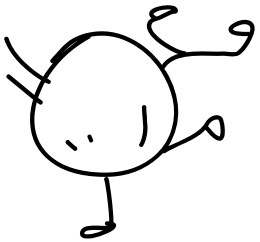


Entscheidungsbäume

Vorlesung 186.844

11.12.2014

TECHNISCHE
UNIVERSITÄT
WIEN
Vienna University of Technology

Überblick

I. Was sind Entscheidungsbäume?
II. Erstellen von Entscheidungsbäumen
III. CART im Detail
IV. Möglichkeiten und Einschränkungen

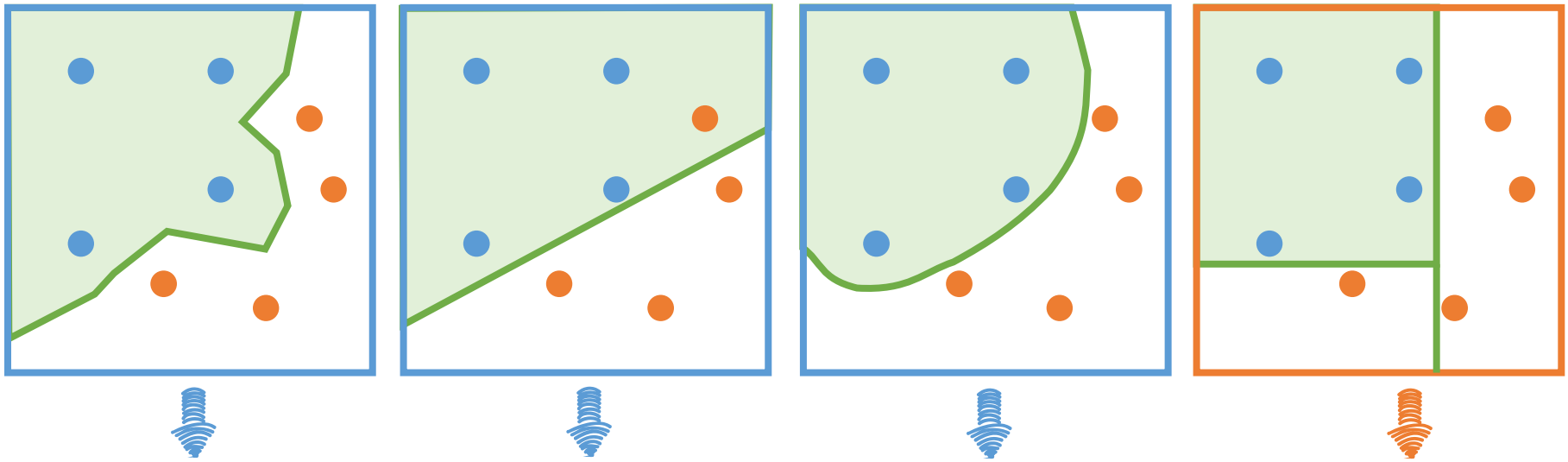
© Nicole M. Artner

Was sind Entscheidungsbäume?

© Nicole M. Artner

Entscheidungsgrenzen

… gefunden oder definiert durch



Nächster Nachbar (Nearest Neighbor)

lineare Diskriminan-tenfunktion

nicht-lineare Diskriminan-tenfunktion

Entscheid-ungsbäume

© Nicole M. Artner

Merkmale

bisher:

- Merkmalsvektoren mit Elementen aus $\mathbb{R}$
- Distanz zwischen Merkmalsvektoren → Metriken, Ähnlichkeitsmaße (similarity measures), etc.

$$\boldsymbol{x} \in \mathbb{R}^{n \cdot m \cdot f}$$

Listen von Eigenschaften
z.B.: Fell = organe

in dieser Vorlesung:

- nominelle Merkmale
- Beschreibungen ohne Ordnung
- keine Metriken zur Distanzmessung

© Nicole M. Artner

Nominelle Merkmale

- Beschreibung eines Musters durch Eigenschaften und ihre Ausprägungen
- Katze
  - Eigenschaften: Fell, Augen, Konstitution
  - Ausprägungen (Merkmalsvektor): orange, grün, fett
- Obst
  - Eigenschaften: Geschmack, Farbe, Form, Größe
  - Ausprägungen (Merkmalsvektor): sauer, rot, rund, klein

© Nicole M. Artner

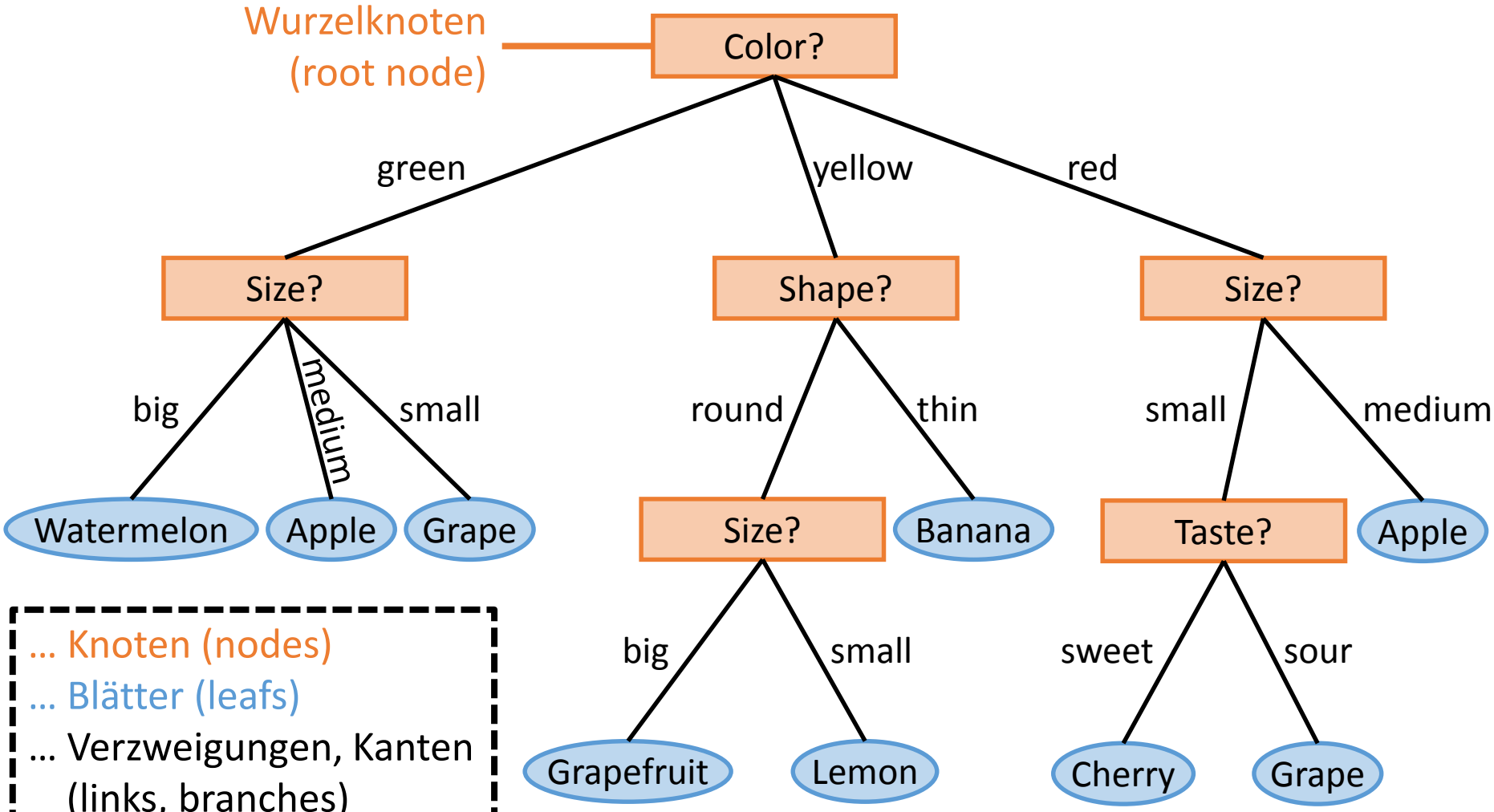
Entscheidungsbäume

… Möglichkeit zur Klassifikation von Mustern mit nominellen Merkmalsvektoren

- intuitive Klassifikation durch eine Abfolge von Fragen
- nächste Frage ist abhängig von der vorherigen Antwort
- Antwort: ja/nein, richtig/falsch oder eine konkrete Ausprägung

© Nicole M. Artner

Entscheidungsbäume

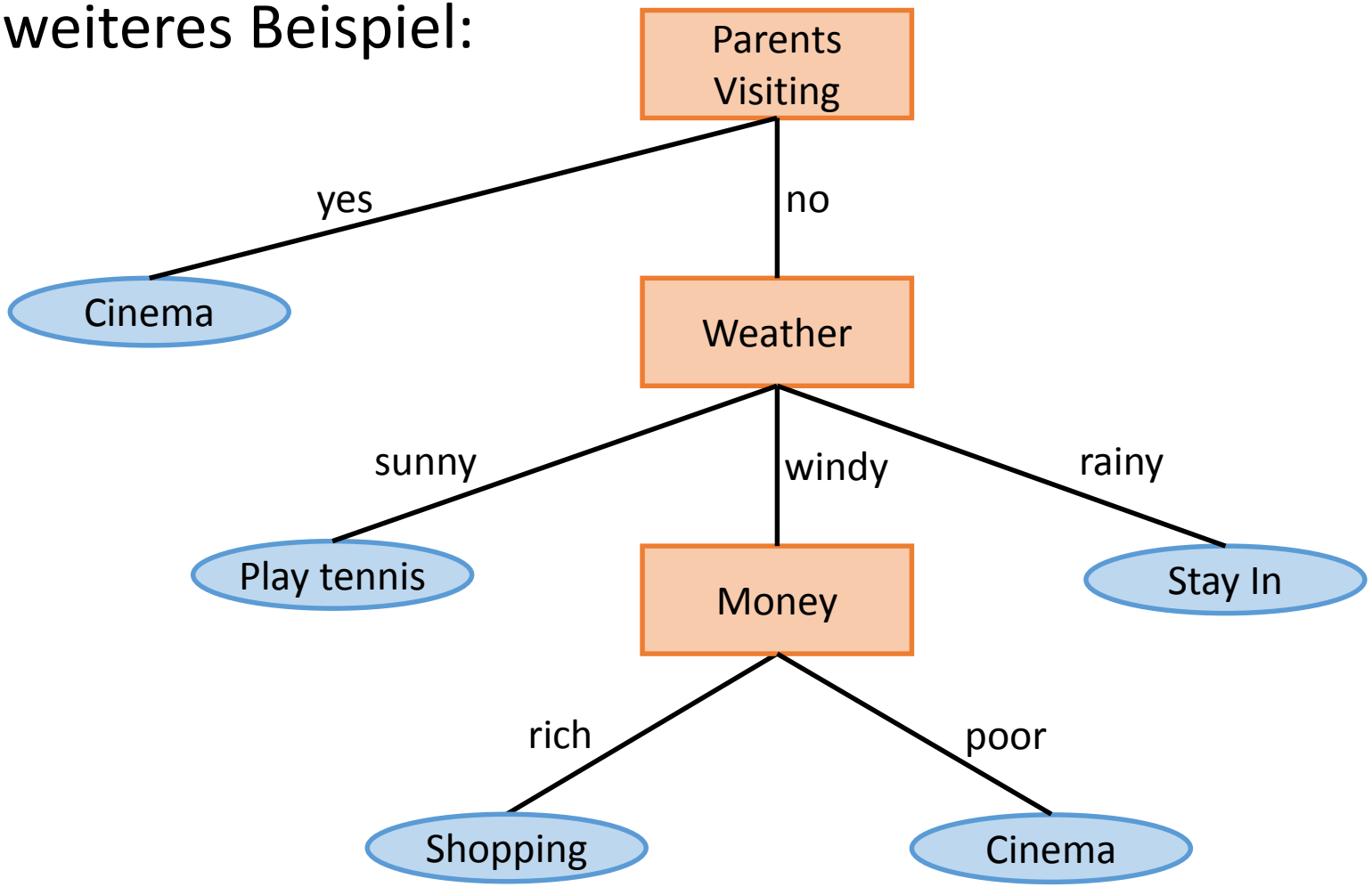


... Knoten (nodes)
... Blätter (leafs)
... Verzweigungen, Kanten (links, branches)

© Nicole M. Artner

Entscheidungsbäume

… weiteres Beispiel:



