Mathematische Grundlagen III

Maschinelles Lernen

Garance PARIS

30. Juni 2011

Maschinelles Lernen

- Künstliche Generierung von Wissen aus Erfahrung
- Erkennung komplexer Muster und Regelmäßigkeiten in vorhandenen Daten
- Ziel: Verallgemeinerung (Generalisierung)
  - Über das Nachschlagen bereits gesehener Beispiele hinausgehen
  - Beurteilung unbekannter Daten
- Beispiele:
  - Die Gleichung einer Geraden an Hand zweier Punkten bestimmen
  - Einen unbekannten Text mit Wortkategorien annotieren

Beispiel

Der Manager eines Golf-Clubs möchte wissen, wann er viele Kunden zu erwarten hat, damit er Studenten als Aushilfe einstellen kann, und wann keiner spielen will, damit er seinen Angestellten freigeben kann.

Zwei Wochen lang führt er Buch darüber, wie das Wetter ist und ob er viele oder wenige Kunden an dem Tag hat.

Er schreibt sich auf:

- ob das Wetter heiter, bewölkt oder regnerisch ist,
- wie warm es ist,
- wieviel Luftfeuchtigkeit es gibt,
- ob der Wind stark weht oder nicht,
- ob er viele Kunden er an dem Tag hat

Beispiel: Der Wetterdatensatz

Das Ergebnis ist für einen Menschen nicht besonders gut durchschaubar...

Outlook	Temp.	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Beispiel: Ein numerischer Datensatz

Outlook	Temp.	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

Beispiel: Ein numerischer Datensatz

Outlook	Temp.	Humidity	Windy	# Cust.
sunny	85	85	false	12
sunny	80	90	true	10
overcast	83	86	false	30
rainy	70	96	false	15
rainy	68	80	false	16
rainy	65	70	true	5
overcast	64	65	true	24
sunny	72	95	false	10
sunny	69	70	false	18
rainy	75	80	false	20
sunny	75	70	true	25
overcast	72	90	true	18
overcast	81	75	false	32
rainy	71	91	true	8

Was kann gelernt werden?

Einige Aufgabetypen:

Klassifikation:

Instanzen vordefinierten Kategorien zuweisen (1. Typ Datensatz)

Clustering:

Instanzen ihrer Ähnlichkeit nach in Gruppen aufteilen (ohne vordefinierte Kategorien)

Numerische Vorhersage: (Regression)

Der numerische Wert eines Merkmals auf Grund der Werte aller anderen Merkmale bestimmen (3. Typ Datensatz)

Klassifikation: Terminologie

Konzept: Das, was gelernt werden soll
Bsp.: Ob unter angegebenen Bedingungen gerne gespielt wird oder nicht

Instanz: Ein einziges Beispiel im Datensatz
Bsp.: Das Wetter an einem der 14 Tagen

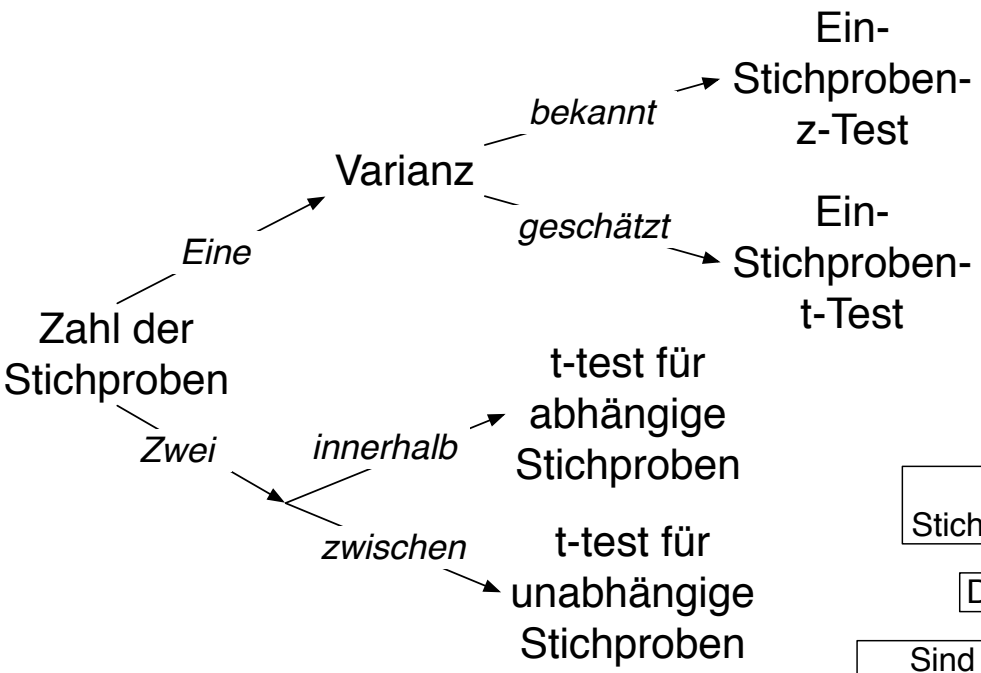
Attribut: Ein Merkmal einer Instanz
Bsp.: *outlook, temperature, humidity, windy*

