Methoden des Maschinellen Lernens

VL 04: Naive Bayes und Entscheidungsbäume

Philipp Viertel Patrick Palsbröker

FH Bielefeld Campus Minden

9. Mai 2018

1. Wahrscheinlichkeitstheorie
   - Bedingte Wahrscheinlichkeiten
   - Bayes Regel
   - Unabhängige Ereignisse
   - Klassifikation mit Naive Bayes

2. Entscheidungsbäume
   - ID3 Algorithmus
   - Beispiel
   - Eigenschaften des ID3

Wahrscheinlichkeitstheorie

Definition:

- Ein Element (world) ω
- Ein Ereignis (sample space) Ω
- Eine Wahrscheinlichkeit $\mathrm{P}(\omega)$
- Eine Proposition ϕ

- Ein Ereignis ist eine endliche Menge an Elementen

$$\Omega = [\omega_1, ..., \omega_n]$$

- An dieser Stelle werden wir nur diskrete Ereignisse mit endlicher Anzahl an Elementen betrachten
- Für Elemente in einem Ereignis gilt:
  - Ausschließend (exclusive) nur ein Ereignis kann eintreten
  - Ausschöpfend (exhaustive) es muss mindestens ein Ereignis eintreten

- Beispielsweise das Ereignis "Wurf eines sechs seitigen Würfels" A:

$$A = [1, 2, 3, 4, 5, 6]$$

- Jedes Element eines Ereignisses hat eine Wahrscheinlichkeit P(ω) wobei:
  - Die Wahrscheinlichkeit für ein Element liegt zwischen 0 und 1
  - Die Summe aller Elemente in einem Ereignis ist 1

$$0 \leq P(\omega) \leq 1 \forall \omega \; ; \sum_{\omega \epsilon \Omega} P(\omega) = 1$$

- Bei einem nicht manipulierten Würfel hat jedes Element die gleiche Wahrscheinlichkeit aufzutreten

$$P(\omega) = \frac{1}{6} \forall \omega \epsilon A$$

- In der Wahrscheinlichkeitstheorie sucht man meistens nach Gruppen von Elementen, diese nennt man Proposition

$$\text{Für jede Proposition } \phi \text{ gilt, } P(\phi) = \sum_{\omega \epsilon \phi} P(\omega)$$

bei gleich-verteilten Wahrscheinlichkeiten auch:

$$P(\phi) = \frac{|\phi|}{|\Omega|} = \frac{\text{Anzahl der Fälle in denen } \phi \text{ hält}}{\text{Anzahl der möglichen Fälle}}$$

- Hier am Beispiel der Proposition "Wurf einer geraden Zahl"

$$"geradeZahl"\, A = [2, 4, 6]\ ;\ P(A) = (\tfrac{1}{6}) + (\tfrac{1}{6}) + (\tfrac{1}{6}) = \tfrac{1}{2}$$

- Außerdem lässt sich ein Zusammenhang zwischen der Wahrscheinlichkeit einer Proposition und seiner Negation herleiten:

$$\begin{aligned} P(\neg A) &= \textstyle\sum_{\omega \epsilon \neg A} P(\omega) \\ &= \textstyle\sum_{\omega \epsilon \neg A} P(\omega) + \sum_{\omega \epsilon A} P(\omega) - \sum_{\omega \epsilon A} P(\omega) \\ &= \textstyle\sum_{\omega \epsilon \Omega} P(\omega) - \sum_{\omega \epsilon A} P(\omega) \\ &= 1 - P(A) \end{aligned}$$

- Somit lässt sich die Proposition "Wurf einer ungeraden Zahl" $\neg A$ folgendermaßen lösen:

$$P(\neg A) = 1 - P(A) = \tfrac{1}{2}$$

[Russell and Norvig(2010)]

Bedingte Wahrscheinlichkeiten

- Bisher nur unbedingte Wahrscheinlichkeit (keine weiteren Informationen verfügbar)
- Jetzt hinzuziehen von bereits eingetretenen Propositionen
- Das Eintreten von Proposition von A nachdem Proposition B bereits eingetreten ist:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

