Big Data Anwendungen

Aufgabenblatt 2

Agenda

- Aufgabenblatt 1 – Deskriptive Methoden zur Datenexploration

- Aufgabenblatt 2 – Klassifikation
  - Aufgabe 1 – Entscheidungsbaum I
  - Aufgabe 2 – Entscheidungsbaum II
  - Aufgabe 3 – Verständnisfragen

- Aufgabenblatt 3 – Recommender Systems
- Aufgabenblatt 4 – Clusteringverfahren
- Aufgabenblatt 5 – Stream Mining
- Aufgabenblatt 6 – Social Network Analysis

Aufgabe 1 (a) - Aufgabenstellung

- Gegeben seien folgende Trainingsdaten:

Churn	Anruf im CallCenter	Beschwerde	Zahlungsverzug
Ja	Ja	Ja	Ja
Ja	Nein	Nein	Nein
Ja	Nein	Ja	Ja
Nein	Ja	Nein	Ja
Nein	Nein	Ja	Nein
Nein	Nein	Nein	Ja

- Leiten Sie das Attribut Churn mit Hilfe eines Entscheidungsbaums ab. Nutzen Sie hierfür als Splitkriterium den Gini-Index und entwickeln Sie solange neue Knoten bis entweder kein Splitattribut mehr verfügbar ist oder in einem Knoten alle Daten in der gleichen Klasse sind. **(10 Punkte)**

Wiederholung Vorlesung: Gini Index

- Idee
  Minimierung der „Heterogenität“ innerhalb der neuen Knoten

- Mathematisch
  Wahrscheinlichkeit, dass wiederholtes Ziehen mit Zurücklegen...
  ... von Fällen im Knoten zu unterschiedlichen Klassen führt

- Formal
  - Allgemein: $1 - \sum_{i=1}^{n} p_i^2$ mit n ist Anzahl der Klassen
  - 2-Klassen: $1 - p_0^2 - p_1^2$

- Wertebereich:
  - $\left[0 \ \ldots 1 - \frac{1}{k}\left[= 1 - k\left(\frac{1}{k}\right)^2\right]\right]$

- Interpretation
  - Gini Index = 0: Perfekte Klassifikation
  - Gini Index = $1 - \frac{1}{k}$: Alle Klassen gleich häufig vertreten

Aufgabe 1 (a) – Lösung (Split mit Gini-Index)

- Vorgehen:
  - Ermittlung Gini-Index für alle Splits
  - Wahl des Splits mit min. Gini-Index

Churn	CallCenter	Beschwerde	Zahlungsverzug
Ja	Ja	Ja	Ja
Ja	Nein	Nein	Nein
Ja	Nein	Ja	Ja
Nein	Ja	Nein	Ja
Nein	Nein	Ja	Nein
Nein	Nein	Nein	Ja

- Jetzt: Gini-Index für Split CallCenter

Klasse im Teilbaum	Gini-Index	Gewicht
Ja	$1-\left(\frac{\mathbf{1}}{\mathbf{2}}\right)^2-\left(\frac{1}{2}\right)^2=\frac{1}{2}$	$\frac{1}{3}$
Nein	$1-\left(\frac{2}{4}\right)^2-\left(\frac{2}{4}\right)^2=\frac{1}{2}$	$\frac{2}{3}$
Gewichtete Summe		$\frac{1}{2}\cdot\frac{1}{3}+\frac{1}{2}\cdot\frac{2}{3}=0{,}5$

Aufgabe 1 (a) – Lösung (Split mit Gini-Index)

- Vorgehen:
  - Ermittlung Gini-Index für alle Splits
  - Wahl des Splits mit min. Gini-Index

- Bisher
  - Gini-Index CallCenter: 0,50

Churn	CallCenter	Beschwerde	Zahlungsverzug
Ja	Ja	Ja	Ja
Ja	Nein	Nein	Nein
Ja	Nein	Ja	Ja
Nein	Ja	Nein	Ja
Nein	Nein	Ja	Nein
Nein	Nein	Nein	Ja

- Jetzt: Gini-Index für Split Beschwerde

Klasse im Teilbaum	Gini-Index	Gewicht
Ja	$1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 \approx 0{,}44$	$\frac{1}{2}$
Nein	$1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 \approx 0{,}44$	$\frac{1}{2}$
Gewichtete Summe		$0{,}44 \cdot \frac{1}{2} + 0{,}44 \cdot \frac{1}{2} = 0{,}44$

Aufgabe 1 (a) – Lösung (Split mit Gini-Index)

- Vorgehen:
  - Ermittlung Gini-Index für alle Splits
  - Wahl des Splits mit min. Gini-Index
- Bisher
  - Gini-Index CallCenter: 0,50
  - Gini-Index Beschwerde: 0,44
- Jetzt: Gini-Index für Split Zahlungsverzug

Churn	CallCenter	Beschwerde	Zahlungsverzug
Ja	Ja	Ja	Ja
Ja	Nein	Nein	Nein
Ja	Nein	Ja	Ja
Nein	Ja	Nein	Ja
Nein	Nein	Ja	Nein
Nein	Nein	Nein	Ja

Klasse im Teilbaum	Gini-Index	Gewicht
ja	$1-\left(\frac{2}{4}\right)^2-\left(\frac{2}{4}\right)^2=\frac{1}{2}$	$\frac{2}{3}$
nein	$1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2=\frac{1}{2}$	$\frac{1}{3}$
Gewichtete Summe		$\frac{1}{2}\cdot\frac{2}{3}+\frac{1}{2}\cdot\frac{1}{3}=0{,}5$

Aufgabe 1 (a) – Lösung (Split mit Gini-Index)

- Vorgehen:
  - Ermittlung Gini-Index für alle Splits
  - Wahl des Splits mit min. Gini-Index
- Bisher
  - Gini-Index CallCenter: 0,50
  - Gini-Index Beschwerde: 0,44
  - Gini-Index Zahlungsverzug; 0,50

Churn	CallCenter	Beschwerde	Zahlungsverzug
Ja	Ja	Ja	Ja
Ja	Nein	Nein	Nein
Ja	Nein	Ja	Ja
Nein	Ja	Nein	Ja
Nein	Nein	Ja	Nein
Nein	Nein	Nein	Ja

⇒ Das erste Splitattribut ist gemäß Gini-Index “Beschwerde”.

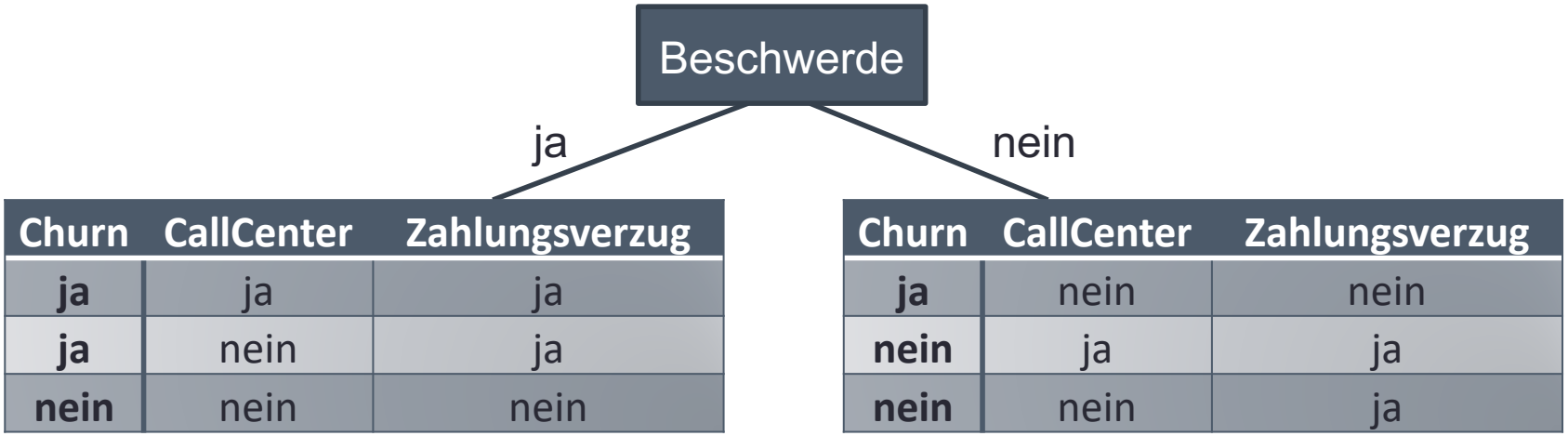
- Nächste Schritte
  - Aufmalen des Wurzelknotens
  - Aufteilen der Beobachtungen auf die beiden Kindknoten
  - Rekursives Durchführen der Schritte in den Kindknoten

Aufgabe 1 (a) – Lösung (Split mit Gini-Index)

- Bisher
  1. Split nach Beschwerde

Churn	CallCenter	Beschwerde	Zahlungsverzug
Ja	Ja	Ja	Ja
Ja	Nein	Nein	Nein
Ja	Nein	Ja	Ja
Nein	Ja	Nein	Ja
Nein	Nein	Ja	Nein
Nein	Nein	Nein	Ja