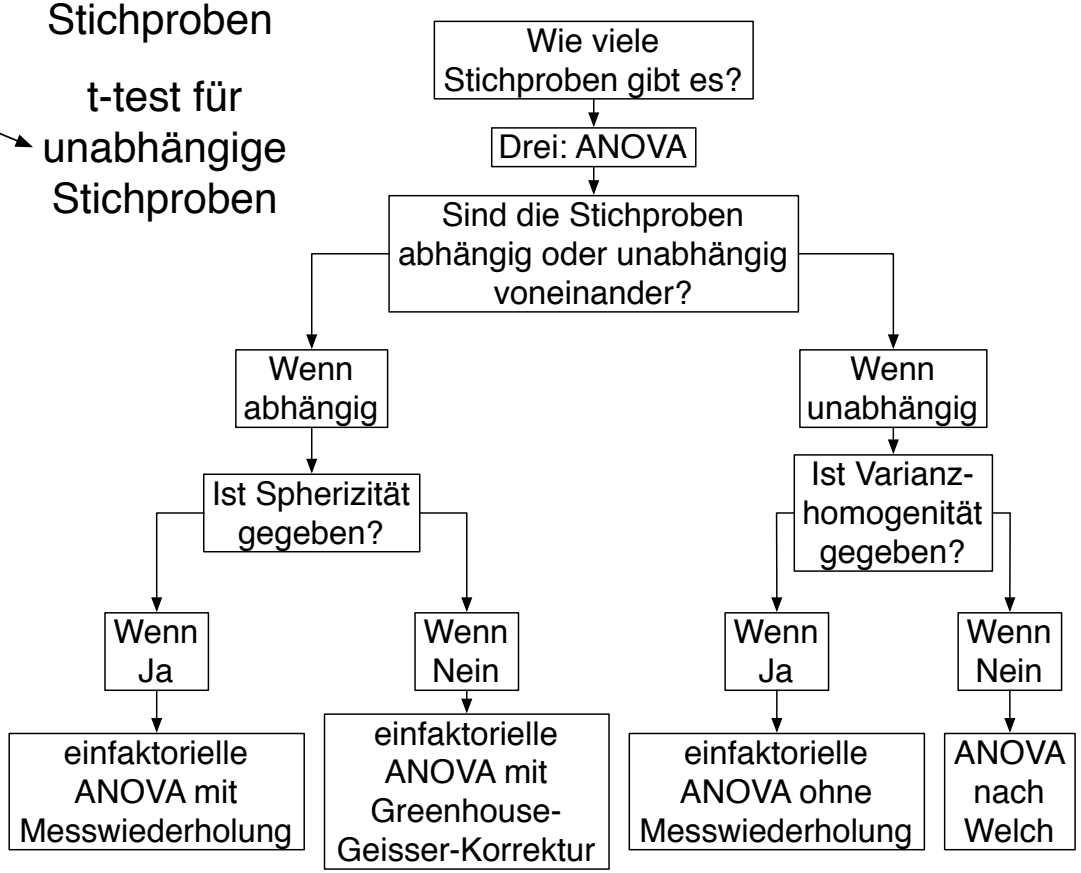
Wert: Wert eines Attributs
Bsp.: Für *outlook*: sunny, overcast, rainy

