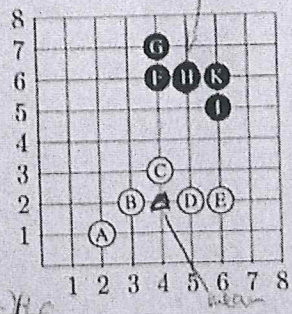


For k-Means, let $k = 2$, and the initial cluster centres are $C_1 = B$, and $C_2 = H$. Specify the cluster assignments of the initial iteration.

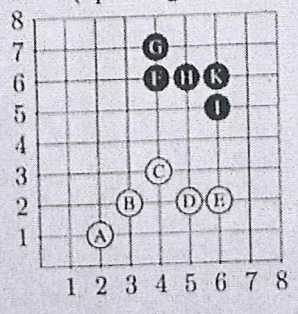
Cluster C_1	A B C D G mit $C_1 = B$
Cluster C_2	H K I F G

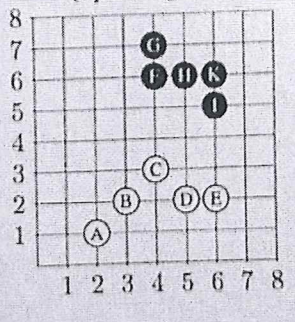
(b) The next step of k-Means would be the re-assignment. Given are the following new points with their assignment to clusters (black and white). Again using k-Means, draw the updated cluster means into the plot. (2P)

(Spare Figure)

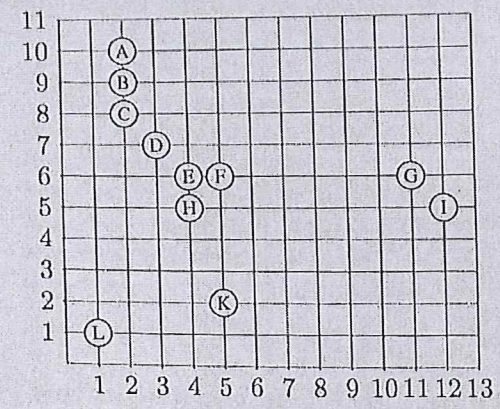


(Spare Figure)





(c) Given the following data points (2P)



For k-Means, choose a value for k, and give k initial centroids such that

i. After the initial assignment, every cluster is non-empty.

ii. After updating the means, and computing the next assignment, at least one cluster gets empty.

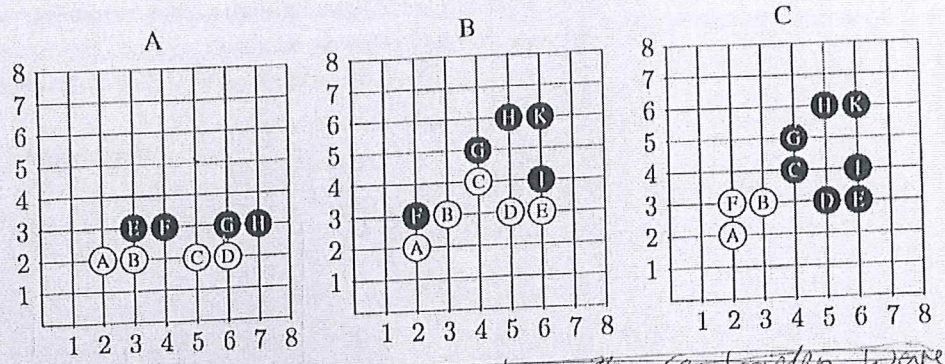
$k = 3 + 1$

Cent1 ( )

Cent2 ( )

Cent3 ( )

(d) Can the following partitions into two classes, black and white, be a final result of k-Means? If not, briefly justify your answer. (3P)



A ~~nein nach Update der Centroiden wäre~~ konvergiert nicht so, (???vermutlich d wäre dann schwarz???)

