Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Thomas Seidl
Andrea Maldonado, Florian Richter, Yunpu Ma

WS 2020/21
08.04.2021

Exam II
Knowledge Discovery in Databases I

General Information:

- In addition to this exam file, also download the textfile "solution.txt". It is a template to turn in your answers for this exam.
- Only use the "solution.txt" as a submission file. We do not accept other file formats.
- Upload your solution via the Uni2Work system as "solution_[MatrNo]", where [MatrNo] is you immatriculation number.
- For each task give only the numbers of correct statements in your solution. Do not explain anything. Stick to the example. Use "," as a delimiter. Do not alter the template format.
- You have 180 minutes for the exam. This is more time than necessary to solve all tasks. It includes additional time for downloading and 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.
- Exam-Hotline: 089 / 2180 9313
- If your exam regulations allow voiding exams and you want to do so, insert "entwerten" as the first line of your 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- I do not publish any contents of this exam like tasks or review data.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"?
1 K

2 A

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 4 tasks.

Aufgabe	mögliche Punkte	erreichte Punkte
1.Clustering	28	
2.Frequent Itemset Mining	22	
3.Decision Tree	16	
4.Data Handling	22	
Summe:	88	
Note:		

Aufgabe 1 Clustering (28 Punkte)

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

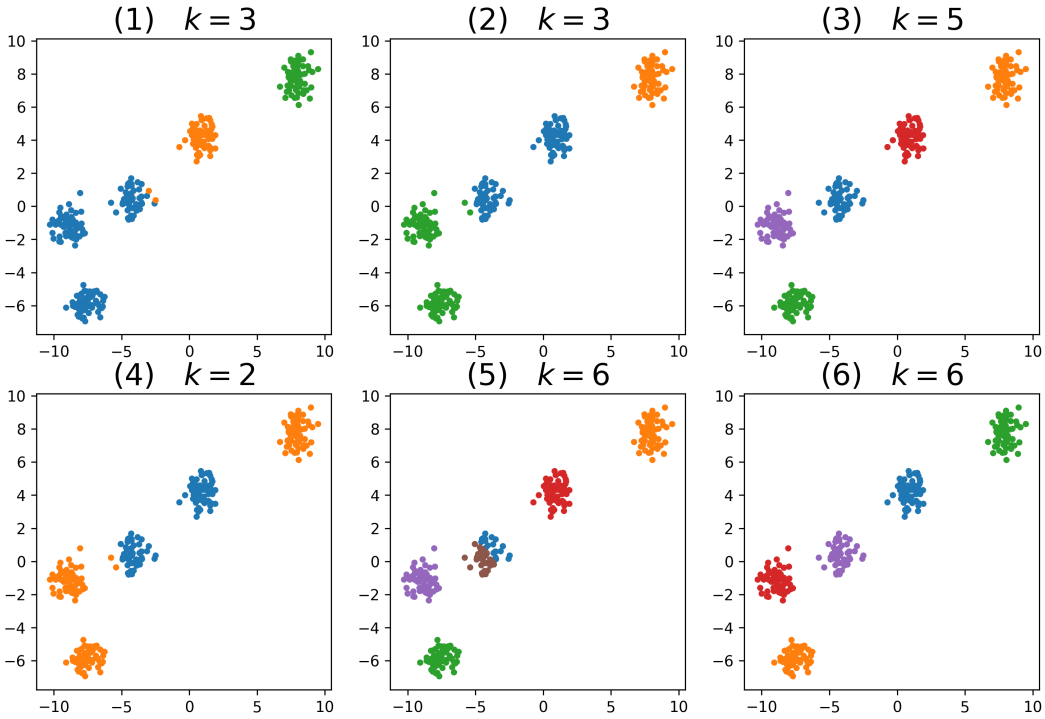
1 k-Means is a parameter-free clustering technique.

2 k-Means solves an optimization problem.

3 k-Means terminates even without a given iteration limit.

4 k-Means is invariant against the choice of initial means.

(b) Which clusterings are valid outputs of k-Means?



(c) Which statements are true regarding DBSCAN?

