Klausur

APPLIED MACHINE LEARNING FUNDAMENTALS

Data Science, WWI 21DS B, DHBW Mannheim

Matrikelnummer:

3. Februar 2023, 10:00 Uhr - 11:00 Uhr

Hinweise:

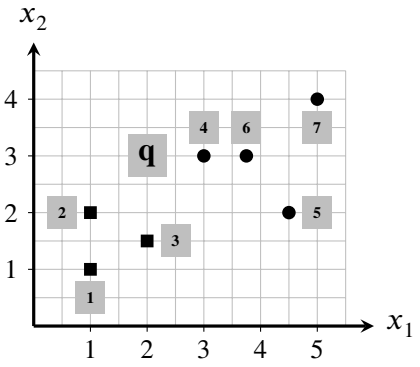
1. Bitte überprüfen Sie, ob Sie **alle 14 Aufgabenblätter** (ausgenommen des Deckblatts) erhalten haben, **bevor** Sie mit der Bearbeitung der Klausur beginnen. Bitte wenden Sie sich an die Prüfungsaufsicht, falls Ihr Druckexemplar unvollständig sein sollte.
2. Vergessen Sie nicht, Ihre Matrikelnummer auf der Klausur anzugeben. **Bitte verwenden Sie nicht Ihren Namen (Anonymisierung)!**
3. Notieren Sie Ihre Antworten direkt auf den Aufgabenblättern. Benutzen Sie gegebenenfalls die leeren Rückseiten oder die letzte Seite dieser Klausur.
4. Es steht Ihnen frei, die Fragen entweder auf Englisch oder auf Deutsch zu beantworten. Bitte übersetzen Sie keine technischen Begriffe, um Verwirrung zu vermeiden.
5. Die Klausur besteht aus 60 Punkten und ist **innerhalb von 60 Minuten** zu lösen. Die maximal zu erreichenden Punkte pro Aufgabe sind jeweils angegeben. Nutzen Sie sie als Hinweis darauf, wie umfangreich Ihre Antworten sein sollten.
6. Bitte schalten Sie alle Kommunikationsgeräte aus. Die folgenden Hilfsmittel sind erlaubt: ❶ Nicht programmierbarer Taschenrechner ❷ Zweiseitig handbeschriebenes Cheat Sheet
7. **Verstöße gegen die Prüfungsordnung werden als Täuschungsversuch gewertet!**

Aufgabe	**1**	**2**	**3**	**4**	**5**	**gesamt**
Punkte	/15	/15	/13	/7	/10	/60

1 k-Nearest Neighbors

1.1 Ihnen liegt ein Datensatz bestehend aus den beiden Klassen ■ und ● vor. Sie möchten den unbekannten Datenpunkt $q := (2, 3)^\intercal$ mithilfe des k-Nearest Neighbors Algorithmus klassifizieren. Sie wählen $k = 3$.

Berechnen Sie die Klassifikation a) mithilfe der **Manhattan-Distanz** und b) mithilfe der **Euklidischen Distanz**. **(6 p)**



Zeile	x_1	x_2	Klasse
1	1.00	1.00	■
2	1.00	2.00	■
3	2.00	1.50	■
4	3.00	3.00	●
5	4.50	2.00	●
6	3.75	3.00	●
7	5.00	4.00	●

1.2 Angenommen, Sie hätten $k = 7$ in Aufgabe 1.1 gewählt. Welche Klasse wäre vorhergesagt worden? An welchem Problem leidet Ihr Klassifikator nun? **(2 p)**

1.3 Skizzieren Sie zwei Möglichkeiten eines *Tie-Breaks*, falls beide Klassen in der Nachbarschaft von q gleich häufig vertreten sind. **(4 p)**

1.4 Kreuzen Sie die richtigen Aussagen zum k-Nearest Neighbors Algorithmus an. **(3 p)**

- [ ] Der k-Nearest Neighbors Algorithmus ist ein modellbasierter Algorithmus.
- [ ] k kann z. B. auf dem *Validation Set* ermittelt werden.
- [ ] Die Wahl von k hat keinen nennenswerten Einfluss auf die Vorhersagen.
- [ ] Ein zu großes k führt zu *Overfitting*.
- [ ] Der Algorithmus fällt in die Kategorie des *Lazy Learnings*.
- [ ] Die Lernphase ist rechen- und zeitintensiv.
- [ ] Die Vorhersage der Klassen unbekannter Punkte ist rechen- und zeitintensiv.
- [ ] k sollte auf dem *Train Set* ermittelt werden.

