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

WS 2020/21
27.02.2021

Exam I
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- I do not publish any contents of this exam like tasks or review data.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

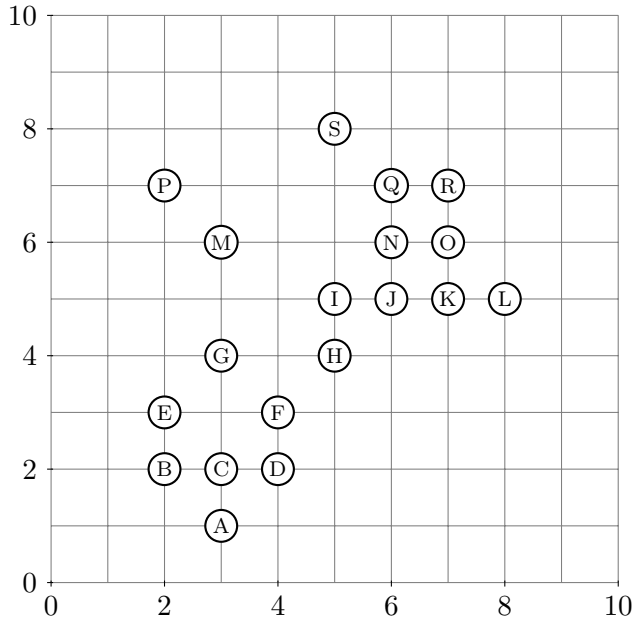
The correct answer "Example: 1,3" would yield three points. One point is given for "Example: 1,2,3", "Example: 1", or "Example: 3". The remaining possibilities yield zero points.

The exam contains 5 tasks.

Aufgabe	mögliche Punkte	erreichte Punkte
1.DBSCAN	18	
2.Data Handling	27	
3.Privacy	12	
4.Kernels	24	
5.Information Gain and Decision Tree	21	
Summe:	102	
Note:		

Aufgabe 1 DBSCAN (18 Punkte)

Given the following 2-dimensional dataset, which contains 19 objects, determine the correct possibilities for the stated 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 above. Outliers do not count as clusters.



(a)
- 1 Q is a core object with $\varepsilon = 2$ and MinPts $= 5$.
- 2 Q is a core object with $\varepsilon = 2$ and MinPts $= 6$.
- 3 Q is a core object with $\varepsilon = 2$ and MinPts $= 7$.
- 4 A is border object with $\varepsilon = 1$ and MinPts $= 4$.
- 5 A is border object with $\varepsilon = 2$ and MinPts $= 6$.
- 6 A is border object with $\varepsilon = 2$ and MinPts $= 8$.

(b)
- 1 A is an outlier $\implies$ $N_\varepsilon(A)$ does not contain B.
- 2 If J is a border object, then $\varepsilon > 1$.
- 3 A is a core object with MinPts $= 2$ $\implies$ $N_\varepsilon(A)$ contains only core objects.
- 4 MinPts $= 2$ and $\varepsilon = 0$ $\implies$ DBSCAN outputs 19 clusters.
- 5 MinPts $= 2$ and $\varepsilon \geq 1$ $\implies$ DBSCAN outputs 5 clusters.
- 6 MinPts $= 2$ and $\varepsilon \geq 2$ $\implies$ DBSCAN outputs only 1 cluster.

(c) $\boxed{1}$ If MinPts $= 1$, then there are infinitely many clusters.

$\boxed{2}$ If MinPts $= 1$, then there are only outliers.

$\boxed{3}$ If MinPts $= 1$, then there are only core objects.

$\boxed{4}$ Let MinPts $= 2$. DBSCAN finds only one cluster $\implies \varepsilon \geq 1$.

$\boxed{5}$ Let MinPts $= 2$. DBSCAN finds only one cluster $\implies \varepsilon \geq 2$.

$\boxed{6}$ Let MinPts $= 2$. DBSCAN finds only one cluster $\implies \varepsilon \geq 3$.

Aufgabe 2 Data Handling (27 Punkte)

Given the following dataset, answer the questions below by selecting true statements.

Id	Name	Gender	Age	Income	Skill Score	Security Clearance
1	Leia McDaniel	f	29	70k	9	high
2	Eloise Sweet	f	34	50k	4	None
3	Efan Draper	m	26	60k	5	low
4	Yvonne Bullock	f	21	55k	2	low
5	Lily-Grace Shea	f	41	60k	4	low
6	Saeed Bryan	m	22	65k	6	medium
7	Kelan Reyes	m	37	55k	3	None
8	Donald Kim	m	31	60k	1	medium
9	Saniya Watt	f	27	65k	6	high

(a) Which attributes are categorical?

1 Gender is a categorical attribute.

2 Age is a categorical attribute.

3 Income is a categorical attribute.

4 Skill Score is a categorical attribute.

5 Security Clearance is a categorical attribute.

(b) Which attributes are ordinal?