Kategorie oder Klasse:
Ziel der Klassifikation, Bsp.: yes, no

Entscheidungsbäume

30. Juni 2011


Zahl der Stichproben
Eine
Varianz
bekannt
Ein-Stichproben-z-Test
geschätzt
Ein-Stichproben-t-Test
Zwei
innerhalb
t-test für abhängige Stichproben
zwischen
t-test für unabhängige Stichproben


Wie viele Stichproben gibt es?
Drei: ANOVA
Sind die Stichproben abhängig oder unabhängig voneinander?
Wenn abhängig
Wenn unabhängig
Ist Sphärizität gegeben?
Ist Varianz-homogenität gegeben?
Wenn Ja
Wenn Nein
Wenn Ja
Wenn Nein
einfaktorielle ANOVA mit Messwiederholung
einfaktorielle ANOVA mit Greenhouse-Geisser-Korrektur
einfaktorielle ANOVA ohne Messwiederholung
ANOVA nach Welch

Entscheidungsbäume

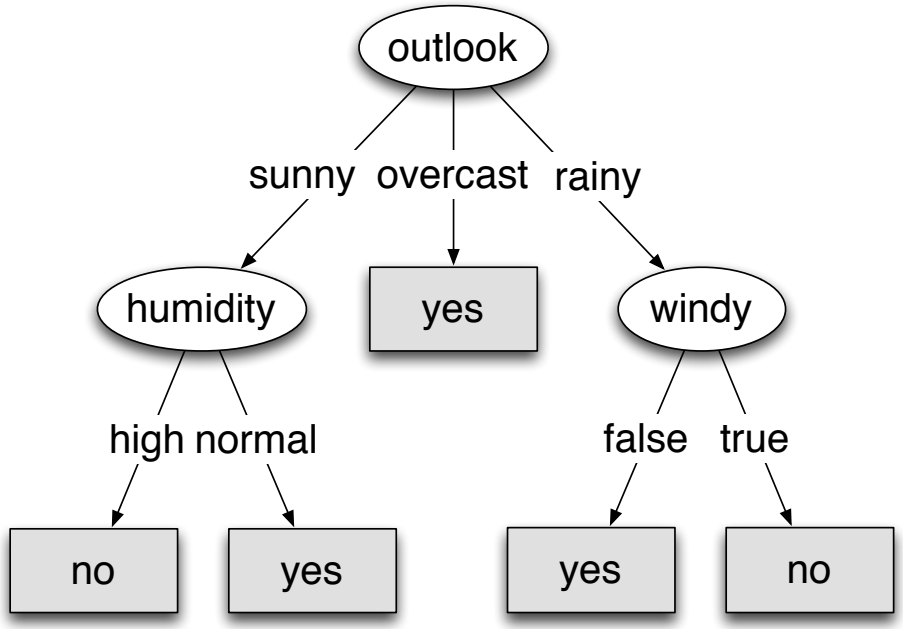
Allgemein

Ein von einem Experten erstellter Baum,
um aufeinanderfolgende, hierarchische Entscheidungen
zu veranschaulichen

Data Mining

Ein auf Basis verfügbarer Daten erstelltes Modell,
das benutzt werden kann, um Voraussagen über
neue Daten zu machen

Baum für den Wetter-Datensatz



- Nicht-terminale Knoten testen ein Attribut
- Baumzweige entsprechen den Werten, die ein Attribut annehmen kann
- Blätter weisen Instanzen einer Kategorie zu

ID3: Intuition

- Bestimmen, welches Attribut die Daten am besten klassifiziert
- Dieses Attribut als Wurzel des Baums verwenden
- Einen Zweig für jeden Wert erstellen, den das Attribut annehmen kann
- Dieses Prozess für jeden Baumzweig wiederholen, jeweils mit der Untermenge der zu klassifizierenden Daten
- Rekursiv weitermachen, bis die Klassifikation unter allen Blättern eindeutig ist

Auswahl des klassifizierenden Attribut

Frage: Wie wird das beste Attribut bestimmt?

Antwort: Es wird das Attribut gewählt, das

...die meiste Information beiträgt

Auswahl des klassifizierenden Attribut

Frage: Wie wird das beste Attribut bestimmt?

