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

WS 2021/22
02.03.2022

Exam I - Open Book-Prüfung
Knowledge Discovery in Databases I

General Information:

- You have 90 minutes for the exam. Additionally you get 5 minutes extra time for downloading and 5 minutes extra time for uploading. You have plenty of time to react to technical issues. Also, you should upload preliminary versions to avoid major uploading issues. New submissions overwrite previous submissions.
- In addition to this exam file, also download the textfile "answers.txt". It is a template to turn in your answers for this exam.
- Upload your solution via the Uni2Work system also as "answers.txt". We do not accept other file formats.
- All places where you are required to insert content are marked with three dots ("..."). Always substitute these three dots with your content. No explanation is needed. If you fill in your answers in different places than these three dots, they will not be graded. Do not alter the template format.
- For each task, there may be additional instructions for filling in your answers.
- For definitions we refer to the script.
- If your exam regulations allow voiding exams (Entwertung) and you want to do so, please do not upload a submission.

By submitting a solution you accept the following conditions:

- I prepared the solution on my own without third-party assistance.
- I am the legitimate owner of this Uni2Work account and do not prepare the solution for somebody else.
- I am currently enrolled as a student and certified to take part in this exam. I am able to prove this at any state of this exam.
- I do not publish any contents of this exam like tasks or review data.
- I regularly update my solution to decrease the chance of potential technical problems at the end of the exam submission time. The last submission is graded. Be careful: Uni2Work will close your session after some minutes of inactivity.

Scoring of Multiple Choice:
Each task in this exam is identified with letters and roman numbers and has a corresponding line in the template. If you think that a statement is true, insert the corresponding number of the statement into this line. If you think that a statement is false, leave it out in the solution. Regarding the examination regulations (Prüfungsordnungen), correctly given true statements and correctly skipped false statements yield one point. Incorrectly given false statements and incorrectly omitted true statements decrease the score by one point. Bonus and malus points are accounted within one question block. Each block yields at least zero points, so you do not accumulate malus points with skipped tasks or tasks you could not solve sufficiently.

Example: Which letters are used in "KDD"?

$\boxed{1}$ K

$\boxed{2}$ A

$\boxed{3}$ D

The correct answer "Example: 1,3" would yield three points. One point is given for "Example: 1,2,3", "Example: 1", or "Example: 3". The remaining possibilities yield zero points.

The exam contains 3 tasks:

Task	Achievable Points
1.Clustering	25
2.Classification	28
3.Frequent Pattern Mining	22
Sum:	75

Aufgabe 1 Clustering (4+6+4+8+3 Punkte)

(a) Which statements are true regarding k-Means?

$\boxed{1}$ After running k-Means, every point has a positive Silhouette coefficient.

$\boxed{2}$ k-Means is in general not deterministic.

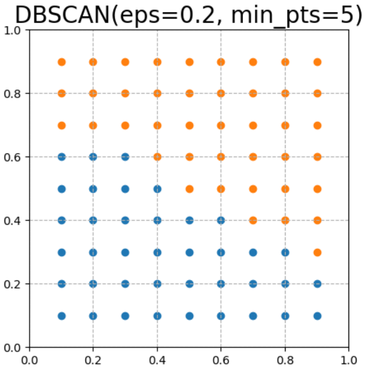
$\boxed{3}$ k-Means requires the computation of distances between all pairs of points in the dataset.

$\boxed{4}$ k-Means is sensitive to outliers.

(b) Which of the following clusterings (encoded in color, black corresponds to points classified as noise) is a valid output for the algorithm specified in the title of the plot?

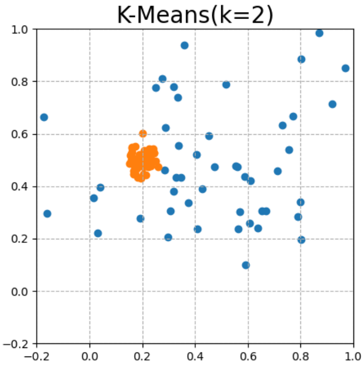
[1]

DBSCAN(eps=0.2, min_pts=5)



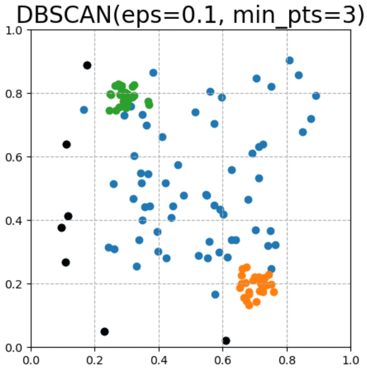
[2]