Maximal erreichbare Punkte für Aufgabe 1: **15 Punkte**

2 Gradient Descent / Gradientenabstieg

2.1 Berechnen Sie die Gradienten der folgenden Funktionen. **Hinweis zu b):** Mit $\|\boldsymbol{x}\|_2$ ist die Euklidische Norm des Vektors $\boldsymbol{x}$ gemeint. **Punkteverteilung:** a) 2 p / b) 3 p / c) 3 p **(8 p)**

a) $\begin{cases} f : \mathbb{R}^3 \to \mathbb{R} \\ f(x, y, z) = 3x^2 - 5y^2 + 2z^2 \end{cases}$ b) $\begin{cases} g : \mathbb{R}^m \to \mathbb{R} \\ g(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2 \end{cases}$

c) $\begin{cases} h : \mathbb{R}^3 \to \mathbb{R} \\ h(\boldsymbol{x}) = \sum_{i=1}^{2} \left(100 \cdot \left(x_{i+1} - x_i^2 \right)^2 + \left(x_i - 1 \right)^2 \right) \end{cases}$

2.2 Betrachten Sie erneut die Funktion f aus Aufgabe 2.1 a). Nehmen Sie nun an, Sie befinden sich im Punkt $\boldsymbol{p} := \left(\frac{1}{3}, \frac{1}{5}, \frac{1}{2}\right)^\intercal$. In welcher Richtung befindet sich der **steilste Abstieg** und in welcher der **steilste Aufstieg**? **(2 p)**

2.3 Führen Sie nun **eine Iteration** des *Gradient Descent* Algorithmus für die Funktion f aus Aufgabe 2.1 a) und dem Startpunkt $\boldsymbol{p}$ (siehe Aufgabe 2.2) aus. Als Lernrate benutzen Sie bitte $\alpha = 0.05$. **(2 p)**

2.4 Beschreiben Sie in wenigen Worten, was unter einer **konvexen Funktion** zu verstehen ist. Welchen entscheidenden Vorteil hat eine konvexe Kostenfunktion bei der Anwendung des *Gradient Descent* Algorithmus? **(3 p)**

Maximal erreichbare Punkte für Aufgabe 2: **15 Punkte**

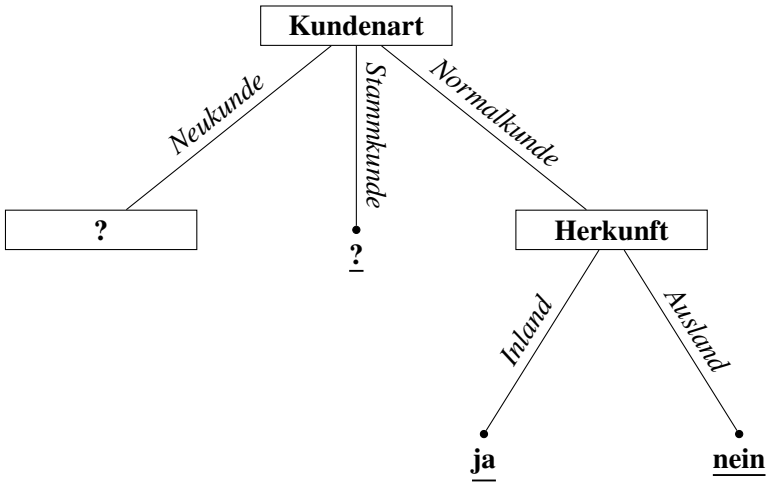
3 Entscheidungsbäume und Ensemblemethoden

3.1 Der folgende Kundenstamm besteht aus 14 Kunden. Für jeden dieser Kunden wurden die Merkmale **Kundenart (KA)**, **Zahlungsgeschwindigkeit (ZG)**, **Kauffrequenz (KF)** und **Herkunft (H)** erfasst. Die letzte Spalte gibt Auskunft darüber, ob dem jeweiligen Kunden ein Kauf auf Rechnung gestattet wird oder nicht.

