

Data Mining und Maschinelles Lernen

Prof. Kristian Kersting
Zhongjie Yu
Johannes Czech



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Sommersemester 2020
14. Juli 2020
Bonusübungsblatt 1

Die Abgabefrist dieser Bonusübung ist am **15.07.2020 um 23:59 Uhr**.

Die Bonusübung wird am **16.07.2020 um 13:30 Uhr** besprochen.

Aufgabe	1	2	3	4	5	6	7	8
Maximal Punktzahl	16	4	3	4	15	19	6	7
Erreichte Punktzahl								

Gruppe A	Nachnahme	Vorname	Matrikelnummer
1	Nowak	Patrick	2455200
2	Göller	Nicolas	2244770
3	Helm	Falko	2310138
4	Cakaj	Rinor	2713683

Benötigte Dateien

Alle benötigten Datensätze und Skriptvorlagen finden Sie in unserem Moodle-Kurs:

<https://moodle.informatik.tu-darmstadt.de/course/view.php?id=937>

Gruppeneinteilung

Bearbeiten Sie diese Übung in Dreier- oder Vierergruppen. Es steht Ihnen frei die Gruppen selbst zu bilden. Nutzen Sie hierfür die Gruppeneinteilung und das Forum in Moodle. Sehen Sie bitte aufgrund der aktuellen Coronakrise davon ab, sich lokal zu treffen und nutzen Sie stattdessen digitale Kommunikationskanäle. Zur gemeinsamen Bearbeitung der Abgabe können sie beispielsweise <https://overleaf.com> nutzen.

Theoretische Aufgaben

Bei theoretischen Übungsaufgaben, sind wir Ihnen sehr dankbar, wenn Sie diese in \LaTeX formatieren und als PDF einreichen. Nutzen Sie hierfür die \LaTeX -Vorlage und die vorgesehene Blöcke:

```
\begin{solution}
% Geben sie hier ihre Antwort an.
\end{solution}
```

Geben Sie dabei ihre Gruppenmitglieder und Gruppennummer in der Datei `group_members.tex` an.

Wenn Sie mit \LaTeX nicht ausreichend vertraut sind, können Sie auch einen hochauflösenden Scan einer handgeschriebenen Lösung einreichen. Bitte schreiben Sie ordentlich und leserlich.

Programmieraufgaben

Bei Aufgaben, die mit einem `</>` versehen sind, handelt es sich um Programmieraufgaben. Bearbeiten Sie in diesem Fall die vorgegebene Programmierzvorlage. Verwenden Sie bevorzugt **Python 3.7**, da wir diese Version zum Testen ihrer Lösung benutzen. Benennen Sie die Funktionsdateien nicht um und ändern Sie die angegebenen Funktionssignaturen nicht. Wenn Sie das Gefühl haben, dass es ein Fehler bei den Zuweisungen, fragen Sie uns auf Moodle.

Formalien zur Abgabe

Bitte laden Sie Ihre Lösungen in der entsprechenden Rubrik auf Moodle hoch. Sie müssen **nur eine Lösung pro Gruppe** einreichen. Wenn Sie keinen Zugang zu Moodle haben, setzen Sie sich bitte so schnell wie möglich mit uns in Verbindung. Laden Sie alle Ihre Lösungsdateien (die PDF-Abgabe und .py-Dateien) als eine einzige .zip-Datei hoch. Bitte beachten Sie, dass wir keine anderen als die angegebenen Dateiformate akzeptieren. **Laden Sie den gegebenen Datensatz zur Programmieraufgabe nicht in der Abgabe hoch.** Nutzen Sie folgende Namensgebung:

```
dmml_bonus1_group<groupid>.zip
└── dmml_bonus1.pdf
└── 02_dt_classification.py
└── 03_dt_regression.py
└── 04_random_forest.py
```

Verspätete Abgaben

Verspätete Abgaben werden akzeptiert, aber für jeden Tag, an dem die Abgabefrist überschritten wird, werden 25 % der insgesamt erreichbaren Punkte abgezogen. Nachdem die Übung offiziell besprochen wurde, können Sie die Aufgabe nicht mehr einreichen.

Bewertungsfaktoren

Die Bewertung dieser Übung hängt von den folgenden Faktoren ab:

- Richtigkeit der Antwort
- Klarheit der Präsentation der Ergebnisse
- Schreibstil

Wenn Sie bei einer Aufgabe nicht weiterkommen, versuchen Sie zu erklären warum und beschreiben Sie die Probleme, auf die Sie gestoßen sind, da Sie dafür Teilpunkte erhalten können.

Umgang mit Plagiaten

Sie dürfen gerne kursbezogene Themen in der Vorlesung oder in unseren Moodle Foren diskutieren. Sie sollten allerdings keine Lösungen mit anderen Gruppen teilen, und alles, was Sie einreichen, muss Ihre eigene Arbeit sein. Es ist Ihnen auch nicht gestattet, Material aus dem Internet zu kopieren. Sie sind verpflichtet, jede Informationsquelle, die Sie zur Lösung der Übungsaufgabe verwendet haben (d.h. andere Materialien als die Vorlesungsmaterialien), anzuerkennen. Zitierungen haben keinen Einfluss auf Ihre Note. Nicht anerkennen einer Quelle, die Sie verwendet haben, ist dagegen ein klarer Verstoß gegen die akademische Ethik. Beachten Sie, dass die Universität sehr ernst mit Plagiaten umgeht.

Aufgabe 1.1: 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

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	?

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

1.1a) ID3 Algorithmus (10 Punkte)

Erstellen Sie den Entscheidungsbauum mittels des ID3 Algorithmus. Berechnen Sie dabei die **Entropie** und den **Informationsgewinn** (engl. *gain*) der Attribut-Selektion für jeden Schritt.