© Nicole M. Artner

Entscheidungsbäume

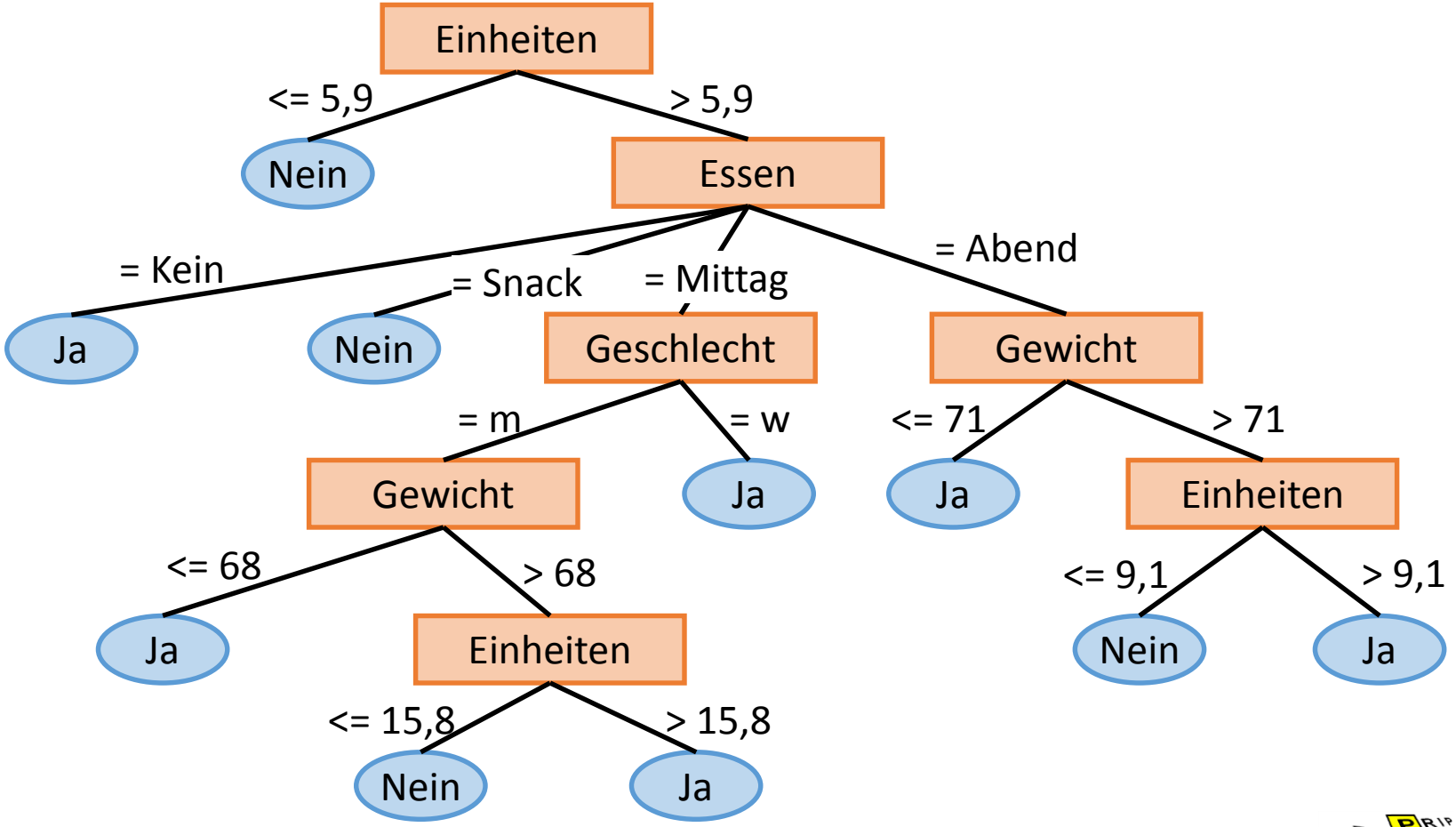
… Besoffenheitstest nach Padraig Cunningham

Eigenschaft	Ausprägung
Alter	in Jahren
Geschlecht	männlich / weiblich
Pause	in Minuten
Dauer	in Stunden
Gewicht	in kg
Größe	in cm
Essen	letztes Essen: Kein, Snack, Mittag, Abend
Einheiten	wieviel getrunken wurde

© Nicole M. Artner

Entscheidungsbäume

… Besoffenheitstest nach Padraig Cunningham



© Nicole M. Artner

Klassifikation

… mit Entscheidungsbäumen

- Klassifikation eines Musters beginnt immer bei **Frage des Wurzelknotens**
- abhängig von Antwort (Ausprägung) folgt man entsprechender **Verbindung zu einem Folgeknoten**
- es muss **eine Entscheidung** getroffen werden → nur einer Verbindung wird gefolgt
- dann wird Entscheidung für **„Frage“ des Folgeknoten** gefällt, usw…
- bis man bei einem **Blattknoten** ankommt → keine weiteren Fragen
- Blattknoten ist einer **Klasse** zugeordnet → Muster wird zugewiesen

© Nicole M. Artner

Vorteile

… von Entscheidungsbäumen

- einfach zu **interpretieren**
- logische **Strukturen** → Pfad vom Wurzelknoten zum Blatt → Klassifikation eines Musters
- Interpretation der **Klassen** selbst
  - z.B.: Apfel=(grün UND mittel) oder (rot UND mittel)
- in der Regel **schnelle** Klassifikation
- einfache Integration von **Expertenwissen**

© Nicole M. Artner

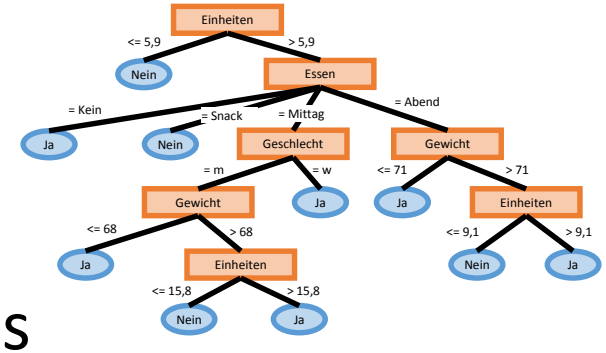
Erstellen von Entscheidungsbäumen

© Nicole M. Artner

Wie wachsen Entscheidungsbäume?

Training

- Trainingsdatensatz $\mathcal{D}$ mit Labels → überwachtes Lernen
- Satz von Eigenschaften/Fragen



Ziel des Trainings

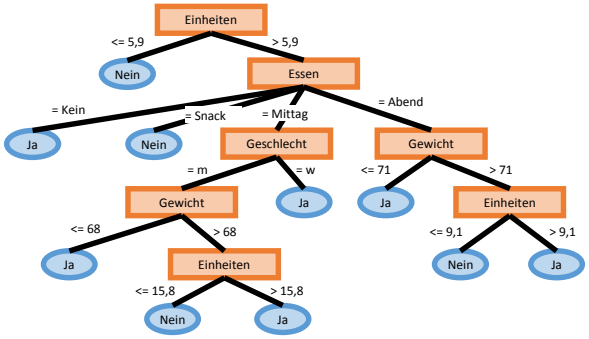
- **Automatische** Erzeugung eines Baums
- Unterteilung der Trainingsdaten in immer kleiner werdende **Untermengen**
- **Idealfall**: jede Untermenge ist rein → beinhaltet Trainingsdaten von nur einer Klasse

© Nicole M. Artner

Wie wachsen Entscheidungsbäume

Herausforderung

- normalerweise sind **Untermengen nicht rein**
- Untermengen bestehen aus einer **Mischung von Klassen**
- in jedem Zweig:
  - Soll **Verzweigung beendet** werden? Ist Fehler klein genug? → Teilung/Wachstum wird gestoppt
  - Soll eine **weitere Eigenschaft** gewählt werden? → Baum wächst weiter



© Nicole M. Artner

Beispiel: Restaurantauswahl

Eigenschaften für eine Restaurantauswahl:

Eigenschaft	Erklärung
Alternative	Gibt es eine geeignete Restaurantalternative in der Nähe?
Bar	Gibt es eine bequeme Bar im Restraunt?
Fri/Sa	Ist es Freitag oder Samstag?
Hungrig	Ist man hungrig oder nicht?
Kunden	Wie viele andere Leute sind im Restaurant?
Preis	Preisspanne
Regen	Regnet es draußen?
Reservierung	Hat man eine Reservierung?
Typ	Art des Restaurants (französisch, thailändisch, Fast Food, italienisch)
Verhalten	Bleiben oder gehen?

© Nicole M. Artner

Beispiel: Restaurantauswahl

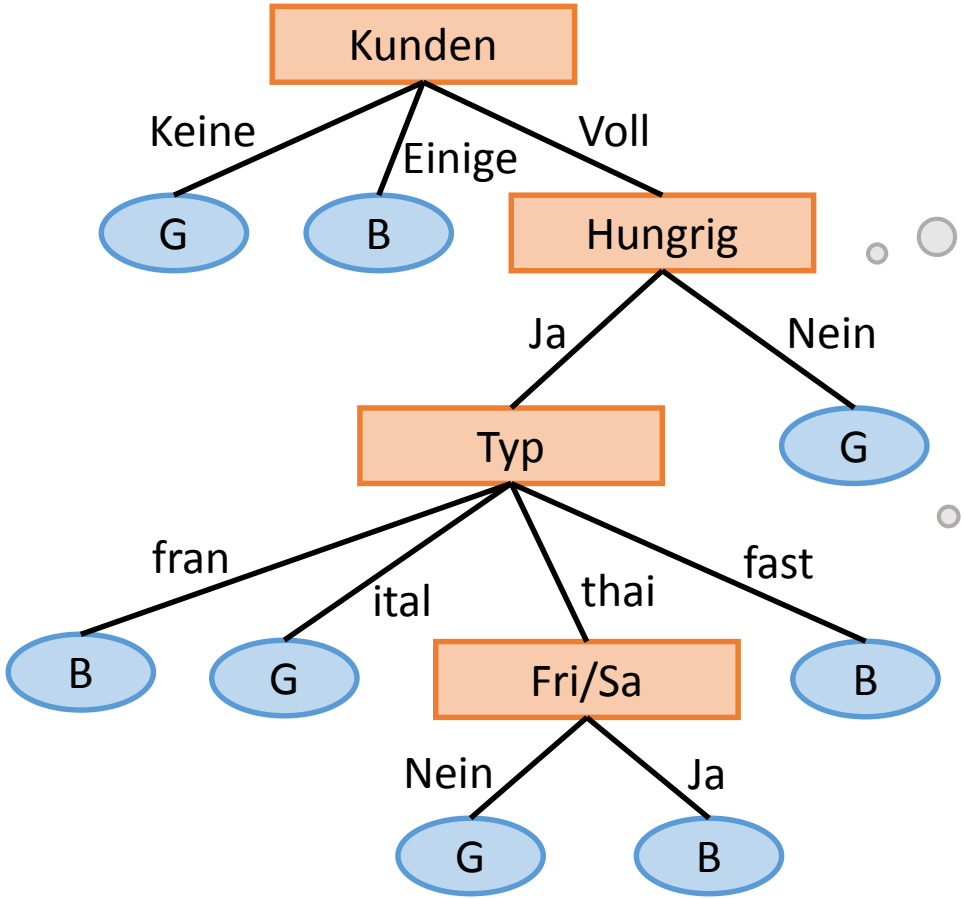
Trainingsdaten:

Fall	Alt.	Bar	Fri/Sat	Hun.	Kun.	Preis	Reg.	Res.	Typ	Ver.
1	Ja	Nein	Nein	Ja	Einige	$$$	Nein	Ja	fran	Ja
2	Ja	Nein	Nein	Ja	Voll	$	Nein	Nein	thai	Nein
3	Nein	Ja	Nein	Nein	Einige	$	Nein	Nein	fast	Ja
4	Ja	Nein	Ja	Ja	Voll	$	Nein	Nein	thai	Ja
5	Ja	Nein	Ja	Nein	Voll	$$$	Nein	Ja	fran	Nein
6	Nein	Ja	Nein	Ja	Einige	$$	Ja	Ja	ital	Ja
7	Nein	Ja	Nein	Nein	Keine	$	Ja	Nein	fast	Nein
8	Nein	Nein	Nein	Ja	Einige	$$	Ja	Ja	thai	Ja
9	Nein	Ja	Ja	Nein	Voll	$	Ja	Nein	fast	Nein
10	Ja	Ja	Ja	Ja	Voll	$$$	Nein	Ja	ital	Nein
11	Nein	Nein	Nein	Nein	Keine	$	Nein	Nein	thai	Nein
12	Ja	Ja	Ja	Ja	Voll	$	Nein	Nein	fast	Ja

© Nicole M. Artner

Beispiel: Restaurantauswahl

Automatisch erzeugter Entscheidungsbaum:



Beschreibt die Daten perfekt.