- Resultierender Baum

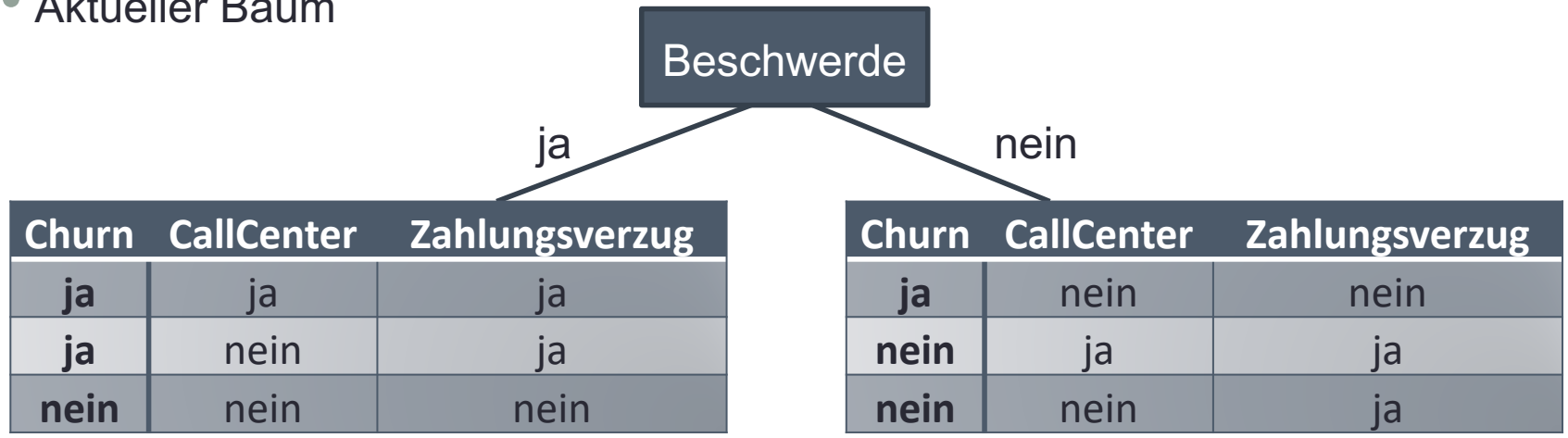


Churn	CallCenter	Zahlungsverzug
ja	ja	ja
ja	nein	ja
nein	nein	nein

Churn	CallCenter	Zahlungsverzug
ja	nein	nein
nein	ja	ja
nein	nein	ja

Aufgabe 1 (a) – Lösung (Split mit Gini-Index)

- Aktueller Baum



Churn	CallCenter	Zahlungsverzug
ja	ja	ja
ja	nein	ja
nein	nein	nein

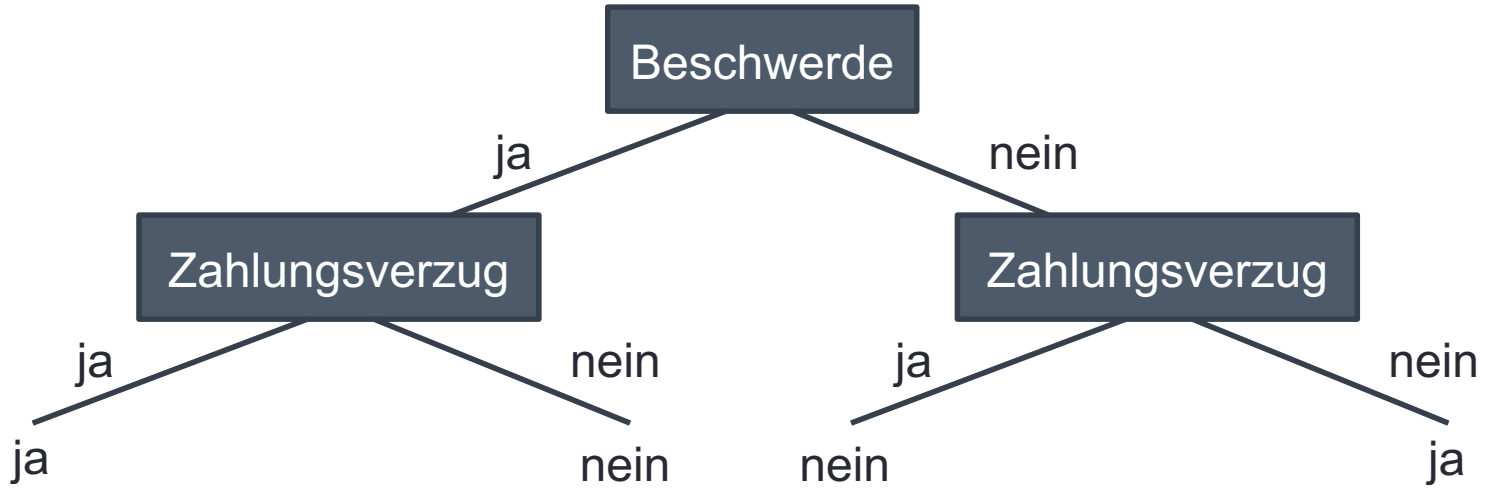
Churn	CallCenter	Zahlungsverzug
ja	nein	nein
nein	ja	ja
nein	nein	ja

- Prinzipielles Vorgehen
  Ermittlung des Gini-Index für beide Splits (CallCenter, Zahlungsverzug)…
  … im linken und im rechten Teilbaum

- Besonderheit hier:
  - Nach „ja“: Zahlungsverzug sagt Churn exakt vorher ⇒ Gini-Index: 0
  - Nach „nein“: Zahlungsverzug sagt Churn exakt vorher ⇒ Gini-Index: 0
  - ⇒ Finaler Baum direkt darstellbar

Aufgabe 1 (a) – Lösung (Split mit Gini-Index)

- Ergebnis



Aufgabe 1 (b) - Aufgabenstellung

- Gegeben seien folgende Trainingsdaten:

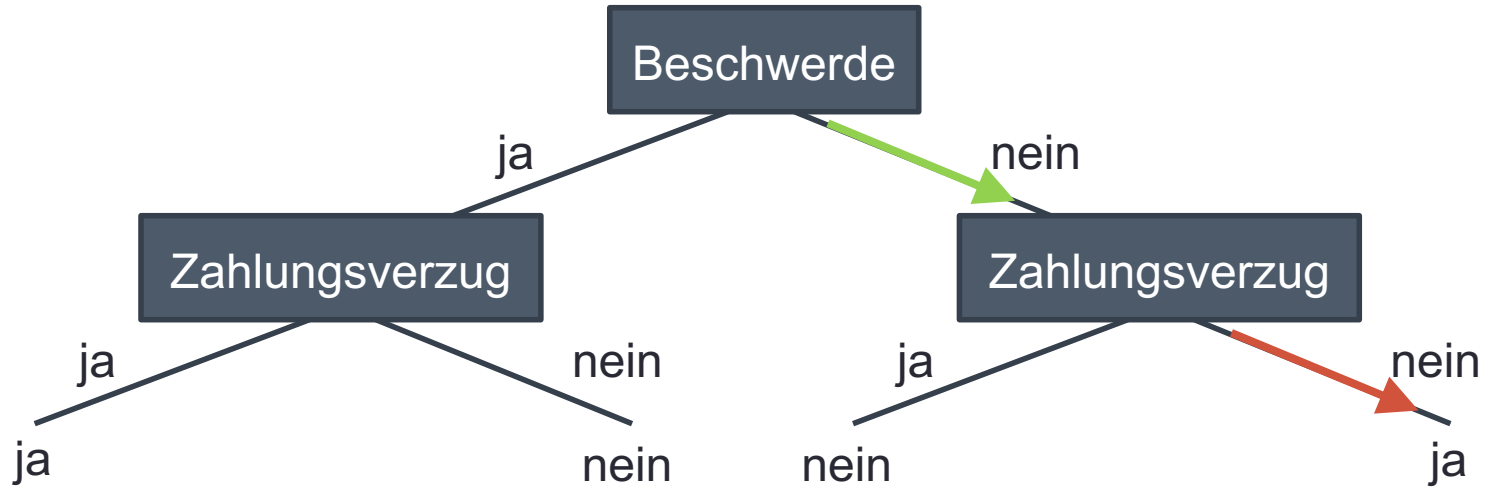
Churn	Anruf im CallCenter	Beschwerde	Zahlungsverzug
Ja	Ja	Ja	Ja
Ja	Nein	Nein	Nein
Ja	Nein	Ja	Ja
Nein	Ja	Nein	Ja
Nein	Nein	Ja	Nein
Nein	Nein	Nein	Ja

- Sagen Sie mit Hilfe des Entscheidungsbaums aus (a) die Empfehlung für folgende Beobachtungen voraus (Ja, Nein, Nein) und (Ja, Ja, Nein).