Lösungsvorschlag:

Wir filtern im ersten Schritt nach Attributen und schätzen p_+ und p_- für jede Ausprägung des Attributs durch Zählen.
Es ergibt sich

Tabelle 3: Attribut: Ausblick

Ausprägung	Anzahl	davon +	davon -	Entropy
Sonnig	5	2	3	0.971
Bewölkt	4	4	0	0
Regen	5	3	2	0.971

Tabelle 4: Attribut: Temperatur

Ausprägung	Anzahl	davon +	davon -	Entropy
Warm	4	2	2	1
Mild	6	4	2	0.918
Kühl	4	3	1	0.811

Tabelle 5: Attribut: Luftfeuchtigkeit

Ausprägung	Anzahl	davon +	davon -	Entropy
Hoch	7	3	4	0.985
Normal	7	6	1	0.591

Tabelle 6: Attribut: Wind

Ausprägung	Anzahl	davon +	davon -	Entropy
Stark	6	3	3	1
Schwach	8	6	2	0.811

Pro Ausprägung setzen wir nun

$$p_+ = \frac{\text{davon} +}{\text{Anzahl}} \quad \text{und} \quad p_- = \frac{\text{davon} -}{\text{Anzahl}}$$

und berechnen die Entropien mittels $I(p_+, p_-) = (-p_+ \cdot \log p_+) + (-p_- \cdot \log p_-)$. Wobei der Logarithmus zur Basis 2 benutzt wurde. Die Ergebnisse haben wir an die Tabelle angefügt. Letztlich finden wir für jedes Attribut Att. den Informationsgehalt mittels

$$\text{Information(Att.)} = - \sum_{\text{Ausprägung A von Att.}} \frac{\text{Anzahl(A)}}{\text{ges}} \cdot \text{Entropy(A)}$$

Es ergibt sich (ges=14),c

Tabelle 7: Informationsgehalt

Attribut	Information
Ausblick	-0.694
Temperatur	-0.911
Luftfeuchtigkeit	-0.788
Wind	-0.892

, sodass wir uns als erstes Kriterium für das Attribut Ausblick entscheiden. Bei Bewölkung sind bereits alle Tage in einer Kategorie (+), weshalb wir hier ein Blatt mit einem + erstellen. Betrachten wir nun also alle Tage an denen es Sonnig ist und erstellen hier rekursiv einen Teilbaum. Um die Informationsgewinne zu berechnen, benötigen wir zuerst die Gesamt-Entropie des Datensatzes. Dazu schauen wir uns an wie oft Baseball gespielt wurde und wie oft nicht. Es folgt $H(S) = \frac{-9}{14} \cdot \log_2(\frac{9}{14}) - \frac{5}{14} \cdot \log_2(\frac{5}{14}) = 0.94$. Der Informationsgewinn in der 1. Iteration erhalten wir indem wir von $H(S) = 0.94$ (Gesamt-Entropie des Datensatzes) die Informationsgehalte der einzelnen Attribute abziehen:

Tabelle 8: Informationsgewinn

Attribut	Informationsgewinn
Ausblick	0.24645
Temperatur	0.029
Luftfeuchtigkeit	0.1516
Wind	0.0479

Wir fassen nochmal alle Daten, welche wir in diesem Knoten gegeben haben zusammen:

Tabelle 9: Trainingsdatensatz, nur Sonnig

Tag	Ausblick (A)	Temperatur (T)	Luftfeuchtigkeit (L)	Wind (W)	Spielt Baseball (B)
T1	Sonnig	Warm	Hoch	Schwach	Nein
T2	Sonnig	Warm	Hoch	Stark	Nein
T8	Sonnig	Mild	Hoch	Schwach	Nein
T9	Sonnig	Kühl	Normal	Schwach	Ja
T11	Sonnig	Mild	Normal	Stark	Ja

Wir zählen, so wie oben und erhalten:

Tabelle 10: nur Sonnig, Attribut: Temperatur

Ausprägung	Anzahl	davon +	davon -	Entropy
Warm	2	0	2	0
Mild	2	1	1	1
Kühl	1	1	0	0

Tabelle 11: nur Sonnig, Attribut: Luftfeuchtigkeit

Ausprägung	Anzahl	davon +	davon -	Entropy
Hoch	3	0	3	0
Normal	2	2	0	0

Tabelle 12: nur Sonnig, Attribut: Wind

Ausprägung	Anzahl	davon +	davon -	Entropy
Schwach	3	1	2	0.918
Stark	2	1	1	1

Man sieht direkt, dass Attribut L auf den Trainingsdaten eine optimalen Informationsgewinn hat. Dies verifiziert man durch Berechnen der Informationsgehalte(ges=5)):

Tabelle 13: Informationsgehalt, nur Sonnig

Attribut	Information
Temperatur	-0.4
Luftfeuchtigkeit	0
Wind	-0.95

Wir schauen uns noch die Informationsgewinne nach der 2. Iteration an. Die Gesamtentropie ist nun $H(S) = -\frac{2}{5} \cdot \log_2(\frac{2}{5}) - \frac{3}{5} \cdot \log_2(\frac{3}{5}) = 0.97$. Daraus folgt,

Tabelle 14: Informationsgewinn

Attribut	Informationsgewinn
Temperatur	0.57
Luftfeuchtigkeit	0.9709
Wind	0.0209

Unter dem Sonnig-Ast erzeugen wir also einen Knoten 'Luftfeuchtigkeit' mit 2 Blättern. Ast 'hoch' führt zu - und Ast 'normal' zu +.

Nun müssen wir noch den Teilbaum finden, der an den Ast 'Regen' anhängt. Filtern wir also nach diesem Attribut und zählen dann wieder, so erhalten wir:

Tabelle 15: Trainingsdatensatz, nur Regen

Tag	Ausblick (A)	Temperatur (T)	Luftfeuchtigkeit (L)	Wind (W)	Spielt Baseball (B)
T4	Regen	Mild	Hoch	Schwach	Ja
T5	Regen	Kühl	Normal	Schwach	Ja
T6	Regen	Kühl	Normal	Stark	Nein
T10	Regen	Mild	Normal	Schwach	Ja
T14	Regen	Mild	Hoch	Stark	Nein

Tabelle 16: nur Regen, Attribut: Temperatur

Ausprägung	Anzahl	davon +	davon -	Entropy
Mild	3	2	1	0.918
Kühl	2	1	1	1

Tabelle 17: nur Regen, Attribut: Luftfeuchtigkeit

Ausprägung	Anzahl	davon +	davon -	Entropy
Hoch	2	1	1	1
Normal	3	2	1	0.918

Tabelle 18: nur Regen, Attribut: Wind

Ausprägung	Anzahl	davon +	davon -	Entropy
Schwach	3	3	0	0
Stark	2	0	2	0

Wir schauen uns noch die Informationsgewinne nach der 3. Iteration an. Die Gesamtentropie ist nun $H(S) = 0$. Daraus folgt,