K-Means(k=2)



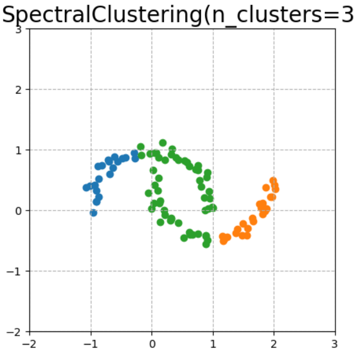
[3]

DBSCAN(eps=0.1, min_pts=3)



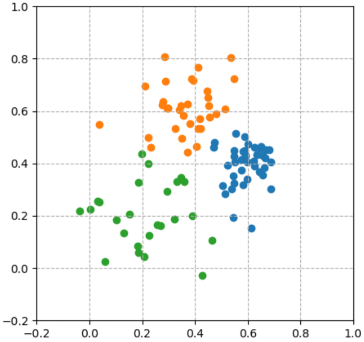
[4]

SpectralClustering(n_clusters=3)



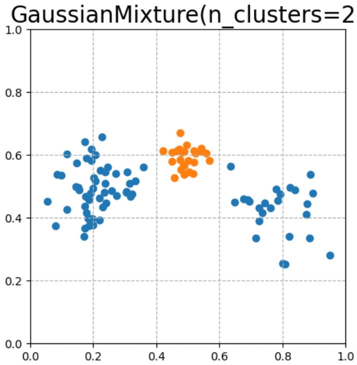
[5]

MeanShift



[6]

GaussianMixture(n_clusters=2)



(c) Which statements are true regarding DBSCAN?

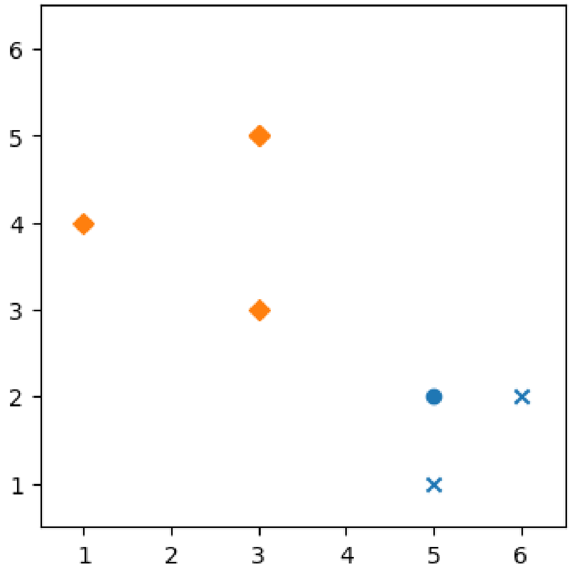
$\boxed{1}$ A cluster might contain no border points.

$\boxed{2}$ For any two points p, q in the same cluster, either p is directly density-reachable from q or q is directly density-reachable from p.

$\boxed{3}$ When running DBSCAN multiple times with the same hyperparameters on the same dataset, every point is either always or never classified as noise.

$\boxed{4}$ If $min\_points = 4$, every cluster needs to have at least two core points.

(d) The following scatter plot shows a dataset D consisting of 6 points in two dimensions. A clustering $\mathcal{C}$ is specified by the different symbols of the points. The ground-truth partition $\mathcal{G}$ is specified by the colors of the points.



Which of the following statements are true in this setting?

$\boxed{1}$ $|P| = 36$, where $P = \{(o, p) \in D \times D | o \neq p\}$.

$\boxed{2}$ $|S_{\mathcal{G}}| = 12$, where $S_{\mathcal{G}} = \{(o, p) \in P | \exists G_i \in \mathcal{G} : \{o, p\} \subseteq G_i\}$

$\boxed{3}$ The number of true positives is 5.

$\boxed{4}$ $H(\mathcal{C}) = \frac{\log_2 2}{2} + \frac{\log_2 6}{6} + \frac{\log_2 3}{3}$, where $H(\cdot)$ denotes the entropy.