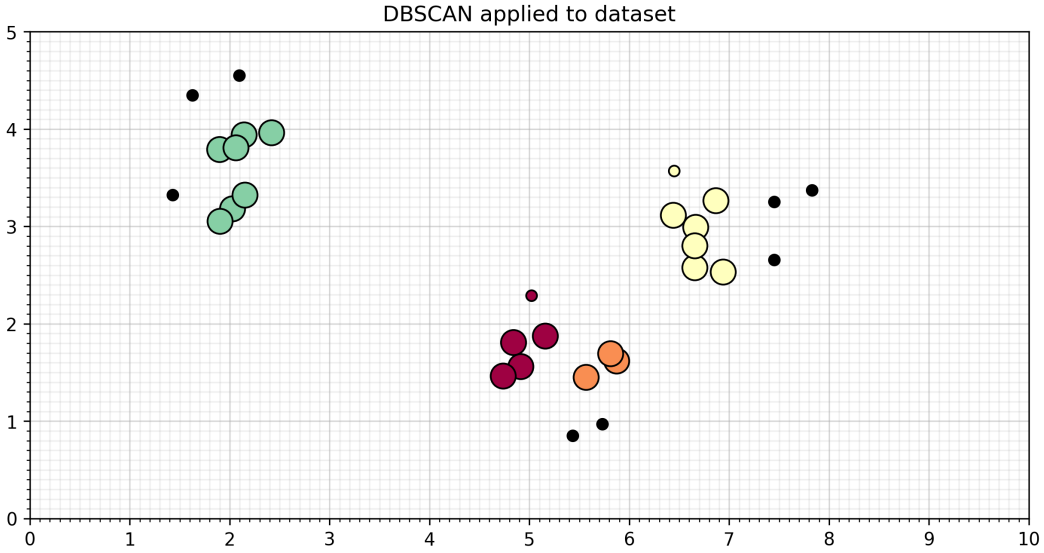
1 DBSCAN yields always the same cluster assignment for same ε and $MinPts$.

2 DBSCAN classifies objects into three different object classes.

3 Each object in a cluster C is density-reachable from all other objects in C.

4 A cluster might contain only core points.

(d) Which statements about the following clustering result achieved with DBSCAN are true? Similarly colored points belong to the same cluster. Black points are labeled as noise, small colored points are border points and large colored points are core points.



1 There are only three clusters in the data.

2 DBSCAN assigns all remaining black nodes to the closest clusters.

3 $0 < \varepsilon \leq 0.3$.

4 $0.3 < \varepsilon \leq 0.7$.

5 $0.7 < \varepsilon \leq 1.0$.

6 $MinPts = 1$.

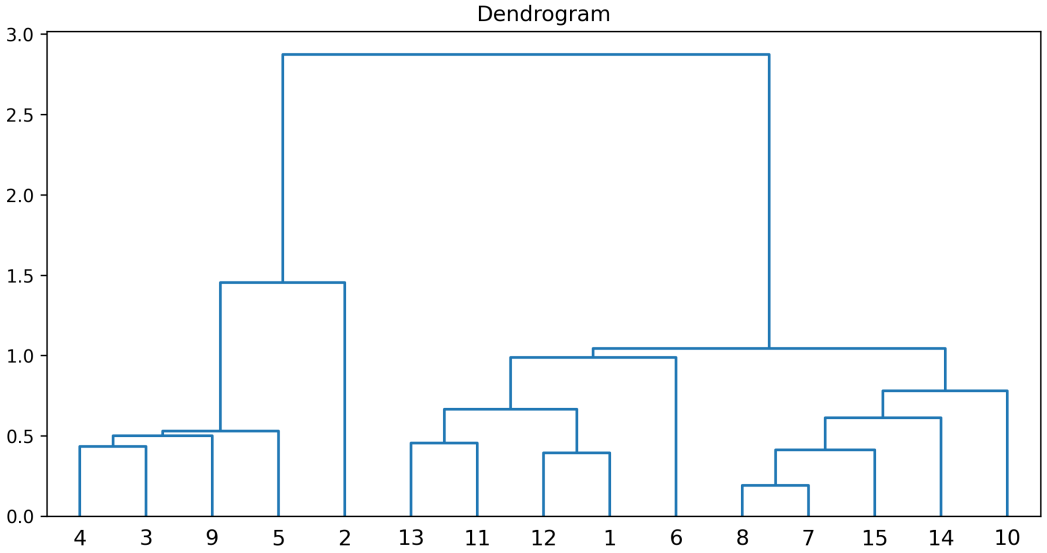
7 $MinPts = 2$.