(5 Punkte)

Aufgabe 1 (b) – Lösung (Vorhersage)

- Entscheidungsbaum aus (a)

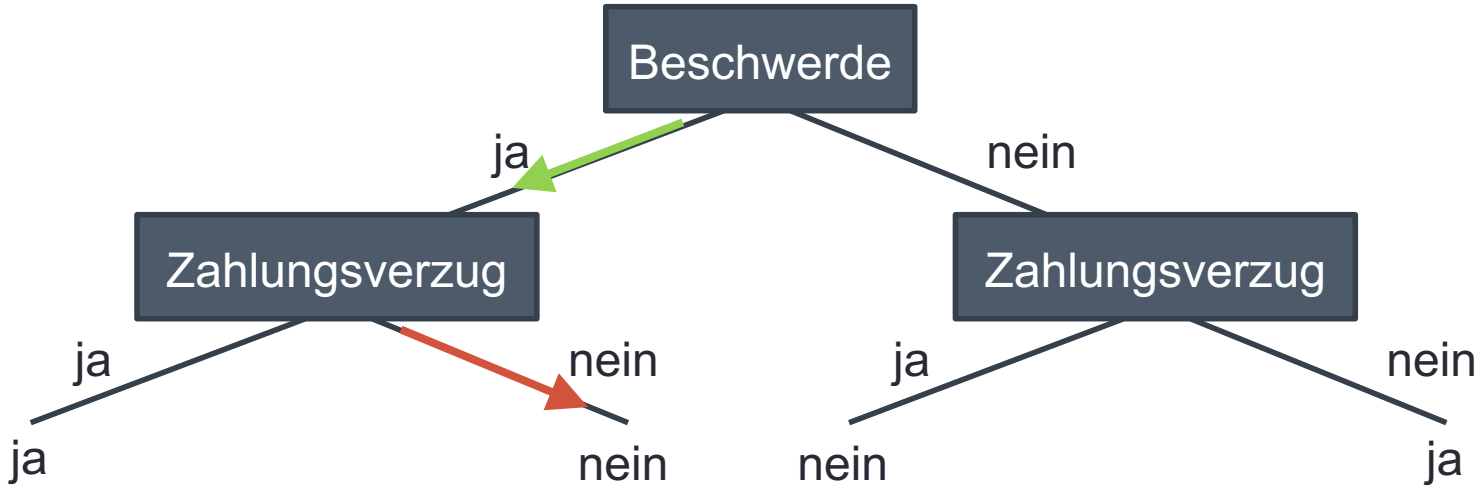


- Erste Vorhersage

Churn	Anruf im CallCenter	Beschwerde	Zahlungsverzug
Ja	Ja	Nein	Nein
	Ja	Ja	Nein

Aufgabe 1 (b) – Lösung (Vorhersage)

- Entscheidungsbaum aus (a)



- Zweite Vorhersage

Churn	Anruf im CallCenter	Beschwerde	Zahlungsverzug
Ja	Ja	Nein	Nein
Nein	Ja	Ja	Nein

Aufgabe 1 (c) - Aufgabenstellung

- Gegeben seien folgende Trainingsdaten:

Churn	Anruf im CallCenter	Beschwerde	Zahlungsverzug
Ja	Ja	Ja	Ja
Ja	Nein	Nein	Nein
Ja	Nein	Ja	Ja
Nein	Ja	Nein	Ja
Nein	Nein	Ja	Nein
Nein	Nein	Nein	Ja

- Leiten Sie das Attribut Churn wie in Aufgabe (a) ab, nutzen Sie jetzt das X^2-Maß zur Auswahl der Splits. **(10 Punkte)**

Wiederholung Vorlesung: χ^2 – Maß

- Idee
  - Werte lassen sich besonders gut in Klassen unterteilen, wenn abhängig
  - Zwei Verteilungen sind unabhängig, wenn $p(x_i \cap y_j) = p(x_i) \cdot p(y_j)$
  - Prüfung typischerweise mit χ^2-Test

- Mathematisch
  Hypothesentest selbst wird nicht ausgeführt, ...
  ... sondern nur die Teststatistik ermittelt und für Splits verglichen

- Formal
  - Prüfung über χ^2-Teststatistik: $\chi^2 = \sum_{j=1}^{k} \frac{(n_j - \overline{n_j})^2}{\overline{n_j}}$

    mit k ist Zahl der (Klasse, Attributwert)-Kombinationen
    und n_j ($\overline{n_j}$) Zahl der Beob. (erwartet) einer (Klasse, Attributwert)-Kombi.

- Interpretation
  - Je größer die χ^2-Teststatistik umso zuverlässiger kann davon... ausgegangen werden, dass Größen nicht unabhängig

Aufgabe 1 (c) – Lösung (Split mit χ^2 – Maß)

- Vorgehen:
  - Ermittlung χ^2 -Teststatistik für Splits
  - Wahl des Splits mit max. χ^2

Churn	CallCenter	Beschwerde	Zahlungsverzug
Ja	Ja	Ja	Ja
Ja	Nein	Nein	Nein
Ja	Nein	Ja	Ja
Nein	Ja	Nein	Ja
Nein	Nein	Ja	Nein
Nein	Nein	Nein	Ja

- Jetzt: χ^2 -Teststatistik für Split CallCenter

		Call Center		
		Ja	**Nein**	**Summe**
Churn	**Ja**	**1**	**2**	**3**
	Nein	**1**	**2**	**3**
	Summe	**2**	**4**	**6**

- $$X^2_{CallCenter} = \frac{\left(1-\frac{3\cdot 2}{6}\right)^2}{\frac{3\cdot 2}{6}} + \frac{\left(2-\frac{3\cdot 4}{6}\right)^2}{\frac{3\cdot 4}{6}} + \frac{\left(1-\frac{2\cdot 3}{6}\right)^2}{\frac{2\cdot 3}{6}} + \frac{\left(2-\frac{3\cdot 4}{6}\right)^2}{\frac{3\cdot 4}{6}} = 0+0+0+0 = 0$$

Aufgabe 1 (c) – Lösung (Split mit χ^2 – Maß)

- Vorgehen:
  - Ermittlung χ^2 -Teststatistik für Splits
  - Wahl des Splits mit max. χ^2

- Bisher
  - χ^2 für Split CallCenter: 0,00

Churn	CallCenter	Beschwerde	Zahlungsverzug
Ja	Ja	Ja	Ja
Ja	Nein	Nein	Nein
Ja	Nein	Ja	Ja
Nein	Ja	Nein	Ja
Nein	Nein	Ja	Nein
Nein	Nein	Nein	Ja

- Jetzt: χ^2 -Teststatistik für Split Beschwerde

		Beschwerde		
		Ja	**Nein**	**Summe**
Churn	**Ja**	**2**	**1**	**3**
	Nein	**1**	**2**	**3**
	Summe	**3**	**3**	**6**

- $X^2_{Beschwerde} = \frac{\left(2-\frac{3\cdot3}{6}\right)^2}{\frac{3\cdot3}{6}} + \frac{\left(1-\frac{3\cdot3}{6}\right)^2}{\frac{3\cdot3}{6}} + \frac{\left(1-\frac{3\cdot3}{6}\right)^2}{\frac{3\cdot3}{6}} + \frac{\left(2-\frac{3\cdot3}{6}\right)^2}{\frac{3\cdot3}{6}} = \frac{1}{4}\cdot\frac{6}{9} + \frac{1}{4}\cdot\frac{6}{9} + \frac{1}{4}\cdot\frac{6}{9} + \frac{1}{4}\cdot\frac{6}{9} = \frac{2}{3}$

Aufgabe 1 (c) – Lösung (Split mit χ^2 – Maß)

- Vorgehen:
  - Ermittlung χ^2 -Teststatistik für Splits
  - Wahl des Splits mit max. χ^2
- Bisher
  - χ^2 für Split CallCenter: 0,00
  - χ^2 für Split Beschwerde: 0,67
- Jetzt: χ^2 -Teststatistik für Split Zahlungsverzug

Churn	CallCenter	Beschwerde	Zahlungsverzug
Ja	Ja	Ja	Ja
Ja	Nein	Nein	Nein
Ja	Nein	Ja	Ja
Nein	Ja	Nein	Ja
Nein	Nein	Ja	Nein
Nein	Nein	Nein	Ja

		Zahlungsverzug		
		Ja	**Nein**	**Summe**
Churn	**Ja**	**2**	**1**	**3**
	Nein	**2**	**1**	**3**
	Summe	**4**	**2**	**6**

- $$X^2_{Beschwerde} = \frac{\left(2-\frac{3\cdot4}{6}\right)^2}{\frac{3\cdot4}{6}} + \frac{\left(1-\frac{3\cdot2}{6}\right)^2}{\frac{3\cdot2}{6}} + \frac{\left(2-\frac{3\cdot4}{6}\right)^2}{\frac{3\cdot4}{6}} + \frac{\left(1-\frac{3\cdot2}{6}\right)^2}{\frac{3\cdot2}{6}} = 0+0+0+0 = 0$$

Aufgabe 1 (c) – Lösung (Split mit χ^2 – Maß)

- Vorgehen:
  - Ermittlung χ^2 -Teststatistik für Splits
  - Wahl des Splits mit max. χ^2
- Bisher
  - χ^2 für Split CallCenter: 0,00
  - χ^2 für Split Beschwerde: 0,67
  - χ^2 für Split Zahlungsverzug; 0,00

Churn	CallCenter	Beschwerde	Zahlungsverzug
Ja	Ja	Ja	Ja
Ja	Nein	Nein	Nein
Ja	Nein	Ja	Ja
Nein	Ja	Nein	Ja
Nein	Nein	Ja	Nein
Nein	Nein	Nein	Ja

⇒ Das erste Splitattribut ist gemäß χ^2-Teststatistik “Beschwerde”.

