Hochschule München University of Applied Sciences

Prüfung Maschinelles Lernen (DC) - AUFGABENSAMMLUNG

Sommersemester 2023

Prof. Dr. Sarah Brockhaus

Bearbeitungzeit: 90 min

Hilfsmittel: nicht-programmierbarer Taschenrechner, ein beidseitig handbeschriebenes DIN A4 Blatt

Die Angabe ist vollständig wieder abzugeben.

Tragen Sie Ihre Rechnungen und Ergebnisse auf dieser Angabe ein.

Bitte kontrollieren Sie, ob Sie eine vollständige Aufgabenstellung mit 9 Seiten erhalten haben.

Name:	Vorname:
Matrikelnummer:	Platznummer:
Studiengang: ◯ Informatik ◯ Data Science & Scientific Computing ◯ Wirtschaftsinformatik	
erreichte Punktzahl:	Note:
Unterschrift Prüfer:	Zweitprüfer:

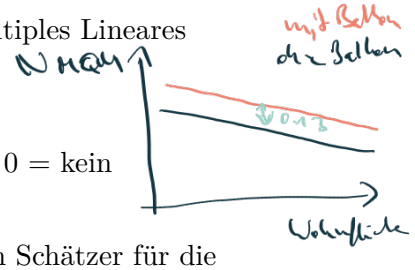
Aufgabe	1	2	3	4	5	6	$\sum$
Punkte	15	15	15	15	15	15	90
Erreichte Punkte							

Aufgabe 1: **(15 Punkte)**

Gegeben sind Daten aus dem Münchner Mietspiegel von 2019. Wir beschränken uns hier auf Wohnungen, die nach 1966 gebaut wurden und in durchschnittlicher Wohnlage liegen. Damit erhalten wir einen Datensatz mit Angaben zu $n = 839$ Wohnungen.
Wir möchten nun die Nettomiete je Quadratmeter (in Euro) über folgendes Multiples Lineares Regressionsmodell modellieren:

$$E(y|x) = w_0 + w_1 x_1 + w_2 x_2,$$

mit x_1: Wohnfläche in Quadratmetern und x_2: 1 = Balkon/Terrasse vorhanden, 0 = kein Balkon/Terrasse.



a) Nach welchem Kriterium werden im linearen Regressionsmodell die optimalen Schätzer für die Parameter $w_0, \ldots, w_p$ bestimmt?

Kleinste-Quadrate-Kriterium

b) Sie erhalten die folgenden Schätzer für die Parameter: $w_0 = 16.79$, $w_1 = -0.03$, $w_2 = 0.13$. Interpretieren Sie w_1 und w_2. Ist der Intercept w_0 hier sinnvoll interpretierbar? Begründen Sie Ihre Antwort kurz.

w_1: NMQM ist um 3 cent billiger, je QM, den die Wohng größer ist, gegeben alle anderen Features bleiben gleich (gegeben keine Änderung bei Balkon/Terrasse)

w_2: NMQM ist für eine Wohnung mit Balkon oder Terrasse im Schnitt um 13 cent teuer, als eine Wohnung ohne Balkon/Terrasse, gegeben die Wohnfläche ist gleich.

→ Nein, denn Intercept wäre NMQM für eine Wohng mit 0 Quadratmeter und ohne Balkon/Terrasse; aber Wohng mit 0 QM existiert nicht.

c) Sagen Sie die Nettomiete je Quadratmeter für eine Wohnung mit Balkon vorher, die eine Wohnfläche von 55 Quadratmetern hat.

$\hat{y} = 16.79 - 0.03 * 55 + 0.13 * 1 = 15,27 €$

d) Sie erhalten ein R^2 von 0.19. Interpretieren Sie diesen Wert. Welchen Wertebereich hat R^2 im Allgemeinen?

Wertebereich von R^2 i.d.R $[0,1]$
Das Modell erklärt 19% der Varianz in der Zielgröße (und ist damit verbesserungswürdig).

e) Erklären Sie die Grundidee von Ridge-Regression in ein bis zwei Sätzen.

Bei der Ridge-Regression optimiert man ein penalisiertes KQ-Kriterium. Der Penalisierungsterm bestraft die Größe der Parameter w_j mit $\lambda \cdot \sum w_j$ mit λ kann man die Modellkomplexität mit einem Parameter stufenlos einstellen.

penalisiertes KQ-Kriterium:

$$\frac{1}{2}\sum_{i=1}^{n}\left(y^{(i)} - (w_0 + w_1 x_1^{(i)} + \ldots + w_p x_p^{(i)})\right)^2 + \lambda \sum_{j=0}^{p} w_j^2$$

bestraft

λ steuert Kompromiss aus Datenanpassung & Modellkomplexität.