Antwort: Es wird das Attribut gewählt, das

...die meiste Information beiträgt

...die Entropie im Datensatz am meisten reduziert

Informationsgewinn

Für jeden Attribut A:

$$Gain(S, A) = E(S) - \sum_{v \in A} \frac{|S_v|}{|S|} E(S_v)$$

S	:	Datensatz
$E(S)$	:	Entropie im Datensatz oder Untermenge davon
S_v	:	Untermenge von S für die A den Wert v hat
$\|S\|$	:	Mächtigkeit von S
$\|S_v\|$	:	Mächtigkeit von S_v

Der Informationsgewinn $Gain(S, A)$ ist die Reduzierung der Entropie, die man dadurch erwartet, dass man den Wert von Attribut A kennt

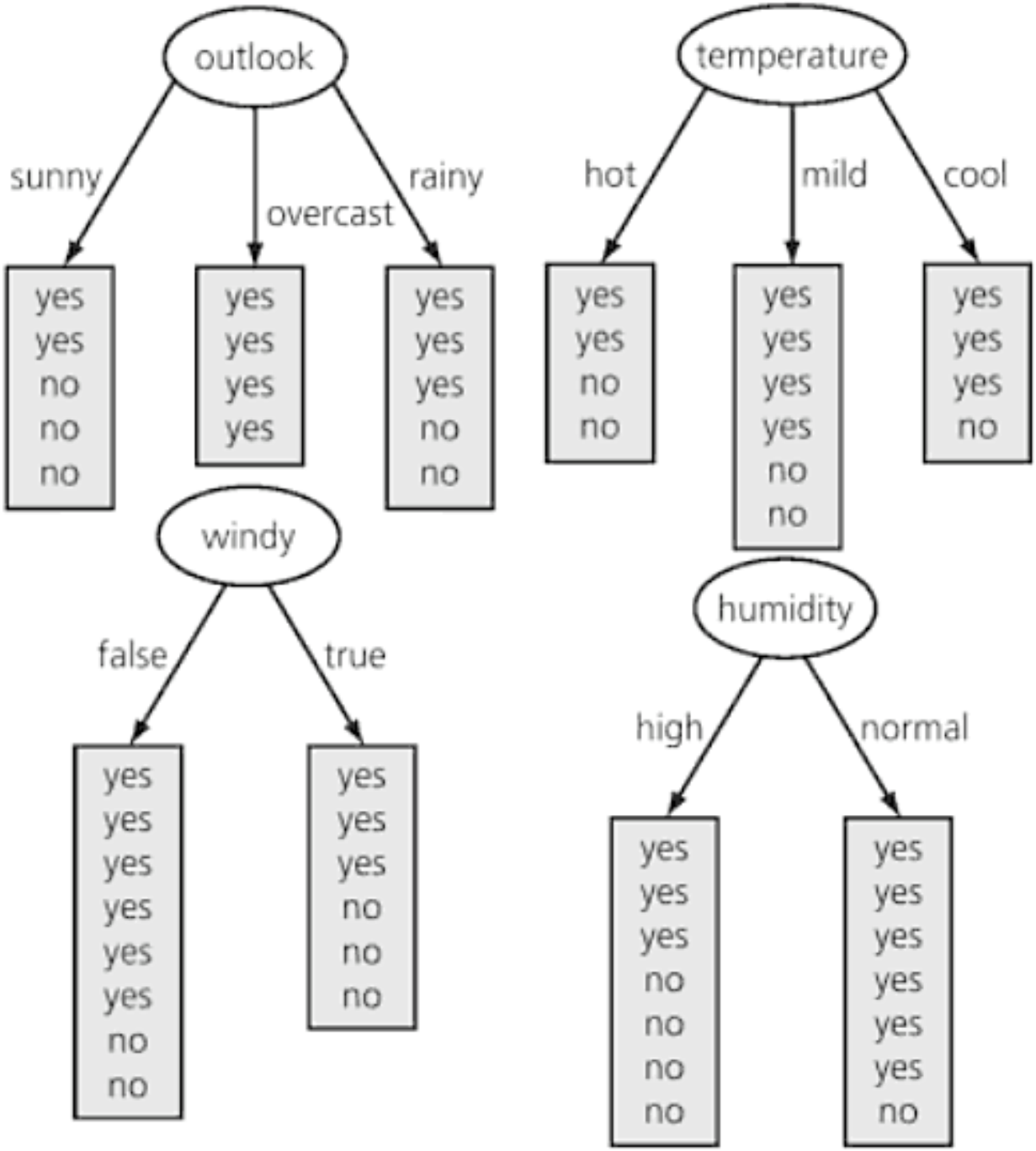
Entropie im Datensatz

$$E(S) = -\sum_{k=1}^{c} p_k \log_2 p_k$$

c: Anzahl Kategorien
p_k: Anteil der Instanzen in S, die Kategorie k angehören

Informationstheoretische Interpretation

- $E(S)$: Anzahl Bits, die benötigt werden, um ein beliebiges Element aus S zu kodieren
- $E(S)$ ist gleich null, wenn alle Instanzen von S der selben Kategorie angehören
- Bei einer binären Klassifikation ist $E(S)$ gleich eins, wenn S gleich viele Instanzen aus jeder Klasse enthält



Garance PARIS

Beispiel für den Wetterdatensatz