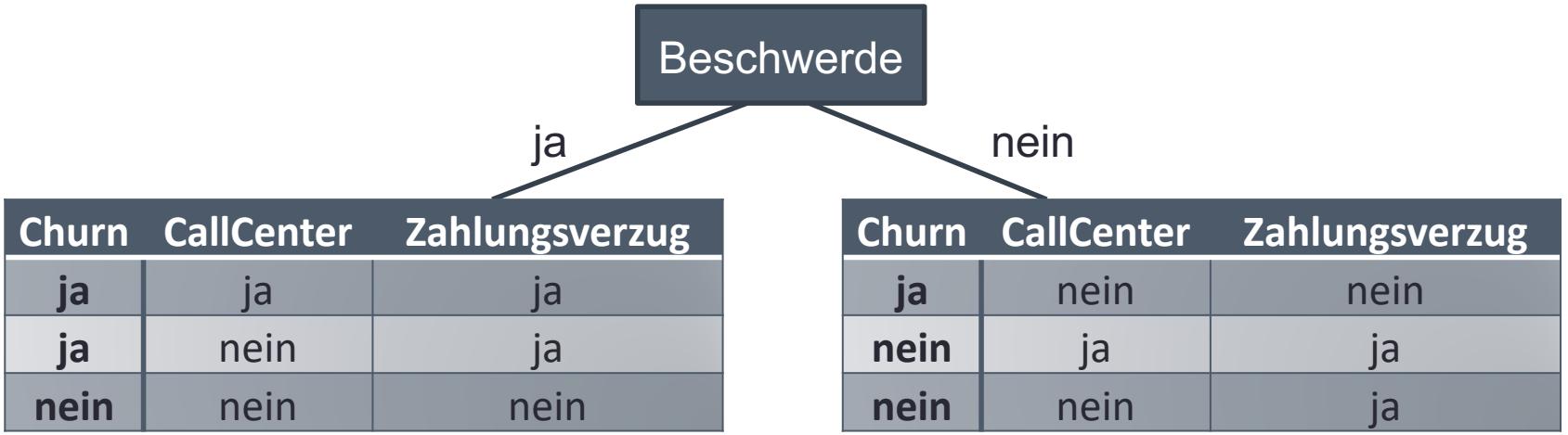
- Nächste Schritte
  - Aufmalen des Wurzelknotens
  - Aufteilen der Beobachtungen auf die beiden Kindknoten
  - Rekursives Durchführen der Schritte in den Kindknoten

Aufgabe 1 (c) – Lösung (Split mit χ^2 – Maß)

- Bisher
  1. Split nach Beschwerde

Churn	CallCenter	Beschwerde	Zahlungsverzug
Ja	Ja	Ja	Ja
Ja	Nein	Nein	Nein
Ja	Nein	Ja	Ja
Nein	Ja	Nein	Ja
Nein	Nein	Ja	Nein
Nein	Nein	Nein	Ja

- Resultierender Baum (vgl. Aufgabe 1 (a))

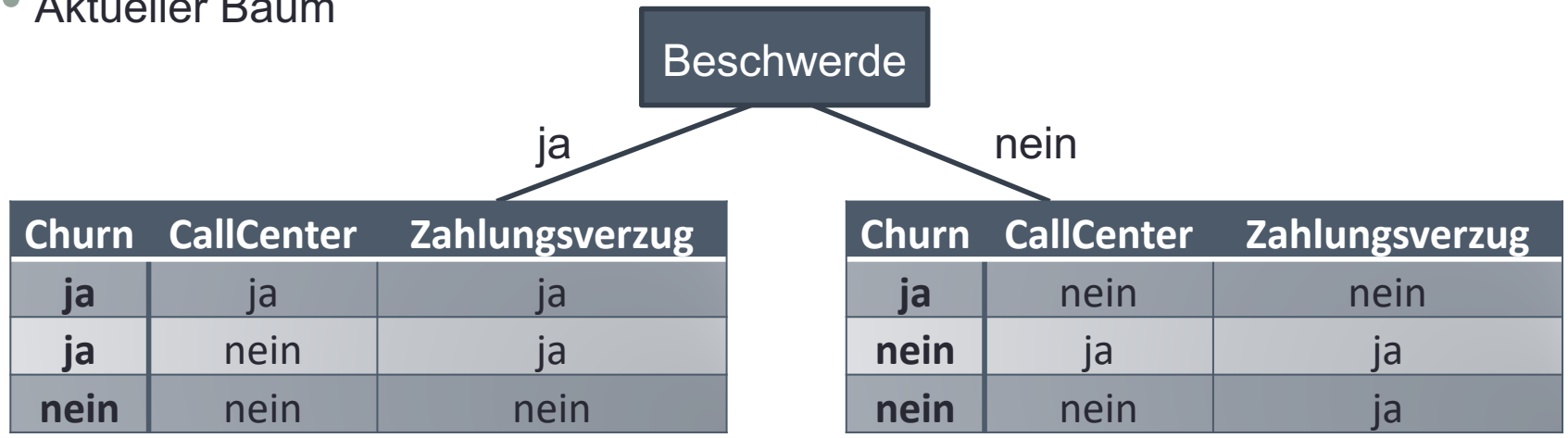


Churn	CallCenter	Zahlungsverzug
ja	ja	ja
ja	nein	ja
nein	nein	nein

Churn	CallCenter	Zahlungsverzug
ja	nein	nein
nein	ja	ja
nein	nein	ja

Aufgabe 1 (c) – Lösung (Split mit χ^2 – Maß)

- Aktueller Baum



Churn	CallCenter	Zahlungsverzug
ja	ja	ja
ja	nein	ja
nein	nein	nein

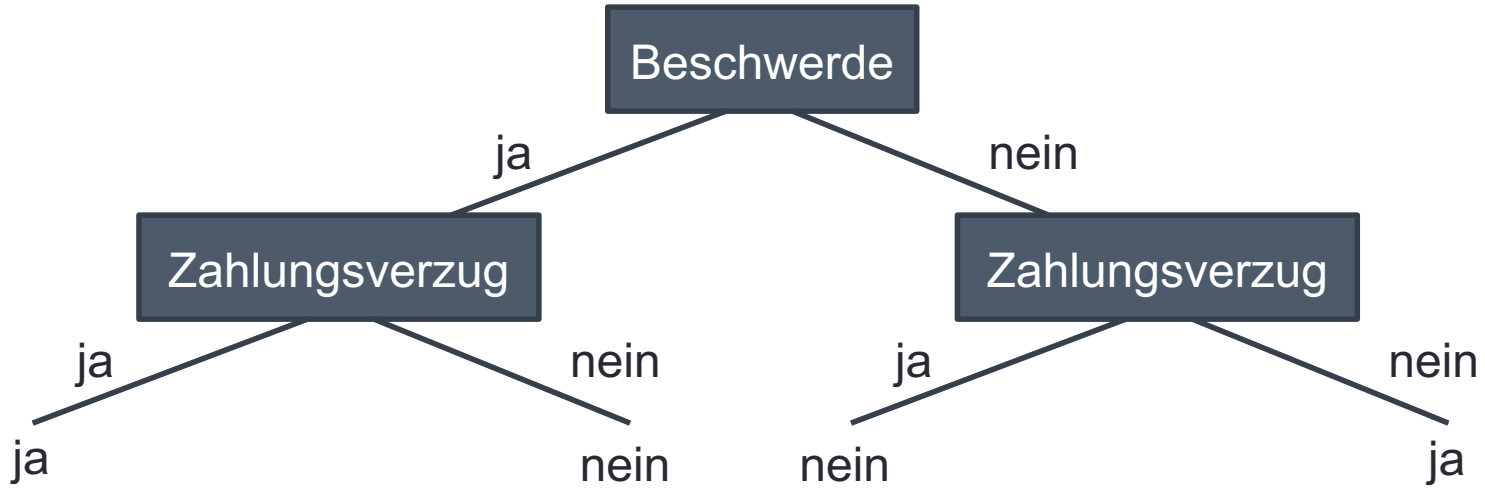
Churn	CallCenter	Zahlungsverzug
ja	nein	nein
nein	ja	ja
nein	nein	ja

- Prinzipielles Vorgehen
  Ermittlung der χ^2-Teststatistik für Splits (CallCenter, Zahlungsverzug)…
  … im linken und im rechten Teilbaum

- Besonderheit hier (vgl. Aufgab 1 (a)):
  - Nach „ja“: Zahlungsverzug sagt Churn exakt vorher ⇒ max. χ^2-Maß
  - Nach „nein“: Zahlungsverzug sagt Churn exakt vorher ⇒ max. χ^2-Maß
  - ⇒ Finaler Baum direkt darstellbar

Aufgabe 1 (c) – Lösung (Split mit χ^2 – Maß)

