Data Mining und Maschinelles Lernen

Prof. Kristian Kersting
Zhongjie Yu
Johannes Czech

TECHNISCHE UNIVERSITÄT DARMSTADT

Sommersemester 2021
30. 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*.

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

Lösungsvorschlag:

Wahl des Wurzelknotens. **Spielt Baseball (B)** besteht aus 9 Ja- und 5 Nein-Beispielen, also

$$Entropy\,(B) = -\frac{9}{9+5}\log_2\frac{9}{9+5} - \frac{5}{9+5}\log_2\frac{5}{9+5} = 0.940 \tag{1}$$

Wir beginnen mit dem Attribut Ausblick. $Ausblick = Sonnig$ hat 2 positive und 3 negative Proben.

$$Entropy\left(B_{\text{Ausblick=Sonnig}}\right) = -\frac{2}{2+3}\log_2\frac{2}{2+3} - \frac{3}{2+3}\log_2\frac{3}{2+3} = 0.971 \tag{2}$$

Analog dazu:

$$Entropy\left(B_{\text{Ausblick=Regen}}\right) = -\frac{3}{3+2}\log_2\frac{3}{3+2} - \frac{2}{3+2}\log_2\frac{2}{3+2} = 0.971 \tag{3}$$

$$Entropy\left(B_{\text{Ausblick=Bewölkung}}\right) = -\frac{4}{4}\log_2\frac{4}{4} = 0 \tag{4}$$

$$\begin{aligned} Gain\,(B, Ausblick) = {} & Entropy\,(B) \\ & - \frac{B_{\text{Ausblick=Sonnig}}}{B} \cdot Entropy\left(B_{\text{Ausblick=Sonnig}}\right) \\ & - \frac{B_{\text{Ausblick=Regen}}}{B} \cdot Entropy\left(B_{\text{Ausblick=Regen}}\right) \\ & - \frac{B_{\text{Ausblick=Bewölkung}}}{B} \cdot Entropy\left(B_{\text{Ausblick=Bewölkung}}\right) \\ = {} & 0.940 - 5/14 \cdot 0.971 - 5/14 \cdot 0.971 - 4/14 \cdot 0 \\ = {} & 0.246 \end{aligned} \tag{5}$$

Nach dem gleichem Vorgehen erhalten wir:

$$\begin{aligned} &Gain\,(B, Temperatur) = 0.029 \\ &Gain\,(B, Luftfeuchtigkeit) = 0.151 \\ &Gain\,(B, Wind) = 0.048 \end{aligned} \tag{6}$$

Ausblick wird als Entscheidungsattribut im Wurzelknoten verwendet, da er den höchsten Informationsgewinn hat. Wir fahren mit den Unterteilungen “Sonnig”, “Regen” und “Bewölkung” des Wurzelknotens fort.

Für “Ausblick=Sonnig” erhalten wir für B_{Sonnig} 2 positive und 3 negative Beispiele.

$$Entropy\,(B_{\text{Sonnig}}) = -\frac{2}{2+3}\log_2\frac{2}{2+3} - \frac{3}{2+3}\log_2\frac{3}{2+3} = 0.971 \tag{7}$$

$$Entropy\,(B_{\text{Sonnig, Temperatur=Warm}}) = -\frac{2}{2}\log_2\frac{2}{2} = 0 \tag{8}$$

$$Entropy\left(B_{\text{Sonnig, Temperatur=Mild}}\right) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1 \tag{9}$$

$$Entropy\left(B_{\text{Sonnig, Temperatur=Kühl}}\right) = -\frac{1}{1}\log_2\frac{1}{1} = 0 \tag{10}$$

Danach,

$$Gain\,(B_{\text{Sonnig}}, Temperatur) = 0.971 - 2/5 \cdot 1 = 0.571 \tag{11}$$

Also,

$$\begin{aligned} &Gain\,(B_{\text{Sonnig}}, Luftfeuchtigkeit) = 0.971 \\ &Gain\,(B_{\text{Sonnig}}, Wind) = 0.019 \end{aligned} \tag{12}$$

Die Luftfeuchtigkeit wird als Entscheidungsattribut des Knotens gewählt, und alle Daten werden perfekt klassifiziert.