Sie möchten einen Entscheidungsbaum auf diesen Daten trainieren. Ein Teil des Baumes liegt Ihnen bereits vor. Vervollständigen Sie ihn unter Verwendung der **Entropie**! **(7 p)**

Zeile	KA	ZG	KF	H	Rechnung
1	Neukunde	niedrig	niedrig	Inland	nein
2	Neukunde	niedrig	niedrig	Ausland	nein
3	Stammkunde	niedrig	niedrig	Inland	ja
4	Normalkunde	mittel	niedrig	Inland	ja
5	Normalkunde	hoch	hoch	Inland	ja
6	Normalkunde	hoch	hoch	Ausland	nein
7	Stammkunde	hoch	hoch	Ausland	ja
8	Neukunde	mittel	niedrig	Inland	nein
9	Neukunde	hoch	hoch	Inland	ja
10	Normalkunde	mittel	hoch	Inland	ja
11	Neukunde	mittel	hoch	Ausland	ja
12	Stammkunde	mittel	niedrig	Ausland	ja
13	Stammkunde	niedrig	hoch	Inland	ja
14	Normalkunde	mittel	niedrig	Ausland	nein

$$\begin{aligned} \textbf{KA} &\in \{\text{Neukunde, Normalkunde, Stammkunde}\} \\ \textbf{ZG} &\in \{\text{niedrig, mittel, hoch}\} \\ \textbf{KF} &\in \{\text{niedrig, hoch}\} \\ \textbf{H} &\in \{\text{Inland, Ausland}\} \end{aligned}$$


Kundenart
Neukunde
Stammkunde
Normalkunde
?
?
Herkunft
Inland
Ausland
ja
nein

3.2 Kreuzen Sie die richtigen Aussagen zu Entscheidungsbäumen an. **Hinweis:** Ein Entscheidungsstumpf ist ein Entscheidungsbaum mit nur einer Ebene. **(2 p)**

- ☐ Entropie und Gini-Index führen immer auf denselben Entscheidungsbaum.
- ☐ Ein Entscheidungsstumpf hat eine hohe *Variance*.
- ☐ Ein Entscheidungsstumpf hat einen hohen *Bias*.
- ☐ Entscheidungsbäume haben den Nachteil einer schlechten Interpretierbarkeit.
- ☐ Bei zwei Klassen ist 1 der maximale Wert des Gini-Index.

3.3 Beschreiben Sie kurz und bündig, was man unter einem *Random Forest* versteht. Welche Schritte werden in der Trainingsphase ausgeführt? Was ist der Vorteil gegenüber eines einzelnen Entscheidungsbaums? **(4 p)**

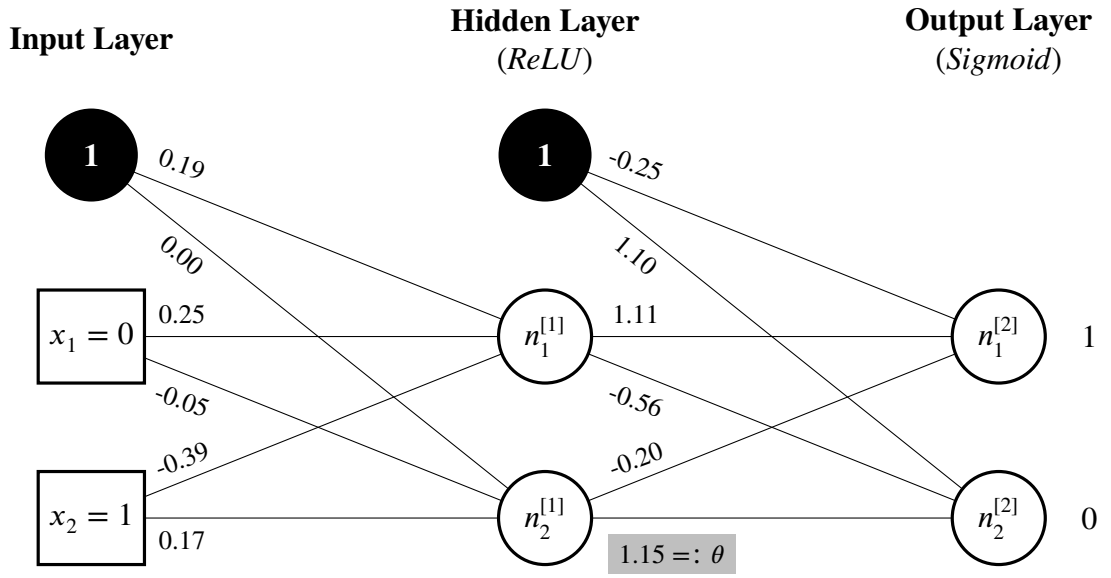
Maximal erreichbare Punkte für Aufgabe 3: **13 Punkte**