- Ergebnis



Agenda

- Aufgabenblatt 1 – Deskriptive Methoden zur Datenexploration

- Aufgabenblatt 2 – Klassifikation
  - Aufgabe 1 – Entscheidungsbaum I
  - Aufgabe 2 – Entscheidungsbaum II
  - Aufgabe 3 – Verständnisfragen

- Aufgabenblatt 3 – Recommender Systems
- Aufgabenblatt 4 – Clusteringverfahren
- Aufgabenblatt 5 – Stream Mining
- Aufgabenblatt 6 – Social Network Analysis

Aufgabe 2 (a) - Aufgabenstellung

- Gegeben seien folgende Trainingsdaten:

Empfehlung	Form	Farbe	Material
Kaufen	Formumspielend	rot	Leder
Kaufen	Körperbetont	gelb	Baumwolle
Kaufen	Körperbetont	rot	Leder
Nicht kaufen	Formumspielend	gelb	Leder
Nicht kaufen	Körperbetont	rot	Baumwolle
Nicht kaufen	Körperbetont	gelb	Leder

- Leiten Sie das Attribut Empfehlung mit Hilfe eines Entscheidungsbaums ab. Nutzen Sie hierfür als Splitkriterium den Informationsgewinn und entwickeln Sie solange neue Knoten bis entweder kein Splitattribut mehr verfügbar ist oder in einem Knoten alle Daten in der gleichen Klasse sind. **(10 Punkte)**

Wiederholung Vorlesung: Informationsgewinn

- Idee
  - Bewertung der Reduktion der „Unordnung“ durch Split
  - „Shannon Entropie“ als Maß für „Ordnung“ eines Systems

- Ursprung Informationstheorie
  - Durch „Kodierung“ von Daten mit mehr Parametern: mehr Information

- Formal
  - $I = \text{Entropie}(\text{Gesamtdaten}) - \text{Entropie}(\text{Knoten nach Split})$
    mit Entropie: $-\sum_{i=1}^{n}(p_i \cdot \log_2 p_i)$ und n ist Anzahl der Klassen

- Interpretation
  - Je größer der Informationsgewinn...
    ... umso besser das Splitattribut

Aufgabe 2 (a) – Lösung (Split mit Information Gain)

- Vorgehen:
  - Ermittlung Entropie gesamt
  - Ermittlung Entropie für alle Splits
  - Wahl des Splits mit max. Entropiereduktion (Information Gain)

Empfehlung	Form	Farbe	Material
Kaufen	Formum.	rot	Leder
Kaufen	Körperbet.	gelb	Baumwolle
Kaufen	Körperbet.	rot	Leder
Nicht kaufen	Formum.	gelb	Leder
Nicht kaufen	Körperbet.	rot	Baumwolle
Nicht kaufen	Körperbet.	gelb	Leder

- $E_{ges} = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = -(-0{,}5 - 0{,}5) = 1$
- Jetzt: Entropie für Split Form

Klasse im Teilbaum	Entropie	Gewicht
formumspielend	$-\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$	$\frac{2}{6}$
körperbetont	$-\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = 1$	$\frac{4}{6}$
Gewichtete Summe		$\frac{2}{6}\cdot 1 + \frac{4}{6}\cdot 1 = 1$

- Information Gain: $E_{ges} - E_{Form} = 1 - 1 = 0$

Aufgabe 2 (a) – Lösung (Split mit Information Gain)

- Bisher
  - $E_{ges} = 1$
  - $I_{Form} = 0{,}000$

Empfehlung	Form	Farbe	Material
Kaufen	Formum.	rot	Leder
Kaufen	Körperbet.	gelb	Baumwolle
Kaufen	Körperbet.	rot	Leder
Nicht kaufen	Formum.	gelb	Leder
Nicht kaufen	Körperbet.	rot	Baumwolle
Nicht kaufen	Körperbet.	gelb	Leder

- Jetzt: Information Gain für Split Farbe

Klasse im Teilbaum	Entropie	Gewicht
rot	$-\left(\frac{1}{3}\log_2\frac{1}{3}+\frac{2}{3}\log_2\frac{2}{3}\right) \approx 0{,}918$	$\frac{3}{6}$
gelb	$-\left(\frac{1}{3}\log_2\frac{1}{3}+\frac{2}{3}\log_2\frac{2}{3}\right) \approx 0{,}918$	$\frac{3}{6}$
Gewichtete Summe		$\frac{3}{6}\cdot 0{,}918+\frac{3}{6}\cdot 0{,}918 = 0{,}918$

- Information Gain: $E_{ges} - E_{Farbe} = 1 - 0{,}918 = 0{,}082$

Aufgabe 2 (a) – Lösung (Split mit Information Gain)

- Bisher
  - $E_{ges} = 1$
  - $I_{Form} = 0{,}000$
  - $I_{Farbe} = 0{,}082$

Empfehlung	Form	Farbe	Material
Kaufen	Formum.	rot	Leder
Kaufen	Körperbet.	gelb	Baumwolle
Kaufen	Körperbet.	rot	Leder
Nicht kaufen	Formum.	gelb	Leder
Nicht kaufen	Körperbet.	rot	Baumwolle
Nicht kaufen	Körperbet.	gelb	Leder

- Jetzt: Information Gain für Split Material

Klasse im Teilbaum	Entropie	Gewicht
Baumwolle	$-\left(\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{2}\log_2\frac{1}{2}\right)=1$	$\frac{2}{6}$
Leder	$-\left(\frac{2}{4}\log_2\frac{2}{4}+\frac{2}{4}\log_2\frac{2}{4}\right)=1$	$\frac{4}{6}$
Gewichtete Summe		$\frac{2}{6}\cdot 1+\frac{4}{6}\cdot 1=1$

- Information Gain: $E_{ges} - E_{Material} = 1 - 1 = 0$

Aufgabe 2 (a) – Lösung (Split mit Information Gain)

- Vorgehen:
  - Ermittlung Information Gain für Splits
  - Wahl des Splits mit max. Info. Gain

Empfehlung	Form	Farbe	Material
Kaufen	Formum.	rot	Leder
Kaufen	Körperbet.	gelb	Baumwolle
Kaufen	Körperbet.	rot	Leder
Nicht kaufen	Formum.	gelb	Leder
Nicht kaufen	Körperbet.	rot	Baumwolle
Nicht kaufen	Körperbet.	gelb	Leder

- Bisher
  - $I_{Form} = 0{,}000$
  - $I_{Farbe} = 0{,}082$
  - $I_{Material} = 0{,}000$

⇒ Das erste Splitattribut ist gemäß Information Gain “Farbe”.

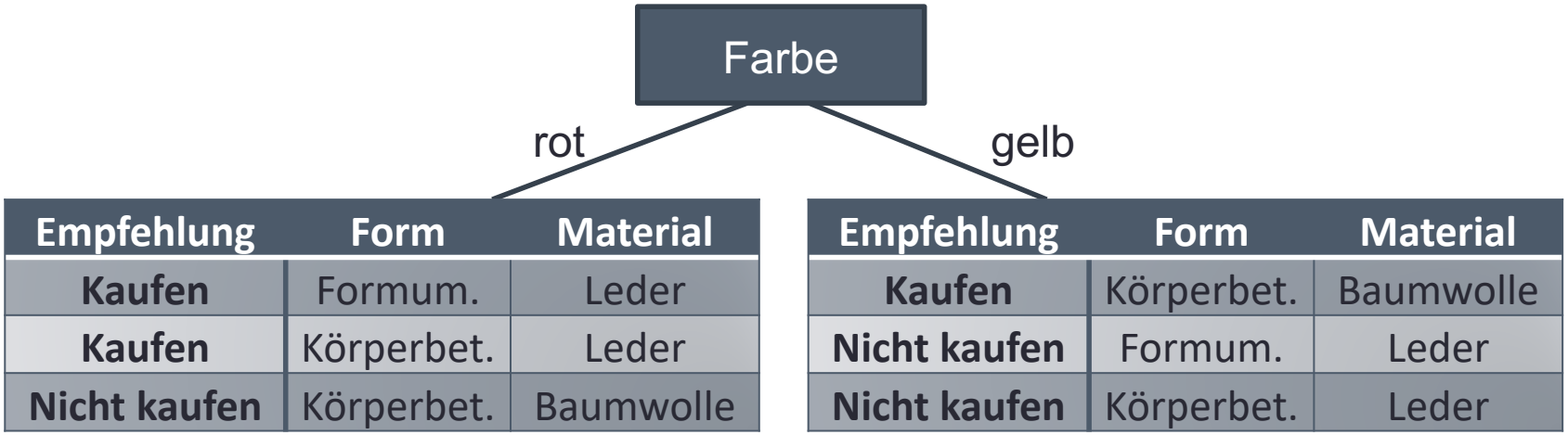
- Nächste Schritte
  - Aufmalen des Wurzelknotens
  - Aufteilen der Beobachtungen auf die beiden Kindknoten
  - Rekursives Durchführen der Schritte in den Kindknoten

Aufgabe 2 (a) – Lösung (Split mit Information Gain)