Wenden wir uns dem Knoten “Regen” zu, der 3 Ja- und 2 Nein-Beispiele hat.

$$Entropy\,(B_{Regen}) = -\frac{3}{2+3}\log_2\frac{3}{2+3} - \frac{2}{2+3}\log_2\frac{2}{2+3} = 0.971 \tag{13}$$

$$Entropy\,(B_{Regen,Temperatur=Warm}) = 0 \tag{14}$$

$$Entropy\,(B_{Regen,Temperatur=Mild}) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.918 \tag{15}$$

$$Entropy\,(B_{Regen,Temperatur=Khl}) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1 \tag{16}$$

Anschließend,

$$Gain\,(B_{Regen}, Temperatur) = 0.971 - 3/5 \cdot 0.918 - 2/5 \cdot 1 = 0.020 \tag{17}$$

Also,

$$\begin{aligned} &Gain\,(B_{Regen}, Luftfeuchtigkeit) = 0.020 \\ &Gain\,(B_{Regen}, Wind) = 0.971 \end{aligned} \tag{18}$$

Wind wird als Entscheidungsattribut des Knotens gewählt, und alle Daten werden perfekt klassifiziert. Für den Knoten “Bewölkung” werden alle Daten perfekt klassifiziert.

7.1b) Visualisierung (3 Punkte)

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

Lösungsvorschlag:

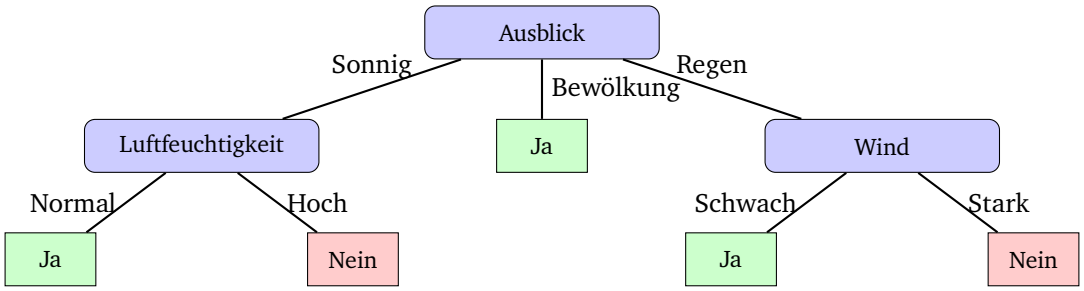


Abbildung 1: Entscheidungsbaum des ID3 Algorithmus.

7.1c) Vorhersage (3 Punkte)

Geben Sie anhand ihres Entscheidungsbaumes eine Vorhersage für die Tage 15 bis 17 aus Tabelle 3, ob Baseball gespielt wird.

Lösungsvorschlag:

Tabelle 3: Vorhersage-Datensatz, ob Baseball gespielt wird.

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

Aufgabe 7.2: AdaBoost (15)

In dieser Aufgabe werden Sie AdaBoost auf die gegebenen Trainingsbeispiele aus der Tabelle 4 anwenden.

Tabelle 4: 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:

$$\alpha_i = \frac{1}{2} \log \left(\frac{1 - err_i}{err_i} \right) \tag{19}$$

zur Berechnung von α_i.

Verwenden Sie die Formel

$$w_n^{(i+1)} = w_n^{(i)} \exp\{-\alpha_i \, t_n \, y_i(\mathbf{x}_n)\} \tag{20}$$

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.

Lösungsvorschlag:

Erste Iteration
Zunächst müssen wir den ersten Entscheidungsstumpf finden. Zu Beginn hat jeder Datenpunkt ein Gewicht von $\frac{1}{10}$.