Tabelle 19: Informationsgewinn

Attribut	Informationsgewinn
Temperatur	0.02
Luftfeuchtigkeit	0.02

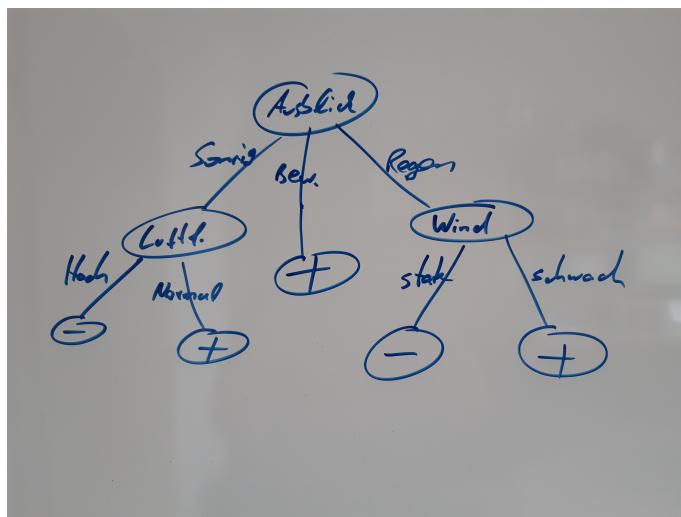
Wir sehen wieder direkt die perfekte Klassifizierung beim Attribut Wind (Informationswert=0). Die anderen beiden haben jeweils einen Informationswert von -0.951. Wir erstellen hier also einen Knoten 'Wind' mit 2 Blättern '+' für Ausprägung 'schwach' und '-' für Ausprägung 'stark'.

1.1b) Visualisierung (3 Punkte)

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

Lösungsvorschlag:

Es ergab sich insgesamt der folgende Entscheidungsbaum:



1.1c) Vorhersage (3 Punkte)

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

Lösungsvorschlag:

Wenn wir dem erlernten Entscheidungsbaum folgen, ergibt sich

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	-

Wir spielen also nur an Tag 16.

Aufgabe 1.2: </> Entscheidungsbäume - Klassifikation (4)

Im Institut für Produktionstechnik und Umformmaschinen, kurz PtU¹, der TU Darmstadt gibt es die Aufgabe eines Scherschneideverfahrens. Dabei wird mit Hilfe eines Stempels ein Loch in einen metallischen Werkstoff gestanzt, siehe Abbildung 1. Es gibt zwei Werkstoffe im Versuch: CuSn6 und 16MnCr5. Die Dicke des Materials beträgt entweder 0.4 mm oder 0.5 mm. Die Geschwindigkeit des Stempels liegt in den drei Stufen 100, 200 und 300 vor.

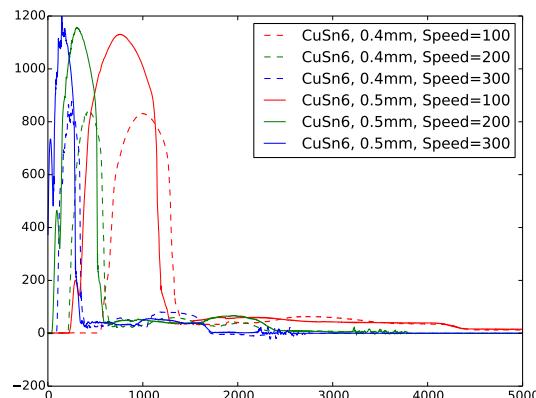
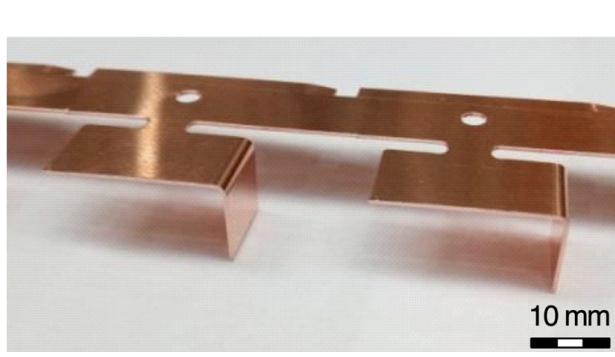


Abbildung 1: Links: Veranschaulichung des PtU Prozess. Rechts: Beispiel des Kraftsignals.

Es gibt mehrere Sensoren, die den Status des Stanzen messen:

- Kraft Sensor / Force sensor (Fz)
- Verschiebungssensor / Displacement sensor (w)
- Beschleunigungssensor / Acceleration sensor (acc)

Im Experiment messen die Sensoren den Status des Stanzen von Beginn bis Ende jedes Prozesses. Die Zeitreihendaten von jedem Sensor werden als 1D-Vektor dargestellt. Das Interessante an dieser Aufgabe ist, dass der Status des Stanzen von der Art und Dicke des Materials abhängt. Andererseits ist es uns möglich, aus den Zeitreiheninformationen von den Sensoren auf die Art und Dicke des Materials und die Geschwindigkeit des Stanzen zu schließen.

- **Filename_Fz_raw.csv**: Enthält Daten aus dem Kraftsensor. Jeder Zeilenvektor repräsentiert das Kraftsignal. Die Anzahl der Zeilen in der Datei beträgt 2787, was der Anzahl der Experimente im Datensatz entspricht.
- **Filename_Speed.csv**: Enthält die Geschwindigkeit des Schlags der 2787 Proben.
- **Filename_thickness.csv**: Enthält die Materialstärke der 2787 Proben.

In dieser Übung verwenden wir die Daten des Kraftsensors, um die Proben nach der Geschwindigkeit des Stempels zu klassifizieren.

1.2a) </> 02_dt_classification.py (3 Punkte)

Laden Sie die Daten des Kraftsensor aus **Filename_Fz_raw.csv** und die Geschwindigkeit des Stanzen aus **Filename_Speed.csv** und vervollständigen Sie den Code in **02_dt_classification.py**:

- </> Trainieren Sie einen Entscheidungsbaum zur Klassifikation in der Methode `fit_dt_classifier()` mithilfe von `sklearn` unter der Verwendung der Standardparameter.
- </> Ermitteln Sie die Testgenauigkeit in der Methode `get_test_accuracy()`.