- Bisher
  1. Split nach Farbe

Empfehlung	Form	Farbe	Material
Kaufen	Formum.	rot	Leder
Kaufen	Körperbet.	gelb	Baumwolle
Kaufen	Körperbet.	rot	Leder
Nicht kaufen	Formum.	gelb	Leder
Nicht kaufen	Körperbet.	rot	Baumwolle
Nicht kaufen	Körperbet.	gelb	Leder

- Resultierender Baum

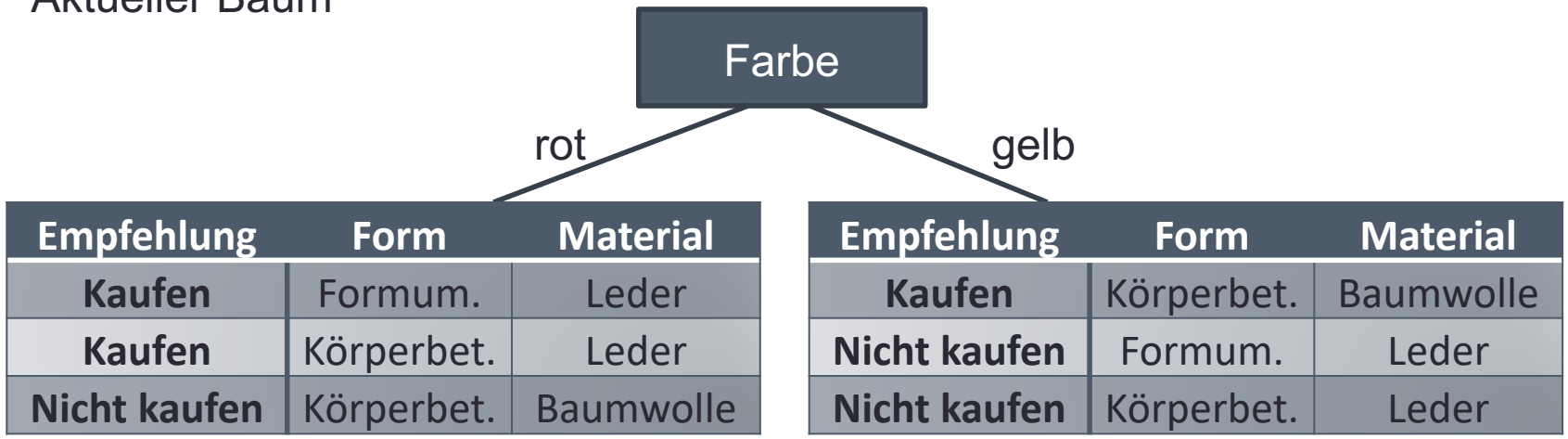


Empfehlung	Form	Material
Kaufen	Formum.	Leder
Kaufen	Körperbet.	Leder
Nicht kaufen	Körperbet.	Baumwolle

Empfehlung	Form	Material
Kaufen	Körperbet.	Baumwolle
Nicht kaufen	Formum.	Leder
Nicht kaufen	Körperbet.	Leder

Aufgabe 2 (a) – Lösung (Split mit Information Gain)

- Aktueller Baum



Empfehlung	Form	Material
Kaufen	Formum.	Leder
Kaufen	Körperbet.	Leder
Nicht kaufen	Körperbet.	Baumwolle

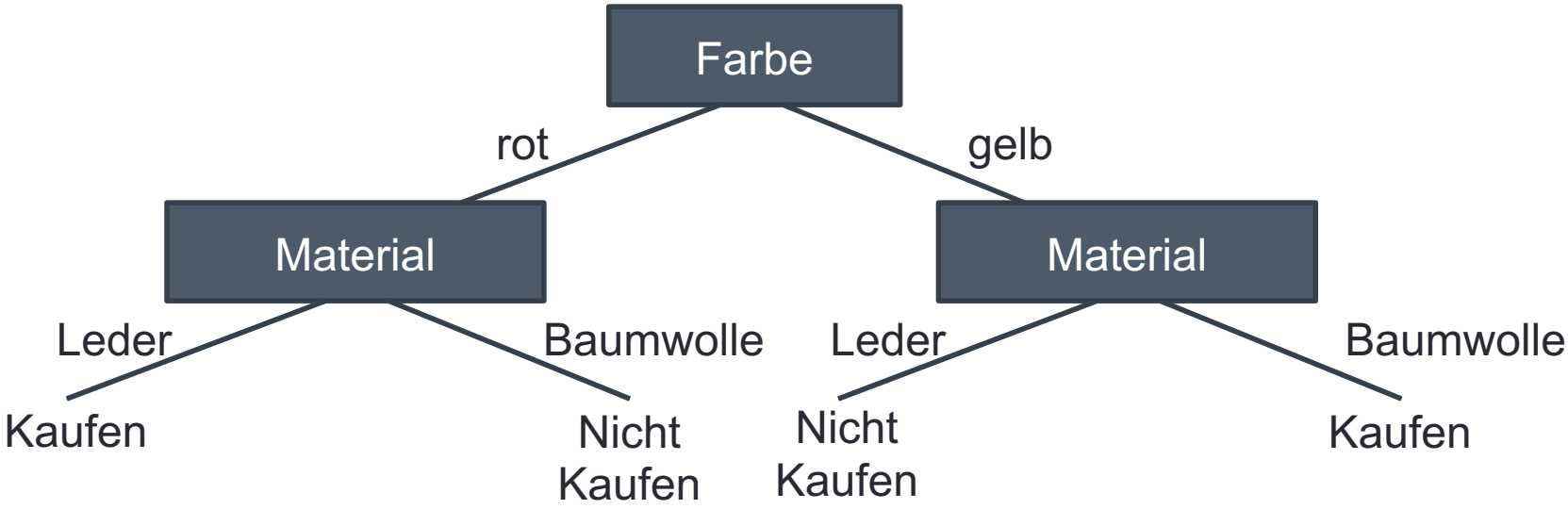
Empfehlung	Form	Material
Kaufen	Körperbet.	Baumwolle
Nicht kaufen	Formum.	Leder
Nicht kaufen	Körperbet.	Leder

- Prinzipielles Vorgehen
  Ermittlung des Information Gain für Splits (Form, Material)…
  … im linken und im rechten Teilbaum

- Besonderheit hier:
  - Nach „rot“: Material sagt Empfehlung vorher ⇒ max. Information Gain
  - Nach „gelb“: Material sagt Empfehlung vorher ⇒ max. Information Gain
  - ⇒ Finaler Baum direkt darstellbar

Aufgabe 2 (a) – Lösung (Split mit Information Gain)

- Ergebnis



Aufgabe 2 (b) - Aufgabenstellung

- Gegeben seien folgende Trainingsdaten:

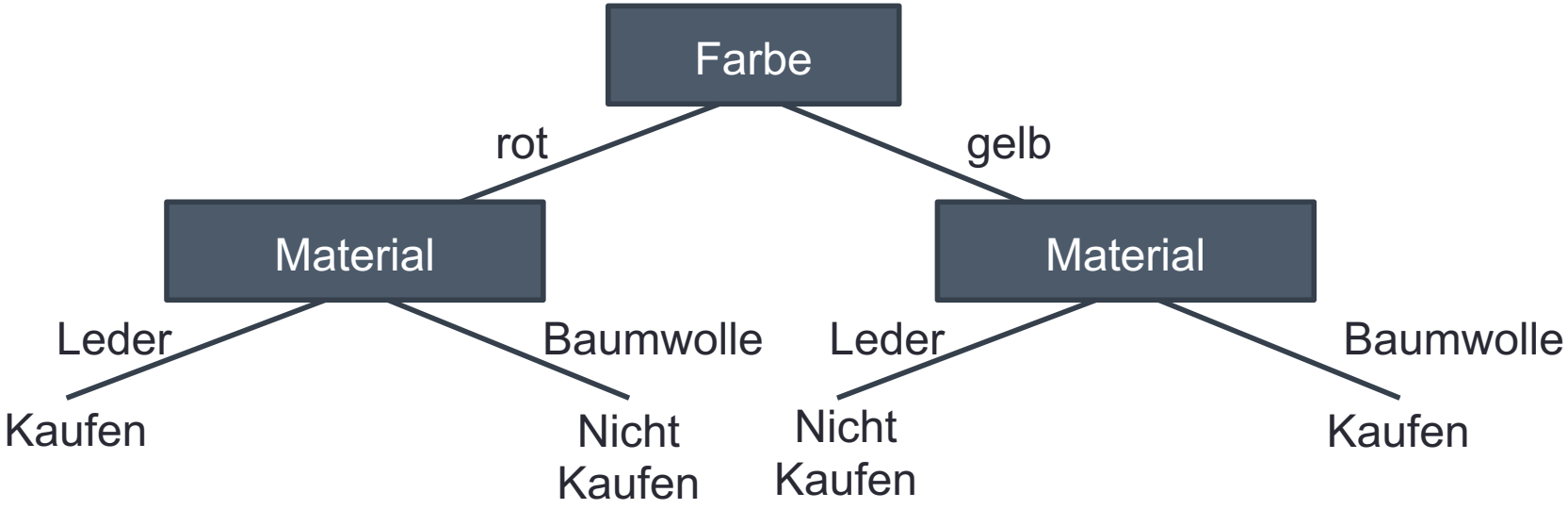
Empfehlung	Form	Farbe	Material
Kaufen	Formumspielend	rot	Leder
Kaufen	Körperbetont	gelb	Baumwolle
Kaufen	Körperbetont	rot	Leder
Nicht kaufen	Formumspielend	gelb	Leder
Nicht kaufen	Körperbetont	rot	Baumwolle
Nicht kaufen	Körperbetont	gelb	Leder

- Überführen Sie den Entscheidungsbaum aus (a) in eine Menge von Regeln.