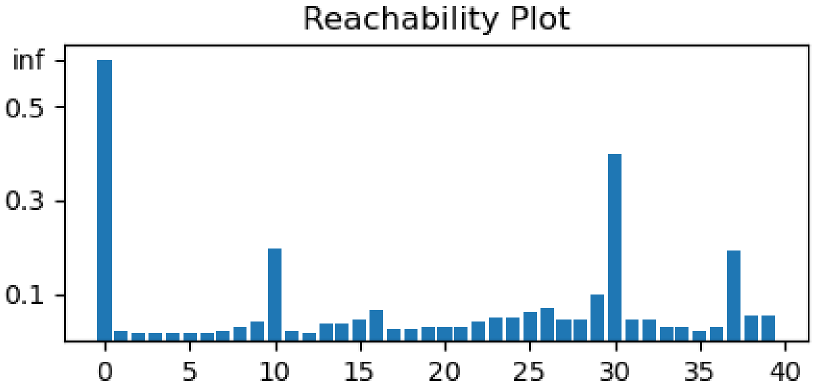
$\boxed{5}$ $H(\mathcal{G}|\mathcal{C}) = 0$, where $H(\cdot|\cdot)$ denotes the conditional entropy.

$\boxed{6}$ Out of all partitionings of D into exactly two non-empty subsets, $\mathcal{G}$ has the maximal entropy.

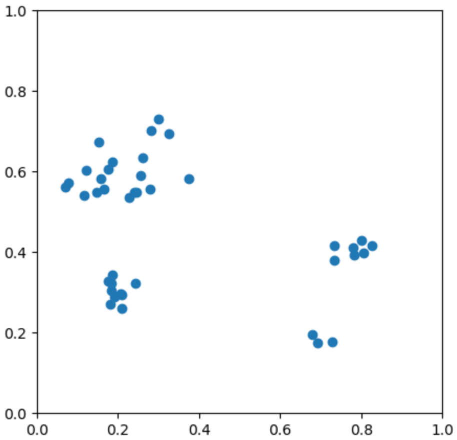
$\boxed{7}$ With respect to $\mathcal{C}$ and the Manhattan distance, the point $(6, 2)$ has a positive Silhouette coefficient.

$\boxed{8}$ $\mathcal{C}$ can be produced by DBSCAN.

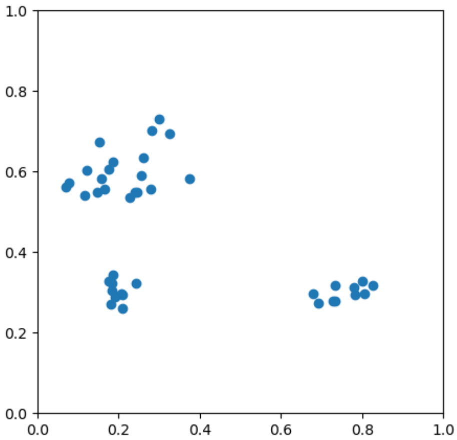
(e) The following reachability plot was obtained after running OPTICS. Which of the datasets below was used to create this reachability plot?
Please fill in the number of the right dataset in the solution.



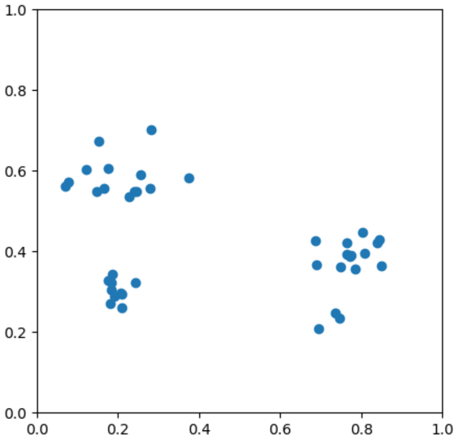
[1]



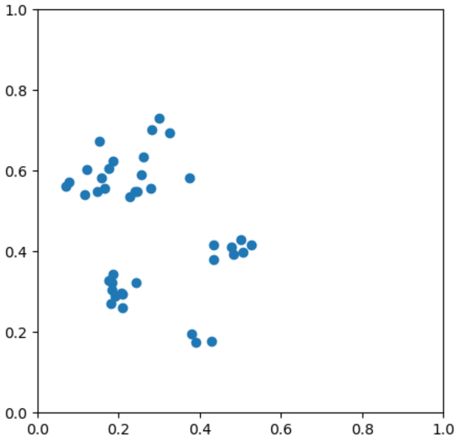
[2]



[3]