¹<https://www.ptu.tu-darmstadt.de/>

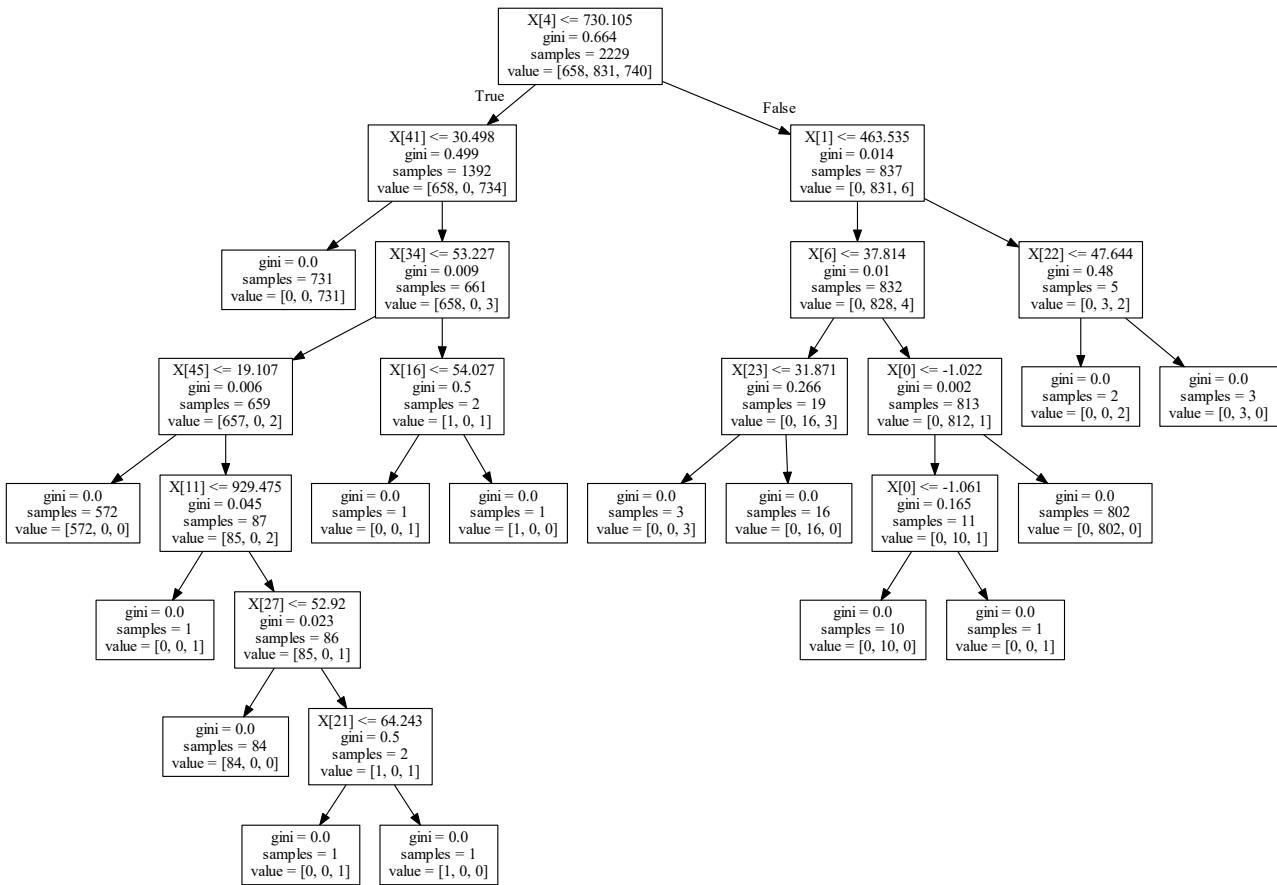
</> Plotten Sie den resultierenden Entscheidungsbaum in `export_tree_plot()`. Sie können dabei die Funktion `tree.export_graphviz()` verwenden.

1.2b) Visualisierung (1 Punkt)

Zeigen Sie die erstelle Visualisierung des Entscheidungsbaumes zur Klassifikation aus Unteraufgabe a).

Lösungsvorschlag:

Es ergab sich der folgende Entscheidungsbaum:



Aufgabe 1.3: Entscheidungsbäume - Regression (3)

Ein Entscheidungsbaum zur Regression kann auf den PtU Datensatz (s. Abb. 1) angewendet werden, um auf die Dicke des Materials zu schließen, welche ein kontinuierlicher Wert ist.

1.3a) </> 03_dt_regression.py (2 Punkte)

Laden Sie die Daten des Kraftsensor aus `Filename_Fz_raw.csv` und die Materialstärken aus `Filename_thickness.csv` und vervollständigen Sie den Code in `03_dt_regression.py`:

</> Trainieren Sie einen Entscheidungsbaum zur Regression in `fit_dt_regressor()`.

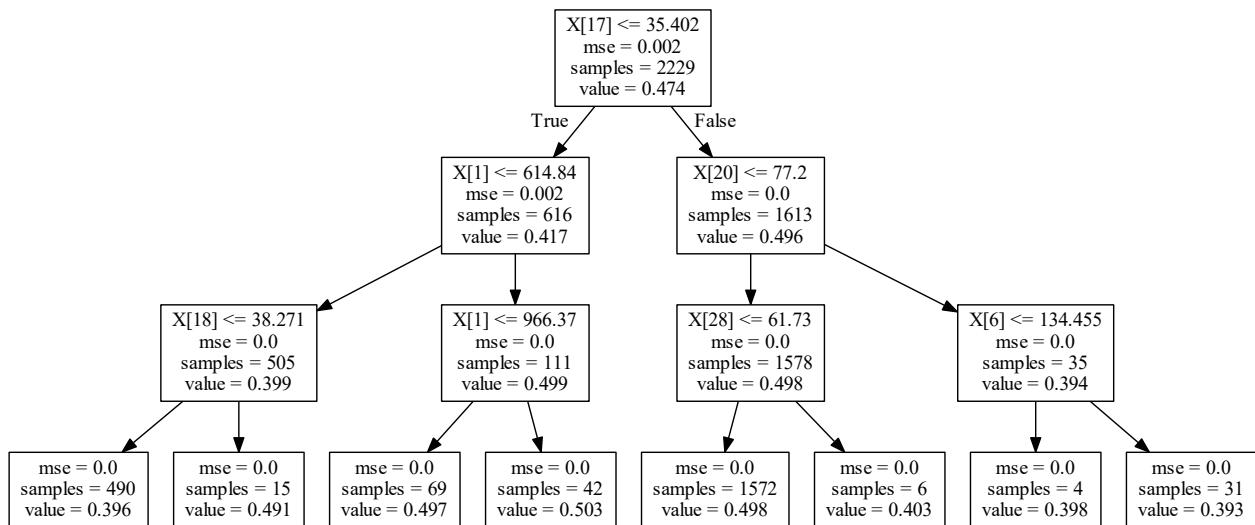
</> Berechnen Sie den mittleren quadratischen Fehler für den Testdatensatz in der Funktion `get_test_mse()`.

1.3b) Visualisierung (1 Punkt)

Zeigen Sie die erstelle Visualisierung des Entscheidungsbaumes zur Regression aus Unteraufgabe a).

Lösungsvorschlag:

Es ergab sich der folgende Entscheidungsbaum:



Aufgabe 1.4: </> Zufallswälder (4)

In dieser Aufgabe verwenden wir den PtU Datensatz (s. Abb. 1), um mit einem Zufallswald (engl. Random Forest) die Geschwindigkeit des Einschlags aus den Zeitreihen zu klassifizieren.

