Data Mining und Maschinelles Lernen

Prof. Kristian Kersting
Zhongjie Yu
Johannes Czech

TECHNISCHE UNIVERSITÄT DARMSTADT

Sommersemester 2021
18. Juni 2021
Übungsblatt 7

Diese Übung wird am **24.06.2021** um **13:30 Uhr** besprochen und **nicht bewertet**.

Aufgabe	1	2	3	4	5
Maximal Punktzahl	16	15	19	6	7
Erreichte Punktzahl					

Benötigte Dateien
Alle benötigten Datensätze und Skriptvorlagen finden Sie in unserem Moodle-Kurs:
https://moodle.informatik.tu-darmstadt.de/course/view.php?id=1058

Theoretische Aufgaben
Bei theoretischen Übungsaufgaben können Sie Ihren Lösungsvorschlag in LaTeX formatieren. Nutzen Sie hierfür die LaTeX-Vorlage und die vorgesehene Blöcke.

```
\begin{solution}
% your solution goes here
\end{solution}
```

Aufgabe 7.1: Entscheidungsbäume - ID3 Algorithmus (16)

Die folgende Tabelle zeigt die Entscheidung, ob Baseball gespielt wird, basierend auf vier Wetterattributen.

Tabelle 1: Trainingsdatensatz, ob Baseball gespielt wird basierend auf der Wetterlage.

Tag	Ausblick (A)	Temperatur (T)	Luftfeuchtigkeit (L)	Wind (W)	Spielt Baseball (B)
T1	Sonnig	Warm	Hoch	Schwach	Nein
T2	Sonnig	Warm	Hoch	Stark	Nein
T3	Bewölkung	Warm	Hoch	Schwach	Ja
T4	Regen	Mild	Hoch	Schwach	Ja
T5	Regen	Kühl	Normal	Schwach	Ja
T6	Regen	Kühl	Normal	Stark	Nein
T7	Bewölkung	Kühl	Normal	Stark	Ja
T8	Sonnig	Mild	Hoch	Schwach	Nein
T9	Sonnig	Kühl	Normal	Schwach	Ja
T10	Regen	Mild	Normal	Schwach	Ja
T11	Sonnig	Mild	Normal	Stark	Ja
T12	Bewölkung	Mild	Hoch	Stark	Ja
T13	Bewölkung	Warm	Normal	Schwach	Ja
T14	Regen	Mild	Hoch	Stark	Nein

Die Aufgabe ist es folgende Frage zu beantworten: *Unter welchen Bedingungen wir Baseball gespielt?*

Tabelle 2: Vorhersage-Datensatz, ob Baseball gespielt wird.

Tag	Ausblick (A)	Temperatur (T)	Luftfeuchtigkeit (L)	Wind (W)	Spielt Baseball (B)
T15	Sonnig	Mild	Hoch	Schwach	?
T16	Bewölkung	Mild	Normal	Schwach	?
T17	Regen	Kühl	Normal	Stark	?

7.1a) ID3 Algorithmus (10 Punkte)

Erstellen Sie den Entscheidungsbaum mittels des ID3 Algorithmus. Berechnen Sie dabei die **Entropie** und den **Informationsgewinn** (engl. *gain*) der Attribut-Selektion für jeden Schritt. Verwenden Sie bei der Berechnung der Entropie den Logarithmus zur Basis 2, Logarithmus-Dualis.

Hinweis: Sie können **Spielt Baseball (B)** mit B kennzeichnen. Der Informationsgewinn ist nach Vorlesung wie folgt definiert: *Differenz zwischen den Informationen der Beispiele mit und ohne die Aufteilung durch* X_j.

Bsp. Informationsgewinn für Aufteilung des Wurzelknotens nach dem Merkmal *Ausblick*.

$$\begin{aligned} Gain\left(B, Ausblick\right) = Entropy\left(B\right) \\ &- \frac{B_{\text{Ausblick=Sonnig}}}{B} \cdot Entropy\left(B_{\text{Ausblick=Sonnig}}\right) \\ &- \frac{B_{\text{Ausblick=Regen}}}{B} \cdot Entropy\left(B_{\text{Ausblick=Regen}}\right) \\ &- \frac{B_{\text{Ausblick=Bewölkung}}}{B} \cdot Entropy\left(B_{\text{Ausblick=Bewölkung}}\right) \end{aligned}$$

7.1b) Visualisierung (3 Punkte)

Erstellen Sie eine Visualsierung (Plot oder eingefügte Zeichnung) des Entscheidungsbaumes aus Aufgabenteil a).