B nein, weil F hier ~~ein outlier~~ eigentlich als weiß klassifiziert sein müsste, wenn das das finale Ergebnis

C: Ja

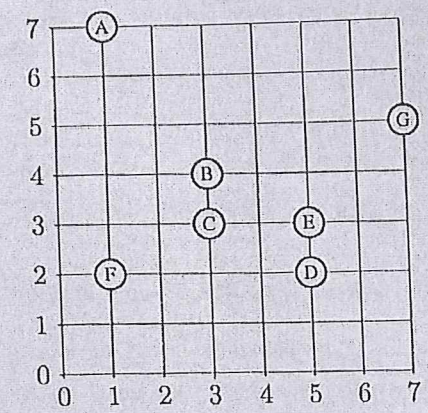
(e) Give the definitions of single-link, complete-link and average-link distances. (3P)

single-link: minimale Distanz zwischen Punkten/Clustern

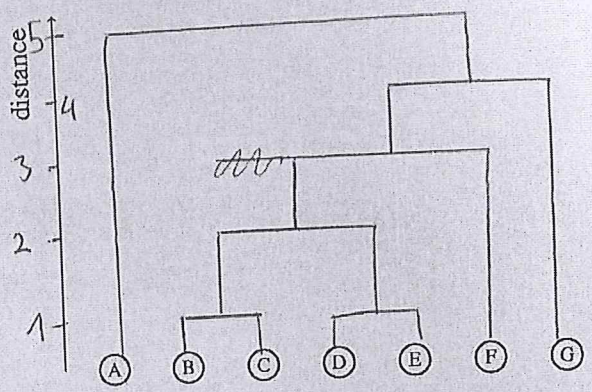
complete-link: maximale Distanz zwischen Punkten/Clustern

Average-Link: Durchschnittliche Distanz aller Punkte in Cluster1 zu allen Punkten in Cluster2

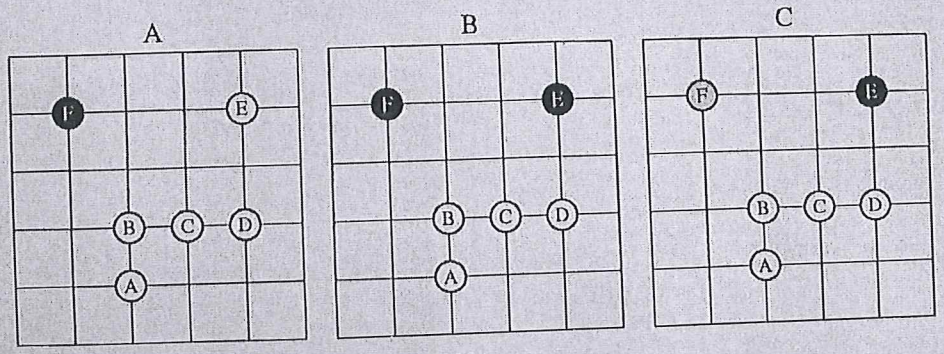
(f) Apply agglomerative hierarchical clustering using single-link cluster distances on the 2-dimensional data set from the figure below using the Manhattan distance as the basic object distance. Draw the corresponding binary dendrogram on the right. Augment each merging step with the corresponding distance value. (3P)



Dendrogram



(g) Can the following clustering results be obtained using agglomerative hierarchical clustering with single-link and Manhattan distance as ground measure? If not, justify your answer. The class labels are given through the three colours: white, grey, and black. (2P)



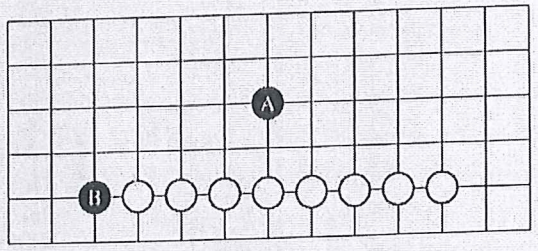
A: Ja

B Nein: E würde ~~dbt~~ zu weiß klassifiziert werden, weil da die dist(E,D) = 2 aber dist(E,F) = 3

C Ja

Task 6 Outlier Detection (2+1+1+4 Points)

(a) Consider the following dataset (only white points). Without computation, determine whether A or B, when added, would be regarded as a stronger outlier regarding the angle-based outlier score ABOD. Briefly explain your decision. (2P)