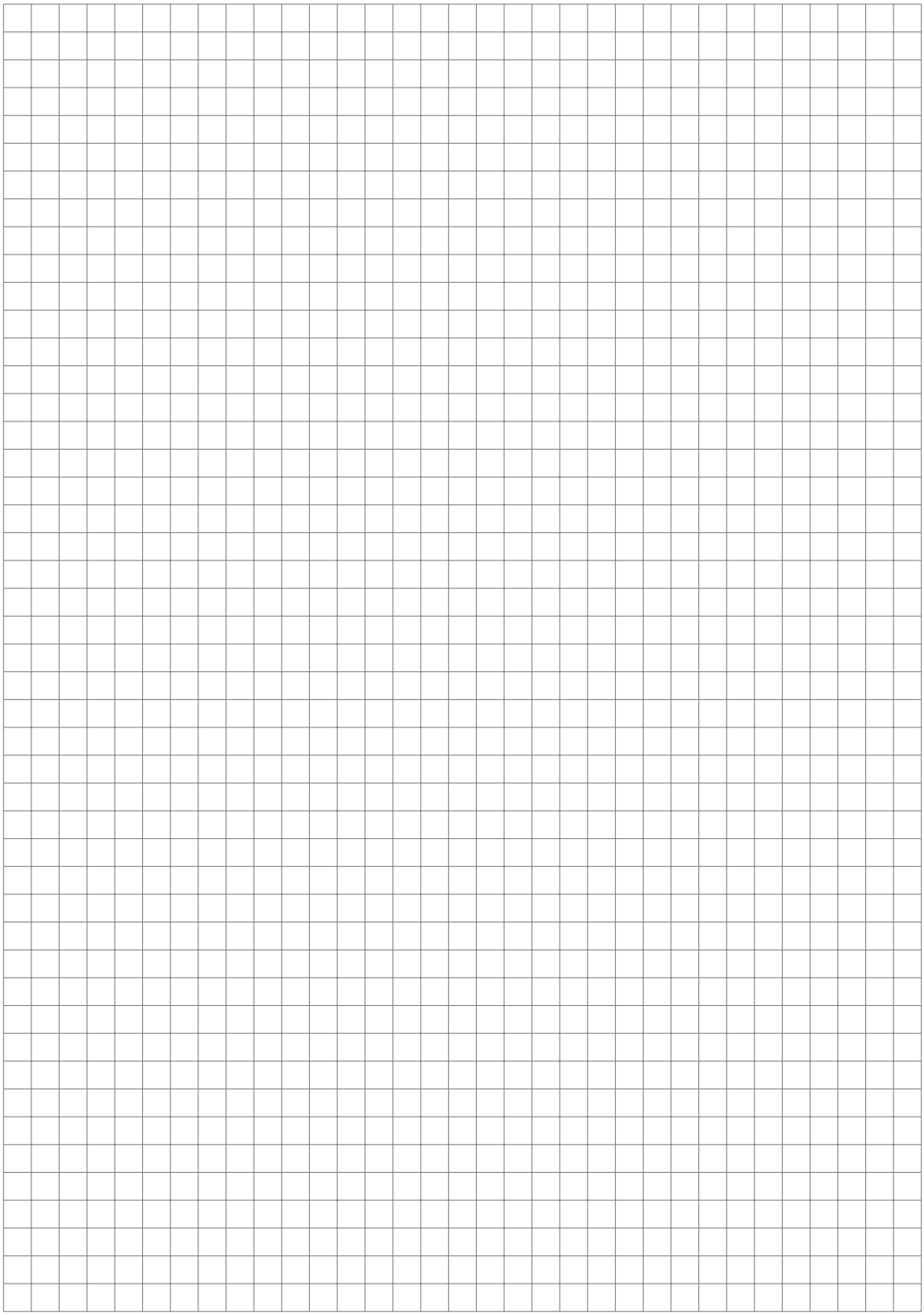
4 Künstliche Neuronale Netze

4.1 Sie möchten das abgebildete neuronale Netz mit dem stochastischen *Gradient Descent* Algorithmus trainieren. Das aktuelle Trainingsbeispiel ist durch den Vektor $\boldsymbol{x} = (0, 1)^\intercal$ gegeben. Das dazugehörige Label ist *one-hot* kodiert und lautet $\boldsymbol{y} = (1, 0)^\intercal$.