7.1c) Vorhersage (3 Punkte)

Geben Sie anhand ihres Entscheidungsbaumes eine Vorhersage für die Tage 15 bis 17 aus Tabelle 2, ob Baseball gespielt wird.

Übungsblatt 7, Data Mining und Maschinelles Lernen
Nachname, Vorname: ____________________ Matrikelnummer: ␣␣␣␣␣␣␣

Aufgabe 7.2: AdaBoost (15)

In dieser Aufgabe werden Sie AdaBoost auf die gegebenen Trainingsbeispiele aus der Tabelle 3 anwenden.

Tabelle 3: Datensatz mit zwei Merkmalen und zwei Zielklassen.

$\mathbf{x_1}$	$\mathbf{x_2}$	**Klasse**
1	5	+
2	2	+
5	8	+
6	10	+
8	7	+
3	1	-
4	6	-
7	4	-
9	3	-
10	9	-

Entscheidungsstümpfe mit ganzzahligem Schwellwert (z.B. $\mathbf{x_1} \leq T \Rightarrow +$ oder $\mathbf{x_1} > T \Rightarrow +$) sollen als Basis-Lerner verwendet werden. Der Basis-Lerner minimiert die Summe der Gewichtungen der falsch klassifizierten Beispiele aus allen möglichen Aufteilungen. Für ein Unentschieden wählen Sie die erste gefundene Übereinstimmung, beginnend mit Entscheidungsstümpfen für $\mathbf{x_1}$ und dann $\mathbf{x_2}$.

Verwenden Sie die Formel:

$$\alpha_i = \frac{1}{2} \log \left(\frac{1 - err_i}{err_i} \right) \tag{1}$$

zur Berechnung von α_i.

Verwenden Sie die Formel

$$w_n^{(i+1)} = w_n^{(i)} \exp\{-\alpha_i \; t_n \; y_i(\mathbf{x}_n)\} \tag{2}$$

für das Update der Gewichte w_n, wobei i den Iterationindex, n den Datenindex und t das Groundtruthlabel $\in \{-1, +1\}$ beschreibt.

Hinweis: Mit $\log(\dots)$ ist hier der Logarithmus zur Basis e, $\ln(\dots)$, gemeint.

7.2a) Algorithmus (12 Punkte)

Zeigen Sie die Ausführung des Adaboost Algorithmus für die **ersten beiden** Iterationen. Geben Sie dabei die **Fehler** (Summe der Gewichtungen der falsch klassifizierten Beispiele) für die möglichen Entscheidungsgrenzen von 1 bis 10 an, sowie die **Gewichtung** jedes Datenpunktes vor und nach Normalisierung an.

7.2b) Gesamtmodell (3 Punkte)

Geben Sie das Gesamtmodell $f(x)$ nach zwei Iterationen an.

Aufgabe 7.3: Naïve Bayes (19)

In dieser Aufgabe verwenden wir wieder den Baseball-Datensatz (s. Tabelle 1 und 2) und einen Naïve Bayes Klassifikator, um zu entscheiden ob Baseball gespielt wird oder nicht.

7.3a) Formel für Merkmalsausprägung (4 Punkte)

Zeigen Sie die Formel für $P(B = Ja \mid Merkmal)$ und $P(B = Nein \mid Merkmal)$ für den gegebenen Datensatz.

7.3b) Wahrscheinlichkeiten (6 Punkte)

Bestimmen Sie angesichts der oben genannten Trainingsdaten alle Wahrscheinlichkeiten, die erforderlich sind, um den Naïve Bayes Klassifikator für beliebige Vorhersagen, ob Baseball gespielt wird, anzuwenden.

7.3c) Vorhersage (9 Punkte)

Treffen Sie Vorhersagen nach Naïve Bayes für die Tage 15 bis 17 aus Tabelle 2, ob Baseball gespielt wird. Geben Sie dabei den Rechenweg an.