Guter Baum → wenige Knoten, nicht tief

© Nicole M. Artner

CART

CART (= **C**lassification **A**nd **R**egression **T**rees) ist eine **iterative Methode** um Entscheidungsbäume zu erstellen/wachsen zu lassen. Abhängig von den Daten in einem Knoten wird: (1) der **Knoten als Blatt deklariert** und eine Klasse gewählt oder (2) ein Merkmal (Eigenschaft) gefunden, um die Daten in **kleinere Untermengen** zu teilen.

© Nicole M. Artner

CART

… wirft sechs grundlegende Fragen auf:

1. Wie viele **Verzweigungen** sollen von einem Knoten ausgehen? Wie viele Entscheidungsausgänge/Teilungen gibt es?
2. Welches **Kriterium** soll an einem Knoten getestet werden?
3. Wann soll ein Knoten **in ein Blatt umgewandelt** werden?
4. Wie kann man einen **Baum verkleinern** oder vereinfachen, wenn er zu „groß“ ist (Pruning)?
5. Welcher **Klasse** soll ein Blatt zugeordnet werden, wenn es nicht rein ist?
6. Wie geht man mit **fehlenden Daten** um?

© Nicole M. Artner

CART im Detail

© Nicole M. Artner

1. Verzweigungen

Jeder **Entscheidungsausgang** an einem Knoten wird als **Verzweigung** (split) bezeichnet, weil die Trainingsdaten aufgespalten/aufgeteilt werden.

Die Anzahl von Verzweigungen (Kanten) ausgehend von einem Knoten wird häufig als **Verteilungsfaktor** (branching factor) oder **Verteilungsverhältnis** (branching ratio) bezeichnet.

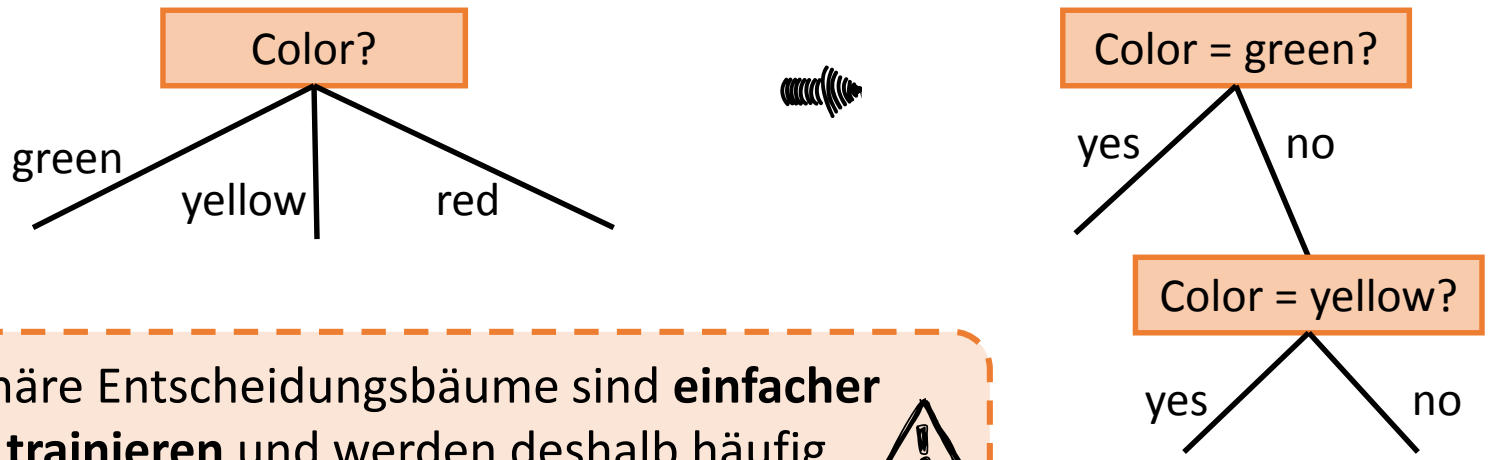
Verteilungsfaktor:

- abhängig vom Design des Baumes
- nicht notwendigerweise konstant in einem Baum

© Nicole M. Artner

1. Verzweigungen

Jeder Baum kann als **binärer Entscheidungsbaum** dargestellt werden → Verteilungsfaktor $B = 2$. Solch ein binärer Baum führt zum **gleichen Klassifikationsergebnis** wie der ursprüngliche Baum mit beliebigem Verteilungsfaktor B an den Knoten.

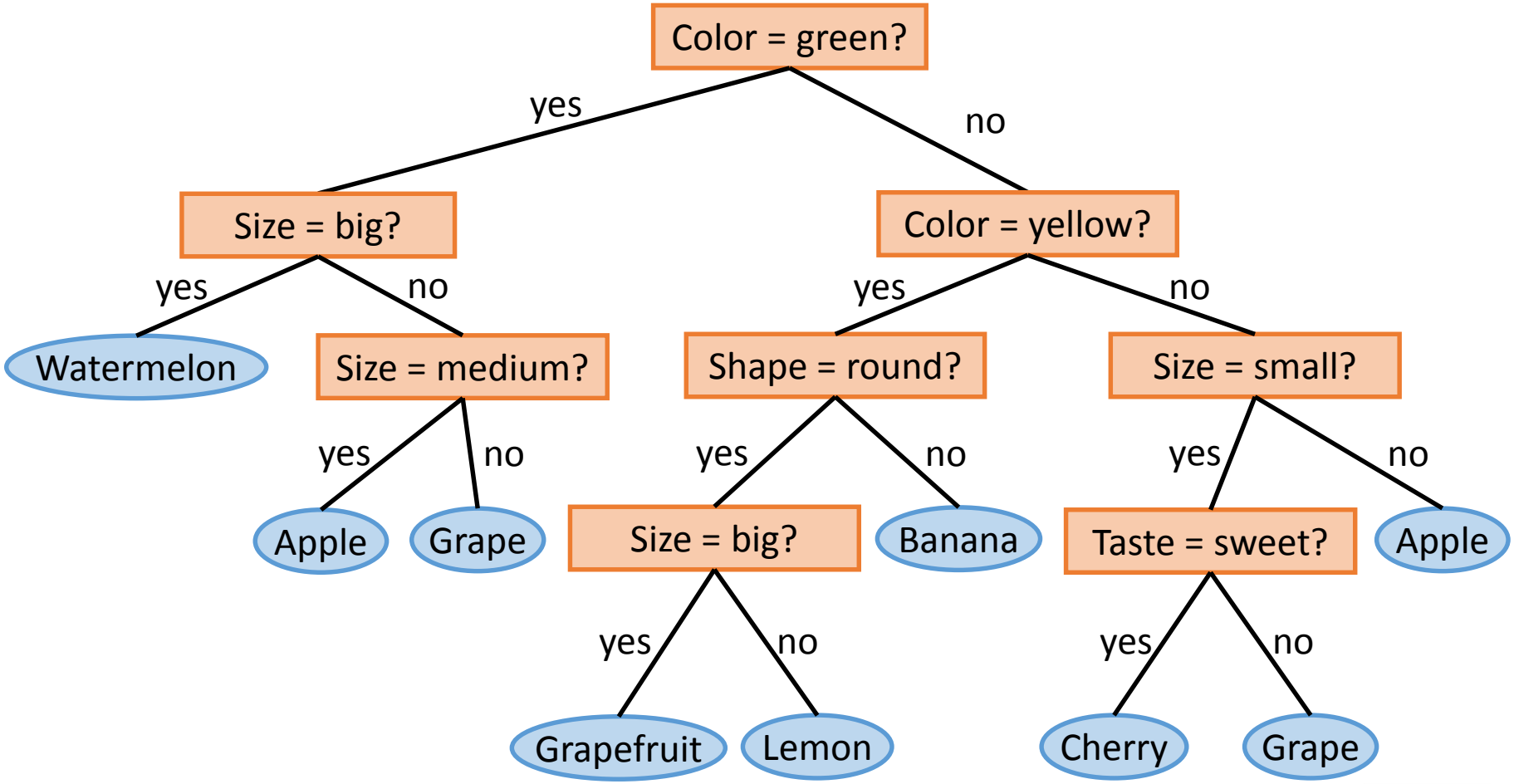


Binäre Entscheidungsbäume sind **einfacher zu trainieren** und werden deshalb häufig verwendet.

© Nicole M. Artner

1. Verzweigungen

… Baum von Folie 8 als Binärbaum



© Nicole M. Artner

2. Kriterium

Ein Kriterium besteht aus einem oder mehreren **Eigenschaften** (Merkmalen) und einer **Regel** (Heuristik) wie die Daten dadurch aufgeteilt werden.

Ziel: Entscheidungsbaum soll so einfach wie möglich sein
→ wenige Knoten, flach

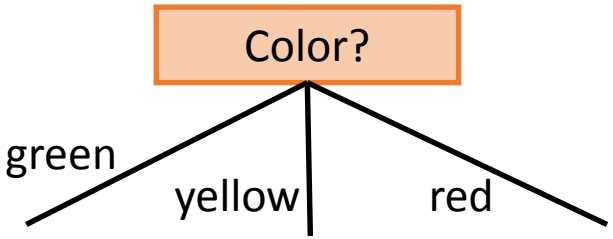
Welche Kriterien sollen in den Knoten verwendet werden?

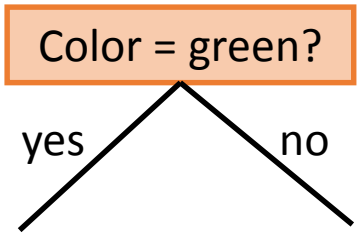
In welcher Reihenfolge?

© Nicole M. Artner

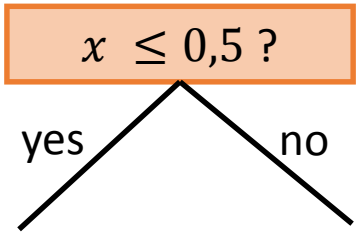
2. Beispiele für Kriterien

- nominell





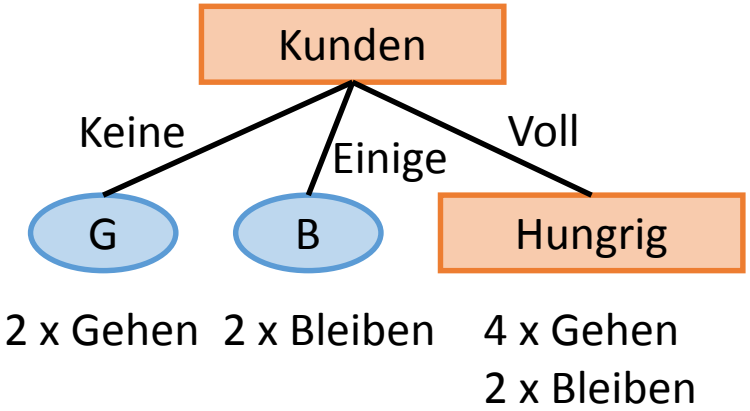
- numerisch



© Nicole M. Artner

2. Das perfekte Kriterium

Ein **perfektes Kriterium** führt zu Knoten die rein sind. Das bedeutet, dass die resultierenden Teilmengen nur Mitglieder einer Klasse beinhalten.



Wie kann man die Reinheit eines Knotens messen?

© Nicole M. Artner

2. Unreinheit

In der Praxis wird die **Unreinheit** (impurity) eines Knoten definiert, weil es einfacher ist.

Sei $i(N)$ die Unreinheit eines Knotens N. Dann ist $i(N) = 0$, wenn alle Muster (Daten) die in N ankommen in die gleiche Klasse fallen. $i(N)$ ist am größten, wenn alle Klassen gleich häufig vorkommen.