KQ Anpassung an Daten

Penalisierung Modellkomplexität

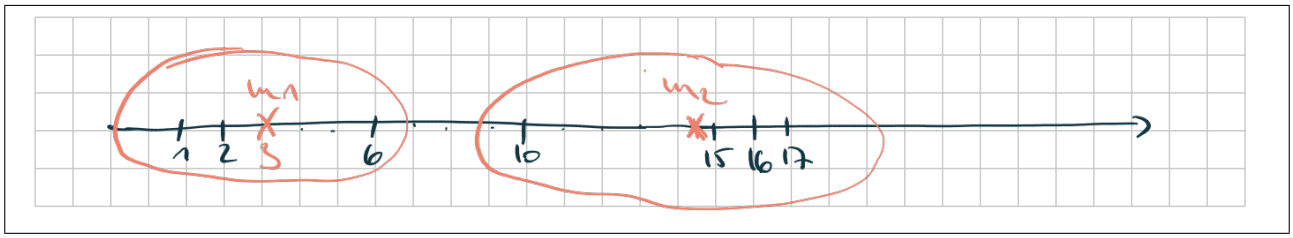
Aufgabe 2: **(15 Punkte)**

a) Was ist die Grundidee von Clustering?

Finden von Gruppen in den Daten, sodass die Beobachtung innerhalb jeder Gruppe (= Cluster) möglichst ähnlich zueinander sind und die Beobachtungen zwischen den Gruppen sich möglichst stark unterscheiden.

Für ein Clustering in zwei Clustern sind die Zahlen 1, 2, 6, 10, 15, 16, 17 gegeben.

b) Zeichnen Sie die Daten auf einen Zahlenstrahl ein (eine Einheit = ein Kästchen).



c) Bestimmen Sie mit Hilfe des k-Means Algorithmus (Abstandsfunktion ist absolute Differenz; Clusterzentren sind arithmetisches Mittel) die beiden Cluster mit den Anfangszentren 2 und 9. Zeichnen Sie die beiden entstehenden Cluster mit ihren Zentren in Ihre obige Skizze ein.

x	Abstand zu 2	Abstand zu 9	Zuordnung zu Cluster ⊛	Abstand zu 1,5	Abstand zu 12,8	Zuordnung zu Cluster ⊛⊛	Abstand zu 3	Abstand zu 14,5	Zuordnung Startzt.
1	1	8	1			1			1
2	0	7	1	0.5		1	1		1
6	4	3	2	4,5	6,8	1	3		1
10	8	1	2	8,5	2,16	2	7	4,5	2
15		6	2			2		0,5	2
16			2			2			2
17			2			2			2

⊛ → neue Clusterzentren: $\frac{1+2}{2} = 1.5$

$\frac{6+10+15+16+17}{5} = \frac{64}{5} = 12{,}8$

⊛⊛ → neue Clusterzentren: $\frac{1+2+6}{3} = 3$

$\frac{10+15+16+17}{4} = 14{,}5$

d) Ist k-Means-Clustering ein deterministischer Algorithmus? Begründen Sie Ihre Antwort kurz.

Nein, weil die initiale Zuordnung zufällig erfolgt und damit auch unterschiedliche Cluster ergeben können.

e) Warum können Sie das Ergebnis von k-Means-Clustering nicht in einem Dendrogram darstellen?

Weil beim k-Means-Clustering keine hierarchische Struktur zwischen den Clustern ist.

Zum Üben: Verwenden Sie Anfangszentren 8 und 14 für einen k-Means-Algorithmus wie in (c).

Aufgabe 3: **(15 Punkte)**

a) Entscheiden Sie ob die folgenden Aussagen über binäre **Entscheidungsbäume** je richtig oder falsch sind und begründen Sie jeweils kurz Ihre Antwort.

1. Entscheidungsbäume sind robust gegenüber kleinen Veränderungen in den Trainingsdaten.

Falsch, - greedy Algorithmus; lokale Entscheidungen.

2. Streng monotone Transformationen ~~der~~ von stetigen Features führen zu äquivalenten Entscheidungsbäumen. Insbesondere bleiben die Prognosen für die Zielgröße gleich.