1 Gender is an ordinal attribute.

2 Age is an ordinal attribute.

3 Income is an ordinal attribute.

4 Skill Score is an ordinal attribute.

5 Security Clearance is an ordinal attribute.

(c) Which attributes are numerical?

1 Gender is a numerical attribute.

2 Age is a numerical attribute.

3 Income is a numerical attribute.

4 Skill Score is a numerical attribute.

5 Security Clearance is a numerical attribute.

(d) As a first data mining step, we transform some dimensions into 3-bin histograms. Which statements are true?

1 Equi-width binning on Skill Score yields one empty bin.

2 Equi-height and equi-width binning on Skill Score put 4 and 5 into the same bin.

3 Using equi-width binning on Skill Score, all bins contain exactly three elements.

4 Equi-width and equi-height binning yield the same binning on Age.

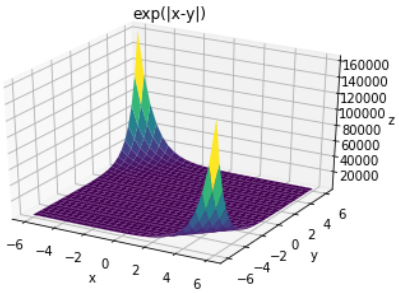
5 34 and 37 are in the same Age bins for equi-width binning.

6 Equi-height binning on Age has a bin that contains only odd numbers.

(e) Which statements are true regarding the following function? Select the correct statements.

$$d(x, y) = e^{|x-y|}$$

Hint:



1 $\forall x, y \in \mathbb{R} : x = y \implies d(x, y) = 0.$

2 $\forall x, y \in \mathbb{R} : d(x, y) = 0 \implies x = y.$

3 $\forall x, y \in \mathbb{R} : d(x, y) = d(y, x).$

4 $\forall x, y \in \mathbb{R} : d(x, y) \neq d(y, x).$

5 d fulfills the triangle equality.

6 d is a metric distance function.

Aufgabe 3 Privacy (12 Punkte)

Given the following dataset where (Id, Name) is the key, (Gender, Age, Income) is the quasi-identifier and (Security Clearance) is the sensitive attribute. Answer the following questions.

Id	Name	Gender	Age	Age-Range	Income	Security Clearance
1	Leia McDaniel	f	29	10-29	6*k	Level 1
2	Eloise Sweet	f	34	30-49	5*k	Level 2
3	Efan Draper	m	26	10-29	6*k	Level 2
4	Yvonne Bullock	f	21	10-29	5*k	Level 1
5	Lily-Grace Shea	f	41	30-49	6*k	Level 3
6	Saeed Bryan	m	22	10-29	5*k	Level 3
7	Kelan Reyes	m	37	30-49	5*k	Level 1
8	Donald Kim	m	31	30-49	6*k	Level 2
9	Saniya Watt	f	27	10-29	6*k	Level 1

(a) Which statements about k-anonymity are true for this dataset?

1 (Gender, Age) is 2-anonymous.

2 (Gender, Age-Range) achieves 2-anonymity by suppression.

3 (Age-Range) results by generalization.

4 (Gender, Income) is 3-anonymous.

5 k-anonymity is independent of keys and sensitive attributes.

6 (Gender) contains suppressed data.

(b) Which statements about l-diversity are true for this dataset?

1 l is bounded by the number of distinct sensitive attribute values.

2 Due to Level 3 occuring two times, each quasi-identifier is at most 2-divers.

3 (Gender) is 2-divers.

4 (Gender) is 3-divers.

5 (Gender, Age-Range) is 1-divers.

6 (Gender = f, Income = 6*k) is 2-divers.

Aufgabe 4 Kernels (24 Punkte)

In machine learning, kernel machines are a class of algorithms that use kernel functions. In this exercise, we study the properties of kernel functions.

(a) Which of the following properties need to be fulfilled for a Mercer kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$?

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

2 Negative definiteness of the kernel matrix.

3 The kernel matrix is elementwise non-negative.

4 Anti-symmetry, i.e., $\kappa(x, y) \neq \kappa(y, x)$.

5 Positive semi-definiteness of the kernel matrix.

6 The kernel matrix is elementwise positive.

(b) Given a $n \times n$ positive semi-definite kernel matrix K, how the positive semi-definiteness is defined?

1 For all $x \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, K satisfies $x^T K x > 0$.

2 For all $x \in \mathbb{R}^n$, K satisfies $x^T K x \geq 0$.

3 For all $x \in \mathbb{R}^n$, K satisfies $x^T K x > 0$.

4 For all $x \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, K satisfies $x^T K x \leq 0$.

5 For all $x \in \mathbb{R}^n$, K satisfies $x^T K x \leq 0$.

6 For all $x \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, K satisfies $x^T K x \geq 0$.

(c) Knowing the definition of a Mercer kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which of the following functions are Mercer kernels for $x, y \in \mathcal{X} = \mathbb{R}^n$?

1 $\kappa(x, y) := c$ with constant $c \in \mathbb{R}$ and $c \geq 0$.

2 $\kappa(x, y) := \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}$.