- Dies hält solange P(B) > 0

- Hieraus abgeleitet die Kettenregel:

$$P(A, B) = P(A|B)P(B)$$

$$\begin{aligned} P(A_1, A_2, ..., A_n) &= P(A_n, ..., A_2, A_1) \\ &= P(A_n|A_{n-1}, ..., A_1)P(A_{n-1}, ..., A_1) \\ &= P(A_n|A_{n-1}, ..., A_1)P(A_{n-1}|A_{n-2}, ..., A_1)P(A_{n-2}, ..., A_1) \\ &= ... \\ &= P(A_n|A_{n-1}, ..., A_1) \ ... \ P(A_2|A_1)P(A_1) \\ &= \prod_i P(A_i|A_1, ..., A_{i-1}) \end{aligned}$$

- Wichtig für Sprachmodellierung (Wahrscheinlichkeit für Wortfolge)

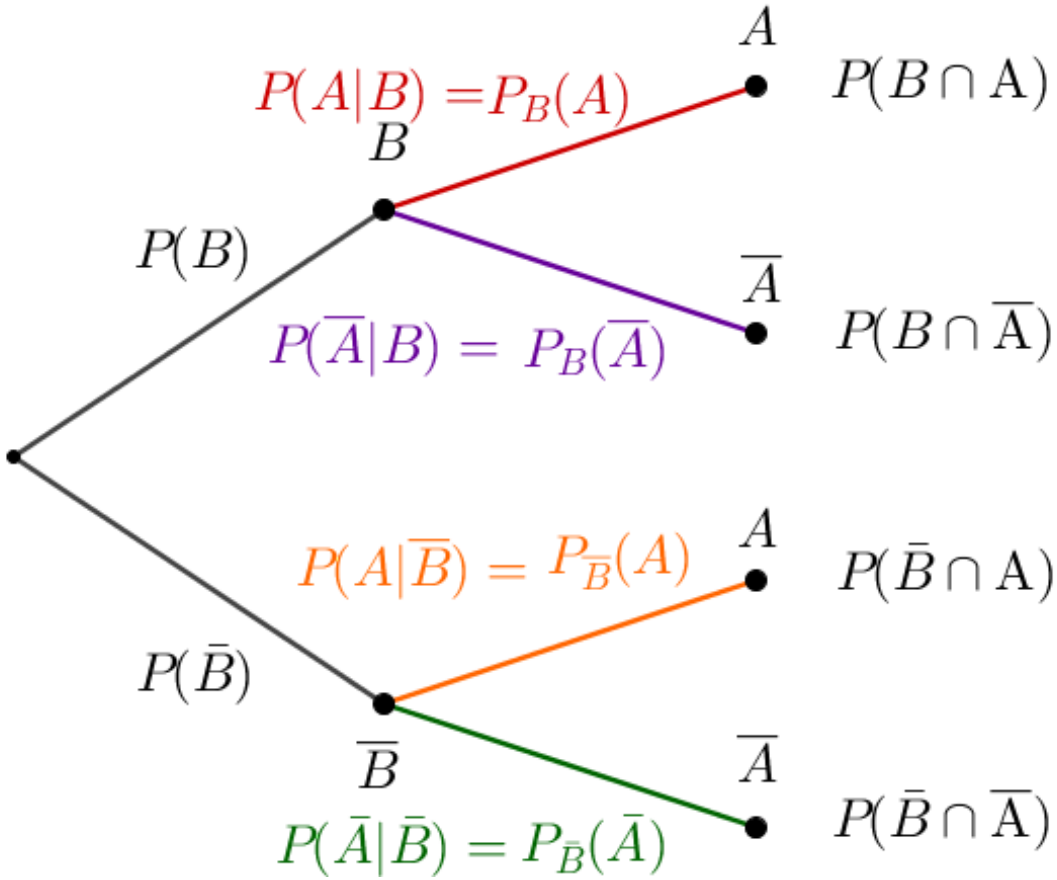


Abbildung: Beispiel Entscheidungsbaum [Serlo(2018)]

Beispiel

Diagnose beim Zahnarzt

- $P(A, B) = P(B, A) =$ Wahrscheinlichkeit, dass A und B gleichzeitig auftreten (Verbundwahrscheinlichkeit)

	Zahnschmerzen	¬Zahnschmerzen
Karies	0.04	0.06
¬Karies	0.01	0.89

- Die Wahrscheinlichkeit, dass ein beliebiger Patient Karies hat beträgt: $P(K, Z) = 0.04$ oder $P(K, \neg Z) = 0.06$ also eine Gesamtwahrscheinlichkeit von 10%