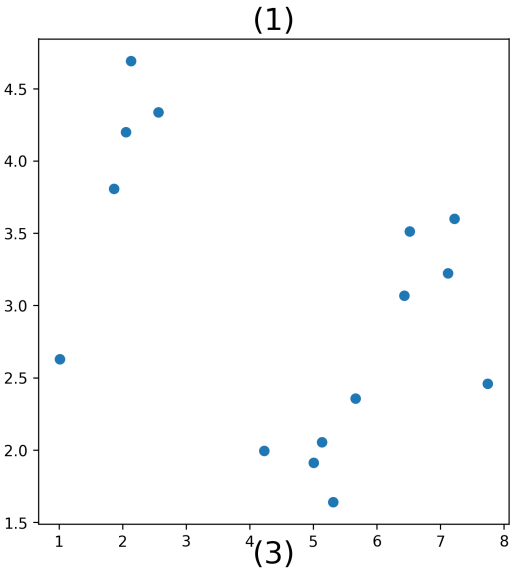
8 $MinPts = 3$.

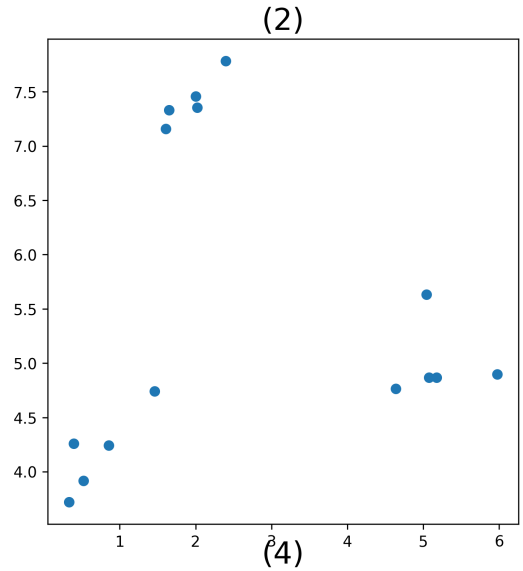
9 For this dataset, DBSCAN with these same parameters will always yield the same clustering.

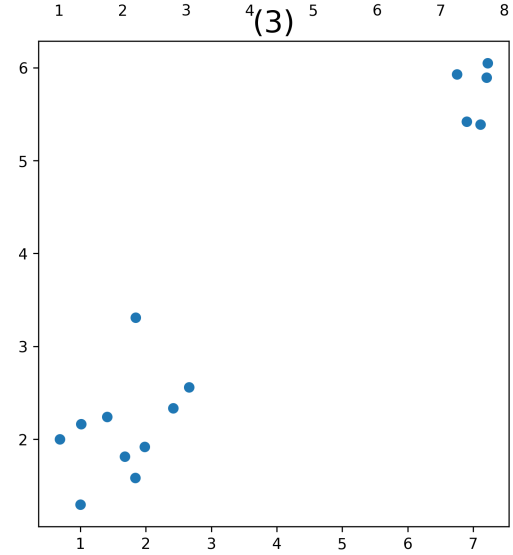
10 For this dataset, one particular object is potentially switched into another cluster in another DBSCAN instance with these same parameter settings.

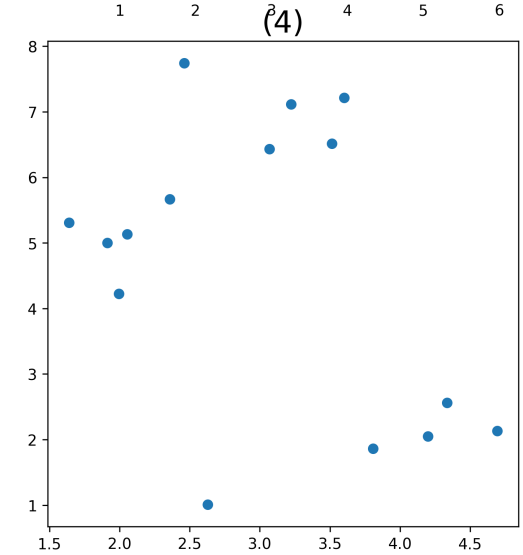
(e) The following dendrogram has to be matched to the corresponding datasets. Which datasets below yield the agglomerative hierarchical clustering (single-link on Euclidean distance) dendrogram? Give all numbers of fitting datasets.