A als stärkerer outlier, weil B vom Winkel aus ähnlicher zu den weißen Punkten

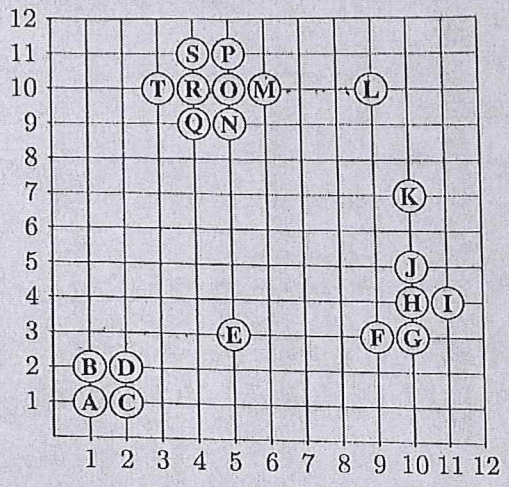
B stärkerer outlier, hat zu allen den gleichen Winkel, A hat eine höhere Varianz

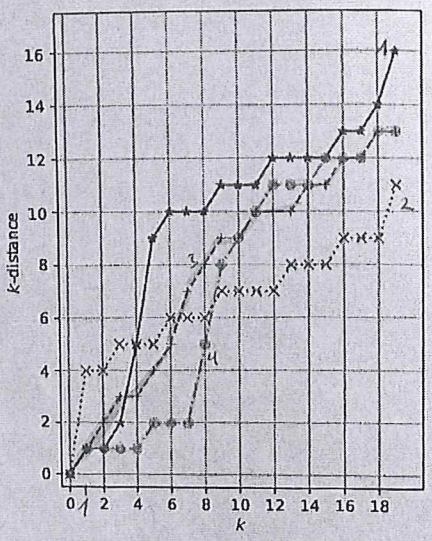
(b) Given a database $D \subseteq \mathcal{U}$ ($|D| < \infty$), a distance measure $dist : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$, a point $p \in D$, and an integer $k \in \mathbb{N}$, let $nndist(D, p, k)$ be the k-distance of p in D, i.e. the distance to the k-nearest neighbour in D. Count p itself as the 0-nearest neighbour. Let τ be the threshold used for determining whether a point is a kNN-based outlier, and let $D' \subseteq \mathcal{U}$ be another database ($|D'| < \infty$). Prove the following statements:

i. If $p \in D \cap D'$ is an outlier in D, it is also an outlier in $D \setminus D'$. (1P)

ii. If p is not an outlier in D, it is also not an outlier in $D \cup D'$. (1P)

(c) Given is the following data set with a plot of k-distance using the Manhattan distance. The point itself is counted as 0-th nearest neighbour. (4P)





Identify which data points correspond to the four lines, and justify your answer.

Line	Point	Justification
Solid 1	C/ B✓	findet Nachbarn A & D mit Dist = 1, 3. Nachbarn (B) dist = 2 (E) (A,D) 4. Nachbarn (E) dist = 5 ✓
Dotted 2	E	findet den ersten Nachbarn (D oder ?) mit Distanz 4, auch zweiter Nachbar gleiche Distanz
Dot-Dashed 4 1	R✓	1-4 Nachbarn dist 1 5-7 -"- dist 2 8. Nachbar dist 5 (L)
~~Dot~~-Dashed 3 4	? ✓	erster Nachbar dist=1 (G) H: dist 2 J, I: dist 3 nicht J weil ~~nah~~ sonst Nachbarn 2&3 gleiche dist, aber es haben 3&4 gleiche dist ✓

4

17

Task 7 Classification

(5+2+2+2+4+2 Points)

Norbert relies on his friends Harold and Gretchen for book recommendations. Since their opinions on books differ frequently, Norbert decides to train a Naïve Bayes Classifier on the combinations of recommendations he has received so far. He has collected the following training dataset:

Book	Harold	Gretchen	Read?
1	r	d	yes
2	r	r	yes
3	d	r	yes
4	r	r	yes
5	d	r	no
6	d	d	no

where r = *recommend* and d = *don't recommend*.