1.4a) </> 04_random_forest.py (2 Punkte)

Laden Sie die Daten des Kraftsensor aus `Filename_Fz_raw.csv` und die Geschwindigkeit des Stanzen aus `Filename_Speed.csv` und vervollständigen Sie den Code in `04_random_forest.py`:

</> Trainieren Sie einen Zufallswald aus 100 Entscheidungsstämpfen in der Methode `fit_tree_stump_forest()`.

</> Trainieren Sie einen einzelnen Entscheidungsstumpf in der Methode `fit_tree_stump()`.

1.4b) Konfusionsmatrix (2 Punkte)

Geben Sie die Konfusionsmatrizen des Trainings- und Testdatensatzes mit absoluten Werten für den Zufallswald an.

Lösungsvorschlag:

Es ergab sich für den Trainingsdatensatz die confusion matrix:

658	0	0
0	831	0
11	6	723

und für den Testdatensatz:

166	0	0
0	207	0
2	4	179

```
X_data.shape: (2787, 4999)
y_data.shape: (2787,)
y_pred: [100. 300. 300. 100. 200. 100. 200. 300. 200. 300.] ...
y_test: [100. 300. 300. 100. 200. 100. 200. 300. 200. 300.] ...
Train Confusion Matrix:
[[658  0  0]
 [ 0 831  0]
 [ 11  6 723]]
Test Confusion Matrix:
[[166  0  0]
 [ 0 207  0]
 [ 2  4 179]]
Random Forest Test Accuracy: 0.989247311827957
Tree Stump Tree Test Accuracy: 0.6953405017921147
```

Aufgabe 1.5: AdaBoost (15)

In dieser Aufgabe werden Sie AdaBoost auf die gegebenen Trainingsbeispiele aus der Tabelle 20 anwenden.

Tabelle 20: Datensatz mit zwei Merkmalen und zwei Zielklassen.

x₁	x₂	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. $x_1 \leq T \Rightarrow +$ oder $x_1 > T \Rightarrow -$) 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 x_1 und dann x_2 .

Verwenden Sie die Formel:

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

zur Berechnung von α_i .

1.5a) 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:

Im ersten Schritt gilt für alle Gewichte $w_i = 0.1$, für $i = 1, \dots, 10$. Also müssen wir nur zählen, wie viele falsch klassifiziert werden. Betrachte dazu Spalten 3 und 4 in Tabelle 21. Da wir bei einem Unentschieden das erste Minimum beginnend mit x_1 wählen sollen, gilt

$$f_1((x_1, x_2)^T) = \begin{cases} +1 & x_1 \leq 2 \\ -1 & \text{sonst} \end{cases}$$

als Stumpf für die erste Iteration. Damit gilt nach Zeile 5 der Tabelle 21 $err_1 = 0.3$. Weiter ist

$$\alpha_1 = \frac{1}{2} \ln \left(\frac{1 - err_1}{err_1} \right) = 0.4236.$$

Behalten wir die Nummerierung in der Aufgabenstellung bei (beginnend bei 1), dann klassifizieren wir die Punkte 3, 4 und 5 falsch. Dementsprechend erhalten wir für die neuen (noch nicht normalisierten) Gewichte

$$w'_i = 0.1 \text{ für } i = 1, 2, 6, 7, 8, 9, 10$$

$$w'_i = 0.1 e^{\alpha_1} = 0.1527 \text{ für } i = 3, 4, 5.$$

Kriterium	Entscheidung	erste Iteration		zweite Iteration		Iteration			
		Fehler x_1	Fehler x_2	x_1 schlecht	x_1 gut	Fehler x_1	x_2 schlecht	x_2 gut	Fehler x_2
$x_i \leq 0$	pro +	0.5	0.5	3	2	0.5821	3	2	0.5821
$x_i \leq 0$	pro -	0.5	0.5	0	5	0.4315	0	5	0.4315
$x_i \leq 1$	pro +	0.4	0.6	3	1	0.4958	3	3	0.6684
$x_i \leq 1$	pro -	0.6	0.4	0	6	0.5178	0	4	0.3452
$x_i \leq 2$	pro +	0.3	0.5	3	0	0.4095	3	2	0.5821
$x_i \leq 2$	pro -	0.7	0.5	0	7	0.6041	0	5	0.4315
$x_i \leq 3$	pro +	0.4	0.6	3	1	0.4958	3	3	0.6684
$x_i \leq 3$	pro -	0.6	0.4	0	6	0.5178	0	4	0.3452
$x_i \leq 4$	pro +	0.5	0.7	3	2	0.5821	3	4	0.7547
$x_i \leq 4$	pro -	0.5	0.3	0	5	0.4315	0	3	0.2589
$x_i \leq 5$	pro +	0.4	0.6	2	2	0.4456	3	3	0.6684
$x_i \leq 5$	pro -	0.6	0.4	1	5	0.568	0	4	0.3452
$x_i \leq 6$	pro +	0.3	0.7	1	2	0.3091	3	4	0.7547
$x_i \leq 6$	pro -	0.7	0.3	2	5	0.7045	0	3	0.2589
$x_i \leq 7$	pro +	0.4	0.6	1	3	0.3954	2	4	0.6182
$x_i \leq 7$	pro -	0.6	0.4	2	4	0.6182	1	3	0.3954
$x_i \leq 8$	pro +	0.3	0.5	0	3	0.2589	1	4	0.4817
$x_i \leq 8$	pro -	0.7	0.5	3	4	0.7547	2	3	0.5319
$x_i \leq 9$	pro +	0.4	0.6	0	4	0.3452	1	5	0.568
$x_i \leq 9$	pro -	0.6	0.4	3	3	0.6684	2	2	0.4456
$x_i \leq 10$	pro +	0.5	0.5	0	5	0.4315	0	5	0.4315
$x_i \leq 10$	pro -	0.5	0.5	3	2	0.5821	3	2	0.5821

Tabelle 21: Die Ergebnisse für Aufgabe 1.5. Die Tabelle liest sich am Beispiel der Zeile 3 folgendermaßen: entscheiden wir uns bei $x_1 \leq 1$ für + dann machen wir 4 Fehler im ersten Schritt (also Fehlermaß 0.4) und 3 schlechte (d.h. höher gewichtete) Fehler bzw. einen guten (d.h. niedriger gewichteten) im zweiten Schritt. Daraus errechnet sich das Fehlermaß. Entscheiden wir uns bei $x_2 \leq 1$ für +, dann machen wir 6 Fehler im ersten Schritt und sowohl 3 schlechte als auch 3 gute im zweiten Schritt. Daraus ergibt sich dann das Fehlermaß.