- Fragt man einen Patient ob dieser Zahnschmerzen hat, kann man dieses Wissen in die Prognose einfließen

	Zahnschmerzen	¬Zahnschmerzen
Karies	0.04	0.06
¬Karies	0.01	0.89

$$P(\text{karies}|\text{zahnschmerzen}) = \frac{P(K,Z)}{P(Z)} = \frac{0.04}{0.04+0.01} = 0.8$$

$$P(\text{karies}|\neg\text{zahnschmerzen}) = \frac{P(K,\neg Z)}{P(\neg Z)} = \frac{0.06}{0.06+0.89} = 0.063$$

Einschub Popositions Notation

- Variablen heißen Zufallsvariable (Zahnschmerzen), diese beginnen immer mit einem Großbuchstaben
- Jede Zufallsvariable hat eine Domäne aus möglichen Werten 2,3,4,true,false,sonne,regen,schnee
- die Größe einer Domäne muss nicht zwingend endlich sein
- Propositionen können mit den Mitteln der Aussagenlogik verknüpft werden

$$P(karies|\neg zahnschmerzen \wedge jugendlich) = 0.1$$

- Anstelle von:

$$P(Wetter = sonnig) = 0.6$$
$$P(Wetter = regen) = 0.1$$
$$P(Wetter = wolkig) = 0.29$$
$$P(Wetter = schnee) = 0.01$$

- wird abgekürzt mit:

$$\mathbf{P}(Wetter) = \langle 0.6, 0.1, 0.29, 0.01 \rangle$$

Marginalisierung und Conditionierung

	Zahnschmerzen	¬Zahnschmerzen	Summe
Karies	0.04	0.06	0.1
¬Karies	0.01	0.89	0.9
Summe	0.05	0.95	1

- $P(K) = P(K, Z) + P(K, \neg Z)$
- Seien $B_1, ..., B_n$ Elementarereignisse mit $\sum_i B_i = \Omega$. Dann ist
  $P(A) = \sum_i P(A, B_i)$ (Marginalisierung)
  $= \sum_i P(A|B_i)P(B_i)$ (Conditionierung)

	Zahnschmerzen	¬Zahnschmerzen	Summe
Karies	0.04	0.06	0.1
¬Karies	0.01	0.89	0.9
Summe	0.05	0.95	1

- Mit diesen Regeln können durch das Aufsummieren ihrer Wahrscheinlichkeiten, Variablen aus der Funktion entfernt werden

Bayes Regel

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- $P(A)$ nennt man auch "Prior" oder "A-priori-Wahrscheinlichkeit":
  Die Wahrscheinlichkeit für A ohne weiteres Wissen
- $P(B|A)$ nennt man auch "Likelihood":
  Mutmaßung für das Auftreten von B, gegeben A, keine Wahrscheinlichkeit
- $P(A|B)$ nennt man auch "Posterior" oder "A-posteriori-Wahrscheinlichkeit„"
  Wie wahrscheinlich ist A, wenn B eingetreten ist?
- $P(B)$ ist die Evidenz und wird als Normierungsfaktor verwendet.

- In der Medizin hat sucht man i.d.R. die Ursache für beobachtete Symptome:

$$P(\text{Ursache}|\text{Symptome}) = \frac{P(\text{Symptome}|\text{Ursache})P(\text{Ursache})}{P(\text{Symptome})}$$

- Aus der A-priori-Wahrscheinlichkeit für bestimmte Krankheiten und der Likelihood der Symptome (wie wahrscheinlich sind Symptome, gegeben einer Krankheit) kann man die Wahrscheinlichkeit für das Vorliegen einer Erkrankung gegeben bestimmter Symptome berechnen.

Beispiel

- Bei Meningitis wird oft ein steifer Hals beobachtet: $P(S|M) = 0.8$
- Eine von 10000 Personen hat Meningitis: $P(M) = 0.0001$
- Eine von 10 Personen hat einen steifen Hals: $P(S) = 0.1$

Ich habe einen steifen Hals. Habe ich Meningitis?

$$P(M|S) = \frac{P(S|M)P(M)}{P(S)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008 = 0.08\%$$

- Bei einem steifen Hals liegt die Wahrscheinlichkeit, an Meningitis erkrankt zu sein, bei nur 0.08%. Also kein Grund zur Sorge in diesem Fall

Beispiel aus Wolfgang Ertel: "Grundkurs Künstliche Intelligenz", Springer Vieweg, 2016 (Seite 148)[Ertel(2013)]

- Alarmanlage meldet jeden Einbruch mit einer Sicherheit von 99%
- Wahrscheinlichkeit für Einbruch: 0.1%
- Wird auch von Tieren o.ä. ausgelöst, Alarm-Wahrscheinlichkeit 10%

Wie wahrscheinlich ist ein Einbruch, wenn der Alarm ertönt?

