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?

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.

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

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

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

Aufgabe 2: **(15 Punkte)**

a) Was ist die Grundidee von Clustering?

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.

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

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

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.

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

3. Entscheidungsbäume können Interaktionen zwischen den Features gut modellieren.

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

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 führen zu äquivalenten Regressionsmodellen. Insbesondere bleiben die Prognosen für die Zielgröße gleich.

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.

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

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.

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

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.

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.

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

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.

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

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

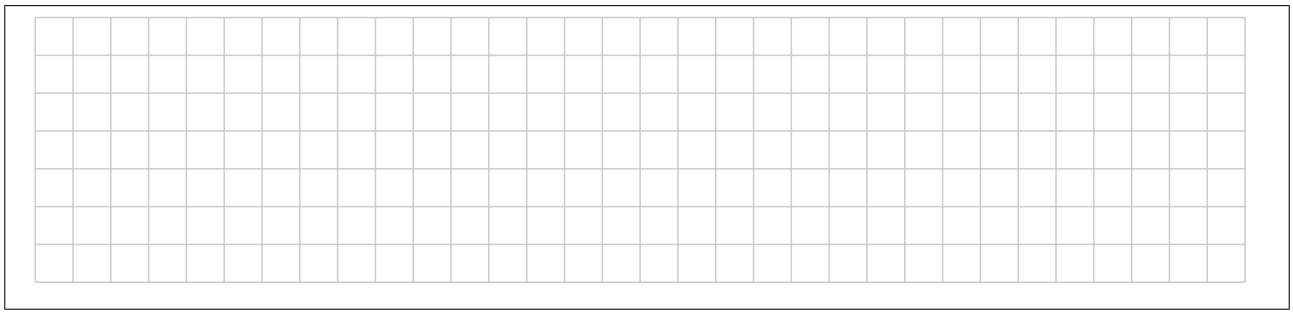
Aufgabe 4: **(15 Punkte)**

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 $e^{w_1} = e^{0.53} = 1.70$.

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?

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

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.

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

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.

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.

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.