Aufbau des Netzes: Das Netz besteht aus einem *Hidden Layer* mit *ReLU*-Aktivierung und einem *Output Layer* mit Sigmoid-Aktivierung. Als Fehlerfunktion wird der quadratische Fehler benutzt.

Berechnen Sie bitte den Gewichtsgradienten $\frac{\partial \mathcal{J}}{\partial \theta}$ für θ (unten grau hinterlegt) und geben Sie den neuen Wert des Gewichts θ bei einer Lernrate von $\alpha = 0.5$ an. **(7 p)**





Maximal erreichbare Punkte für Aufgabe 4: **7 Punkte**

5 Evaluation von Machine Learning Modellen

Sie evaluieren einen binären Klassifikator auf einem Testdatensatz. Die Tabelle zeigt sowohl die Vorhersagen des Modells, als auch die korrekten Labels. Mit $\oplus$ wird die positive Klasse, mit $\ominus$ die negative Klasse bezeichnet.

Beispiel	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Wahrscheinlichkeit	0.40	0.95	0.60	0.45	0.75	0.55	0.70	0.20	0.52	0.15
Vorhersage	$\ominus$	$\oplus$	$\oplus$	$\ominus$	$\oplus$	$\oplus$	$\oplus$	$\ominus$	$\oplus$	$\ominus$
Gold Label	$\oplus$	$\oplus$	$\ominus$	$\ominus$	$\oplus$	$\oplus$	$\ominus$	$\oplus$	$\ominus$	$\ominus$

5.1 Vervollständigen Sie die folgende Konfusionsmatrix! **(2 p)**

Konfusionsmatrix		**vorhergesagt**	
		$\oplus$	$\ominus$
gold	$\oplus$		
	$\ominus$		

5.2 Berechnen Sie die *Accuracy* Ihres Modells. **(1 p)**

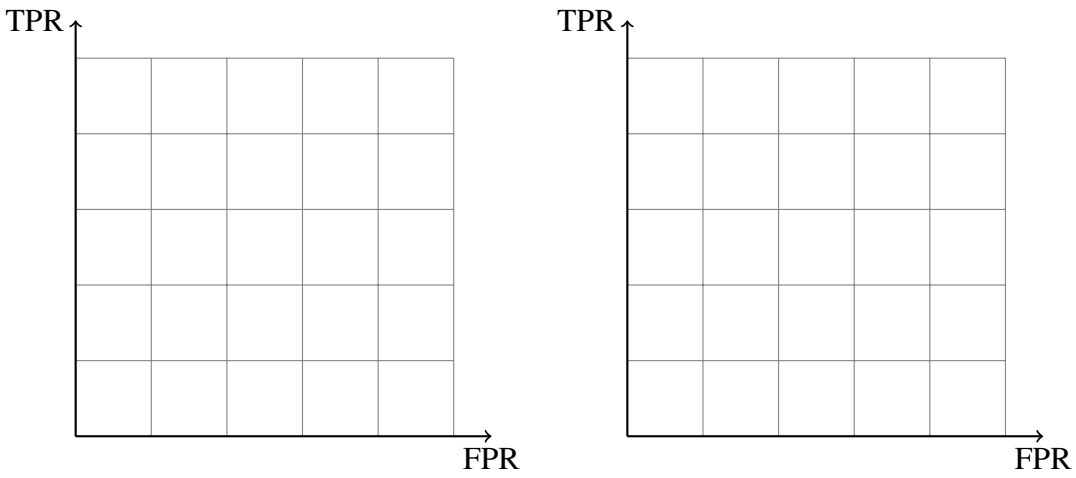
5.3 Wie hoch ist der *Recall* Ihres Modells? **(1 p)**

5.4 Wie hoch ist die *Precision* Ihres Modells? **(1 p)**

5.5 Bestimmen Sie den *F1-Score*! **(1 p)**

5.6 Zeichnen Sie die *ROC*-Kurve und berechnen Sie die *AUC*.

Das rechte Koordinatensystem ist für den Fall, dass das Zeichnen der ROC-Kurve nicht auf Anhieb funktioniert. **(3 p)**



5.7 Welchen Vorteil hat der *F1-Score* gegenüber der *Accuracy*? **(1 p)**

Maximal erreichbare Punkte für Aufgabe 5: **10 Punkte**

Zusätzlicher Platz für Anmerkungen:

Maximal erreichbare Punkte für die Klausur: **60 Punkte**