- Gegeben: $P(A) = 0.1,\ P(E) = 0.001,\ P(A|E) = 0.99$
- Gesucht: $P(E|A)$

$$P(E|A) = \frac{P(A|E)P(E)}{P(A)} = \frac{0.99 \times 0.001}{0.1} = 0.0099 = 0.99\%$$

- Bei Ertönen des Alarms liegt die Wahrscheinlichkeit für einen Einbruch bei nur knapp 1 Prozent. Diese Alarmanlage ist vielleicht nicht sehr hilfreich ...

- Frage: Wie wahrscheinlich ist ein Alarm ohne Einbruch, also $P(A|\neg E)$?
- Mit Marginalisierung:
  $P(A) = P(A|E)P(E) + P(A|\neg E)P(\neg E)$, d.h.
  $0.1 = 0.99 \times 0.001 + P(A|\neg E) \times (1 - 0.001) =$
  $0.00099 + P(A|\neg E) \times 0.999$
- $P(A|\neg E) = 0,0991$

In knapp 10 Prozent der Fälle wird der Alarm ohne Einbruch ausgelöst ...

Unabhängige Ereignisse

$$P(\text{Zahnschmerzen}, \text{Regen}) = P(\text{Regen}|\text{Zahnschmerzen})P(\text{Zahnschmerzen})$$

$$P(\text{Regen}|\text{Zahnschmerzen}) = \ ?$$

Unabhängige Ereignisse

$$P(\text{Zahnschmerzen}, \text{Regen}) = P(\text{Regen}|\text{Zahnschmerzen})P(\text{Zahnschmerzen})$$

$$P(\text{Regen}|\text{Zahnschmerzen}) = \ ? \ = P(\text{Regen})$$

- Zwei Ereignisse A und B sind unabhängig, wenn

$$P(A|B) = P(A)$$

- Daraus folgt:

$$P(A, B) = P(A|B)P(B) = P(A)P(B)$$

Bedingte Unabhängigkeit:

- X und Y sind bedingt unabhängig (gegeben Z), wenn $P(X|Y,Z) = P(X|Z)$ bzw. $P(Y|X,Z) = P(Y|Z)$
- Daraus folgt:

$$P(X,Y|Z) = P(X|Y,Z)P(Y|Z) = P(X|Z)P(Y|Z)$$

Klassifikation mit Naive Bayes

- Naive Bayes ist eine Verallgemeinerung der Bayes Regel

$$P(H|D_1, ..., D_n) = \frac{P(D_1,...,D_n|H)P(H)}{P(D_1,...,D_n)}$$

- Ist "naiv" in der Annahme, dass alle Daten bedingt unabhängig sind:

$$P(D_1, ..., D_n|H) = P(D_1|H) \cdot ... \cdot P(D_n|H) = \prod_i P(D_i|H)$$

- Beobachtung: $P(D_1, ..., D_n)$ für alle Hypothesen $h \in H$ gleich

- Naive Bayes Klassifikator bzw. MAP("Maximum a Posteriori")

$$h_{MAP} = \text{argmax}_{h \in H} P(h|D_1, ..., D_n) = \text{argmax}_{h \in H} P(h) \prod_i P(D_i|h)$$

- Naive Bayes: Wähle die plausibelste Hypothese, die von den Daten unterstützt wird.

