Klausur (SS 2042)
Übungsklausur Informationstechnik II

Institut für Technik der Informationsverarbeitung
Prof. Dr.-Ing. Klaus Uhr

Klausur: Übungsklausur Informationstechnik II
Datum: 42. Februar 2042

Teilnehmer: **Jim Panse**
Matr.-Nr.: **0815**
ID: **1**

Hörsaal: Homeoffice
Platz: 42

Es gelten die folgenden Regelungen:

- Die Bearbeitungszeit beträgt 90 Minuten.
- Es sind keine Hilfsmittel erlaubt, außer
  - einem doppelseitig und handschriftlich beschriebenen DIN-A4-Blatt und
  - einem Wörterbuch.
- Antworten können Sie in deutscher sowie in englischer Sprache verfassen.
- Nutzen Sie nur **dokumentenechte Schreibgeräte** – keine Bleistifte oder rote Farbe!
- Die Verwendung von eigenem Papier ist nicht zugelassen.
- Vermeiden Sie das Beschreiben der Rückseiten.
- Bei Bedarf erhalten Sie Zusatzblätter von der Aufsicht.
  - Versehen Sie solche Blätter unbedingt mit Ihrer Matrikelnummer.
  - Ordnen Sie jedes zusätzliche Lösungsblatt einer Aufgabe eindeutig zu.

Die vorliegende Klausur besteht aus **17 Blättern**.

Aufgabe 1: Algorithmen I

Aufgabe 1.1: Allgemeine Fragen

A) Kann ein probabilistischer Algorithmus auch deterministisch sein? Begründen Sie Ihre Antwort.

B) Liefern alle probabilistischen Algorithmen immer ein korrektes Ergebnis? Begründen Sie Ihre Antwort.

C) Welche Optimierung kann bei der Sortierung größerer Objekte angewandt werden?

D) Quicksort arbeitet zwar in-place auf den Eingabedaten, benötigt jedoch dank seiner rekursiven Arbeitsweise auch Platz auf dem Stack für seine Ausführung. Wie ist die Platzkomplexität von Quicksort im worst-case Fall? Begründen Sie Ihre Antwort.

E) Gegeben seien zwei Algorithmen mit den folgenden mittleren Laufzeitkomplexitäten. Welcher der beiden Algorithmen ist schneller? Begründen Sie Ihre Antwort. (Bedenken Sie bei Ihrer Antwort die Problemgröße und erwähnen Sie auch die Laufzeitkomplexität).

Algorithmus 1: $3n^2 + 6n + 20$

Algorithmus 2: $n^3 + 8n + 3$

F) Was sagt die 1:50 Regel in der Laufzeitanalyse aus?

Aufgabe 1.2: Bubblesort

A) Führen Sie den Bubblesort Algorithmus auf den folgenden Eingabedaten aus. Gehen Sie von ASCII kodierten Zeichen für die Vergleichsoperation aus. Das Ziel soll dabei eine aufsteigend sortierte Liste sein. Nutzen Sie die bereitgestellten Felder um jeweils das Ergebnis einer kompletten Iteration darzustellen. Es sollen dabei keine Schritte ausgelassen oder frühzeitig abgebrochen werden. (Hinweis: die Felder reichen aus um die komplette Sortierung darzustellen).

F	C	N	B	K	X	P	D
B	C	D	F	K	N	P	X

Aufgabe 1.3: Algorithmenanalyse

A) Im Folgenden ist der Bogo-Sort genannte Sortieralgorithmus in Pseudocode gegeben. Füllen Sie die Tabelle hinsichtlich der Eigenschaften von Bogo-Sort aus. Gewertet wird ein "Häkchen"falls die Eigenschaft zutrifft und ein "Kreuzßofern sie nicht zutrifft. Das Feld kann auch leer gelassen werden und wird in diesem Fall nicht gewertet. (Bitte beachten: Innerhalb dieser Unteraufgabe führen Fehler zu Punktabzug, sie wird allerdings minimal mit 0 Punkten bewertet).