Es gibt 5 Fälle mit dem kleinsten Fehler in der Tabelle, dargestellt in rot. Wir wählen den ersten, der in der Tabelle erscheint: $x \leq 2 \Rightarrow +$. Dann ist der Fehler in der ersten Iteration:

$$err_1 = \frac{3}{10} \tag{21}$$

Für α_1 ergibt sich:

$$\alpha_1 = \frac{1}{2} \log \left(\frac{1 - err_1}{err_1} \right) = \frac{1}{2} \log \left(\frac{7}{3} \right) \approx 0.424 \tag{22}$$

Wert	Fehler	
	$\mathbf{x_1} \leq$ Wert $\Rightarrow +$	$\mathbf{x_1} >$ Wert $\Rightarrow +$
1	4/10	6/10
2	3/10	7/10
3	4/10	6/10
4	5/10	5/10
5	4/10	6/10
6	3/10	7/10
7	4/10	6/10
8	3/10	7/10
9	4/10	6/10
10	5/10	5/10

Wert	Fehler	
	$\mathbf{x_2} \leq$ Wert $\Rightarrow +$	$\mathbf{x_2} >$ Wert $\Rightarrow +$
1	6/10	4/10
2	5/10	5/10
3	6/10	4/10
4	7/10	3/10
5	6/10	4/10
6	7/10	3/10
7	6/10	4/10
8	5/10	5/10
9	6/10	4/10
10	5/10	5/10

Anschließend werden die Gewichte aktualisiert:

$$w_i \Leftarrow \begin{cases} w_i \cdot e^{-\alpha_1} \approx 0.0654, & \text{wenn } w_i \text{ korrekt klassifiziert wurde} \\ w_i \cdot e^{\alpha_1} \approx 0.1528, & \text{wenn } w_i \text{ falsch klassifiziert wurde} \end{cases} \tag{23}$$

Danach normalisieren wir w_i mit $w_i = \frac{w_i}{\sum_{k=1}^{10} w_k}$.

$$w_i = \begin{cases} 0.071, & \text{wenn } w_i \text{ korrekt klassifiziert wurde} \\ 0.167, & \text{wenn } w_i \text{ falsch klassifiziert wurde} \end{cases} \tag{24}$$

Nun werden die Gewichte der Datenpunkte aktualisiert:

$\mathbf{x_1}$	$\mathbf{x_2}$	Gewichtung (w)	Klasse
1	5	0.071	+
2	2	0.071	+
5	8	0.167	+
6	10	0.167	+
8	7	0.167	+
3	1	0.071	-
4	6	0.071	-
7	4	0.071	-
9	3	0.071	-
10	9	0.071	-

Beachten Sie, dass die Summe der Gewichte hier 0.998 beträgt, was auf Rundungen zurückzuführen ist.

Zweite Iteration
Hier versuchen wir den zweiten Entscheidungsstumpf mit gewichteten Stichproben zu finden.

Zum Beispiel für $x \leq 1 \Rightarrow +$ erhalten wir
$(2, 2) - -weight = 0.071$,
$(5, 8) - -weight = 0.167$,
$(6, 10) - -weight = 0.167$ und
$(8, 7) - -weight = 0.167$
falsch klassifizierte Beispiele und der gewichtete Fehler ist 0.572.

Es gibt 3 Fälle mit dem kleinsten Fehler in der Tabelle, dargestellt in rot. Wir wählen den ersten, der in der Tabelle erscheint: $x \leq 8 \Rightarrow +$. Dann ist der Fehler in der zweiten Iteration:

$$err_2 = 0.213 \tag{25}$$

Für α_2 ergibt sich:

$$\alpha_2 = \frac{1}{2} \log\left(\frac{1 - err_2}{err_2}\right) = \frac{1}{2} \log\left(\frac{0.787}{0.213}\right) \approx 0.652 \tag{26}$$

Wert	Fehler	
	$\mathbf{x_1} \leq$ Wert $\Rightarrow +$	$\mathbf{x_1} >$ Wert $\Rightarrow +$
1	0.572	0.426
2	0.501	0.497
3	0.572	0.426
4	0.643	0.355
5	0.476	0.522
6	0.309	0.689
7	0.380	0.618
8	0.213	0.785
9	0.284	0.714
10	0.355	0.643

Wert	Fehler	
	$\mathbf{x_2} \leq$ Wert $\Rightarrow +$	$\mathbf{x_2} >$ Wert $\Rightarrow +$
1	0.714	0.284
2	0.643	0.355
3	0.714	0.284
4	0.785	0.213
5	0.714	0.284
6	0.785	0.213
7	0.618	0.380
8	0.451	0.547
9	0.522	0.476
10	0.355	0.643