(7 Punkte)

Aufgabe 2 (b) – Lösung (Ableiten von Regeln)

- Entscheidungsbaum



- Default Regel: Nicht kaufen
- Kaufen, wenn
  - Farbe rot und Material Leder
  - Farbe gelb und Material Baumwolle

Aufgabe 2 (c) - Aufgabenstellung

- Gegeben seien folgende Trainingsdaten:

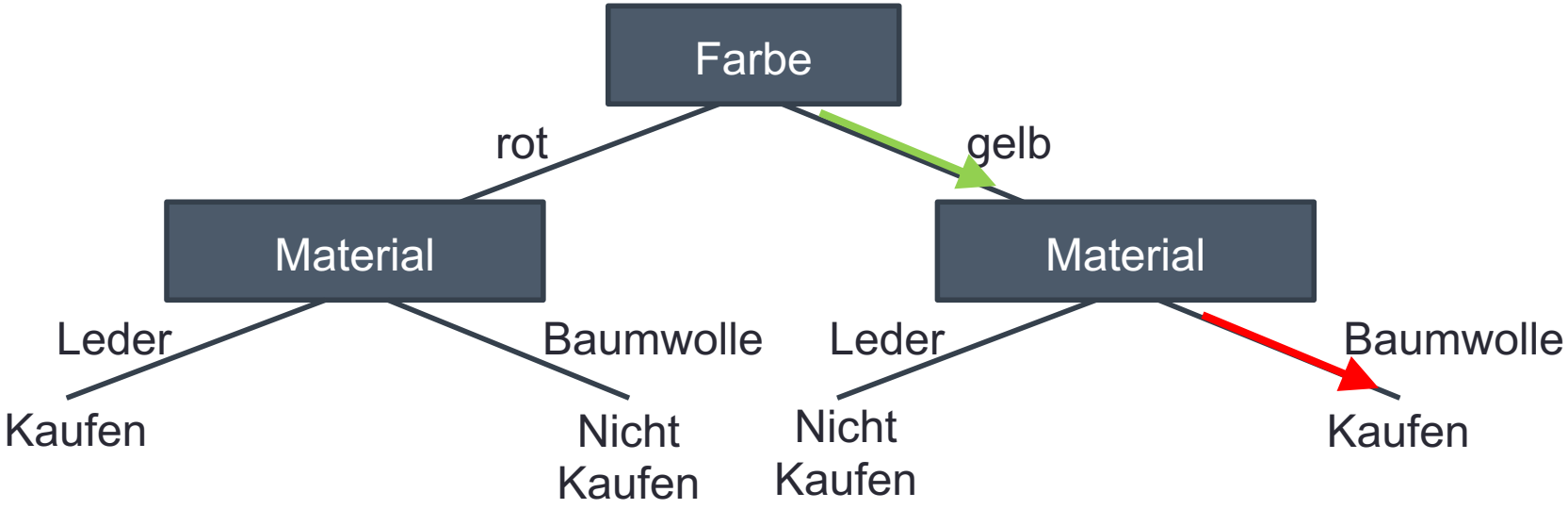
Empfehlung	Form	Farbe	Material
Kaufen	Formumspielend	rot	Leder
Kaufen	Körperbetont	gelb	Baumwolle
Kaufen	Körperbetont	rot	Leder
Nicht kaufen	Formumspielend	gelb	Leder
Nicht kaufen	Körperbetont	rot	Baumwolle
Nicht kaufen	Körperbetont	gelb	Leder

- Sagen Sie mit Hilfe des Entscheidungsbaums aus (a) die Empfehlung für folgende Beobachtung voraus (Formumspielend, gelb, Baumwolle).

(3 Punkte)

Aufgabe 2 (c) – Lösung (Vorhersage)

- Entscheidungsbaum



- Vorhersage

Empfehlung	Form	Farbe	Material
Kaufen	Formumspielend	gelb	Baumwolle

Aufgabe 2 (d) - Aufgabenstellung

- Gegeben seien folgende Trainingsdaten:

Empfehlung	Form	Farbe	Material
Kaufen	Formumspielend	rot	Leder
Kaufen	Körperbetont	gelb	Baumwolle
Kaufen	Körperbetont	rot	Leder
Nicht kaufen	Formumspielend	gelb	Leder
Nicht kaufen	Körperbetont	rot	Baumwolle
Nicht kaufen	Körperbetont	gelb	Leder

- Nutzen Sie einen Naive-Bayes-Klassifikator um für das Tupel aus (c) die Empfehlung vorherzusagen. **(10 Punkte)**

Wiederholung Vorlesung: Bayes Klassifikation

- Gesucht
  - $p(H_\mathrm{i} \mid X) = \frac{p(X|H_\mathrm{i}) \cdot p(H_\mathrm{i})}{p(X)}$ über den Satz von Bayes
- Vereinfachungen
  - $p(X)$ nicht nötig, da Nenner aller Terme
  - Abschätzung Wahrscheinlichkeit Klasse i tritt ein ($p(H_\mathrm{i})$):
    $p(H_i) = \frac{|Fälle\ in\ Klasse\ i|}{|\text{Alle Fälle im Trainingsset}|}$
  - Abschätzung Wahrscheinlichkeikt Fall x tritt ein, wenn H_i eintrat ($p(X \mid H_\mathrm{i})$)
    $p(X \mid \mathrm{H}_i) = \prod_{k=1}^{n} p(x_k \mid H_i)$ mit $p(x_k \mid \mathrm{H}_i) = \frac{|Fälle\ in\ Klasse\ i\ mit\ Attributwert\ x_k|}{|\text{Alle Fälle der Klasse i im Trainingsset}|}$

    (Gilt nur, da Unabhängigkeit der Klassen angenommen)
- Vorgehen
  - Berechnung des Zählers $p(X|H_\mathrm{i}) \cdot p(H_\mathrm{i})$ [von $p(H_\mathrm{i} \mid X)$] für alle Klassen i
  - Wahl der Klasse i mit maximalem $p(X|H_\mathrm{i}) \cdot p(H_\mathrm{i})$

Aufgabe 2 (d) – Lösung (Naive Bayes Klassifikator)

- Vorgehen:
  - Ermittlung $p(X|H_\mathrm{i}) \cdot p(H_\mathrm{i})$ [von $p(H_\mathrm{i} \mid X)$] für alle Klassen i
  - Wahl des Splits mit max. $p(H_\mathrm{i} \mid X)$

Empfehlung	Form	Farbe	Material
Kaufen	Formum.	rot	Leder
Kaufen	Körperbet.	gelb	Baumwolle
Kaufen	Körperbet.	rot	Leder
Nicht kaufen	Formum.	gelb	Leder
Nicht kaufen	Körperbet.	rot	Baumwolle
Nicht kaufen	Körperbet.	gelb	Leder

- Wahrscheinlichkeit der Klassen
  - $p(H_{kaufen}) = \frac{3}{6} = 0{,}5$
  - $p(H_{nicht\ kaufen}) = \frac{3}{6} = 0{,}5$

- Bedingte Wahrscheinlichkeiten Ausprägungen gegeben die Klassen
  - $p(formumspielend \mid H_{kaufen}) = \frac{1}{3}$
  - $p(formumspielend \mid H_{nicht\ kaufen}) = \frac{1}{3}$
  - $p(gelb \mid H_{kaufen}) = \frac{1}{3}$
  - $p(gelb \mid H_{nicht\ kaufen}) = \frac{2}{3}$
  - $p(Baumwolle \mid H_{kaufen}) = \frac{1}{3}$
  - $p(Baumwolle \mid H_{nicht\ kaufen}) = \frac{1}{3}$

Aufgabe 2 (d) – Lösung (Naive Bayes Klassifikator)

- Vorgehen:
  - Ermittlung $p(X|H_\mathrm{i}) \cdot p(H_\mathrm{i})$
    [von $p(H_\mathrm{i} \mid X)$] für alle Klassen i
  - Wahl des Splits mit max. $p(H_\mathrm{i} \mid X)$
- Bisher
  - $p(H_{kaufen}) = \frac{3}{6} = 0{,}5;\ p(H_{nicht\ kaufen}) = \frac{3}{6} = 0{,}5$
  - $p(formumspielend \mid H_{kaufen}) = \frac{1}{3};\ p(formumspielend \mid H_{nicht\ kaufen}) = \frac{1}{3}$
  - $p(gelb \mid H_{kaufen}) = \frac{1}{3}; p(gelb \mid H_{nicht\ kaufen}) = \frac{2}{3}$
  - $p(Baumwolle \mid H_{kaufen}) = \frac{1}{3};\ p(Baumwolle \mid H_{nicht\ kaufen}) = \frac{1}{3}$
- Vorhersage für (Formumspielend, gelb, Baumwolle)
  - $p(H_{kaufen} \mid X) = \frac{0{,}5 \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3}}{P(X)} = \frac{1}{54}$
  - $p(H_{nicht\ kaufen} \mid X) = \frac{0{,}5 \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{3}}{P(X)} = \frac{1}{27}$

  ⇒ Die Vorhersage für das gegebene Tupel ist „nicht kaufen“.