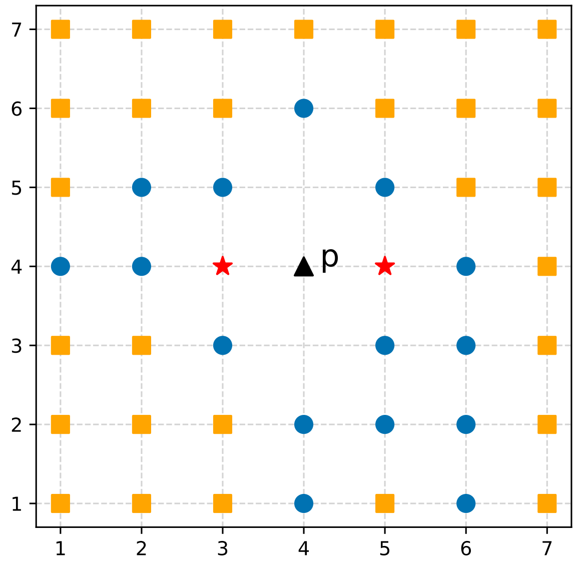
[4]



Aufgabe 2 Classification (11+17 Punkte)

Part I

(a) The following scatter plot shows a dataset D consisting of $n = 46$ training data objects and one query object p in a two-dimensional space. A nearest neighbor classifier on the training data is provided comprising the red stars, blue circles and orange squares corresponding to three different classes respectively using Manhattan distance (L_1 norm) as distance function. Assume that the object p (black triangle) in the middle arrived as new data instance and that the classifier is used to classify p (black triangle) with a given majority vote. Which of the following values for parameter k and classification results could have been produced by the classifier?



□ 1 $k = n - 1$, non-weighted majority vote, p classified as orange square.

□ 2 $k = 2$, non-weighted majority vote, p classified as red star.

□ 3 $k = 4$, non-weighted majority vote, p classified as orange square.

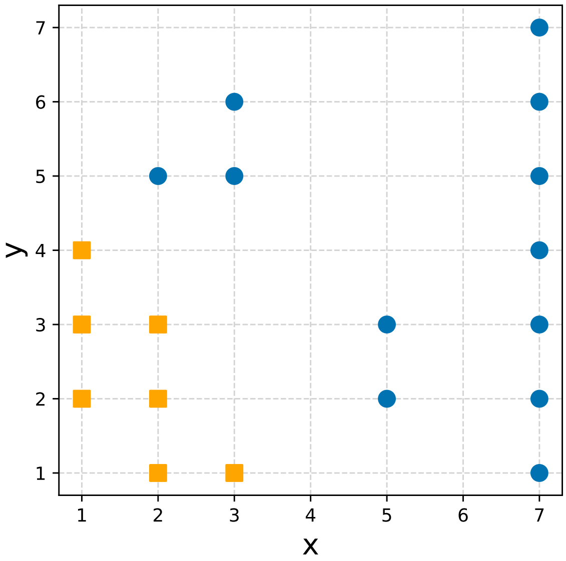
□ 4 $k = 10$, non-weighted majority vote, p classified as blue circle.

□ 5 $k = n$, non-weighted majority vote, p classified as orange square.

□ 6 $k = 10$, class-weighted majority vote, p classified as orange square.

□ 7 $k = 1$, distance-weighted majority vote, p classified as blue circle.

(b) The following scatter plot shows a dataset D with two different classes marked by different symbols. Consider using a linear support vector machine without soft margin. Please give the support vectors using the following format: $[x_1,y_1],[x_2,y_2],\ldots$ where x_i and y_i correspond to the coordinates of a support vector i. Give as many items as you think are needed.



Part II

The following tabular data is given.

Attribute 1	Attribute 2	Attribute 3	Result
Good	Medium	White	1
Bad	Short	White	1
Good	Long	Red	0
Bad	Long	White	1
Bad	Short	Blue	0
Good	Short	Green	0
Bad	Long	Blue	1
Good	Short	White	1

(a) Compute the apriori probability for the Result=1. Please round the numbers to three decimal places. Use the dot (".") as decimal separator (e.g. 0.111).

(b) Compute the apriori probability for the Result=0. Please round the numbers to three decimal places. Use the dot (".") as decimal separator (e.g. 0.111).

(c) Compute the conditional probability for P(Attribute1=Good | Result=1). Please round the numbers to two decimal places. Use the dot (".") as decimal separator (e.g. 0.11).