Anschließend werden die Gewichte aktualisiert:

$$w_i \Leftarrow \begin{cases} w_i \cdot e^{-\alpha_2} \approx w_i \cdot 0.521, & \text{wenn } w_i \text{ korrekt klassifiziert wurde} \\ w_i \cdot e^{\alpha_2} \approx w_i \cdot 1.919, & \text{wenn } w_i \text{ falsch klassifiziert wurde} \end{cases} \tag{27}$$

Die Aktualisierung der Gewicht werden in folgender Tabelle dargestellt:

Alte Gewichtung	Neue Gewichtung	
	korrekt klassifiziert	falsch klassifiziert
0.071	0.037	0.136
0.167	0.087	0.320

Danach normalisieren wir w_i mit $w_i = \frac{w_i}{\sum_{k=1}^{10} w_k}$ und aktualisieren die Gewichtung der Datenpunkte.

x	y	Alte Gewichtung	Neue Gewichtung		Klasse
			Ohne Normalsierung	Nach Normalsierung	
1	5	0.071	0.037	0.045	+
2	2	0.071	0.037	0.045	+
5	8	0.167	0.087	0.106	+
6	10	0.167	0.087	0.106	+
8	7	0.167	0.087	0.106	+
3	1	0.071	0.136	0.166	-
4	6	0.071	0.136	0.166	-
7	4	0.071	0.136	0.166	-
9	3	0.071	0.037	0.045	-
10	9	0.071	0.037	0.045	-

Beachten Sie, dass die Summe der Gewichte hier 0.996 beträgt, was auf Rundungen zurückzuführen ist.

7.2b) Gesamtmodell (3 Punkte)

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

Lösungsvorschlag:

$$f(x) = sign(\alpha_1 \cdot g_1(x) + \alpha_2 \cdot g_2(x))$$

wobei g_i den Entscheidungsstumpf der jeweiligen Iteration darstellt.

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.

Lösungsvorschlag:

$$\begin{aligned} & P(Ja \mid Merkmal) \\ &= \frac{Prior \cdot Likelihood}{P(Merkmal)} \\ &= \frac{P(Ja) \cdot P(Ausblick \mid Ja) \cdot P(Temperatur \mid Ja) \cdot P(Luftfeutchtigkeit \mid Ja) \cdot P(Wind \mid Ja)}{P(Merkmal)} \end{aligned} \tag{28}$$

$$\begin{aligned} & P(Nein \mid Merkmal) \\ &= \frac{Prior \cdot Likelihood}{P(Merkmal)} \\ &= \frac{P(Nein) \cdot P(Ausblick \mid Nein) \cdot P(Temperatur \mid Nein) \cdot P(Luftfeutchtigkeit \mid Nein) \cdot P(Wind \mid Nein)}{P(Merkmal)} \end{aligned} \tag{29}$$

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.

Lösungsvorschlag:

Prior: $P(Ja) = 9/14, \quad P(Nein) = 5/14.$

Ausblick	$P(Ausblick \mid Ja)$	$P(Ausblick \mid Nein)$
Sonnig	2/9	3/5
Regen	3/9	2/5
Bewölkung	4/9	0/5

Temperatur	$P(Temperatur \mid Ja)$	$P(Temperatur \mid Nein)$
Warm	2/9	2/5
Mild	4/9	2/5
Kühl	3/9	1/5

Luftfeuchtigkeit	$P(Luftfeuchtigkeit \mid Ja)$	$P(Luftfeuchtigkeit \mid Nein)$
Hoch	3/9	4/5
Normal	6/9	1/5

Wind	$P(Wind \mid Ja)$	$P(Wind \mid Nein)$
Schwach	6/9	2/5
Stark	3/9	3/5

7.3c) Vorhersage (9 Punkte)

Treffen Sie Vorhersagen nach Naïve Bayes für die Tage 15 bis 17 aus Tabelle 3, ob Baseball gespielt wird. Geben Sie dabei den Rechenweg an.