Aufgabe 2 Frequent Itemset Mining (22 Punkte)

The statements below are partially related to the following dataset:

TID	Itemset
1	'Milk', 'Onion', 'Nutmeg', 'Kidney Beans', 'Eggs', 'Yogurt'
2	'Dill', 'Onion', 'Nutmeg', 'Kidney Beans', 'Eggs', 'Yogurt'
3	'Milk', 'Apple', 'Kidney Beans', 'Eggs'
4	'Milk', 'Unicorn', 'Corn', 'Yogurt'
5	'Corn', 'Onion', 'Onion', 'Kidney Beans', 'Ice cream', 'Eggs'

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

1 The apriori algorithm generates candidates by merging frequent itemsets.

2 The apriori algorithm generates candidates by intersecting infrequent itemsets.

3 The number of distinct frequent itemsets is limited by the number of transactions.

4 The number of distinct frequent itemsets is limited by the number of unique items.

5 Pruning in the apriori algorithm is decided for each generated candidate itemset.

6 A maximal frequent itemset is always also a closed frequent itemset.

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

1 (Kidney Beans, Eggs, Onion) is the largest frequent itemset.

2 There are exactly 9 frequent itemsets.

3 There are exactly 10 frequent itemsets.

4 (Kidney Beans, Eggs) is the most frequent itemset of size 2.

5 No frequent itemset has support 0.8.

6 (Eggs, Yogurt) has support 0.5.

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

1 On this dataset, maximal and closed frequent itemsets are the same.

2 There are 4 closed frequent itemsets.

3 There are no maximal frequent itemsets of size 2.

4 ('Kidney Beans', 'Eggs', 'Onion') is a closed frequent itemset.

5 ('Kidney Beans', 'Eggs', 'Onion') is a maximal frequent itemset.

6 ('Yogurt') is not a closed frequent itemset.

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

1 The rule (Kidney Beans) $\rightarrow$ (Eggs) has support 0.6.

2 The rule (Kidney Beans, Eggs) $\rightarrow$ (Onion) has support 0.8.

3 The rule (Kidney Beans) $\rightarrow$ (Eggs, Onion) has confidence 0.75.

4 The rule (Onion) $\rightarrow$ (Kidney Beans, Eggs) has confidence 0.75.

Aufgabe 3 Decision Tree (16 Punkte)

Below is a dataset containing data about stock investment decisions. The three predictors Momentum (down, stable, up), Risk (low, medium, high), Dividends (no, yes) and Positiontype (long, short) are used to plan an investment strategy. The target variable Action gives the decision whether to invest or to skip.

Momentum	Risk	Dividends	PosType	Action
Stable	Low	No	Long	Invest
Up	Medium	Yes	Short	Skip
Up	Low	No	Short	Invest
Down	Medium	No	Short	Invest
Up	Medium	No	Long	Invest
Stable	Medium	Yes	Long	Invest
Stable	High	No	Short	Invest
Down	Medium	Yes	Long	Skip
Up	High	Yes	Short	Skip
Up	High	Yes	Long	Skip
Stable	High	Yes	Short	Invest
Down	Medium	Yes	Short	Invest
Down	Low	No	Short	Invest
Down	Low	No	Long	Skip

(a) Determine the entropy of Action, $E(Action)$. Which statements are true?

☐ 1 $0.8 < E(Action) \leq 0.9$.

☐ 2 $0.9 < E(Action) \leq 1.0$.

(b) Determine the information gains Gain(T,X) for all four predictors. Which statements are true?

☐ 1 gain(T,Momentum) = 0.25.

☐ 2 gain(T,Risk) = 0.14.

☐ 3 gain(T,Dividends) = 0.38.

☐ 4 gain(T,PosType) = 0.05.

(c) Determine the first split attribute due to information gain. In case of equality, give all candidates. State the corresponding numbers, not the attribute names.

☐ 1 Momentum.

☐ 2 Risk.

☐ 3 Dividends.

☐ 4 PosType.

Aufgabe 4 Data Handling (22 Punkte)

The following dataset (source: Wikipedia, 2021) contains data about all current German federal states. The GDP is the gross domestic product measured in Euro per capita.