(d) Compute the conditional probability for P(Attribute2=Long | Result=1). Please round the numbers to two decimal places. Use the dot (".") as decimal separator (e.g. 0.11).

(e) Compute the conditional probability for P(Attribute2=Long | Result=0). Please round the numbers to two decimal places. Use the dot (".") as decimal separator (e.g. 0.11).

(f) Compute the conditional probability for P(Attribute3=White | Result=1). Please round the numbers to two decimal places. Use the dot (".") as decimal separator (e.g. 0.11).

(g) Calculate the **Gini Index** in the root node for Attribute 1. Please round the numbers to three decimal places. Use the dot (".") as decimal separator (e.g. 0.111).

(h) Calculate the **Gini Index** in the root node for Attribute 2. Please round the numbers to three decimal places. Use the dot (".") as decimal separator (e.g. 0.111).

(i) Construct a decision tree using **Gini Index** as measure for impurity based on the given data. Please fill in the number of the attribute that shall be used for the first split.

(j) Construct a decision tree using **Gini Index** as measure for impurity based on the given data. For which of the following entries would the result obtained from the decision tree be 0?

$\boxed{1}$ Good, Short, Blue

$\boxed{2}$ Bad, Medium, White

$\boxed{3}$ Bad, Medium, Red

$\boxed{4}$ Bad, Short, Green

(k) Calculate the **Information Gain** in the root node for Attribute 3. Please round the numbers to three decimal places. Use the dot (".") as decimal separator (e.g. 0.111).

Aufgabe 3 Frequent Pattern Mining (22 Punkte)

The statements below are partially related to the following dataset:

TID	Itemset
1	'Orange', 'Butter', 'Nutmeg', 'Rice', 'Eggs', 'Chocolate'
2	'Dill', 'Butter', 'Eggs', 'Nutmeg', 'Rice', 'Chocolate'
3	'Orange', 'Apple', 'Ice Cream', 'Eggs'
4	'Orange', 'Angle', 'Corn', 'Chocolate', 'Rice'
5	'Corn', 'Butter', 'Rice', 'Ice cream', 'Eggs'

(a) Which statements are true regarding frequent itemset mining?

$\boxed{1}$ The confidence of an association rule is always smaller than the support of this rule.

$\boxed{2}$ A closed frequent itemset is also always a maximal frequent itemset.

$\boxed{3}$ The number of maximal frequent itemsets is limited by the number of closed frequent itemsets.

$\boxed{4}$ The number of closed frequent itemsets is limited by the power set of unique items.

$\boxed{5}$ The number of closed frequent itemsets is limited by the power set of transactions.

$\boxed{6}$ The apriori algorithm and FP-Growth always yield the same freuqent itemsets as result.

(b) Apply the apriori algorithm on the given dataset with $minSup = 0.5$. Which statements are true?

$\boxed{1}$ ('Eggs', 'Butter', 'Rice') is the largest frequent itemset.

$\boxed{2}$ There are exactly 9 frequent itemsets.

$\boxed{3}$ There are exactly 10 frequent itemsets.

$\boxed{4}$ ('Eggs', 'Orange') is the most frequent itemset of size 2.

$\boxed{5}$ No frequent itemset has support 0.8.

$\boxed{6}$ The support of the itemset ('Butter', 'Rice') is 0.6.

(c) Which statements on the given dataset about maximality and closure are true?

$\boxed{1}$ On this dataset, the number of maximal and closed frequent itemsets are the same.

$\boxed{2}$ There are exactly 6 closed frequent itemsets.

$\boxed{3}$ There are exactly 3 maximal frequent itemsets.

$\boxed{4}$ (’Chocolate’, ’Rice’) is a closed frequent itemset.

$\boxed{5}$ (’Chocolate’, ’Rice’) is a maximal frequent itemset.

$\boxed{6}$ (’Orange’) is not a closed frequent itemset.

(d) We mined association rules on the given dataset. Which statements are true?

$\boxed{1}$ The support of the rule (’Rice’) → (’Chocolate’) is 0.8.

$\boxed{2}$ The support of the rule (’Chocolate’) → (’Rice’) is 0.6.

$\boxed{3}$ The confidence of the rule (’Eggs’) → (’Butter’, ’Rice’) is 0.75.

$\boxed{4}$ The confidence of the rule (’Butter’, ’Rice’) → (’Eggs’) is 0.75.