Richtig, da streng monotone Trafos die möglichen binären Splits nicht ändern.

3. Entscheidungsbäume können Interaktionen zwischen den Features gut modellieren.

Richtig, da Interaktionen automatisch durch die Baumstruktur entstehen.

4. Beim Stutzen stoppt man den Aufbau des Baumes bevor die Blätter minimale Unreinheit erreicht haben.

Falsch, beim Stutzen baut man den Baum zunächst komplett auf und reduziert dann geeignet (=stutzen)

b) Entscheiden Sie ob die folgenden Aussagen über das **Lineare Regressionsmodell** je richtig oder falsch sind und begründen Sie jeweils kurz Ihre Antwort.

1. Streng monotone Transformationen der ~~Features~~ stetigen Features führen zu äquivalenten Regressionsmodellen. Insbesondere bleiben die Prognosen für die Zielgröße gleich.

Falsch, z.B. $\log(x)$ führt zu einem anderen Modell als x zu wundern, da man den logarithmischen Effekt von x auf y modelliert.

2. Beim KQ-Kriterium (Kleinste Quadrate) werden die quadrierten Abstände zwischen den Beobachtungen $(x^{(i)}, y^{(i)})$, $i = 1, \ldots, n$ und der geschätzten Gerade minimiert.

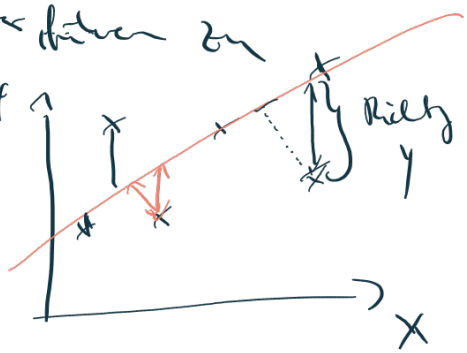
Falsch, man minimiert die quadrierten Abstände in Richtung von y.

3. Der Fit eines multiples lineares Regressionsmodells, d.h. eines Regressionsmodells mit mehr als einem Feature, kann durch eine Hyperebene veranschaulicht werden.

Ja, da der Modell linear in allen (transformierten) Features ist.

4. Lineare Transformationen der stetigen Features führen zu äquivalenten Modellen.

Richtig,



$y = w_0 + w_1 x_1$

$y = w_0 + w_1^* (ax + b) = w_0 + w_1 a x + w_1 b$

c) Entscheiden Sie ob die folgenden Aussagen über **Clustering** je richtig oder falsch sind und begründen Sie jeweils kurz Ihre Antwort.

1. Bei Hierarchischen Clustering-Verfahren muss vor der Schätzung des Modells die Anzahl an Clustern festgelegt werden.

Falsch, da hierarchische Verfahren alle möglichen Anzahlen von Clustern bestimmen.

2. Bei Hierarchischen Clustering-Verfahren kann man je nach zufälliger Initialisierung unterschiedliche Ergebnisse für die Cluster erhalten.

Falsch, hierarchische Clustering-Verfahren sind deterministisch; es gibt keine zufällige Initialisierung.

d) Entscheiden Sie ob die folgenden Aussagen über **KNN** (k-Nearest Neighbors, k-nächste Nachbarn) je richtig oder falsch sind und begründen Sie jeweils kurz Ihre Antwort.

1. In vielen Anwendungen ist es sinnvoll, alle Features zu standardisieren, bevor man KNN anwendet.

Richtig, damit die Abstände über verschiedene Features hinweg vergleichbar sind.

2. Je kleiner die Anzahl k an nächsten Nachbarn im KNN, desto glatter wird die Schätzung, und umso eher läuft man in den Bereich des Underfittings.

Falsch, da je größer k, umso größer die Nachbarschaft, umso glatter die Schätzung.

3. KNN ist robust gegen das Hinzufügen von irrelevanten Features.

Falsch, weil alle Features gleichermaßen zur Berechnung der Nachbarschaft verwendet werden.

e) Entscheiden Sie ob die folgenden Aussagen über **Logistische Regression** je richtig oder falsch sind und begründen Sie jeweils kurz Ihre Antwort.

1. Logistische Regression ist eine Methode aus dem Bereich des Unüberwachten Lernens.

Falsch, logistische Regression modelliert eine binäre Zielgröße.