Abbrechbar	
Heuristisch	
Stabil	
In place	
Inkrementell	
Rekursiv	
Iterativ	

```
Array A = [3,6,33,2,17,5]

while (not_sorted(A))
{
    shuffle(A)
}
```

Aufgabe 2: Big Data und Prozesse

Aufgabe 2.1: Big Data und Prozesse

A) Nennen Sie die 5 V's, welche den Begriff Big Data nach Vorlesung abgrenzen, und beschreiben Sie jeweils kurz was der jeweilige Begriff bedeutet.

Im folgenden wurden Passanten nach ihrem am häufigst benutzten Verkehrsmittel gefragt. Die Ergebnisse sind in Tabelle 2.1 festgehalten.

B) Um welche Datentypen handelt es sich bei den Spalten?

C) Überlegen Sie sich Fragestellungen, welche an den Datensatz gestellt werden könnten.

	Schüler	Studenten	Arbeitnehmer	Rentner	Keine Angabe
Zu Fuß	50	20	5	3	50
Fahrrad	20	90	60	40	120
Skateboard	5	5	2	1	50
ÖPNV	60	130	160	120	40
Kraftfahrzeug	3	30	220	70	60

Tabelle 2.1: Ergebnisse der Passantenbefragung bezüglich Verkehrsmittelnutzung

D) Normieren Sie den Datensatz.

E) Berechnen Sie die in der Vorlesung gestellten Statistikwerte.

F) Zeichnen Sie ein detailliertes Balkendiagramm.

Aufgabe 3: Data preparation

Aufgabe 3.1: Bereinigung der fehlenden Werte im Datensatz

Gegeben seien die folgenden mit Fehler behafteten Sensordaten einer ITIV-Smartwatch. Das Uhrarmband ist während des Laufens fest am Handgelenk angebracht. Der Herzschlagsensor hat manchmal jedoch zu großes Rauschen und deswegen schreibt das System einen NaN (Not-A-Number) im Speicher.

Als Data-Scientist am ITIV bekommen Sie die rohen Herzschlagfrequenzen (in **bpm: Schläge pro Minute**) nach dem Lauf. Sie müssen diese Daten aufbereiten, bevor Ihre Kollegen den Datensatz weiterverwenden können.

A) Theoretische Frage: Gegeben sei der Mittelwert und die Standardabweichung eines Datensatzes, Sie wollen die obere und untere Grenze bei der Anomaliedetektion mit der Standardabweichungsmethode berechnen. Wie lautet die entsprechende Berechnungsvorschrift? Verwenden Sie $1,96\sigma$ für den Akzeptanzbereich der Daten.

B) Finden Sie **zunächst die Ausreißer** im gemessenen Datensatz über die Standardabweichung (Anomaliedetektion) und ersetzen Sie diese mit einem **NaN**. Dafür verwenden Sie die folgenden Werten: **Mittelwert = 184,86 bpm; Standardabweichung = 139,18 bpm; 1,96σ; obere Grenze = 457,65 bpm; untere Grenze = -87,94 bpm**. Füllen Sie die **Spalte "Korrigierte Herzfrequenz ohne Ausreißer (bpm)"** der Tabelle 3.1 mit dem neuen Datensatz aus und **begründen Sie Ihre Antwort**.

Zeitstempel (s)	Gemessene Herzfrequenz (bpm)	Korrigierte Herzfrequenz ohne Ausreißer (bpm)	Korrigierte Herzfrequenz mittels linearer Interpolation (bpm)
30	121		
60	122		
90	500		
120	NaN		
150	137		
180	138		
210	136		
240	140		

Tabelle 3.1: Anomaliedetektion und Ersatz der fehlenden Werten in den gemessenen Herzfrequenzdaten

C) Die Methode über die Standardabweichung für die Anomaliedetektion kann in einigen Fällen nicht alle Ausreißer erkennen. Nennen Sie ein Beispiel dieses Falles **auf Basis der Aufgabe (B)**. **Hint:** macht es Sinn negative Herzfrequenzen zu messen?