3 $\kappa(x, y) := \alpha x^T y + \beta$ for $\alpha, \beta \in \mathbb{R}$ with $\alpha, \beta < 0$.

4 $\kappa(x, y) := (x - 1)^T y$.

5 $\kappa(x, y) := x^T y$.

6 $\kappa(x, y) := \alpha x^T y + \beta$ for $\alpha, \beta \in \mathbb{R}$ with $\alpha, \beta \geq 0$.

(d) Consider the polynomial kernel function

$$K : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}, (x, y) \mapsto (x^T y + \gamma)^p, \text{ with } p = 2, \gamma = 2.$$

Which of the following feature maps $\phi : \mathbb{R}^2 \to \mathbb{R}^6$ satisfy the condition $K(x, y) = \langle \phi(x), \phi(y) \rangle$ for the kernel trick?

1 $\phi : \mathbb{R}^2 \to \mathbb{R}^6, x \mapsto (1, 2x_1, 2x_2, x_1^2, x_2^2)$.

2 $\phi : \mathbb{R}^2 \to \mathbb{R}^6, x \mapsto (1, 2x_1, x_1^2, 2x_2, x_2^2)$.

3 $\phi : \mathbb{R}^2 \to \mathbb{R}^6, x \mapsto (1, 2x_1, 2x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)$.

4 $\phi : \mathbb{R}^2 \to \mathbb{R}^6, x \mapsto (1, 2x_1, 2x_2, x_1^2, x_2^2, 2x_1x_2)$.

5 $\phi : \mathbb{R}^2 \to \mathbb{R}^6, x \mapsto (1, 2x_1, x_1^2, 2x_2, x_2^2, \sqrt{2}x_1x_2)$.

6 $\phi : \mathbb{R}^2 \to \mathbb{R}^6, x \mapsto (1, 2x_1, x_1^2, 2x_2, x_2^2, 2x_1x_2)$.

Aufgabe 5 Information Gain and Decision Tree (21 Punkte)

(a) Let T be a set of $|T|$ training objects with an attribute A and k classes. Let $\{T_i^A \mid i \in \{1, \ldots, m_A\}\}$ be the disjoint, complete partitioning of T produced by a split on the attribute A, where m_A is the number of disjoint values of A. Let $|T_i^A|$ be the number of elements in T_i^A. Suppose that the class membership of T is independently uniformly distributed and independent of the values of the attribute A. Let $I(T, A)$ be the information gain of the partitioning, $H(T)$ the entropy of the dataset, and $H(T_i^A)$, for $i \in \{1 \ldots m_A\}$, the entropy of partitions. Which of the following arguments are correct?

1 $I(T, A)$ can be evaluated just from $H(T_i^A)$, for $i \in \{1 \ldots m_A\}$.

2 $I(T, A)$ can be evaluated from $H(T)$, $|T|$, $H(T_i^A)$, and $|T_i^A|$ for $i \in \{1 \ldots m_A\}$.

3 $I(T, A)$ is evaluated as $I(T, A) = H(T) - \sum_{i=1}^{m_A} H(T_i^A)$

4 $I(T, A)$ is evaluated as $I(T, A) = H(T) + \sum_{i=1}^{m_A} \frac{|T_i^A|}{|T|} H(T_i^A)$

5 $I(T, A)$ is evaluated as $I(T, A) = H(T) + \sum_{i=1}^{m_A} H(T_i^A)$

6 $I(T, A)$ is evaluated as $I(T, A) = H(T) - \sum_{i=1}^{m_A} \frac{|T_i^A|}{|T|} H(T_i^A)$

(b) To evaluate the information gain under the independent uniform data distribution (see above), Which of the following arguments are correct?

1 Entropy $H(T_i^A) = -\log k$ for $i \in \{1 \ldots m_A\}$.

2 Entropy $H(T_i^A) = \log k$ for $i \in \{1 \ldots m_A\}$.

3 Entropy of the dataset is $H(T) = -\log k$.

4 Entropy of the dataset is $H(T) = \log k$.

5 Information gain of the partition is $I(T, A) = 2 \log k$.

6 Information gain of the partition is $I(T, A) = 0$.

7 The value of information gain $I(T, A)$ indicates that split leads to no gain of information.

8 The value of information gain $I(T, A)$ indicates that split leads to a gain of information.

(c) Which of the following arguments are correct when generating a decision tree from the training dataset?

1 Tree pruning is useful for decision tree construction.

2 For splitting an attribute, one can use the entropy as measure.

3 For splitting an attribute, we always use the information gain as measure.

4 An attribute could be chosen for split, when this split corresponds to the lowest information gain.

5 An attribute could be chosen for split, when this split corresponds to the highest information gain.

6 When an attribute A can take enough values, so that no two instances of the training set share the same value of A, the constructed decision tree might lead to insufficient generalization ability on new test data.

7 When an attribute A can take enough values, so that no two instances of the training set share the same value of A, the constructed decision tree might still generalize well on new test data.