Normalisieren (d.h. teilen durch $7 * 0.1 + 0.1527 * 3 = 1.1581$) ergibt

$$w_i^{(1)} = 0.0863 \text{ für } i = 1, 2, 6, 7, 8, 9, 10$$

$$w_i^{(1)} = 0.1365 \text{ für } i = 3, 4, 5.$$

Für den zweiten Iterationsschritt müssen wir darauf achten, dass die Datenpunkte 3, 4 und 5 stärker gewichtet sind, wir also eine erneute Fehlklassifikation vermeiden sollten. Der zu minimierende Fehler ist

$$0.0863(\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4 + \epsilon_8 + \epsilon_9 + \epsilon_{10}) + 0.1365(\epsilon_5 + \epsilon_6 + \epsilon_7)$$

wobei

$$\epsilon_i = \begin{cases} 1 & \text{Datenpunkt } i \text{ wurde falsch klassifiziert} \\ 0 & \text{Datenpunkt } i \text{ wurde korrekt klassifiziert} \end{cases}$$

gilt. Aus Spalte 7 in Tabelle 21 ersehen wir, dass der neue Rumpf

$$f_2((x_1, x_2)^T) = \begin{cases} +1 & x_1 \leq 8 \\ -1 & \text{sonst} \end{cases}$$

ist. Es gilt $\text{err}_2 = 3 * 0.0863 = 0.2589$ und wir klassifizieren die Datenpunkte 6, 7 und 8 falsch. Es folgt

$$\alpha_2 = \frac{1}{2} \ln\left(\frac{1 - \text{err}_2}{\text{err}_2}\right) = 0.5258.$$

Wir müssen die Datenpunkte 6, 7 und 8 neu gewichten, also

$$w'_i = 0.1 * e^{0.5258} = 0.1692 \text{ für } i = 6, 7, 8$$

$$w'_i = 0.0863 \text{ für } i = 1, 2, 9, 10$$

$$w'_i = 0.1365 \text{ für } i = 3, 4, 5.$$

mit. Nach Normalisierung (d.h. teilen durch $3 * 0.1365 + 3 * 0.1692 + 4 * 0.0863 = 1.2623$) folgt

$$w_i^{(2)} = 0.1081 \text{ für } i = 3, 4, 5$$

$$w_i^{(2)} = 0.1340 \text{ für } i = 6, 7, 8$$

$$w_i^{(2)} = 0.0684 \text{ für } i = 1, 2, 9, 10$$

1.5b) Gesamtmodell (3 Punkte)

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

Lösungsvorschlag:

Mit der Notation aus Teilaufgabe a) folgt mit $x = (x_1, x_2)^T \in \mathbb{R}^2$

$$\tilde{f}(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) = \begin{cases} 0.9494 & x_1 \leq 2 \\ 0.1022 & 2 < x_1 \leq 8 \\ -0.9494 & x_1 > 8 \end{cases}$$

Für das Gesamtmodell nach zwei Iterationen ergibt sich also

$$f(x) = sign(\tilde{f}(x)) = \begin{cases} 1 & x_1 \leq 8 \\ -1 & x_1 > 8 \end{cases}$$

Aufgabe 1.6: 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.

1.6a) 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:

Sei Merkmal = (Ausblick, Temperatur, Luftfeuchtigkeit, Wind) ein 4-Tupel aus Untermerkmalen. Dabei kürzen wir Ausblick mit A, Temperatur mit T, Luftfeuchtigkeit mit L und Wind mit W ab. Zur besseren Lesbarkeit sei weiterhin N := 'B = Nein', J := 'B = Ja'. Nach der Unabhängigkeitsannahme des Naive Bayes-Ansatzes gilt

$$\begin{aligned} p(Merkmal|J) &= p(A|J) \cdot p(T|J) \cdot p(L|J) \cdot p(W|J) \text{ und} \\ p(Merkmal|N) &= p(A|N) \cdot p(T|N) \cdot p(L|N) \cdot p(W|N). \end{aligned}$$

Damit können wir rechnen

$$\begin{aligned} p(J|Merkmal) &= \frac{p(Merkmal|J) \cdot p(J)}{p(Merkmal)} \\ &= \frac{p(Merkmal|J) \cdot p(J)}{p(Merkmal|J) \cdot p(J) + p(Merkmal|N) \cdot p(N)} \\ &= \frac{p(A|J) \cdot p(T|J) \cdot p(L|J) \cdot p(W|J) \cdot p(J)}{p(A|J) \cdot p(T|J) \cdot p(L|J) \cdot p(W|J) \cdot p(J) + p(A|N) \cdot p(T|N) \cdot p(L|N) \cdot p(W|N) \cdot p(N)}. \end{aligned}$$

Die erste Gleichheit ist der Satz von Bayes, die zweite die Formel der totalen Wahrscheinlichkeit und die dritte die Unabhängigkeitsannahme von oben. Völlig analog gilt

$$\begin{aligned} p(N|Merkmal) &= \frac{p(Merkmal|N) \cdot p(N)}{p(Merkmal)} \\ &= \frac{p(Merkmal|N) \cdot p(N)}{p(Merkmal|J) \cdot p(J) + p(Merkmal|N) \cdot p(N)} \\ &= \frac{p(A|N) \cdot p(T|N) \cdot p(L|N) \cdot p(W|N) \cdot p(N)}{p(A|J) \cdot p(T|J) \cdot p(L|J) \cdot p(W|J) \cdot p(J) + p(A|N) \cdot p(T|N) \cdot p(L|N) \cdot p(W|N) \cdot p(N)}. \end{aligned}$$

1.6b) 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:

Wir kürzen die Begriffe wie folgt ab: Sonnig - S, Regen - R, Bewölkung - B; Warm - W, Mild - M, Kühl - K; Hoch - H, No - N; Sch - Schwach, St - Stark. Durch einfaches Auszählen ergibt sich $p(B = Ja) = 9/14$, $p(B = Nein) = 5/14$ sowie

Ausblick	Temperatur	Luftfeuchtigkeit	Wind
$p(A = S) = 5/14$	$p(T = W) = 4/14$	$p(L = H) = 7/14$	$p(W = Sch) = 8/14$
$p(A = R) = 5/14$	$p(T = M) = 6/14$	$p(L = No) = 7/14$	$p(W = St) = 6/14$
$p(A = B) = 4/14$	$p(T = K) = 4/14$		
$p(A = S Ja) = 2/9$	$p(T = W Ja) = 2/9$	$p(L = H Ja) = 3/9$	$p(W = Sch Ja) = 6/9$
$p(A = B Ja) = 4/9$	$p(T = M Ja) = 4/9$	$p(L = No Ja) = 6/9$	$p(W = St Ja) = 3/9$
$p(A = R Ja) = 3/9$	$p(T = K Ja) = 3/9$		
$p(A = S Nein) = 3/5$	$p(T = W Nein) = 2/5$	$p(L = H Nein) = 4/5$	$p(W = Sch Nein) = 2/5$
$p(A = B Nein) = 0/5$	$p(T = M Nein) = 2/5$	$p(L = No Nein) = 1/5$	$p(W = St Nein) = 3/5$
$p(A = R Nein) = 2/5$	$p(T = K Nein) = 1/5$		

1.6c) 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.