D) Nun haben Sie keinen Ausreißer mehr, sondern einen Datensatz mit NaN Werte. Ersetzen Sie die fehlenden Werte mittels einer **linearen Interpolation**. Füllen Sie die **Spalte "Korrigierte Herzfrequenz mittels linearer Interpolation (bpm)"** der Tabelle 3.1 mit dem neuen Datensatz aus. **Geben Sie Ihren Rechenweg als Begründung an**.

Aufgabe 4: Maschinelles Lernen

Aufgabe 4.1: Kurzfragen

Sind die folgenden Aussagen **wahr** oder **falsch** ? Begründen Sie Ihre Antwort mit **einem** Satz.

A) Supervised Learning wird nur zur Klassifikation eines Problems genutzt!

B) **Nur** der Decision Tree Algorithmus erreicht einen Trainingsfehler von 0 angewendet auf einen **linear trennbaren** Datensatz!

C) Durch **Overfitting** kann ein Klassifikationsergebnis eines Problems von 100% erreicht werden!

D) **Supervised Learning** nutzt nur gelabelte Daten zur Modellbildung!

E) Ein einzelnes Perzeptron kann eine XOR-Funktion darstellen!

Aufgabe 4.2: k-Nearest-Neighbor-Algorithmus & Cross Validation

Gegeben ist ein Datensatz (siehe Tabelle 4.1) mit den Gesamtpunktzahl und den Noten von 9 Studenten. Vorleistungen (Bonuspunkte) wurden bei der **Notengebung** mitberücksichtigt.

Student	1	2	3	4	5	6	7	8	9
Punkte	53	59	70	79	84	87	91	93	99
Note	B	C	B	B	A	B	A	A	A

Tabelle 4.1: Notenverteilung

A) Verwenden Sie den 1-Nearest-Neighbor um die Arbeit (Note) eines neuen Studenten mit der Punktezahl von 86 zu klassifizieren. Begründen Sie Ihre Antwort.

B) Nun verwenden Sie den 3-Nearest-Neighbor und klassifizieren die Arbeit des gleichen Studenten. Begründen Sie Ihre Antwort.

C) Verwenden Sie den 1-Nearest Neighbor mit einer 3-fold cross validation. Was ist die cross validated accuracy des gegebenen Beispiels von Tabelle 4.1? Die Testsets bestehen aus:

- Set 1: Student 1, 4 und 7
- Set 2: Student 2, 5 und 8
- Set 3: Student 3, 6 und 9

Hinweis: Die Gesamtgenauigkeit ist der Mittelwert der Genauigkeit jedes einzelnen folds!