(a) Determine all probabilities as reduced fractions required for classification of the class variable *Read?* given input variables *Harold* and *Gretchen* with a Naïve Bayes classifier. Remember to not only provide the values, but to also name all probabilities correctly. (5P)

$P(read) = \frac{4}{6} = \frac{2}{3}$ ✓ $P(\neg read) = \frac{2}{6} = \frac{1}{3}$ ✓

	Harold		Gretchen	
	r	d	r	d
read	3/4	1/4	3/4	1/4
¬read	0	2/2 = 1	1/2	1/2

	Harold		Gretchen	
	r	d	r	d
read	P(read | r)	P(read | d)	P(read | r)	P(read | d)
¬read	P(¬read | r)	P(¬read | r)	P(¬read | d)	P(¬read | d)

(b) At lunch, Norbert asks his friends for recommendations regarding a new book 7. Apply the classifier trained in the previous task to determine whether Norbert should read book 7. Provide all necessary computation steps. (2P)

Book	Harold	Gretchen	Read?
7	d	r	?

(...) brauchen wir nicht berechnen, weil beide Brüche durch gleichen Rest geteilt → Ranking gleich, also vergleichen

$P(\text{read} \mid \text{Harold}=d, \text{Gretchen}=r)$

$= \frac{P(\text{read} \mid \text{Harold}=d) \cdot P(\text{read} \mid \text{Gretchen}=r) \cdot P(\text{read})}{P(\text{Harold}=d, \text{Gretchen}=r)}$

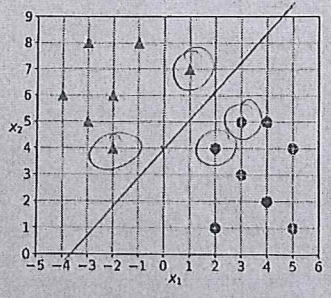
$= \frac{\frac{1}{4} \cdot \frac{3}{4} \cdot \frac{2}{3}}{(\ldots)} = \frac{\frac{3}{24}}{(\ldots)} = \frac{1}{8}$ ✓

$P(\neg\text{read} \mid \text{Harold}=d, \text{Gretchen}=r)$

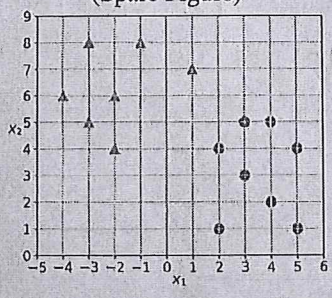
$= \frac{1 \cdot \frac{1}{2} \cdot \frac{1}{3}}{(\ldots)} = \frac{\frac{1}{6}}{(\ldots)}$ ✓

$\Rightarrow \neg$read, weil das größere Rest ist ✓

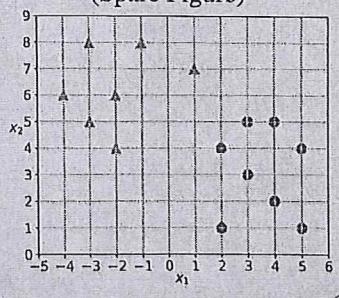
(c) Consider the following dataset consisting of two classes of points in $\mathbb{R}^2$. The *circle* class has label -1, the *triangle* class has label 1. Draw the maximum margin hyperplane inside the figure. No calculations are needed here. (2P)



(Spare Figure)



(Spare Figure)



(d) Which of the points are support vectors? Highlight them in the figure. (2P)

(e) Compute the normalized normal vector $w = (w_1, w_2)^T$ and the corresponding offset w_0 of the maximum margin hyperplane defined by the equation $w^T x + w_0 = 0$. (4P)

4

Vektor Ebene $y = x$

$w_{\text{Steigung}} = -1x + 0$

$w = (-1, 1)^T$

$y = mx + t$

~~$y = -1x$~~

normalize

$\frac{1}{|w|} \cdot w$

x-Achse -4

$0 = \left\langle \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} -4 \\ 0 \end{pmatrix} \right\rangle + b$

$\frac{1}{\sqrt{(-1)^2 + 1^2}} \cdot \begin{pmatrix} -1 \\ 1 \end{pmatrix}$

$= \frac{1}{\sqrt{2}} \cdot \begin{pmatrix} -1 \\ 1 \end{pmatrix}$

nach b auflösen

$b = -\frac{4}{\sqrt{2}}$

$w = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$

Ebene $H(x) = \left\langle \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, x \right\rangle - \frac{4}{\sqrt{2}}$

