Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Thomas Seidl
Janina Sontheim, Maximilian Hünemörder

Knowledge Discovery and Data Mining I

WS 2019/20

Exercise M: Mock Exam

Exercise M-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 may have multiple correct statements.

(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?

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.

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.

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

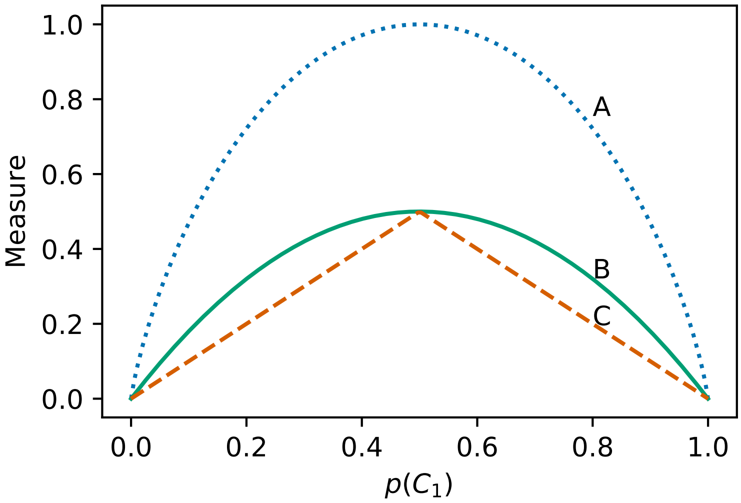
(i)

(ii)

(iii)

(iv)

(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)



A)

B)

C)

Exercise M-2 Data Aggregation

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.

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

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

Exercise M-3 Data Privacy

Given the following table

Key	Quasi-Identifier			Sensitive
Name	Semester	Age	Course	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) 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?

Equivalence Class	Count

Exercise M-4 Frequent Itemset Mining

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:

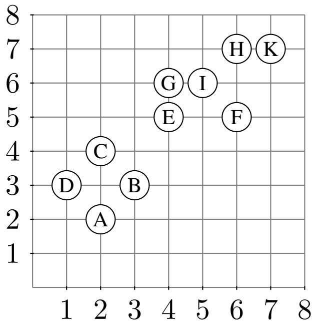
$$L_3 = \{ABC, ABF, ACF, AEF, AEH, AFH, BCE, BCF, BCH, EFH\}$$

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

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

Exercise M-5 Clustering

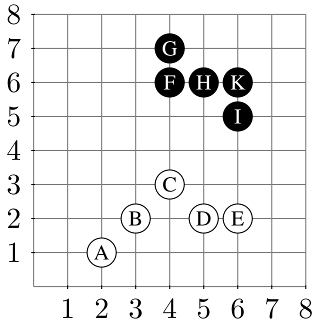
(a) Given are the following points



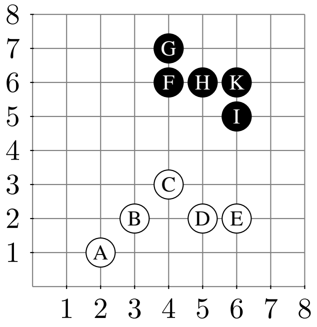
For k-Means, let $k = 2$, and the initial cluster centres are $C_1 = B$, and $C_2 = H$. Specify the cluster assignments of the initial iteration.

Cluster C_1	
Cluster C_2	

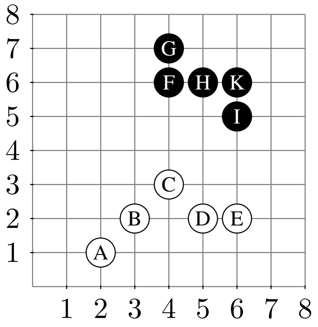
(b) The next step of k-Means would be the re-assignment. Given are the following *new* points with their assignment to clusters (black and white). Again using k-Means, draw the updated cluster means into the plot.



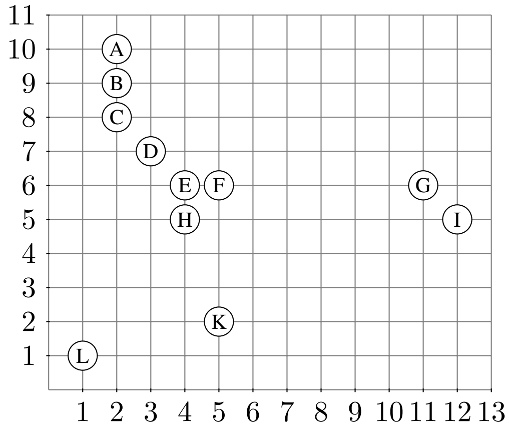
(Spare Figure)



(Spare Figure)