Aufgabe 5: Überwachte Lernverfahren

Aufgabe 5.1: Neuronale Netze

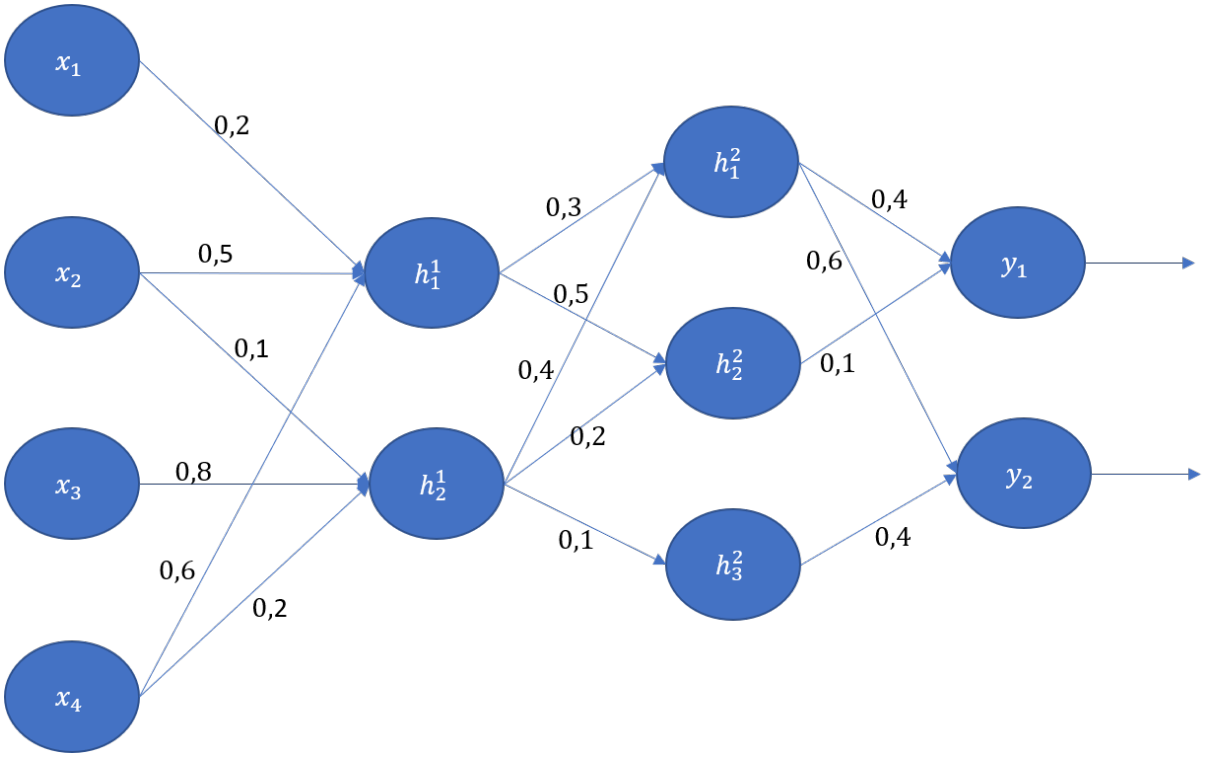


Abbildung 5.1: Neuronales Netz

Gegeben sei das in Abbildung 5.1 gezeigte Neuronale Netz. Nicht eingezeichnete Verbindungen sind mit $w_i = 0$ zu werten.

A) Stellen Sie die Übertragungsfunktion für das gezeigte Netz auf, wenn Sie als Aktivierungsfunktion $y(x) = \begin{cases} 0 & x \geq 0 \\ x & x \geq 0 \end{cases}$ nutzen.

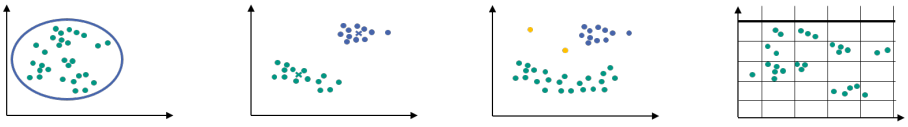
B) Berechnen Sie den Ausgangsvektor $[y_1\ y_2]$ für den Eingangsvektor $[x_1\ x_2\ x_3\ x_4] = [1\ \ 0.5\ \ 2\ \ 0.3]$

C) Welche Aussage folgt aus dem von Ihnen berechneten Ergebnis über die Klassifizierung durch das Neuronale Netz zu diesem Eingangsvektor?

Aufgabe 6: Unüberwachte Lernverfahren

Aufgabe 6.1: Allgemeine Fragen

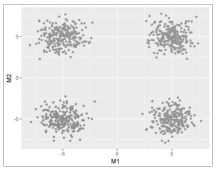
A) Benennen Sie die vier Familien der Clusterverfahren in Abbildung 6.1 und geben Sie je ein Vor- und ein Nachteil an.