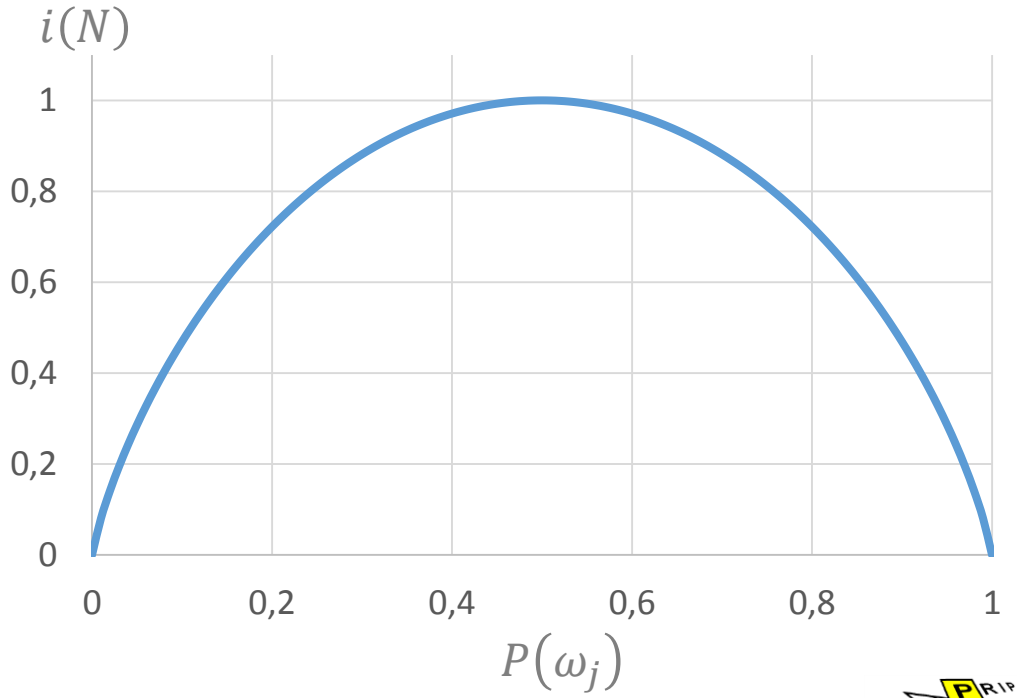
© Nicole M. Artner

2. Entropie

Die Entropie (entropy) ist ein Maß mit dem man die Unreinheit wie folgt bestimmen kann: $i(N) = -\sum_j P(\omega_j) \log_2 P(\omega_j)$

wobei $P(\omega_j)$ die Häufigkeit der Klasse ω_j im Knoten N angibt.

z.B.: für zwei Klassen



© Nicole M. Artner

2. Kriteriumsauswahl

Ziel: Wähle ein Kriterium, sodass die Unreinheit $i(N)$ eines Knotens N möglichst klein wird. Das bedeutet, dass die **Änderung der Unreinheit $\Delta \boldsymbol{i}(\boldsymbol{N})$ maximiert** werden soll.

Für eine **binäre Verzweigung** ist $\Delta i(N) = i(N) - P_L i(N_L) - (1 - P_L) i(N_R)$, wobei N_L und N_R die resultierenden linken und rechten Knoten (Teilmengen) sind. P_L ist der Anteil der Muster der dem linken Knoten zugeordnet wird.

© Nicole M. Artner

2. Beispiel

Ziel: Erstellen eines binären Entscheidungsbaumes für numerische Merkmale

Trainingsdaten:

$n = 16$ Muster

Merkmale x_1 und x_2

Klassen w_1 und w_2

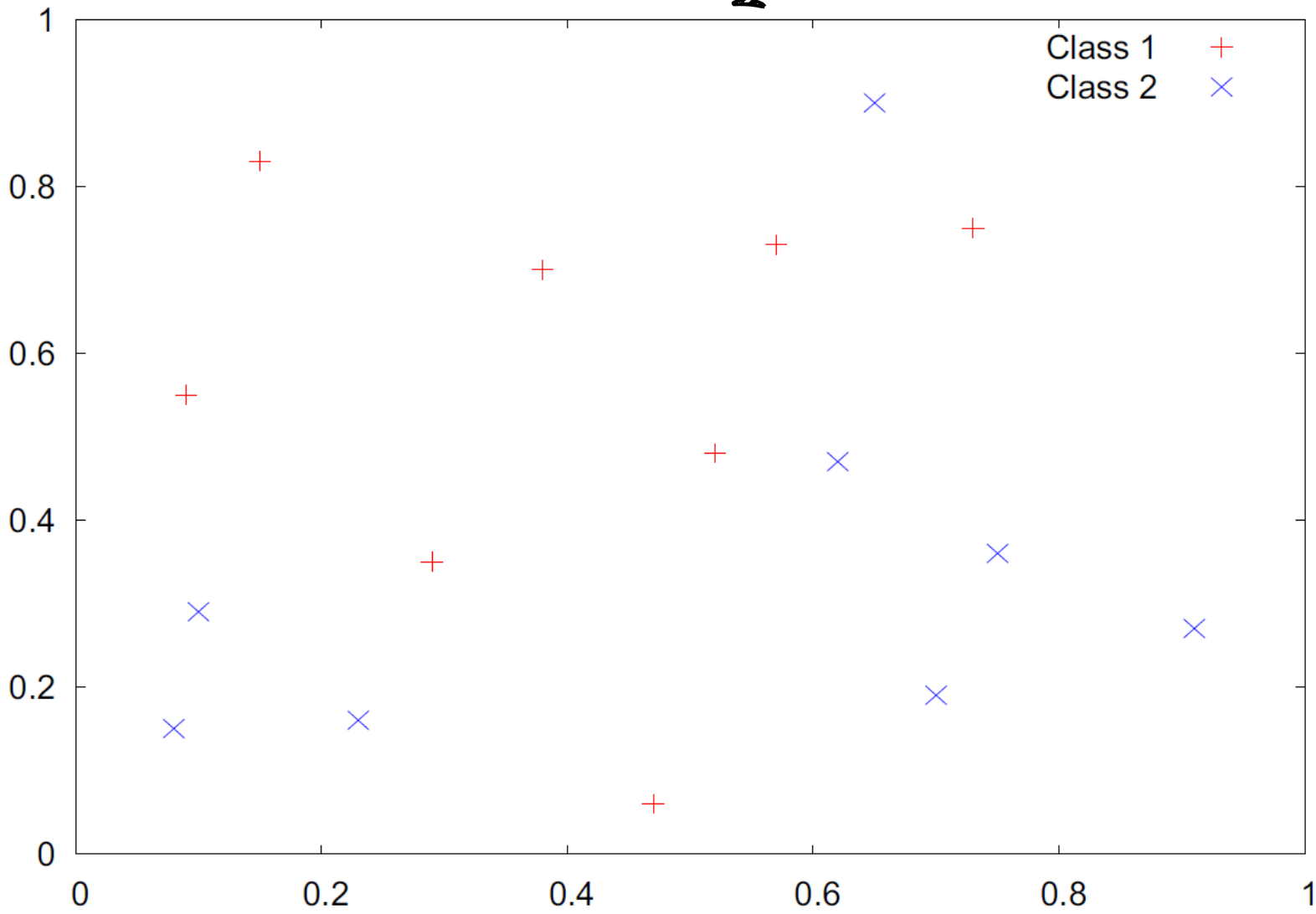
w_1		w_2	
x_1	x_2	x_1	x_2
0,15	0,83	0,10	0,29
0,09	0,55	0,08	0,15
0,29	0,35	0,23	0,16
0,38	0,70	0,70	0,19
0,52	0,48	0,62	0,47
0,57	0,73	0,91	0,27
0,73	0,75	0,65	0,90
0,47	0,06	0,75	0,36

Kriterien:

$x_i \leq x_s$

© Nicole M. Artner

2. Beispiel



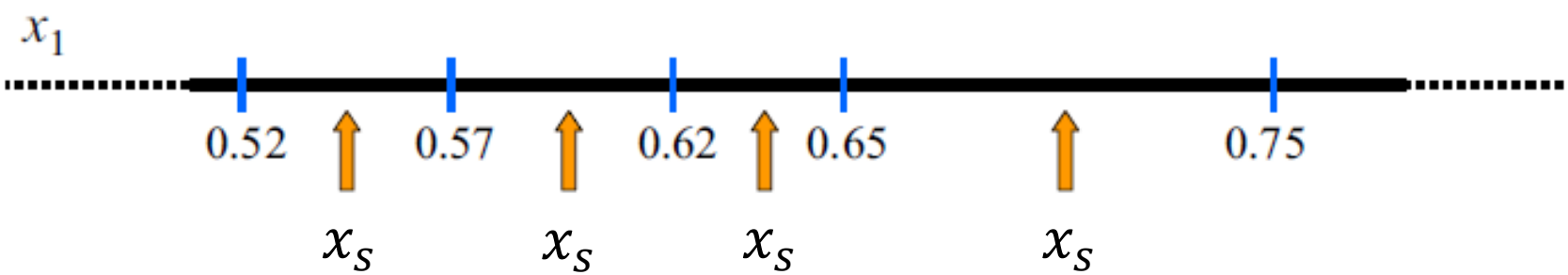
© Nicole M. Artner

2. Beispiel

Berechnung der **Unreinheit des Wurzelknotens**:

$$i(N_0) = -\sum_{i=1}^{2} P(\omega_i) \log_2 P(\omega_i) = -\left[0{,}5 \log_2 0{,}5 + 0{,}5 \log_2 0{,}5\right] = 1{,}0$$

Es gibt jeweils $n - 1 = 15$ Möglichkeiten, um x_1 und x_2 zu teilen.



Teilungsposition?

© Nicole M. Artner

2. Beispiel

- berechne die **Änderung der Unreinheit** $\Delta i(N_0)$ für alle möglichen „Teilungspositionen“
- Teilungsposition x_s überall zwischen x_l und x_r möglich → oft wird **Mittelpunkt** verwendet: $x_s = \frac{x_l + x_r}{2}$

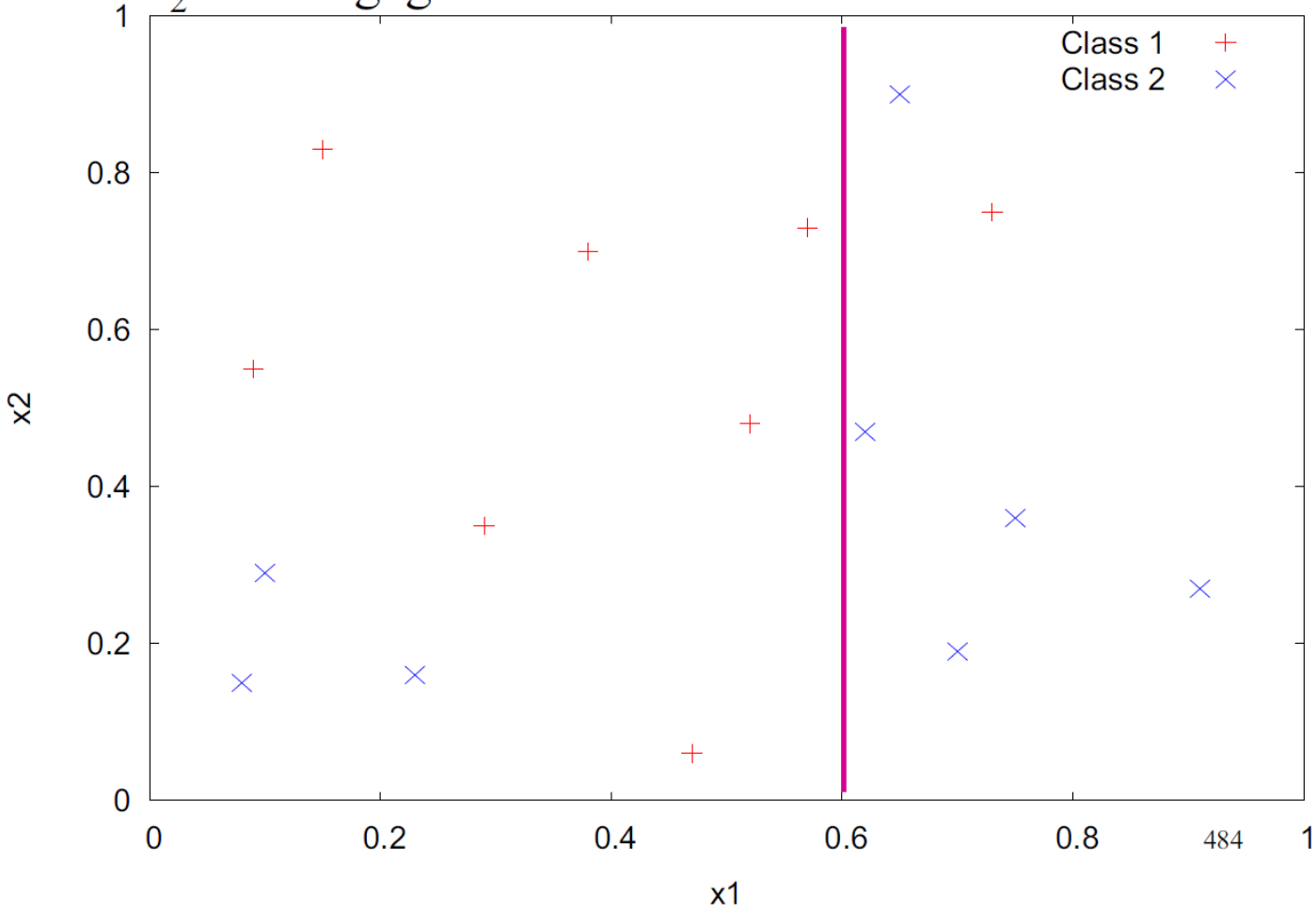
Ergebnis:

- $\Delta i(N_0)$ ist am größten für x_1 zwischen 0,57 und 0,62
- daraus folgt → $x_s = 0{,}60$

© Nicole M. Artner

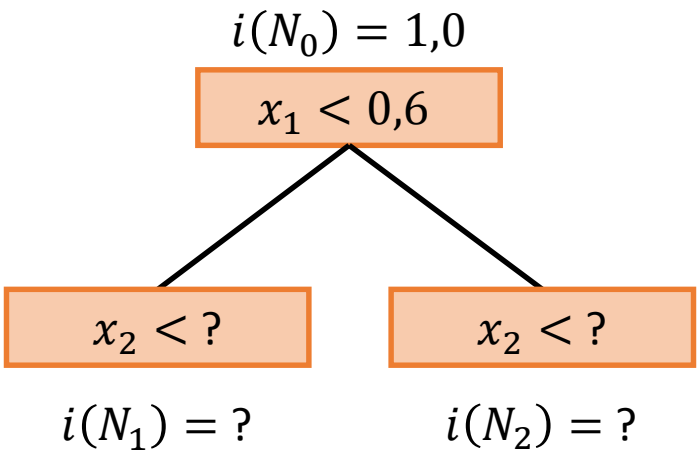
2. Beispiel

- Diese Trennung ist durch eine Linie parallel zur x_2-Achse gegeben.



2. Beispiel

… Ergebnis nach 1. Verzweigung



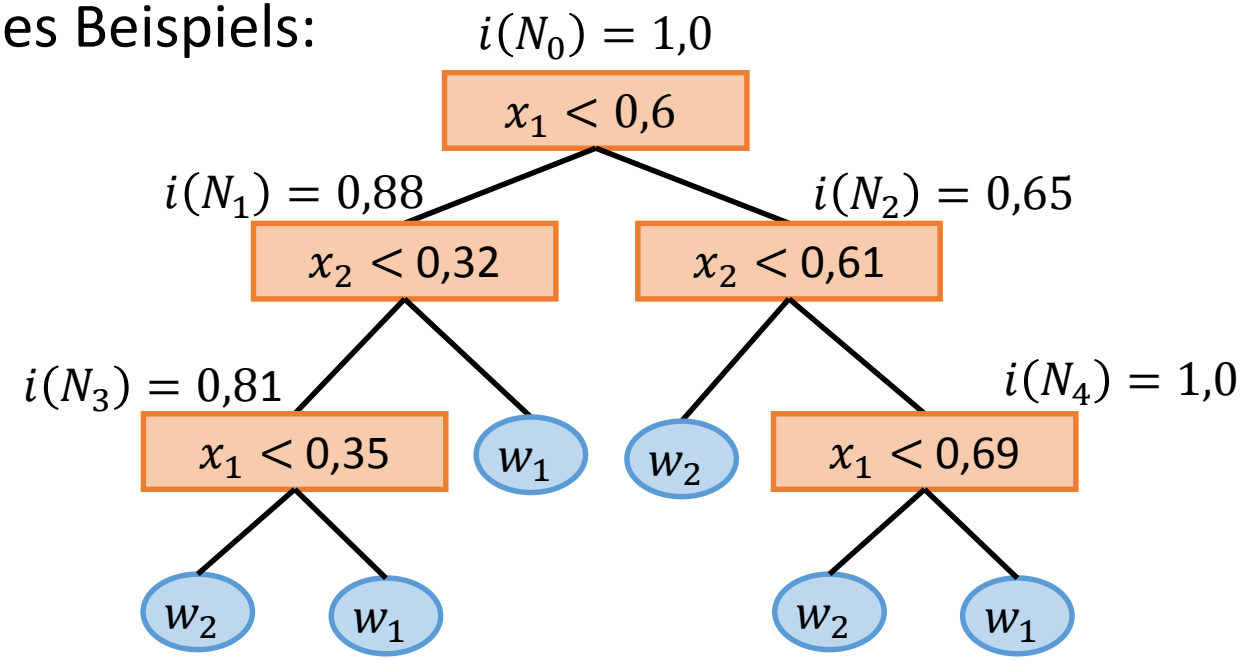
w_1		w_2	
x_1	x_2	x_1	x_2
0,15	0,83	0,10	0,29
0,09	0,55	0,08	0,15
0,29	0,35	0,23	0,16
0,38	0,70	0,70	0,19
0,52	0,48	0,62	0,47
0,57	0,73	0,91	0,27
0,73	0,75	0,65	0,90
0,47	0,06	0,75	0,36

© Nicole M. Artner

2. Beispiel

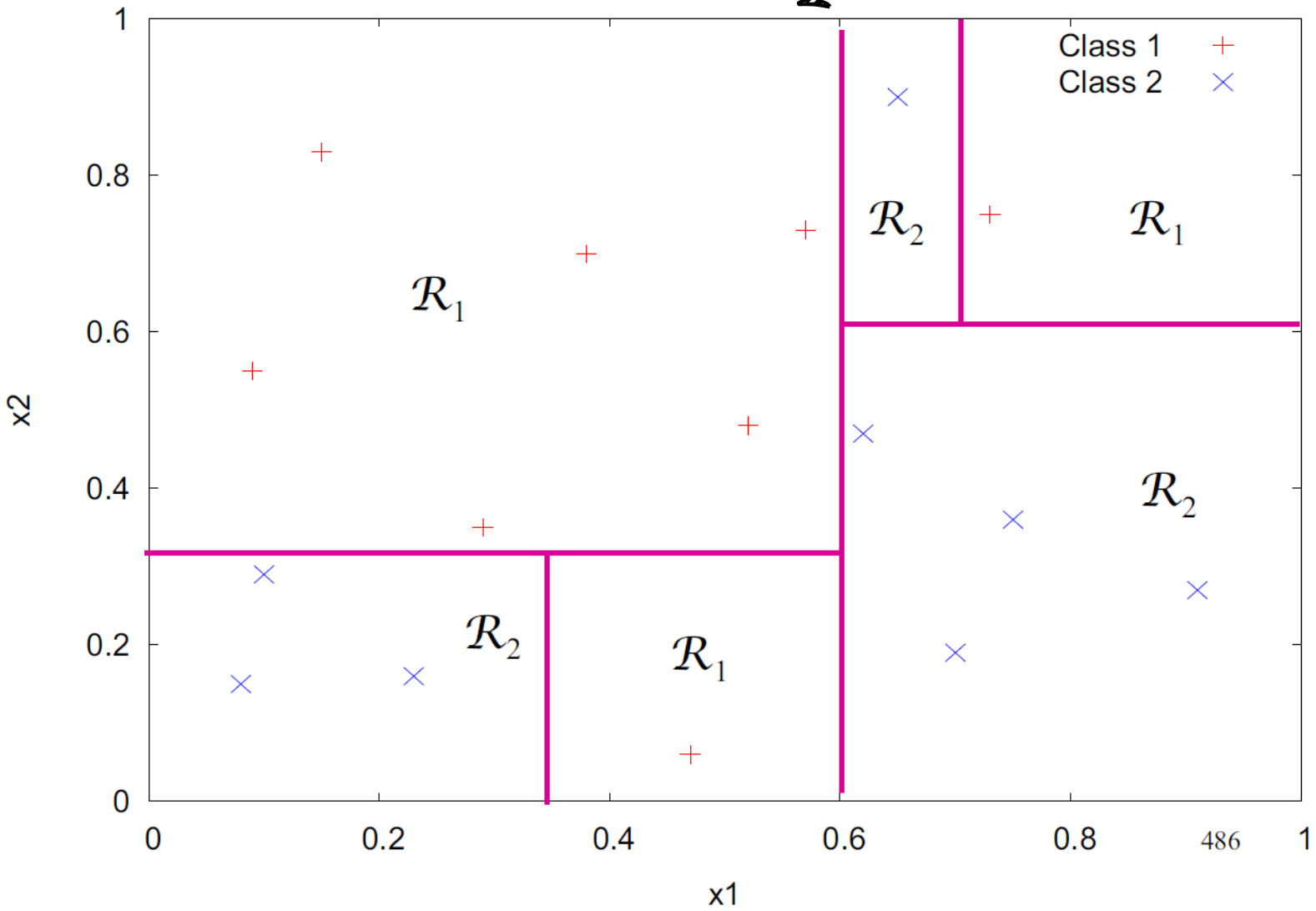
- die Knoten N_1 und N_2 sind nicht rein
- bestimme neue Kriterien für weitere Verzweigungen
- bis die Unreinheit aller Blätter 0 wird

Ergebnis des Beispiels:



© Nicole M. Artner

2. Beispiel



© Nicole M. Artner

3. Stoppen des Trainings

Fragen:

- Wann soll ein Knoten in ein Blatt umgewandelt werden?
- Wann soll die Trennung von Knoten gestoppt werden?

Probleme:

- **Overfitting** auf Trainingsdaten, wenn Training erst gestoppt wird, wenn alle Knoten rein sind
- Training sollte auch **nicht zu früh gestoppt** werden
- Entscheidungsbäume sind oft **sensibel auf kleine Unterschiede** in Trainingsdaten

© Nicole M. Artner

3. Beispiel: Kanu fahren

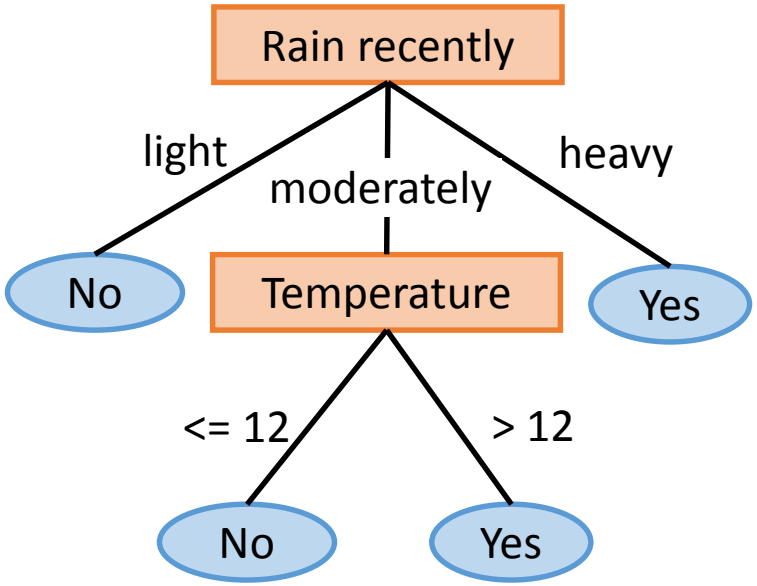
[Quelle: Padraig Cunningham]

Case	Rain Recently	Rain Today	Temp	Windy	Sunshine	Canoeing
C1	Moderate	Moderate	15	Light	Some	Yes
C2	Light	Moderate	14	Moderate	None	No
C3	Moderate	Moderate	7	Gale	None	No
C4	Moderate	Moderate	16	Light	None	Yes
C5	Moderate	Light	12	Light	Some	No
C6	Heavy	Light	12	Moderate	Some	Yes
C7	Light	Light	8	Moderate	Some	No
C8	Moderate	Moderate	8	Gale	Some	No
C9	Heavy	Heavy	15	Moderate	None	Yes
C10	Light	Light	8	Light	Some	No
C11	Light	Light	15	Light	Some	No

© Nicole M. Artner

3. Beispiel: Kanu fahren

…. resultierender Entscheidungsbaum



[Quelle: Padraig Cunningham]