Bayes'sches Lernen

$$h_{MAP} = \text{argmax}_{h \in H} P(h|D_1, ..., D_n) = \text{argmax}_{h \in H} P(h) \prod_i P(D_i|h)$$

Training: Bestimme die Wahrscheinlichkeiten aus Trainingsdaten **S**

- Für jede Klasse h
  - Schätze $P(h) = \frac{|S(h)|}{|S|}$
  - Für jedes Attribut D_i und jede Ausprägung $x \in D_i$:
    - Schätze $P(D_i = x|h) = \frac{|S_{D_i}(x) \cap S(h)|}{|S(h)|}$

Klassifikation: Wähle wahrscheinlichste Klasse h_{MAP} für Vektor $\mathbf{x}$

- $h_{MAP} = \text{argmax}_{h \in H} P(h) \prod_{x \in \mathbf{x}} P(x|h)$

Tafelbeispiel

Beispiel	Himmel	Lufttemp.	Wassertemp	Wind	Schwimmen?
1	sonnig	warm	warm	windstill	ja
2	sonnig	kalt	warm	stürmisch	nein
3	sonnig	warm	warm	brise	ja
4	bewölkt	kalt	kalt	windstill	nein
5	regen	kalt	warm	windstill	ja

- Eingabe: Merkmalsvektor ‘(sonnig, kalt, warm, windstill)‘
- Gesucht: $P(\text{ja})$, $P(\text{nein})$,
  $P(\text{Himmel}{=}\text{sonnig}|\text{ja}), P(\text{Himmel}{=}\text{sonnig}|\text{nein})$, ...

Wähle Klasse

$$h_{MAP} = \text{argmax}_{h \in \{\text{ja, nein}\}} \quad P(h) \cdot P(\text{Himmel}{=}\text{sonnig}|h) \cdot P(\text{Luft}{=}\text{kalt}|h) \cdot P(\text{Wasser}{=}\text{warm}|h) \cdot P(\text{Wind}{=}\text{windstill}|h)$$

Naivität in Naive Bayes

Theoretisch: Unabhängigkeit der Attribute oft nicht gegeben

- $P(D_1, ..., D_n|H) \neq \prod_i P(D_i|H)$
- A-posteriori-Wahrscheinlichkeiten oft unrealistisch nah an 1 oder 0

Praxis: Dennoch häufig sehr gute Ergebnisse

- Wichtig:
  $\text{argmax}_{h \in H} P(h)P(D_1, ..., D_n|h) = \text{argmax}_{h \in H} P(h) \prod_i P(D_i|h)$
- Solange die Maximierung die selben Ergebnisse liefert, müssen die konkreten Schätzungen/Werte nicht exakt stimmen ...

Wenn Attribute nicht (bedingt) unabhängig sind, kann sich der NB verschätzen, d.h. es kommt dann u.U. zu einer höheren Fehlerrate, da bestimmte Eigenschaften in der Trainingsmenge zu hoch gewichtet werden.

Probleme mit Floating Point Underflow

- MAP berechnet Produkt mit vielen Termen
- Problem: Bei kleinen Zahlen kann Floating Point Underflow auftreten!
- Lösung: Logarithmus maximieren (Produkt geht in Summe über)

- Erinnerung: $\log(x \cdot y) = \log(x) + \log(y)$ und Logarithmus streng monoton

$$\begin{aligned} h_{MAP} &= \text{argmax}_{h \in H} P(h|D_1, ..., D_n) \\ &= \text{argmax}_{h \in H} P(h) \prod_i P(D_i|h) \\ &= \text{argmax}_{h \in H} [\log(P(h)) + \sum_i \log(P(D_i|h))] \end{aligned}$$

Laplace-Schätzer

- Problem: Attributsausprägung für bestimmte Klasse nicht in Trainingsmenge:
  - Bedingte Wahrscheinlichkeit ist 0
  - Produkt gleich 0
- Lösung: “Laplace-Schätzer“ (auch “Laplace-Glättung“)

- Statt $P(D_i = x|h) = \frac{|S_{D_i}(x) \cap S(h)|}{|S(h)|}$
- nutze $P(D_i = x|h) = \frac{|S_{D_i}(x) \cap S(h)| + m \cdot p_i}{|S(h)| + m}$
- mit m: frei wählbarer Faktor
- p_i: A-priori-Wahrscheinlichkeit für $P(D_i = x|h)$

“virtuelle“ Trainingsbeispiele (m ist die Zahl der virtuellen Trainingsbeispiele)