- Erst die Gesamtentropie berechnen

  Da von insgesamt 14 Instanzen 5 als *"play=yes"* klassifiziert werden und 9 als *"play=no"* ergibt sich:

$$\begin{aligned} E(S) &= -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} \\ &= \frac{14\,log_2\,14 - 9\,log_2\,9 - 5\,log_2\,5}{14} \\ &\quad \text{(Umformung s. Informationstheorie)} \\ &= 0.94 \end{aligned}$$

- Dann für jedes Attribut und für jeden Wert die Entropie berechnen

  Bsp. für "*windy = false*":
  (8 Instanzen, 6 *"play=yes"* + 2 *"play=no"*)

$$E(S_{false}) = \frac{8\,log_2\,8 - 6, log_2\,6 - 2\,log_2\,2}{8} = 0.81$$

Beispiel für den Wetterdatensatz

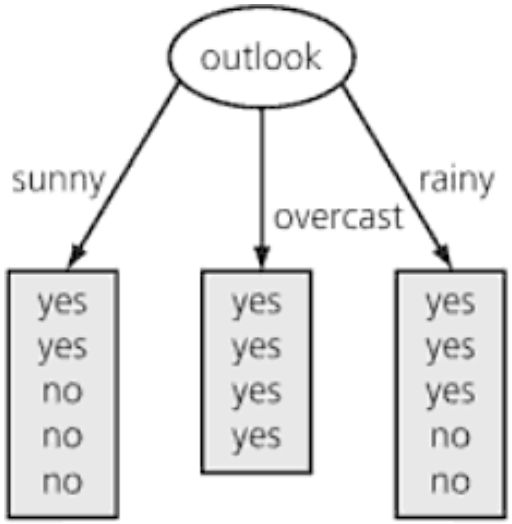
- Bei "*windy* = *true*" kommen dreimal "*play* = *yes*" vor und dreimal "*play* = *no*": Die Entropie ist also gleich eins
- Daraus und aus der Gesamtentropie setzt sich $Gain(S, windy)$ zusammen:

$$\begin{aligned} Gain(S, windy) &= E(S) - \frac{|S_{false}|}{|S|} E(S_{false}) - \frac{|S_{true}|}{|S|} E(S_{true}) \\ &= 0.94 - \frac{8}{14} * 0.81 - \frac{6}{14} * 1 = 0.048 \end{aligned}$$

Beispiel für den Wetterdatensatz

- Ähnlich werden berechnet:

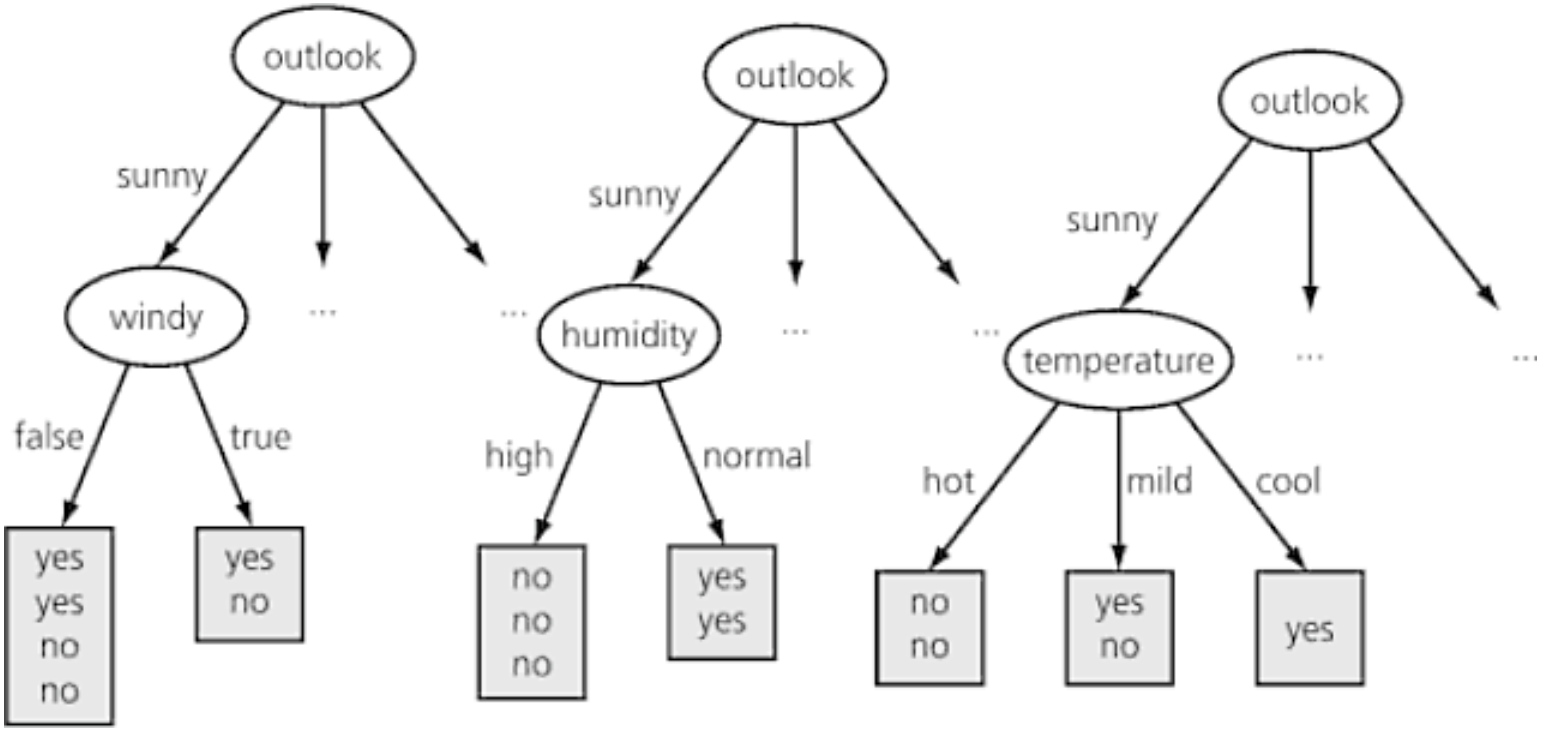
  $Gain(S, windy) = 0.048$
  $Gain(S, outlook) = 0.247$
  $Gain(S, temperature) = 0.029$
  $Gain(S, humidity) = 0.152$



- *Outlook* erzielt den größten Informationsgewinn, daher wird es als Baumwurzel gewählt
- Nächster Schritt:
  Rekursion für jeden Attributwert, z. B. *sunny*,
  5 Instanzen, 2 Mal "*play* = *yes*", 2 Mal "*play* = *no*"