© Nicole M. Artner

3. Beispiel: Kanu fahren

… das letzte Muster des Trainingsdatensatzes wird entfernt

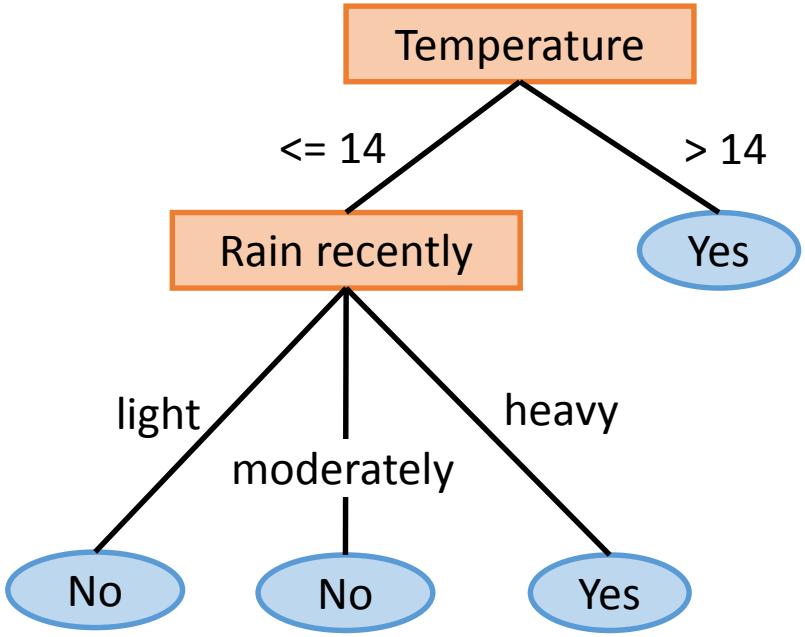
Case	Rain Recently	Rain Today	Temp	Windy	Sunshine	Canoeing
C1	Moderate	Moderate	15	Light	Some	Yes
C2	Light	Moderate	14	Moderate	None	No
C3	Moderate	Moderate	7	Gale	None	No
C4	Moderate	Moderate	16	Light	None	Yes
C5	Moderate	Light	12	Light	Some	No
C6	Heavy	Light	12	Moderate	Some	Yes
C7	Light	Light	8	Moderate	Some	No
C8	Moderate	Moderate	8	Gale	Some	No
C9	Heavy	Heavy	15	Moderate	None	Yes
C10	Light	Light	8	Light	Some	No
~~C11~~	~~Light~~	~~Light~~	~~15~~	~~Light~~	~~Some~~	~~No~~

© Nicole M. Artner

3. Beispiel: Kanu fahren

…. geänderter Entscheidungsbaum

→ Problem: **Overfitting!**



[Quelle: Padraig Cunningham]

© Nicole M. Artner

3. Beispiel: Numerisch

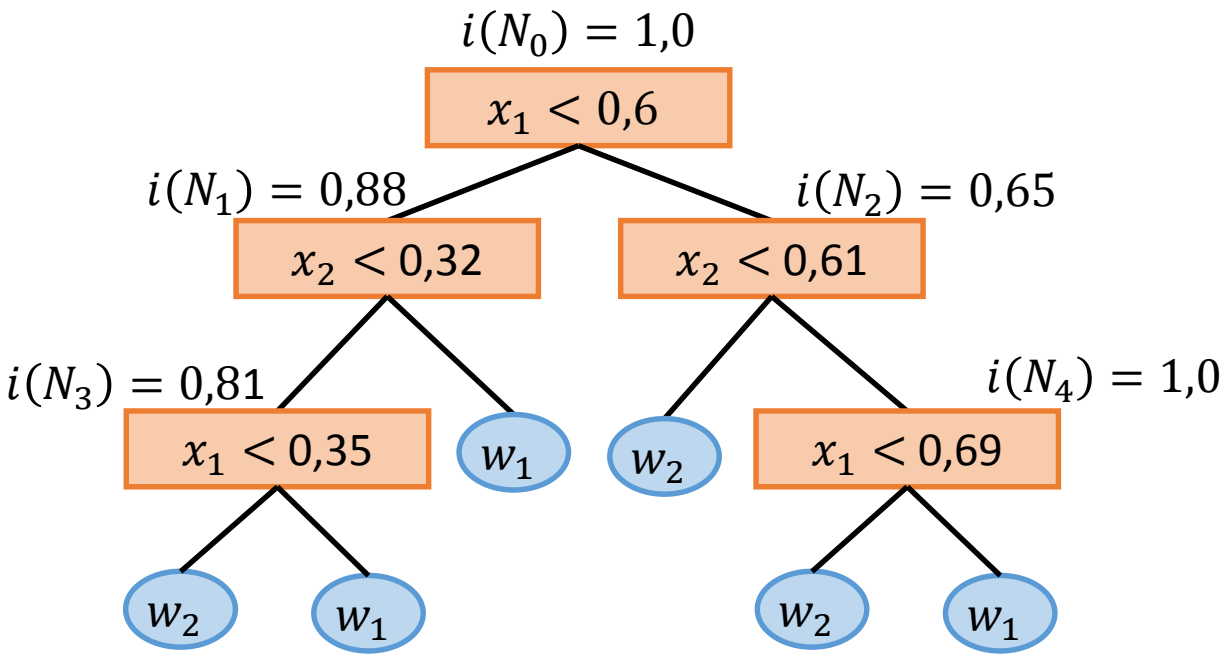
- numerisches Beispiel von vorher mit einem abgeänderten Wert

w_1		w_2	
x_1	x_2	x_1	x_2
0,15	0,83	0,10	0,29
0,09	0,55	0,08	0,15
0,29	0,35	0,23	0,16
0,38	0,70	0,70	0,19
0,52	0,48	0,62	0,47
0,57	0,73	0,91	0,27
0,73	0,75	0,65	0,90
0,47	0,06	0,75	**0,32**