Lösungsvorschlag:

Tag	Ausblick	Temperatur	Luftfeuchtigkeit	Wind	Spielt Baseball
T15	Sonnig	Mild	Hoch	Schwach	Nein

Ergebnis des Entscheidungsbaums: Nein

$$\begin{aligned} &Posterior(Ja) \propto Prior \cdot Likelihood \\ &= P(Ja) \cdot P(Sonnig \mid Ja) \cdot P(Mild \mid Ja) \cdot P(Hoch \mid Ja) \cdot P(Schwach \mid Ja) \\ &= \frac{9}{14} \cdot \frac{2}{9} \cdot \frac{4}{9} \cdot \frac{3}{9} \cdot \frac{6}{9} \\ &= 0.0141 \end{aligned} \tag{30}$$

$$\begin{aligned} &Posterior(Nein) \propto Prior \cdot Likelihood \\ &= P(Nein) \cdot P(Sonnig \mid Nein) \cdot P(Mild \mid Nein) \cdot P(Hoch \mid Nein) \cdot P(Schwach \mid Nein) \\ &= \frac{5}{14} \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{4}{5} \cdot \frac{2}{5} \\ &= 0.0274 \end{aligned} \tag{31}$$

Tag	Ausblick	Temperatur	Luftfeuchtigkeit	Wind	Spielt Baseball
T16	Bewölkung	Mild	Normal	Schwach	Ja

Ergebnis des Entscheidungsbaums: Ja

$$\begin{aligned} &Posterior(Ja) \propto Prior \cdot Likelihood \\ &= P(Ja) \cdot P(Bewoelkung \mid Ja) \cdot P(Mild \mid Ja) \cdot P(Normal \mid Ja) \cdot P(Schwach \mid Ja) \\ &= \frac{9}{14} \cdot \frac{4}{9} \cdot \frac{4}{9} \cdot \frac{6}{9} \cdot \frac{6}{9} > 0 \end{aligned} \tag{32}$$

$$\begin{aligned} &Posterior(Nein) \propto Prior \cdot Likelihood \\ &= P(Nein) \cdot P(Bewoelkung \mid Nein) \cdot P(Mild \mid Nein) \cdot P(Normal \mid Nein) \cdot P(Schwach \mid Nein) \\ &= \frac{5}{14} \cdot \frac{0}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{2}{5} \\ &= 0 \end{aligned} \tag{33}$$

Tag	Ausblick	Temperatur	Luftfeuchtigkeit	Wind	Spielt Baseball
T17	Regen	Kühl	Normal	Stark	Ja

Ergebnis des Entscheidungsbaums: Nein

$$\begin{aligned} &Posterior(Ja) \propto Prior \cdot Likelihood \\ &= P(Ja) \cdot P(Regen \mid Ja) \cdot P(Kuehl \mid Ja) \cdot P(Normal \mid Ja) \cdot P(Stark \mid Ja) \\ &= \frac{9}{14} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{6}{9} \cdot \frac{3}{9} \\ &= 0.0159 \end{aligned} \tag{34}$$

$$\begin{aligned} &Posterior(Nein) \propto Prior \cdot Likelihood \\ &= P(Nein) \cdot P(Regen \mid Nein) \cdot P(Kuehl \mid Nein) \cdot P(Normal \mid Nein) \cdot P(Stark \mid Nein) \\ &= \frac{5}{14} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{1}{5} \cdot \frac{3}{5} \\ &= 0.0034 \end{aligned} \tag{35}$$

Aufgabe 7.4: K-Means (6)

Folgender Datensatz besteht aus 8 Punkten:

$$\begin{aligned} x_1 &= (2,8), \quad x_2 = (2,5), \quad x_3 = (1,2), \quad x_4 = (5,8), \\ x_5 &= (7,3), \quad x_6 = (6,4), \quad x_7 = (8,4), \quad x_8 = (4,7). \end{aligned} \tag{36}$$

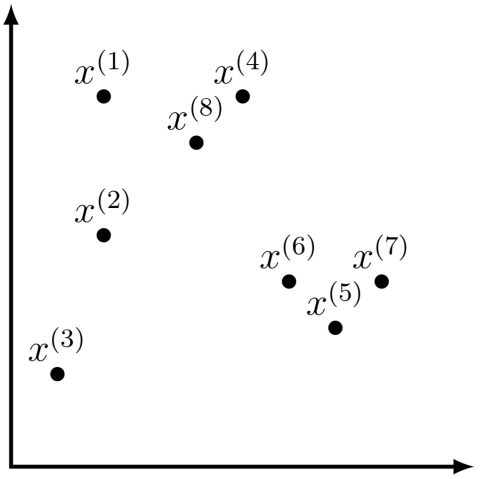


Abbildung 2: Visualisierung des K-Means Datensatzes

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.