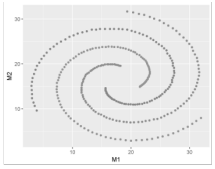


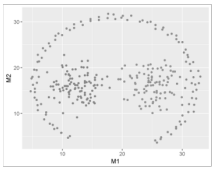
Name:				
Vorteil:				
Nachteil:				

Abbildung 6.1: Familien der Clusterverfahren

B) In Abbildung 6.2 sind drei verschiedene Datensätze abgebildet. Benennen Sie jeweils den Clusteralgorithmus, den Sie verwenden würden um den jeweiligen Datensatz bestmöglich zu clustern. Benennen Sie außerdem die nötigen Hyperparameter, die bestimmt werden müssen, damit der jeweilige Clusteralgorithmus ausgeführt werden kann.







Algorithmus-Vorschlag:			
Welcher Hyperparameter müssen für diesen Alg. bestimmt werden			

Abbildung 6.2: drei verschiedene Datensätze

C) Nennen Sie jeweils zwei Vor- und Nachteile des unüberwachten Lernens (Clustering) gegenüber dem überwachten Lernen (z.B. Klassifikation).

Aufgabe 6.2: Clustering

Bei einer Testfahrt über 100km wurden die Werte verschiedener interner Sensoren des Fahrzeuges aufgezeichnet. Dabei wurde jedoch vergessen, den aktuell eingelegten Gang mit aufzuzeichnen, weswegen Sie gebeten wurden, diesen aus den aufgezeichneten Daten wieder herauszulesen. In Abbildung 6.3 ist die Drehzahl des Motors über die Geschwindigkeit des Fahrzeuges aufgetragen und die Geraden in der Abbildung, sind die gefahrenen Gänge. Ebenfalls zu erkennen sind wenige Fehlmessungen bzw. Ausreißer.

A) Welchen Clusteringalgorithmus schlagen Sie vor um die gestellte Frage zu beantworten? Begründen Sie Ihre Antwort.

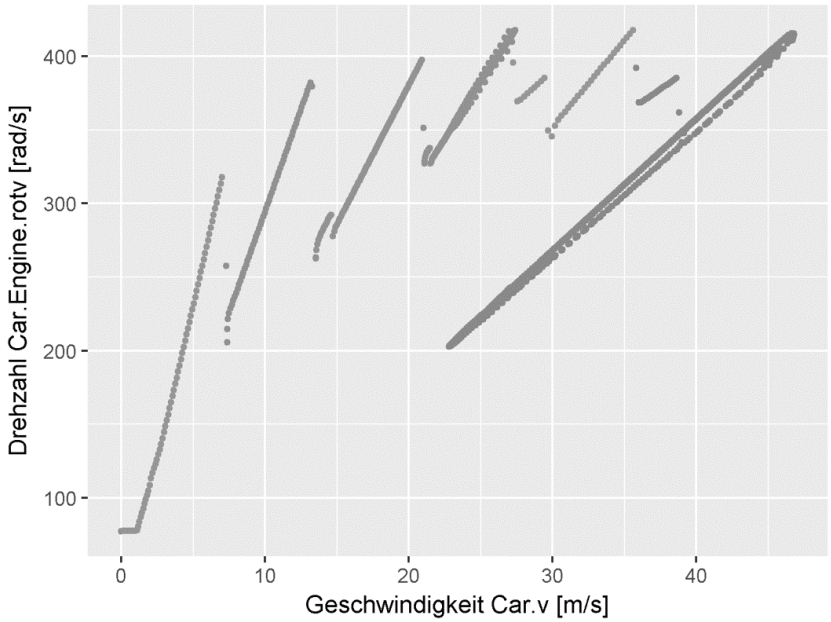


Abbildung 6.3: automotive Daten: Geschwindigkeit zu Motordrehzahl

B) Beschreiben Sie kurz wie Ihr vorgeschlagener Algorithmus funktioniert, welche Hyperparameter für diesen bestimmt werden müssen und welchen Einfluss dieser hat.