© Nicole M. Artner

3. Beispiel: Numerisch

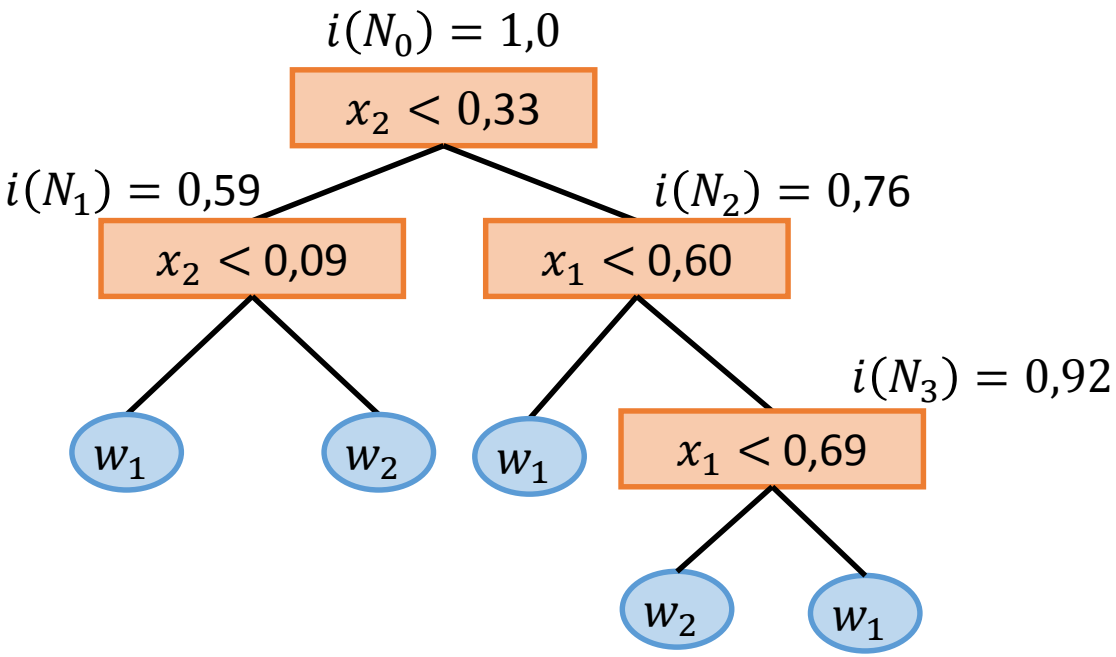
- ursprünglicher Entscheidungsbaum



© Nicole M. Artner

3. Beispiel: Numerisch

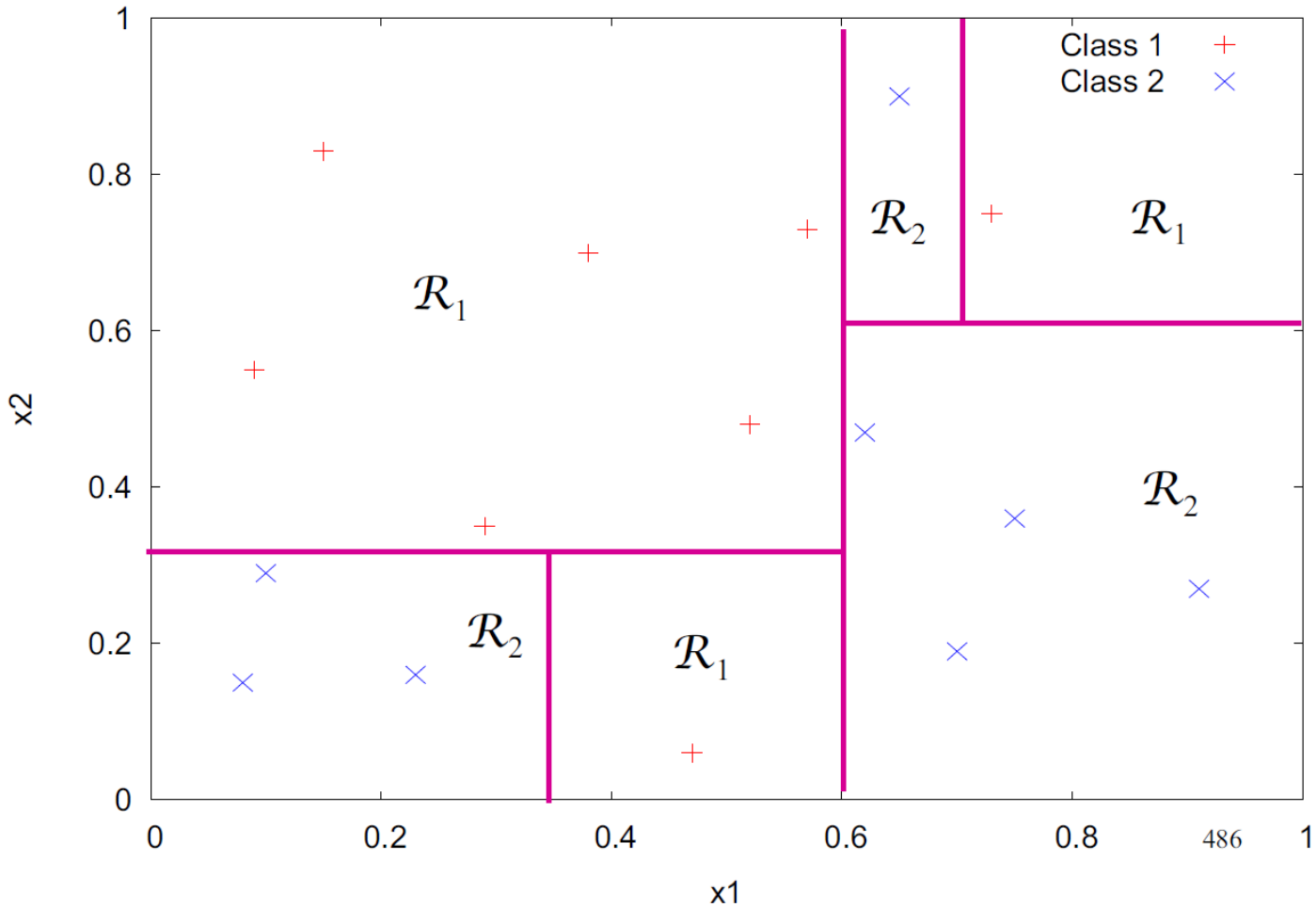
- neuer Entscheidungsbaum



© Nicole M. Artner

3. Beispiel: Numerisch

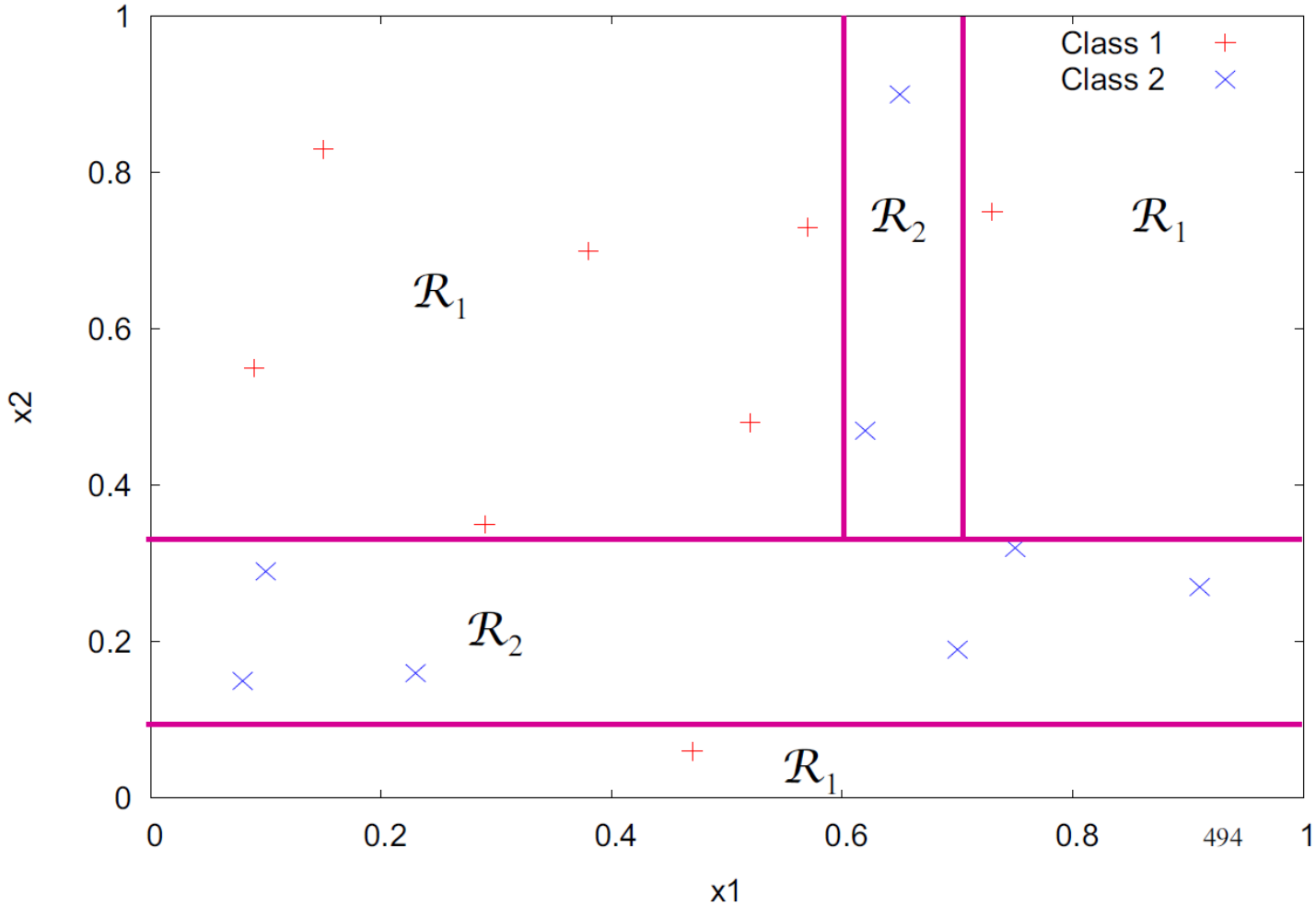
vorher …



© Nicole M. Artner

3. Beispiel: Numerisch

nachher …



© Nicole M. Artner

3. Stoppen des Trainings

… durch verschiedene Abbruchkriterien, z.B.:

Unabhängige Validierungsdaten

- Training wird abgebrochen, wenn Fehler auf Validierungsdaten klein genug ist (Schwellwert)

Änderung der Unreinheit

- Knoten wird zu einem Blatt, wenn $\Delta i(N) < \beta$
- β ist schwierig festzulegen

Anzahl der Muster

- Knoten wird zu einem Blatt, wenn Anzahl der zugehörigen Muster unter einen Schwellwert fällt
- z.B.: 5% der Trainingsdaten

© Nicole M. Artner

4. Pruning

Problem: Training wird auf Grund von **lokalen Abbruchkriterien** gestoppt. Es fehlt der sogenannte Weitblick. Dieses Phänomen wird als „**Horizon Effect**“ bezeichnet.

Lösung: Pruning

- Training wird gestoppt, wenn **Unreinheit aller Knoten null** ist
- Für **alle Paare** von Knoten (mit gemeinsamem Elternknoten):
  - überprüfe den **Einfluss der Knoten** auf die Unreinheit
  - falls **Einfluss klein** (Unreinheit wird nur leicht erhöht)
    → **entferne Knoten**, Elternknoten wird zu Blatt

© Nicole M. Artner

4. Pruning

Informationen zum Pruning:

- Pruning **beginnt** meist mit den **Blatt-Knoten**
- **effiziente Algorithmen** können einen komplexen Teilbaum (subtree) direkt mit einem Blatt ersetzen
- Pruning verhindert den „**Horizon Effect**“
- komplexes Problem:
  - Abbruchkriterien sind vorzuziehen
  - es dauert zu lange bis alle Blätter eine Unreinheit von Null aufweisen
- einfaches Problem:
  - Pruning wird bevorzugt
  - Rechenaufwand vernachlässigbar

© Nicole M. Artner

5. Klassenzugehörigkeit

… von Blättern

Problem: Für welche Klasse soll man sich entscheiden, wenn ein Blatt nicht rein ist?

Lösung: Das Blatt gehört zu jener Klasse die am häufigsten im Knoten vorkommt. D.h. die meisten Muster des Trainingsdatensatzes in dem Blatt waren dieser Klasse zugeordnet.

© Nicole M. Artner

6. Fehlende Daten

… beim Training

Training

- **naive** Lösung: löschen der unvollständigen Muster
  - nur wenn ausreichend Trainingsdaten zur Verfügung stehen

- **„schlauere“** Lösung:
  - Berechnung von $i(N)$ und $\Delta i(N)$ auf den vorhandenen Informationen
  - z.B.: In Knoten N gibt es drei Muster mit zwei Merkmalen, wobei in einem Muster das zweite Merkmal fehlt. Um die beste Teilung zu finden, ermittelt man alle drei möglichen Teilungen basierend auf dem ersten Merkmal und nur zwei Teilungen für das zweite Merkmal.

© Nicole M. Artner

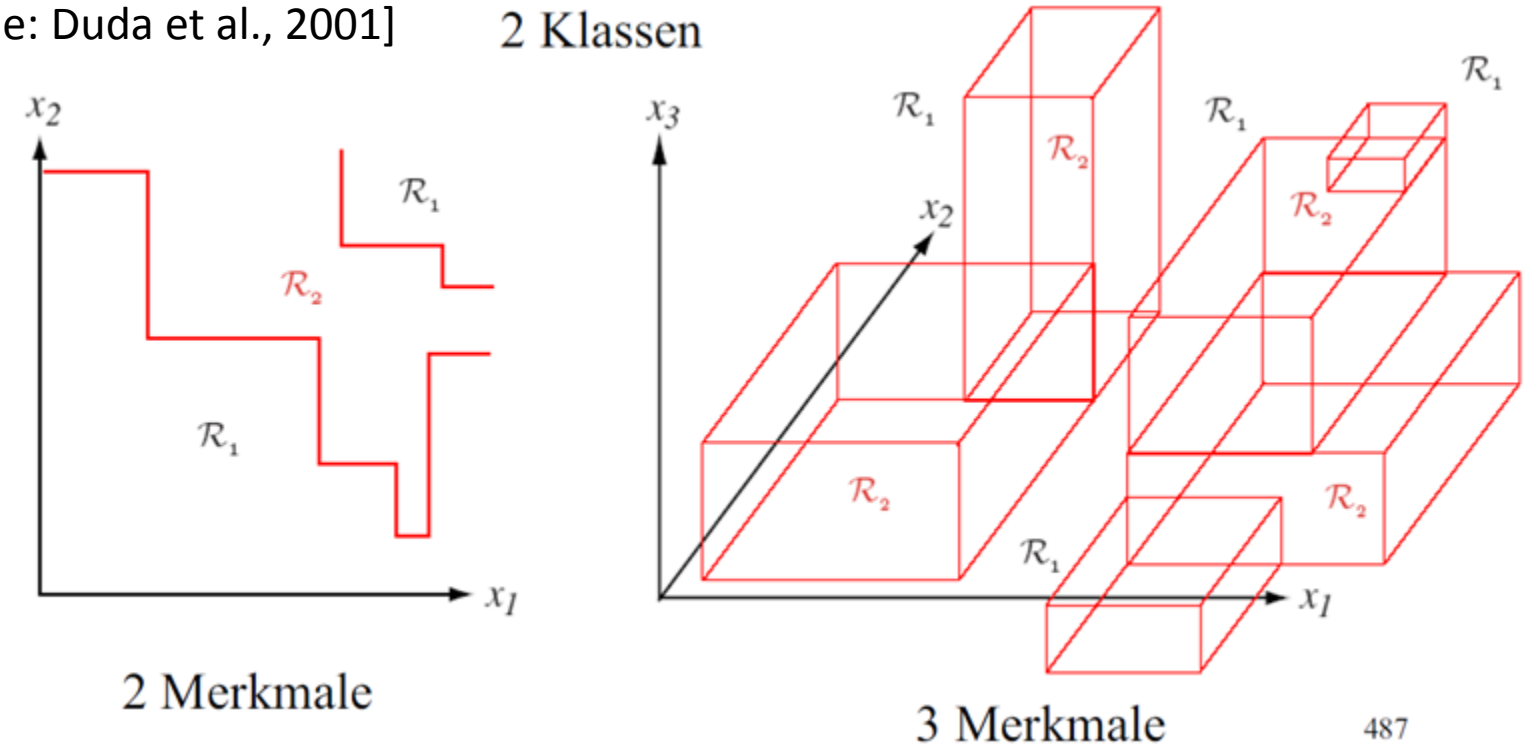
Möglichkeiten und Einschränkungen

© Nicole M. Artner

Entscheidungsgrenzen

Man kann **beliebige Entscheidungsgrenzen approximieren**, wenn die Größe des Entscheidungsbaumes ausreicht.

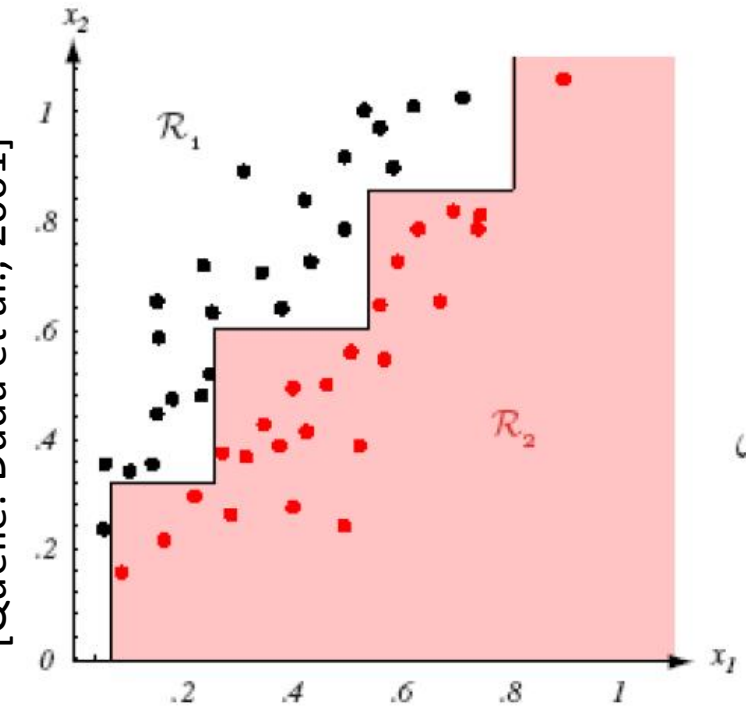
[Quelle: Duda et al., 2001]

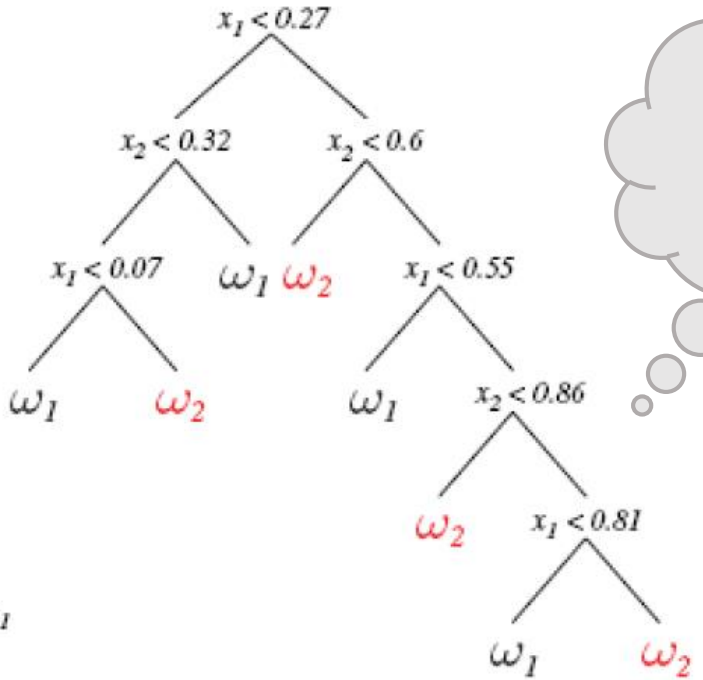


Merkmale und Kriterien

CART und andere Methoden für Entscheidungsbäume funktionieren am besten, wenn **sinnvolle Merkmale und Kriterien** gewählt werden

[Quelle: Duda et al., 2001]



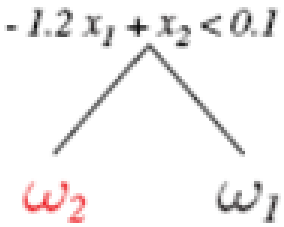


Schlecht gewählte Kritierien!