Lösungsvorschlag:

Wir setzen die Zahlen aus b) in die Formel von a) ein und erhalten für Tag 15

$$\begin{aligned} p(B = Ja|S, M, H, Sch) &= \frac{p(S|Ja)p(M|Ja)p(H|Ja)p(Sch|Ja)p(Ja)}{p(S|Ja)p(M|Ja)p(H|Ja)p(Sch|Ja)p(Ja) + p(S|N)p(M|N)p(H|N)p(Sch|N)p(N)} \\ &= \frac{2/9 \cdot 4/9 \cdot 3/9 \cdot 6/9 \cdot 9/14}{2/9 \cdot 4/9 \cdot 3/9 \cdot 6/9 \cdot 9/14 + 3/5 \cdot 2/5 \cdot 4/5 \cdot 2/5 \cdot 5/14} = \frac{8/567}{8/567 + 24/875} = 0,33967. \end{aligned}$$

Es folgt $p(B = Nein|S, M, H, Sch) = 1 - p(B = Ja|S, M, H, Sch) = 0.66 > 0.5$, sodass an Tag 15 kein Baseball gespielt wird. Tag 16:

$$\begin{aligned} p(B = Ja|B, M, No, Sch) &= \frac{p(B|Ja)p(M|Ja)p(No|Ja)p(Sch|Ja)p(Ja)}{p(B|Ja)p(M|Ja)p(No|Ja)p(Sch|Ja)p(Ja) + p(B|N)p(M|N)p(No|N)p(Sch|N)p(N)} \\ &= \frac{4/9 \cdot 4/9 \cdot 6/9 \cdot 6/9 \cdot 9/14}{4/9 \cdot 4/9 \cdot 6/9 \cdot 6/9 \cdot 9/14 + 0/5 \cdot 2/5 \cdot 1/5 \cdot 2/5 \cdot 5/14} = 1, \end{aligned}$$

also wird an Tag 16 Baseball gespielt.

Für Tag 17 ergibt sich

$$\begin{aligned} p(B = Ja|R, K, No, St) &= \frac{p(R|Ja)p(K|Ja)p(No|Ja)p(St|Ja)p(Ja)}{p(R|Ja)p(K|Ja)p(No|Ja)p(St|Ja)p(Ja) + p(R|N)p(K|N)p(No|N)p(St|N)p(N)} \\ &= \frac{3/9 \cdot 3/9 \cdot 6/9 \cdot 3/9 \cdot 9/14}{3/9 \cdot 3/9 \cdot 6/9 \cdot 3/9 \cdot 9/14 + 2/5 \cdot 1/5 \cdot 1/5 \cdot 3/5 \cdot 5/14} = \frac{1/63}{1/63 + 3/875} = 0,822. \end{aligned}$$

Somit wird an Tag 17 Baseball gespielt.

Aufgabe 1.7: K-Means (6)

Folgender Datensatz besteht aus 8 Punkten:

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

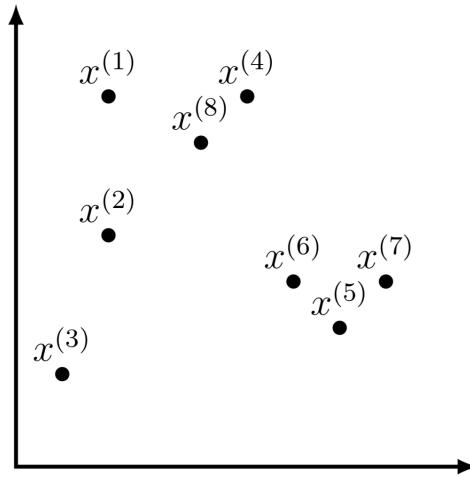


Abbildung 2: Visualisierung des K-Means Datensatzes

1.7a) K-Means Algorithmus (6 Punkte)

Benutzen Sie den K-Means Algorithmus mit der Euklidischen Distanz um diese 8 Datenpunkte in $K = 3$ Cluster einzuteilen. Nehmen Sie dabei an, dass die Clusterzentren mit den Punkten x_2 , x_4 und x_8 initialisiert sind. Führen Sie zwei Iterationen des K-Means Algorithmus durch und geben Sie die Koordinaten der Zentroide der Cluster an.

Lösungsvorschlag:

Wir färben die Cluster ein. Das erste Cluster mit Mittelpunkt x_2 ist grün, das zweite Cluster mit Mittelpunkt x_4 ist lila und das dritte Cluster mit Mittelpunkt x_8 ist blau. Der Abstand zwischen zwei Vektoren $v = (v_1, v_2), u = (u_1, u_2) \in \mathbb{R}^2$ ist gegeben durch $\|v - u\| = \sqrt{v_1 u_1 + v_2 u_2}$. Bezeichne $z_i^{(j)}$ das Zentrum des i -ten Clusters ($i = 1, 2, 3$) im j -ten Iterationsschritt. Wir initialisieren $z_1^{(0)} := x_2$, $z_2^{(0)} := x_4$, $z_3^{(0)} := x_8$. Im ersten Schritt berechnen wir

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
grün:	3	0	$\sqrt{10}$	$\sqrt{18}$	$\sqrt{29}$	$\sqrt{17}$	$\sqrt{37}$	$\sqrt{18}$
lila:	3	$\sqrt{18}$	$\sqrt{52}$	0	$\sqrt{29}$	$\sqrt{17}$	5	$\sqrt{2}$
blau:	$\sqrt{5}$	$\sqrt{8}$	$\sqrt{34}$	$\sqrt{2}$	5	$\sqrt{13}$	5	0

Dabei ist der k -te Eintrag gegeben durch $\|z_k^{(0)} - x_l\|$. Durch spaltenweises Vergleichen erhalten wir, dass x_2 und x_3 grün werden. x_1, x_5, x_6 und x_8 werden blau und x_4 wird lila. Bei x_7 haben wir ein Unentschieden zwischen blau und lila. Wir entscheiden uns für blau, da die beiden Nachbarn von x_7 - x_5 und x_6 - auch blau sind. Würden wir x_7 lila zuordnen, so würden wir die Varianz von lila stark erhöhen, da diese sonst 0 ist. Da der k-means Algorithmus aber die Varianz innerhalb der Cluster minimieren soll, ist blau sicher keine falsche Wahl für x_7 .