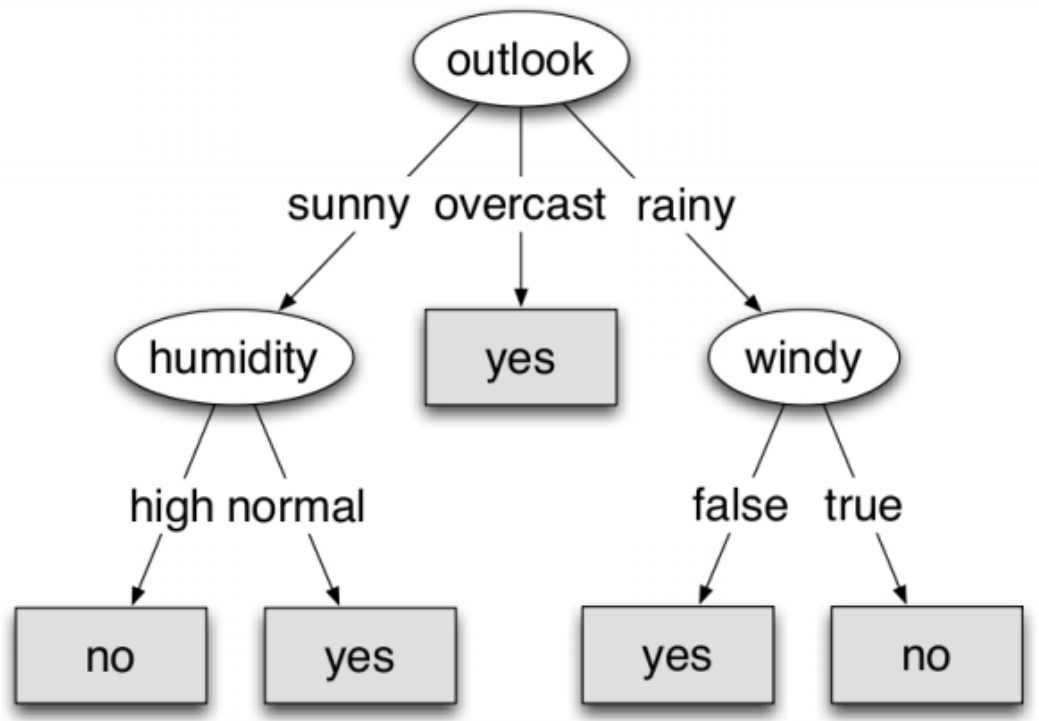
Grenzen von Naive Bayes

- Ungeeignet zum Lernen von Regeln wie:

  Wenn: (Wassertemp=warm $\wedge$ Wind=$\neg$stürmisch) dann: Schwimmen

- Mit Naive Bayes nicht möglich, da angenommen wird das alle Attribute unabhängig sind

Entscheidungsbäume



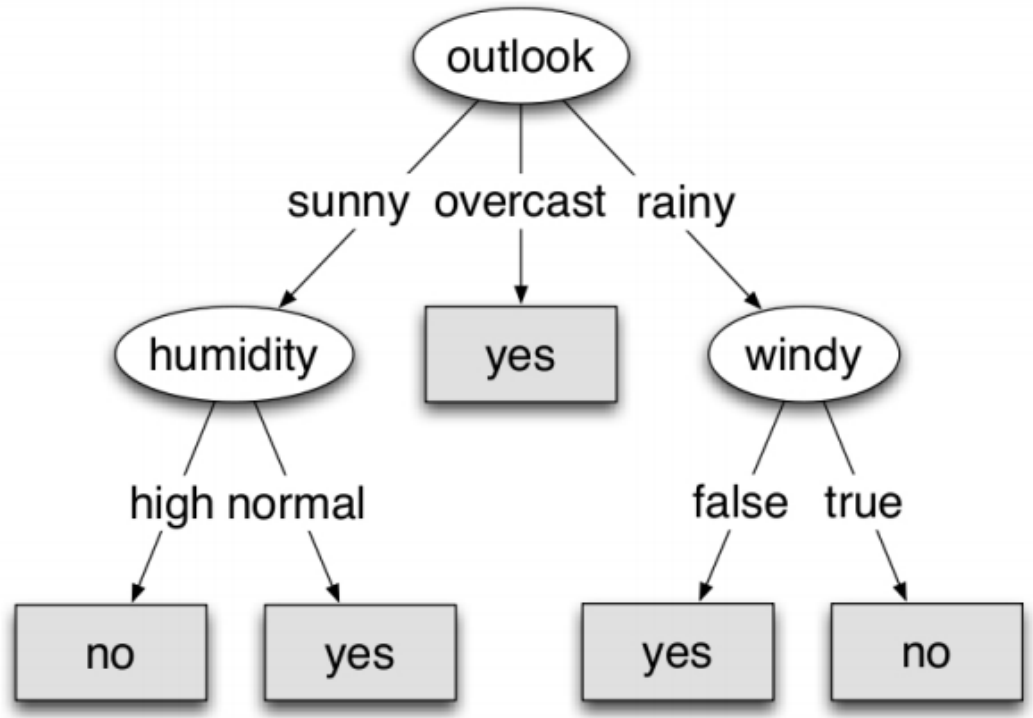
- Eingabe: Instanz $\mathbf{x} \epsilon X$
  - Instanz wird durch Vektor von Attributen repräsentiert