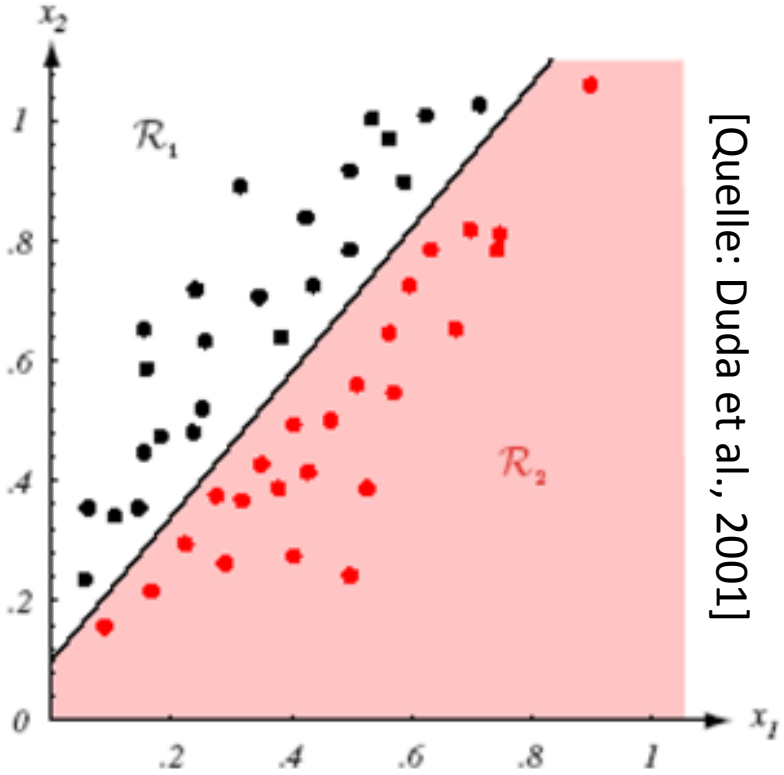
© Nicole M. Artner

Merkmale und Kriterien

- Vorverarbeitung mit PCA
  - auffinden der wichtigsten Achsen
  - optimierte Kriterien in den Knoten



- Beispiel:
  - Entscheidungsgrenze: linear Kombination
  - Entscheidungsbaum: Wurzelknoten + zwei Blätter

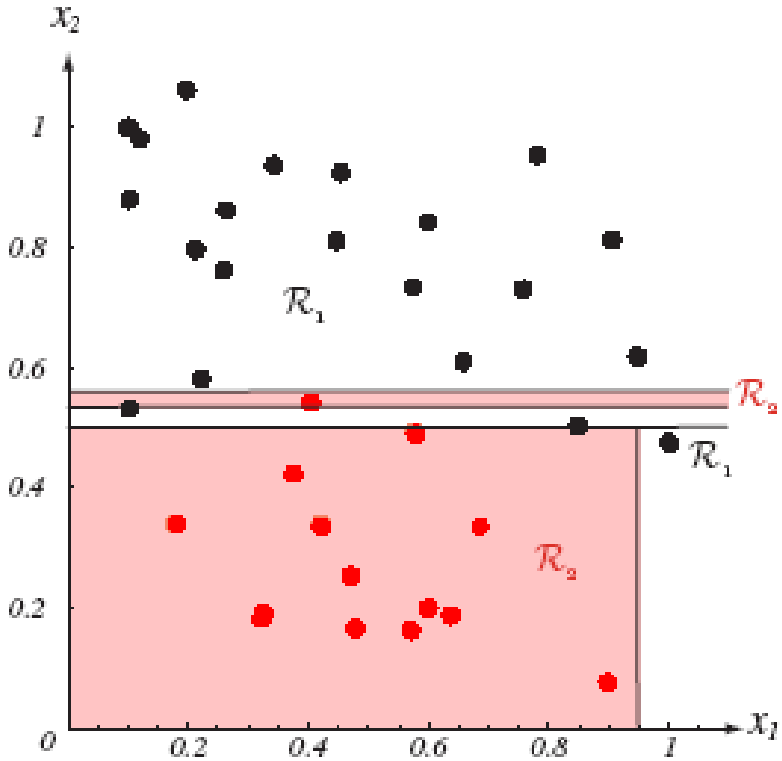


[Quelle: Duda et al., 2001]

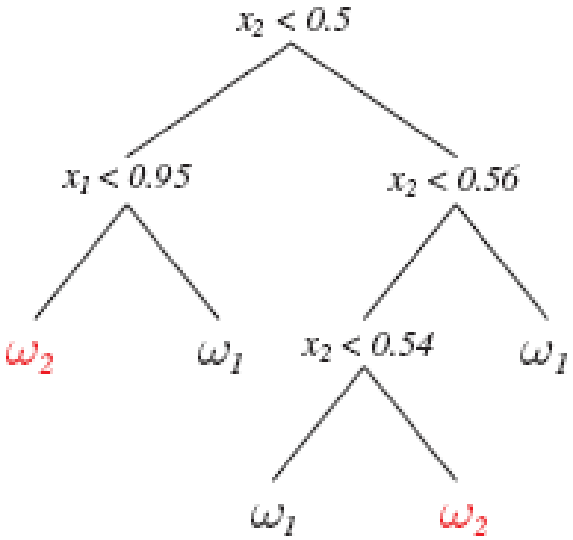
© Nicole M. Artner

Multivariate Bäume

- optimale Entscheidungsgrenzen nicht parallel zu den Achsen
- CART und ähnliche Methoden sind nicht ausreichend



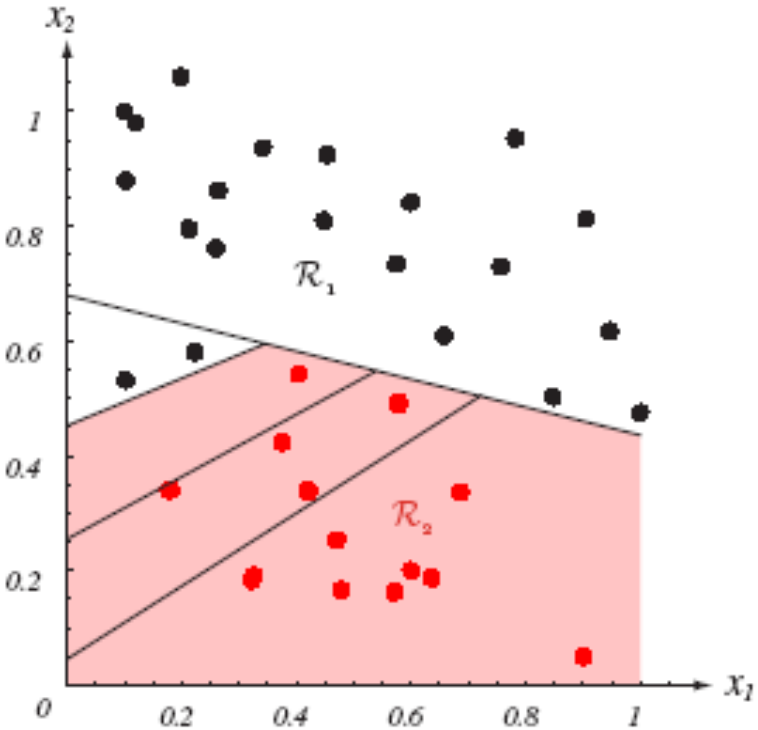
[Quelle: Duda et al., 2001]

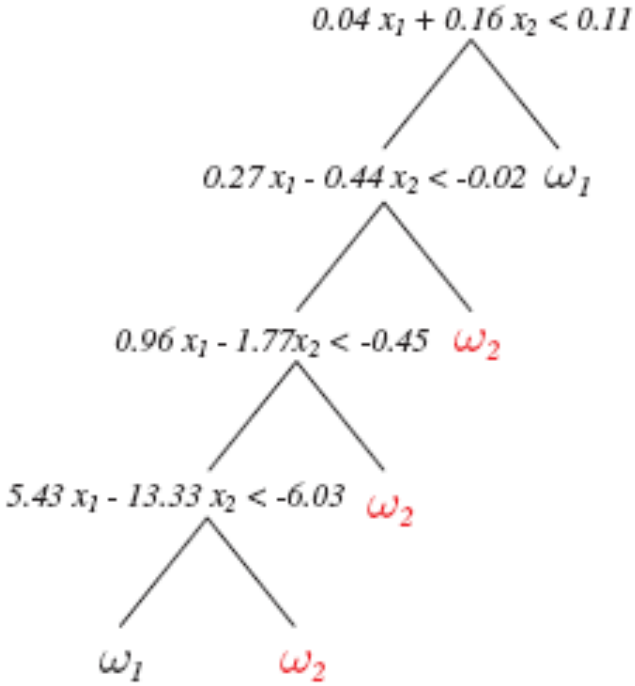


Multivariate Bäume

- Lösung: Verwendung von Entscheidungsgrenzen die nicht parallel zu den Achsen liegen
- z.B. in jedem Knoten linearer Klassifikator, trainiert durch Perceptron o.ä.

[Quelle: Duda et al., 2001]





© Nicole M. Artner

Andere Methoden

… zur Erstellung/Training von Entscheidungsbäumen

ID3

- nur für **nominelle** Daten
- numerische Daten → werden **in Intervalle eingeteilt** die dann als ungeordnete, nominelle Eigenschaften interpretiert werden
- Verzweigungen sind **selten binär**
- Anzahl der Levels (Ebenen) = Anzahl der Trainingsdaten (Muster)
- Baum wächst **bis alle Konten rein** sind
- **Pruning** ist möglich

© Nicole M. Artner

Andere Methoden

… zur Erstellung/Training von Entscheidungsbäumen

C4.5

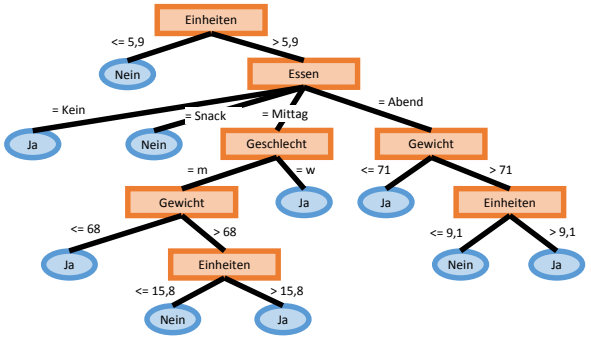
- Nachfolger und Verbesserung von ID3
- populärster Algorithmus für Entscheidungsbäume
- numerische Werte werden wie in CART behandelt
- Heuristiken für Pruning basieren auf statistischer Signifikanz der Teilungen
- Verbesserungen im Umgang mit fehlenden Informationen (im Vergleich zu CART)

© Nicole M. Artner

Zusammenfassung

- Entscheidungsbäume werden zur Klassifikation von Mustern eingesetzt die mit nominellen Merkmalen beschrieben werden

- CART ist ein Algorithmus für die automatische Erstellung von Bäumen



© Nicole M. Artner