Agenda

- Aufgabenblatt 1 – Deskriptive Methoden zur Datenexploration

- Aufgabenblatt 2 – Klassifikation
  - Aufgabe 1 – Entscheidungsbaum I
  - Aufgabe 2 – Entscheidungsbaum II
  - Aufgabe 3 – Verständnisfragen

- Aufgabenblatt 3 – Recommender Systems
- Aufgabenblatt 4 – Clusteringverfahren
- Aufgabenblatt 5 – Stream Mining
- Aufgabenblatt 6 – Social Network Analysis

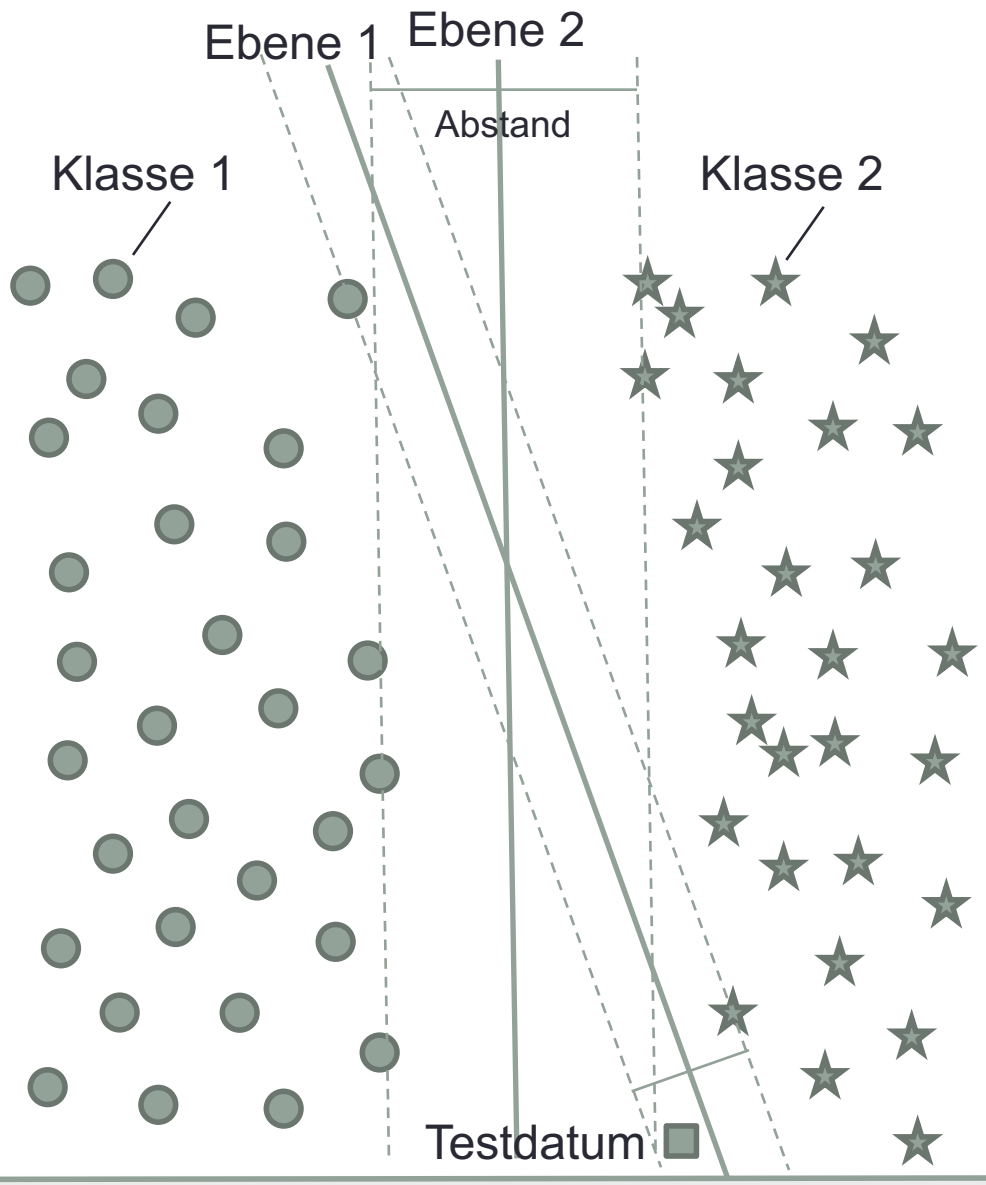
Aufgabe 3 (a) - Aufgabenstellung

- Beschreiben Sie, wie Support Vector Machines funktionieren. **(5 Punkte)**

Wiederholung Vorlesung

- Idee
  - Wahl einer Ebene, ...
    ... die Klassen trennt
  - Berechnung des Abstands...
    ... Zwischen Ebene und nächstem Punkt jeder Seite (Support Vector)
  - Ebene mit höchstem Abstand...
    ... trennt Klassen bestmöglich

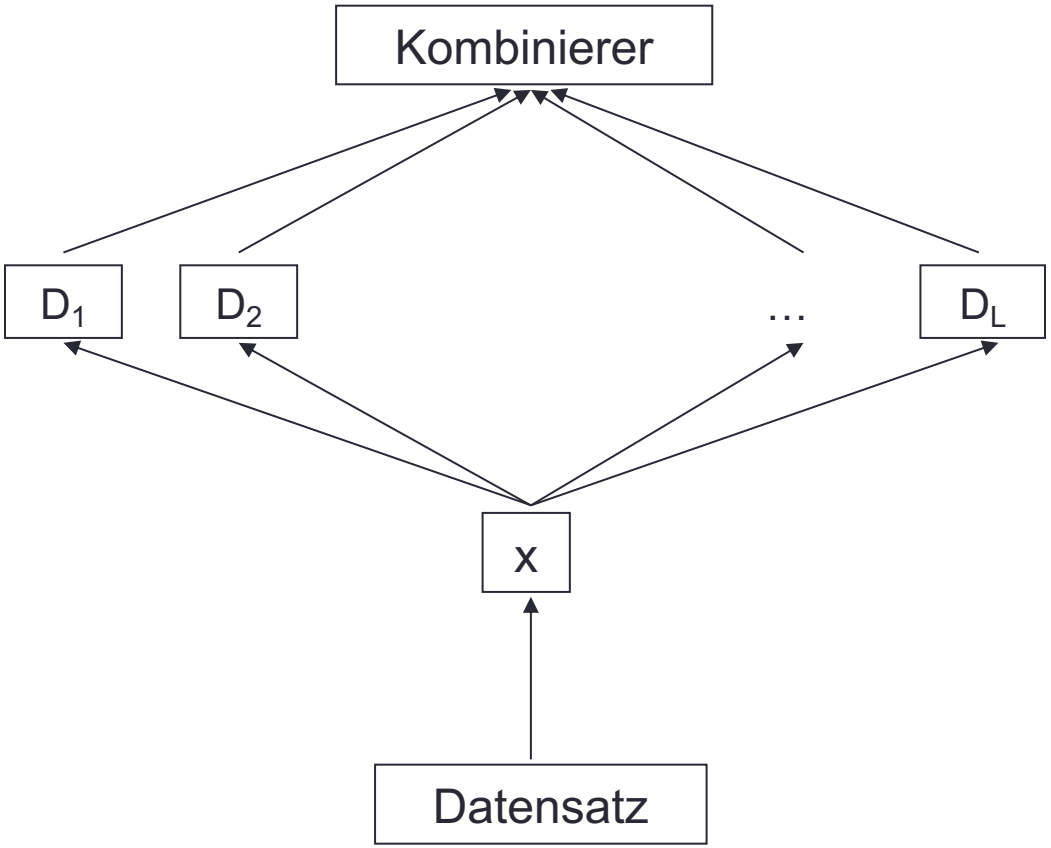
- Prüfung der Idee
  - Betrachtung Testdatum
  - Ebene 1: Klasse 1
  - Ebene 2: Klasse 2
  - Zuordnung zu Klasse 2...
    ... sinnvoller!



Aufgabe 3 (b) - Aufgabenstellung

- Diskutieren Sie, wie die Qualität von Klassifikatoren durch den Benutzer deutlich gesteigert werden kann. **(5 Punkte)**

Wiederholung Vorlesung: Kombinierte Klassifikatoren



Kombinations-Ebene:
Einsatz verschiedener Kombinationstechniken

Klassifikator-Ebene:
Einsatz verschiedener Klassifikatoren

Feature-Ebene:
Einsatz verschiedener Feature-Mengen

Daten-Ebene:
Einsatz verschiedener Teilmengen