Lösungsvorschlag:

1. Initialisierung:

$$\begin{aligned} \mu_A^{(0)} &= x_3 = (1,2) \\ \mu_B^{(0)} &= x_6 = (6,4) \\ \mu_C^{(0)} &= x_7 = (8,4) \end{aligned} \tag{37}$$

2. Bestimmung der euklidischen Distanzen zu Clusterzentroiden:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
A	$\sqrt{37}$	$\sqrt{\mathbf{10}}$	$\mathbf{0}$	$2\sqrt{13}$	$\sqrt{37}$	$\sqrt{29}$	$\sqrt{53}$	$\sqrt{34}$
B	$\mathbf{4\sqrt{2}}$	$\sqrt{17}$	$\sqrt{29}$	$\sqrt{\mathbf{17}}$	$\sqrt{\mathbf{2}}$	$\mathbf{0}$	2	$\sqrt{\mathbf{13}}$
C	$2\sqrt{13}$	$\sqrt{37}$	$\sqrt{53}$	5	$\sqrt{\mathbf{2}}$	2	$\mathbf{0}$	5

3. Zuweisung der Datenpunkte nach geringstem Abstand:

$$\begin{aligned}\text{Cluster}_A^{(0)} &: \{x_2 = (2,5), x_3 = (1,2)\} \\ \text{Cluster}_B^{(0)} &: \{x_1 = (2,8), x_4 = (5,8), x_5 = (7,3), x_6 = (6,4), x_8 = (4,7)\} \\ \text{Cluster}_C^{(0)} &: \{x_7 = (8,4)\}\end{aligned} \tag{38}$$

Da der Punkt x_5 zu $\mu_B^{(0)}$ und $\mu_C^{(0)}$ den gleichen Abstand hat, hätte man ihn auch Cluster C zuordnen können.

4. Aktualisierung der Zentroide:

$$\begin{aligned}\mu_A^{(1)} &= \frac{1}{2}(x_2 + x_3) = (1.5, 3.5) \\ \mu_B^{(1)} &= \frac{1}{5}(x_1 + x_4 + x_5 + x_6 + x_8) = (4.8, 6) \\ \mu_C^{(1)} &= x_7 = (8,4)\end{aligned} \tag{39}$$

Durchführen der zweite Iteration des K-Means-Algorithmus und aktualisieren der Koordinaten der Zentroide.

1. Bestimmung der euklidischen Distanzen zu Clusterzentroiden:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
A	$\frac{\sqrt{82}}{2}$	$\frac{\sqrt{\mathbf{10}}}{\mathbf{2}}$	$\frac{\sqrt{\mathbf{10}}}{\mathbf{2}}$	$\frac{\sqrt{130}}{2}$	$\frac{\sqrt{122}}{2}$	$\frac{\sqrt{82}}{2}$	$\frac{\sqrt{170}}{2}$	$\frac{\sqrt{74}}{2}$
B	$\frac{\mathbf{2\sqrt{74}}}{\mathbf{5}}$	$\frac{\sqrt{221}}{5}$	$\frac{\sqrt{761}}{5}$	$\frac{\sqrt{\mathbf{101}}}{\mathbf{5}}$	$\frac{\sqrt{346}}{5}$	$\frac{2\sqrt{34}}{5}$	$\frac{2\sqrt{89}}{5}$	$\frac{\sqrt{\mathbf{41}}}{\mathbf{5}}$
C	$2\sqrt{13}$	$\sqrt{37}$	$\sqrt{53}$	5	$\sqrt{\mathbf{2}}$	$\mathbf{2}$	$\mathbf{0}$	5