Entscheidungsbäume

Beispiel für den Wetterdatensatz: Rekursion



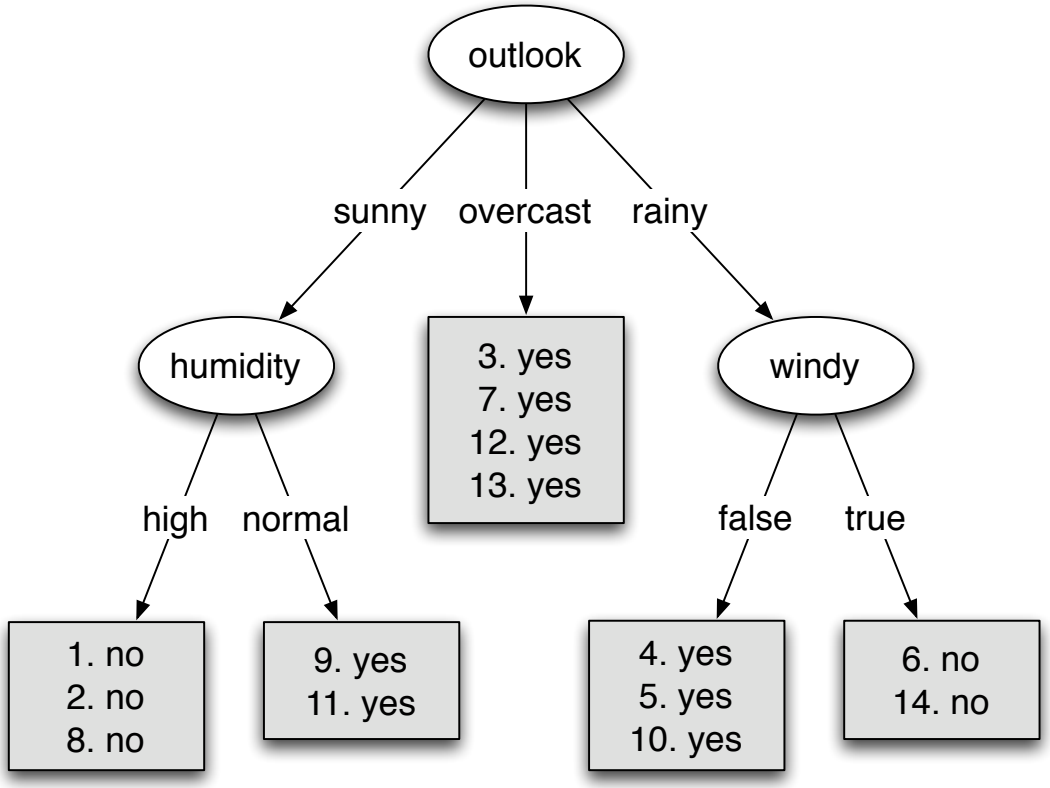
$Gain(S, windy) = 0.020$ $Gain(S, temperature) = 0.571$

$Gain(S, humidity) = 0.971$

Humidity erzielt den größten Informationsgewinn, daher wird es als nächsten Knoten unter "*outlook* = *sunny*" eingesetzt

Entscheidungsbäume

Baum für den Wetter-Datensatz



Alle Instanzen unter einem Blatt werden gleich klassifiziert (Entropie $= 0$)

Entscheidungsbäume

Anmerkung

- Der Pfad zwischen Wurzel und Blatt stellt eine Konjunktion dar
- Jeder Attribut darf auf dem Pfad einmal vorkommen
- Ein Attribut kann in verschiedenen Unterbäume mehrmals vorkommen

Umwandlung in Entscheidungsregeln

`IF (outlook==sunny)` $\wedge$ `(humidity==high)`
`THEN play=no`
`IF (outlook==sunny)` $\wedge$ `(humidity==normal)`
`THEN play=yes`
...

Entscheidungsbäume

Interpretation

- Wenn das Wetter bewölkt ist, spielen Leute immer
- Es gibt Leute, die sogar bei Regen spielen, aber nicht wenn der Wind weht
- Wenn die Sonne scheint, spielen Leute gerne, außer wenn die Luftfeuchtigkeit hoch ist

Schlußfolgerungen:

- Der Manager kann an Tagen, an denen Regen und starken Wind vorausgesagt wird, seinem Personal frei geben
- An sonnigen Tagen, an denen es Luft gibt, kann er sich überlegen, ob er zusätzlich ein paar Studenten einstellt

Eigenschaften des Algorithmus

- Allgemein werden flache Bäume bevorzugt
- Die Suche ist unvollständig, d. h. nicht alle möglichen Bäume werden in Betracht gezogen
- Es ist möglich, dass der “beste” (kleinste) Baum nicht gefunden wird
  (z. B. wenn es einen insgesamt kleineren Baum gibt, der eine andere Wurzel hat, als die, die anfangs den höchsten Informationsgewinn bietet)