Aufgabe 7.4: K-Means (6)

Folgender Datensatz besteht aus 8 Punkten:

$$x_1 = (2,8), \quad x_2 = (2,5), \quad x_3 = (1,2), \quad x_4 = (5,8),$$
$$x_5 = (7,3), \quad x_6 = (6,4), \quad x_7 = (8,4), \quad x_8 = (4,7). \tag{3}$$

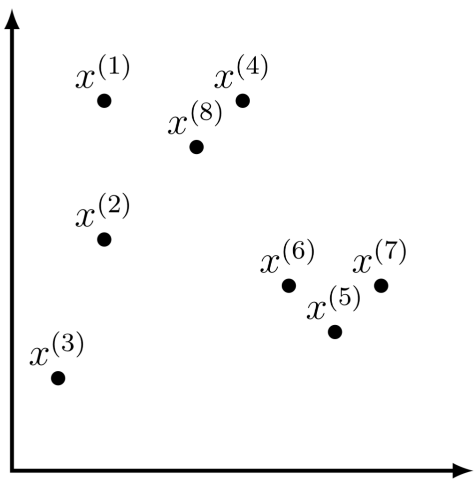


Abbildung 1: Visualisierung des K-Means Datensatzes

7.4a) K-Means Algorithmus (6 Punkte)

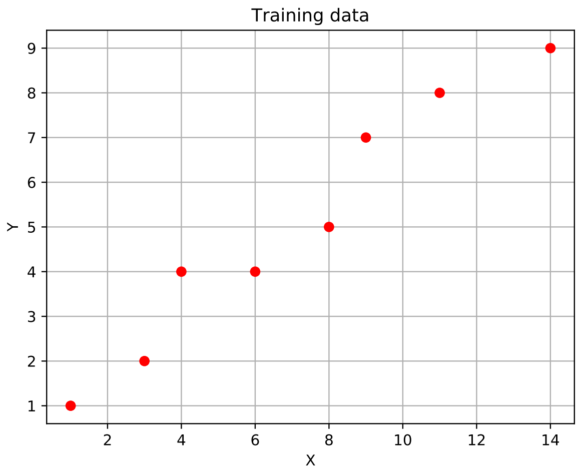
Benutzen Sie den K-Means Algorithmus mit der Euklidischen Distanz um diese 8 Datenpunkte in $K = 3$ Cluster einzuteilen. Die Cluster werden dabei als Cluster A, Cluster B und Cluster C beschrieben. Nehmen Sie dabei an, dass die Clusterzentren $\mu_A^{(0)}, \mu_B^{(0)}, \mu_C^{(0)}$ mit den Punkten x_3, x_6 und x_7 initialisiert sind. Führen Sie zwei Iterationen des K-Means Algorithmus durch und geben Sie die Koordinaten der Zentroide der Cluster an.

Aufgabe 7.5: Regressionsanalyse (7)

Gegeben sind folgende Datenpunkte:

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

Abbildung 2: Veranschaulichung der Datenpunkte zur linearen Regression.



Wir möchten eine Regression nach dem Prinzip der kleinsten Fehlrequadrate erstellen:

$$y = f(x) = \langle W, x \rangle + b\,. \tag{4}$$

Mit der Hilfe eines $(p+1)$-dimensionalen Vektors $\vec{x} = (1, x_1, \cdots, x_p)$ und $x \in \mathbb{R}^{1\times p}$, können wir b in dem Vektor W codieren:

$$y = f(x) = \left\langle W', \vec{x}^T \right\rangle, \tag{5}$$

wobei hier $W' \in \mathbb{R}^{2\times 1}$ und $\vec{x} \in \mathbb{R}^{1\times 2}$.

7.5a) Herleitung (5 Punkte)

Zeigen Sie, dass das optimale W':

$$W' = (\vec{X}^T\vec{X})^{-1}\vec{X}^T Y, \tag{6}$$

entspricht, wobei $\vec{X} \in \mathbb{R}^{n\times 2}$ und $Y \in \mathbb{R}^{n\times 1}$.

7.5b) Parameterbestimmung (2 Punkte)

Berechnen Sie W und b für den gegebenen Punktdatensatz. Die Inverse $(\vec{X}^T\vec{X})^{-1}$ muss dabei nicht manuell berechnet werden.