2. Im Allgemeinen,lässt sich die absolute Größe der geschätzten (Steigungs-)Parameter als Variablenwichtigkeit interpretieren.

Falsch, da die Größe der Parameter auch von der Skala der Features abhängt.

3. Logistische Regressionsmodelle sind geeignet, um metrische Zielgrößen zu modellieren.

Falsch, logistische Regressionsmodelle sind für binären Response geeignet.

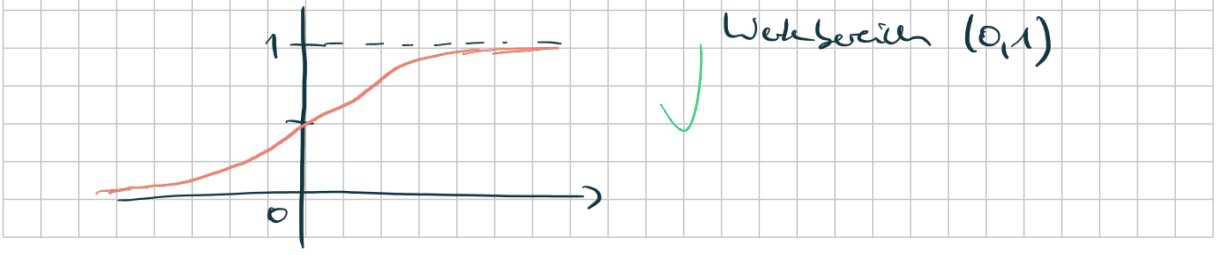
Aufgabe 4: **(15 Punkte)**

Wir betrachten einen Klassifikator, genauer ein logistisches Regressionsmodell,

$$f : \mathbb{R} \to [0, 1], \quad f(x) = \frac{e^{w_1 x + w_0}}{1 + e^{w_1 x + w_0}}$$

bei dem ausgehend von der Anzahl der Länder in denen das Patent gültig ist (x), bestimmt werden soll, ob gegen das Patent Einspruch erhoben wird (Klasse 1) oder nicht (Klasse 0). Nach dem Training erhalten wir $w_0 = -1.2$ und $w_1 = 0.53$.

a) Erstellen Sie eine Skizze der Funktion $g(x) = \frac{e^x}{1+e^x}$. Geben Sie den Wertebereich dieser Funktion explizit an.



b) Interpretieren Sie den geschätzten Parameter $w_1 = 0.53$, sowie den Exponenten dieses Parameters, also $e^{w_1} = e^{0.53} = 1.70$.

Je höher die Anzahl der Länder, in denen das Patent gültig ist, umso höher die Wahrscheinlichkeit dass Einspruch erhoben wird.

Um jedes Land mehr in dem das Patent gültig ist, steigt die Chance für Einspruch multiplikativ um den Faktor 1,7.

c) Angenommen, wir kennen die Anzahl der Länder, in denen ein neues Patent gelten wird. Was ist die Bedeutung des Ausgabewertes $f(x)$ und wie wird dieser Wert für die Klassifikation verwendet?

$f(x)$ ist die geschätzte Wahrscheinlichkeit, dass Einspruch erhoben wird.

Für die Klassifikation verwendet man üblicherweise

(nein) 0, wenn $f(x) < 0.5$

(ja) 1, wenn $f(x) \geq 0.5$.

d) Nennen Sie je einen Vorteil und einen Nachteil von logistischen Regressionsmodellen.

Vorteil: interpretierbares Modell;

Nachteil: Starke Strukturannahme durch linearen Prädiktor

Entropie: $-\sum_j p_j \log_2(p_j)$

Aufgabe 5: **(15 Punkte)**

Gegeben sei der Datensatz

$$\mathcal{D} = \{([\text{klein}, \text{süß}]^T, \text{ja}), ([\text{groß}, \text{süß}]^T, \text{ja}), ([\text{klein}, \text{sauer}]^T, \text{ja}), ([\text{groß}, \text{sauer}]^T, \text{nein})\}.$$

a) Erstellen Sie einen Entscheidungsbaum mit der Entropie als Unreinheitsmaß und berechnen Sie für jeden Knoten die Unreinheit $i(N)$ und für jedes Splitting die Verbesserung der Unreinheit $\Delta i(N)$. Zeichnen Sie den Entscheidungsbaum und geben Sie für jedes Blatt an, welche Entscheidung dort getroffen wird.