(f) For a different training dataset with the same class labels, the following parameters have been learned:

$$w = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad w_0 = -2$$

Classify the following two new points (i.e. determine whether they belong to class *triangle* or to class *circle*):

$$p_1 = \begin{pmatrix} -3 \\ 5 \end{pmatrix}, \quad p_2 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

(2P)

$w^T x + w_0 = 0$

$H(p_1) = \text{sign}\left(\left\langle \begin{pmatrix} -1 \\ 0 \end{pmatrix} \begin{pmatrix} -3 \\ 5 \end{pmatrix} \right\rangle + \begin{pmatrix} -1 \\ 0 \end{pmatrix}\right)$

$\text{sign}\left(3 + \begin{pmatrix} -1 \\ 0 \end{pmatrix}\right) = \text{sign}2 = 1$

$\Rightarrow$ triangle

$H(p_2) = \text{sign}\left(\left\langle \begin{pmatrix} -1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 \\ -2 \end{pmatrix} \right\rangle + \begin{pmatrix} -1 \\ 0 \end{pmatrix}\right)$

$= \text{sign}\left(\begin{pmatrix} -1 \\ 0 \end{pmatrix} + \begin{pmatrix} -1 \\ 0 \end{pmatrix}\right) = \text{sign}(-2)$

$= -1$

$\Rightarrow$ circle

Task 8 **Evaluation** (1+2+2+4 Points)

Given a data set $D = \{o_1, \ldots, o_{13}\}$, let $C(o_i) \in \mathcal{C} = \{A, B\}$ denote the true class of the objects. Furthermore, let K be a classifier, and let $K(o_i) \in \mathcal{C}$ denote the predicted class label. The following table shows the confusion matrix for K.

		$K(o)$ A	$K(o)$ B	
$C(o)$	A	9	0	9
	B	3	1	4
		12	1	/13

(a) Calculate the classification accuracy of K (as reduced fraction). (1P) 1P.

accuracy = $\frac{9+1}{13} = \frac{10}{13}$ ✓

(b) Calculate the recall of K for each class in $\mathcal{C}$ (as reduced fraction). (2P) 2P.

$\text{Recall}_A = \frac{|\{o \in D \mid K(o) = C(o)\}|}{C_i} = \frac{9}{9} = 1$ ✓

$\text{Recall}_B = \frac{1}{4} = 0{,}25$ ✓

(c) Calculate the precision of K for each class in $\mathcal{C}$ (as reduced fraction). (2P) 2P.

$\text{Precision}_A = \frac{|\{o \in D \mid K(o) = C(o)\}|}{K_i}$

$= \frac{9}{12} = \frac{3}{4}$ ✓

$\text{Precision}_B = \frac{1}{1} = 1$ ✓

(d) To evaluate the overall performance of a classifier, one commonly takes the average of the F_1-score over all classes using one of the following two approaches:

i. **Micro Average F_1-Measure**: The values of TP, FP and FN are added up over all classes. Then precision, recall and F_1-measure are computed using these sums.

ii. **Macro Average F_1-Measure**: Precision and recall are computed for each class individually, afterwards the average precision and average recall are used to compute the F_1-measure.

Calculate the Micro- and Macro-Average F_1-measures for the example above (as reduced fractions). What do you observe?

Note The F_1-score is the harmonic mean of precision and recall. The harmonic mean of two values a, b is given by

$$\frac{2 \cdot a \cdot b}{a+b}$$

(4P) 1,5 P.

F1 A micro

~~Rec~~ ~~A micro~~

$9 + 0 + 3 = 12$

F_1 B micro $= \frac{0+3+1=4}{\cancel{9+0+3=12}}$? } identical

F_1 macro Avg recall $= \left(1 + \frac{1}{4}\right) / 2 = \frac{1}{2} \cdot \frac{5}{4} =$

~~F1 B macro~~ ~~F1 macro~~ Avg precision $= \left(\frac{3}{4} + 1\right) \cdot \frac{1}{2} = \frac{7}{4} \cdot \frac{1}{2}$

$$F_1 \text{ macro} = \frac{2 \cdot \frac{5}{8} \cdot \frac{7}{8}}{\frac{5}{8} + \frac{7}{8}}$$