ID	State	Since	Area	Population	GDP (€/cap)
1	Baden-Württemberg	1952	35,752	11,100,394	47,290
2	Bavaria	1949	70,552	13,124,737	48,323
3	Berlin	1990	892	3,669,491	41,967
4	Brandenburg	1990	29,479	2,521,893	29,541
5	Bremen	1949	419	681,202	49,215
6	Hamburg	1949	755	1,847,253	66,879
7	Hesse	1949	21,115	6,288,080	46,923
8	Lower Saxony	1949	47,609	7,993,448	38,423
9	Mecklenburg-Vorpommern	1990	23,180	1,609,675	28,940
10	North Rhine-Westphalia	1949	34,085	17,932,651	39,678
11	Rhineland-Palatinate	1949	19,853	4,084,844	35,457
12	Saarland	1957	2,569	990,509	36,684
13	Saxony	1990	18,416	4,077,937	31,453
14	Saxony-Anhalt	1990	20,446	2,208,321	28,800
15	Schleswig-Holstein	1949	15,799	2,896,712	33,712
16	Thuringia	1990	16,172	2,143,145	29,883

(a) Which statements about data reduction are true?

1 Removing the GDP column to reduce the data is a sampling technique.

2 Replacing Area and Population by their ratio as density is a dimensionality reduction.

3 Replacing values in column Since into '$<$ 1990' and '$\geq$ 1990' is a roll-up.

4 Aggregating states according to their foundation year Since ('$\geq$ 1990') is a binning technique.

(b) We aggregate states into two sets, which are states added before and after 1990. Which statements about this aggregation are true?

ID	State	Since	Area	Population	GDP (€/cap)
1	Old states	$<$ 1990	248,508	66,939,830	43,453
2	New states	$\geq$ 1990	108,585	16,230,462	32,715

1 'Since' can be omitted in this aggregation due to redundancy.

2 'Area' is aggregated as a holistic measure as the sum of all state areas.

3 'Population' is aggregated as a distributive measure as the sum of all state populations.

4 'GDP' is aggregated as a distributive measure by the mean of all state GDPs.

5 The derived population density is an algebraic measure.

6 There are no holistic measures in the dataset.

(c) Perform equi-width and equi-height binning with 4 bins on the 'Area' attribute. Which statements are true?

1 One equi-width bin contains the majority of states.

2 One bin is empty for equi-width binning.

3 Brandenburg and North Rhine-Westphalia are put into the same equi-width bin.

4 Rhineland-Palatine and Saxony-Anhalt are put into the same equi-height bin.

5 Brandenburg and North Rhine-Westphalia are put into the same equi-height bin.

6 Schleswig-Holstein and Thuringia are put into the same equi-height bin.

(d) Which statements about privacy are true for the modified dataset below? Treat (Since) as the sensitive attribute. The quasi-identifiers are (Area), (Population) and (GDP). The coloring carries no additional data but should simplify counting procedures.

ID	State	Since	Area	Population	GDP (k€/cap)
1	Baden-Württemberg	<1990	20k-40k	>10M	>40
2	Bavaria	<1990	>40k	>10M	>40
3	Berlin	1990	<10k	<10M	>40
4	Brandenburg	1990	20k-40k	<10M	<30
5	Bremen	<1990	<10k	<10M	>40
6	Hamburg	<1990	<10k	<10M	>40
7	Hesse	<1990	20k-40k	<10M	>40
8	Lower Saxony	<1990	>40k	<10M	30-40
9	Mecklenburg-Vorpommern	1990	20k-40k	<10M	<30
10	North Rhine-Westphalia	<1990	20k-40k	>10M	30-40
11	Rhineland-Palatinate	<1990	10k-20k	<10M	30-40
12	Saarland	<1990	<10k	<10M	30-40
13	Saxony	1990	10k-20k	<10M	30-40
14	Saxony-Anhalt	1990	20k-40k	<10M	<30
15	Schleswig-Holstein	<1990	10k-20k	<10M	30-40
16	Thuringia	1990	10k-20k	<10M	<30

1 (Area) is 2-anonymous.

2 (Population, GDP) is 3-anonymous.

3 (Area, GDP) is 1-anonymous (i.e. not anonymous).

4 (Area, Population) is 2-anonymous.

5 (GDP) is 2-diverse.

6 There are no diverse quasi-identifiers in this dataset.