2. Zuweisung der Datenpunkte nach geringstem Abstand:

$$\begin{aligned}\text{Cluster}_A^{(1)} &: \{x_2 = (2,5), x_3 = (1,2)\} \\ \text{Cluster}_B^{(1)} &: \{x_1 = (2,8), x_4 = (5,8), x_8 = (4,7)\} \\ \text{Cluster}_C^{(1)} &: \{x_5 = (7,3), x_6 = (6,4), x_7 = (8,4)\}\end{aligned} \tag{40}$$

3. Aktualisierung der Zentroide:

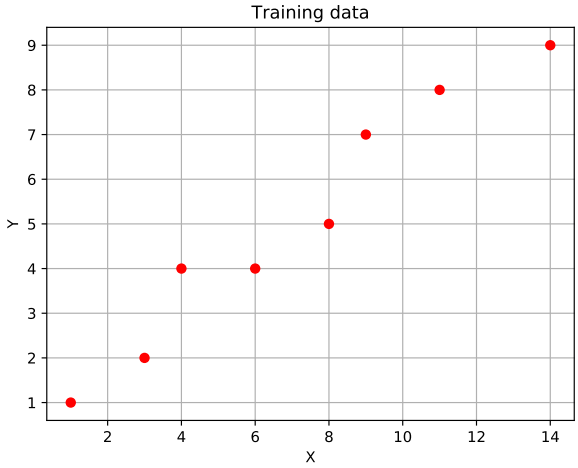
$$\begin{aligned}\mu_A^{(2)} &= \frac{1}{2}(x_2 + x_3) = (1.5, 3.5) \\ \mu_B^{(2)} &= \frac{1}{3}(x_1 + x_4 + x_8) = (\frac{11}{3}, \frac{23}{3}) \\ \mu_C^{(2)} &= \frac{1}{3}(x_5 + x_6 + x_7) = (7, \frac{11}{3})\end{aligned} \tag{41}$$

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 3: Veranschaulichung der Datenpunkte zur linearen Regression.



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

$$y = f(x) = \langle W, x \rangle + b\,. \tag{42}$$

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:

$$y = f(x) = \left\langle W^{'}, \vec{x}^{T} \right\rangle, \tag{43}$$

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 $W^{'}$:

$$W^{'} = (\vec{X}^{T}\vec{X})^{-1}\vec{X}^{T}Y, \tag{44}$$

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

Lösungsvorschlag:

Das Zielfunktion, die wir minimieren wollen, ist

$$O_{LSR}(W^{'}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \left\langle W^{'}, \vec{x_i}^{T} \right\rangle\right)^2 = \frac{1}{n}\left\|Y - \vec{X}W^{'}\right\|^2 \tag{45}$$

Bilden wir die Ableitung in Bezug auf $W^{'}$,

$$\nabla_{W'} O_{LSR} = -\frac{2}{n}\vec{X}^T\left(Y - \vec{X}W^{'}\right) \tag{46}$$

Für ein konvexes Optimierungsproblem stellt $\nabla_{W'} O_{LSR} = 0$ das Optimum dar.

$$-\frac{2}{n}\vec{X}^T\left(Y - \vec{X}W^{'}\right) = 0 \Rightarrow \vec{X}^T Y = \vec{X}^T\vec{X}W^{'} \Rightarrow W^{'} = (\vec{X}^T\vec{X})^{-1}\vec{X}^T Y \tag{47}$$

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.

Lösungsvorschlag:

$$\vec{X}^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 4 & 6 & 8 & 9 & 11 & 14 \end{pmatrix}$$

$$Y^T = \begin{pmatrix} 1 & 2 & 4 & 4 & 5 & 7 & 8 & 9 \end{pmatrix}$$

$$(\vec{X}^T\vec{X}) = \begin{pmatrix} 8 & 56 \\ 56 & 524 \end{pmatrix}$$

$$(\vec{X}^T\vec{X})^{-1} = \begin{pmatrix} 0.4962121 & -0.0530303 \\ -0.0530303 & 0.0075758 \end{pmatrix}$$

$$W^{'} = (\vec{X}^T\vec{X})^{-1}\vec{X}^T Y = \begin{pmatrix} b \\ W \end{pmatrix} = \begin{pmatrix} 0.5455 \\ 0.6364 \end{pmatrix}$$

$$y = f(x) = 0.6364 \cdot x + 0.5455$$

$W = 0.6364$ und $b = 0.5455$.