Nach dem ersten Iterationsschritt erhalten wir also

$$\begin{aligned}\text{grün}_1 &= \{x_2, x_3\} \text{ mit } z_1^{(1)} = (1, 5; 3, 5)^T \\ \text{lila}_1 &= \{x_4\} \text{ mit } z_2^{(1)} = x_4 = (5; 8)^T \\ \text{blau}_1 &= \{x_1, x_5, x_6, x_7, x_8\} \text{ mit } z_3^{(1)} = (5, 4; 5, 2)^T.\end{aligned}$$

Für den zweiten Schritt berechnen wir

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
grün:	4, 53	1, 58	1, 58	5, 70	5, 52	4, 53	6, 52	4, 30
lila:		3	4, 24	7, 21	0	5, 39	4, 12	5
blau:		4, 4	3, 4	5, 44	2, 82	2, 72	1, 34	2, 86

wobei der kl-te Eintrag durch $\|z_k^{(1)} - x_l\|$ gegeben ist. Durch spaltenweises Vergleichen erhalten wir (diesmal gibt es kein Unentschieden)

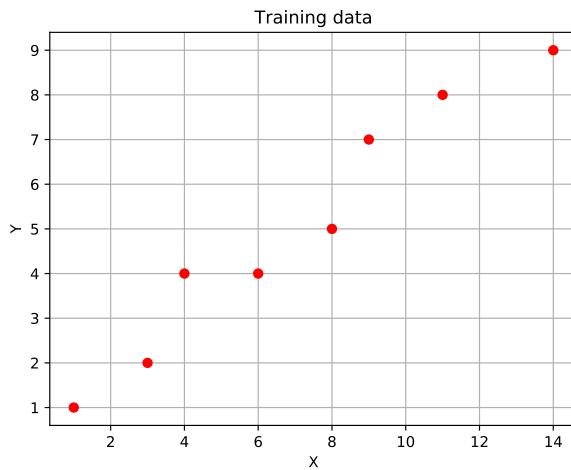
$$\begin{aligned}\text{grün}_2 &= \{x_2, x_3\} \text{ mit } z_1^{(2)} = (1, 5; 3, 5)^T \\ \text{lila}_2 &= \{x_1, x_4, x_8\} \text{ mit } z_2^{(2)} = x_4 = (3, 67; 7, 67)^T \\ \text{blau}_2 &= \{x_5, x_6, x_7\} \text{ mit } z_3^{(2)} = (7; 3, 67)^T.\end{aligned}$$

Aufgabe 1.8: 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 Fehlerequadrate erstellen:

$$y = f(x) = \langle W, x \rangle + b. \quad (3)$$

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

$$y = f(x) = \left\langle W', \vec{x}^T \right\rangle, \quad (4)$$

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

1.8a) Herleitung (5 Punkte)

Zeigen Sie, dass das optimale W' :

$$W' = (\vec{X}^T \vec{X})^{-1} \vec{X}^T Y, \quad (5)$$

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

Lösungsvorschlag:

Wir möchten eine Regression nach dem Prinzip der kleinsten Fehlerquadrate erstellen. Das heißt wir wollen folgende Funktion minimieren:

$$\min_{W,b} S(W, b) = \sum_{i=1}^n (W^T x_i + b - y_i)^2.$$

Dies ist nach den obigen Definitionen äquivalent zur Minimierung folgender Funktion:

$$\min_{W'} S(W') = \sum_{i=1}^n (W'^T \bar{x}_i - y_i)^2.$$

Daraus folgt

$$\begin{aligned} S(W') &= \sum_{i=1}^n (W'^T \bar{x}_i - y_i)^2 = \|\vec{X}W' - Y\|^2 \\ &= (\vec{X}W' - Y)^T (\vec{X}W' - Y) \\ &= W'^T \vec{X}^T \vec{X} W' - Y^T \vec{X} W' - W'^T \vec{X}^T Y + Y^T Y \\ &= W'^T \vec{X}^T \vec{X} W' - 2Y^T \vec{X} W' + Y^T Y \end{aligned}$$

Wir differenzieren nun die Funktion $S(W')$ nach W'

$$\frac{\partial S(W')}{\partial W'} = 2W'^T \vec{X}^T \vec{X} - 2Y^T \vec{X}.$$

Setzen wir die Ableitung gleich 0 erhalten wir

$$\begin{aligned} 0 &= W'^T \vec{X}^T \vec{X} - Y^T \vec{X} \\ W'^T \vec{X}^T \vec{X} &= Y^T \vec{X} \\ \vec{X}^T \vec{X} W' &= \vec{X}^T Y \\ W' &= (\vec{X}^T \vec{X})^{-1} \vec{X}^T Y \end{aligned}$$

unter der Bedingung, dass die Inverse $(\vec{X}^T \vec{X})^{-1}$ existiert.

1.8b) Parameterbestimmung (2 Punkte)

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

Lösungsvorschlag:

Durch ablesen der Koordinaten erhalten wir:

$$\vec{X} = \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \\ 1 & 9 \\ 1 & 11 \\ 1 & 14 \end{pmatrix} \quad \text{und} \quad \vec{Y} = \begin{pmatrix} 1 \\ 2 \\ 4 \\ 4 \\ 5 \\ 7 \\ 8 \\ 9 \end{pmatrix}.$$

Daraus folgt

$$\vec{X}^T \vec{X} = \begin{pmatrix} 8 & 56 \\ 56 & 524 \end{pmatrix} \quad \text{und} \quad (\vec{X}^T \vec{X})^{-1} = \begin{pmatrix} \frac{131}{264} & \frac{-7}{132} \\ \frac{-7}{132} & \frac{1}{132} \end{pmatrix}$$

Damit können wir nun W und b berechnen

$$\begin{aligned} W' &= \begin{pmatrix} \frac{131}{264} & \frac{-7}{132} \\ \frac{-7}{132} & \frac{1}{132} \end{pmatrix} \begin{pmatrix} 40 \\ 364 \end{pmatrix} = (\vec{X}^T \vec{X})^{-1} \vec{X}^T Y \\ &= \begin{pmatrix} \frac{6}{11} & \frac{7}{11} \end{pmatrix} \end{aligned}$$

Daraus folgt $W = \frac{7}{11}$ und $b = \frac{6}{11}$.