(c) Given the following data points

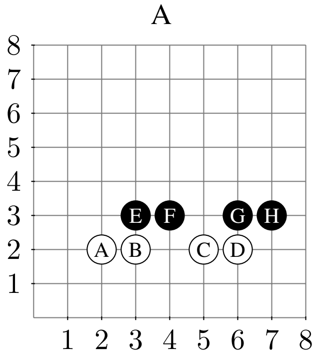


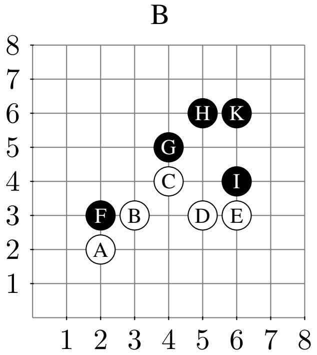
For k-Means, choose a value for k, and give k initial centroids such that

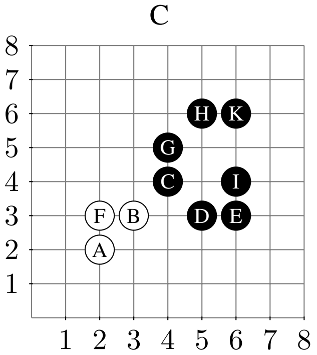
(i) After the initial assignment, every cluster is non-empty.

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

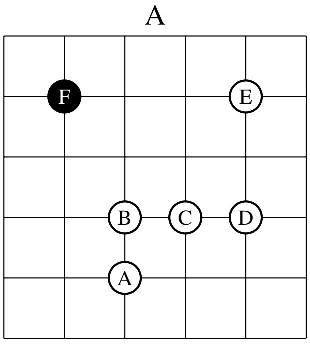
(d) Can the following partitions into two classes, black and white, be a final result of k-Means? If not, briefly justify your answer.

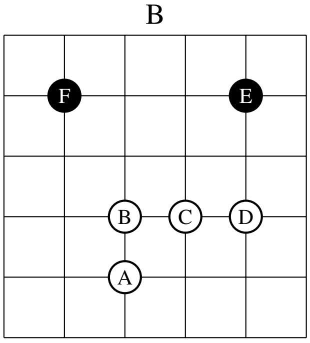


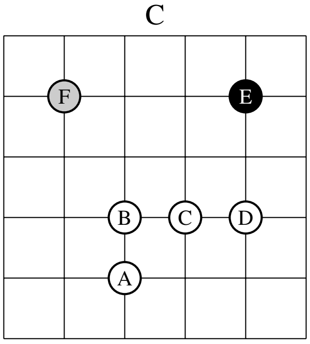




(e) Can the following clustering results be obtained using agglomerative hierarchical clustering with single-link and Manhattan distance as ground measure? If not, justify your answer. The class labels are given through the three colours: white, grey, and black.







Exercise M-6 Classification

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

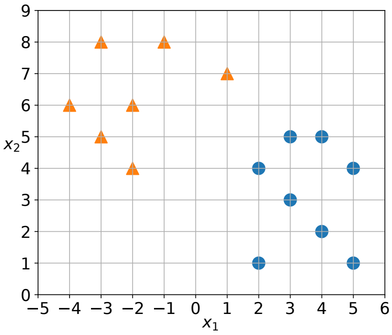
where *r* = *recommend* and *d* = *don't recommend*.

(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.

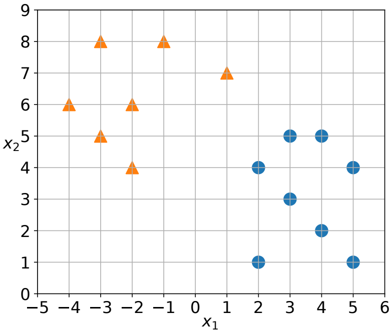
(b) At lunch, Norbert asks his friends for recommendations regarding a new book 7. Apply the classifier trained in the previous task to determine whether Norbert should read book 7. Provide all necessary computation steps.

Book	Harold	Gretchen	Read?
7	d	r	?

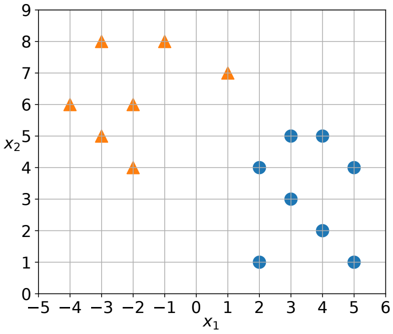
(c) Consider the following dataset consisting of two classes of points in $\mathbb{R}^2$. The *circle* class has label -1, the *triangle* class has label 1. Draw the maximum margin hyperplane inside the figure. No calculations are needed here.



(Spare Figure)



(Spare Figure)



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

(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 $w^T x + w_0 = 0$.

(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*):

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

Exercise M-7 Evaluation

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	B
$C(o)$	A	9	0
	B	3	1

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

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

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

(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

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