klein	süß	ja
groß	süß	ja
klein	sauer	ja
groß	sauer	nein

Unreinheit in Wurzel:

$$i(N) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0{,}811$$

- Split in klein vs. groß:

$$\Delta i(N) = i(N) - P_L\, i(L) - P_R\, i(R) =$$

$$= 0{,}811 - \frac{1}{2}\underbrace{\left(-0\log_2 0 - 1\log_2 1\right)}_{0} - \frac{1}{2}\underbrace{\left(\underbrace{-\frac{1}{2}\log_2\frac{1}{2}}_{1/2} \underbrace{-\frac{1}{2}\log_2\frac{1}{2}}_{1/2}\right)}_{1}$$

$$= 0.811 - 0 - \frac{1}{2} = 0.311$$

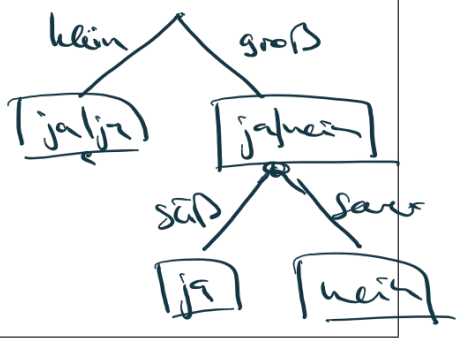
- Split in süß vs sauer:

$$\Delta i(N) = \ldots = 0.811 - 0 - \frac{1}{2} = 0.311$$

→ beide Splits sind gleich gut, wähle Split in klein vs groß:

- im Ast klein: ja
- im Ast mit groß:

Split in süß vs sauer, da einzig möglicher Split & dieser Split führt zu einer Verbesserung.



b) Angenommen Sie möchten Overfitting reduzieren. Erklären Sie anschaulich eine Möglichkeit, dies im Fall eines Entscheidungsbaums zu tun.

Overfitting kann durch frühes Abbrechen nach einem Kriterium, wie minimale Zahl an Beobachtungen je Blatt, reduziert werden.

Aufgabe 6: **(15 Punkte)**

a) Ein Klassifikator klassifiziert die Testdatenpunkte $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$ und $\mathbf{x}^{(3)}$ zu den Klassen -1, $+1$ und $+1$, wobei die echten Klassen $+1$, -1 und $+1$ sind. Berechnen Sie die Genauigkeit, Fehlerrate, Präzision und Trefferquote dieses Klassifikators bzgl. dieser Testdaten.

	wahr	Prognose
$x^{(1)}$	+1	−1
$x^{(2)}$	−1	+1
$x^{(3)}$	+1	+1

Konfusionsmatrix:

wahr \ Prognose		
	tp: 1	fn: 1
	fp: 1	tn: 0

Genauigkeit: $\frac{t_p+t_n}{t_p+f_n+f_p+t_n} = \frac{1}{1+1+1} = \frac{1}{3}$

Fehlerrate: $\frac{f_p}{t_p+f_p} = \frac{1}{1+1} = \frac{1}{2}$

Trefferquote: $\frac{t_p}{t_p+f_n} = \frac{1}{1+1} = \frac{1}{2}$

b) Erklären Sie knapp und präzise die 3-fache Kreuzvalidierung anhand des Beispielsdatensatzes $\mathcal{D} = \{\mathbf{x}^{(i)} \mid 1 \leq i \leq 9\}$. Sie dürfen annehmen, dass die Datenpunkte gut durchmischt gewählt wurden.

Bei der 3-fachen CV* wird der Datensatz in 3 gleich große Blöcke unterteilt, hier z.B.

$\{x^{(1)}, x^{(2)}, x^{(3)}\}$, $\{x^{(4)}, x^{(5)}, x^{(6)}\}$, $\{x^{(7)}, x^{(8)}, x^{(9)}\}$

und dann wird je einer dieser Blöcke als Testdatensatz verwendet und die verbleibenden zwei Blöcke als Trainingsdatensatz, d.h. das Modell wird drei mal neu gefittet und getestet

*CV = Kreuzvalidierung

c) Die folgende Abbildung zeigt eine typische Lernkurve mit dem Fehler über die Zeit. Tragen Sie die fehlenden Beschriftungen ein (zwei Arten von Daten, zwei Arten von Anpassungsproblemen an die Daten) und markieren Sie, wann Sie idealerweise mit dem Training aufhören.