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

  - Wird auch Merkmalsvektor genannt
- Ausgabe: Klasse $y \epsilon Y$
  - z.B. (akzeptiert,abgelehnt), (Spam, kein Spam),(schwimmen, nicht schwimmen)
  - Auch Zielattribut genannt

Entscheidungsbäume sind gut geeignet wenn:

- Instanzen aus Attribut Wert-Paaren mit diskreten Werten bestehen
- Man Kategorien lernen möchte und keine kontinuierliche Funktion
- Man nur unvollständige oder leicht fehlerhafte Trainingsdaten besitzt



- Knoten testen ein Attribut
- Zweige entsprechen den Werten eines Attributs
- Blätter weisen Instanzen einer Kategorie zu

ID3 Algorithmus

Vorgehen:

- Am stärksten klassifizierendes Attribut finden
- Stärkstes Attribut als Wurzel setzen
- Zweig für jeden möglichen Wert des Attributs erstellen
- Rekursiv für jeden Zweig mit der Untermenge der zu klassifizierenden Daten wiederholen bis die Klasse des Zweigs eindeutig ist

Entropie

Gegeben:

- $\mathbf{X}$= Datensatz
- $|\mathbf{Y}|$= Anzahl der Klassen
- p_c= Anzahl der Instanzen in $\mathbf{X}$ die der Klasse c angehören
- H($\mathbf{X}$)= Entropie des Datensatzes

$$H(\mathbf{X}) = -\sum_{c=1}^{\mathbf{Y}} p_c log_2 p_c$$

Informationgain

Gegeben:

- $\mathbf{X}$ = Datensatz
- H($\mathbf{X}$) = Entropie im Datensatz oder einer Untermenge davon
- $\mathbf{X}_V$ = Untermenge von $\mathbf{X}$ für die **x** den Wert v hat
- $|\mathbf{X}|$ = Mächtigkeit $\mathbf{X}$
- $|\mathbf{X}_V|$ = Mächtigkeit von $\mathbf{X}_V$
- Gain($\mathbf{X}$,A) = Erwartete Reduzierung der Entropie bei bekanntem Attribut A

$$Gain(\mathbf{X}, A) = H(\mathbf{X}) - \sum_{v \epsilon A} \frac{|\mathbf{X}_v|}{|\mathbf{X}|} H(\mathbf{X}_v)$$

Beispieldatensatz

Wir wollen lernen bei welchen Wetterbedingungen draußen gespielt wird:

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 aus Mathematische Grundlagen III: Maschinelles Lernen II von Vera Demberg

- Schritt 1: Berechnen der Gesamtentropie
  14 Instanzen. 5 mal “play=no“ und 9 mal “play=yes“

$$H(\mathbf{X}) = -\frac{9}{14} log_2 \frac{9}{14} - \frac{5}{14} log_2 \frac{5}{14} = 0.940286$$

- Schritt 2: Für jedes Attribut den Datensatz nach Werten des Attributs unterteilen und hierfür die Entropie berechnen
  A : “windy = false“:
  8 Instanzen, 6 mal “play=yes“ + 2 mal “play=no“

$$H(\mathbf{X}_{false}) = -\frac{6}{8} log_2 \frac{6}{8} - \frac{2}{8} log_2 \frac{2}{8} = 0.8112781$$

  A : “windy = true“:
  6 Instanzen, 3 mal “play=yes“ + 3 mal “play=no“

$$H(\mathbf{X}_{true}) = -\frac{3}{6} log_2 \frac{3}{6} - \frac{3}{6} log_2 \frac{3}{6} = 1$$

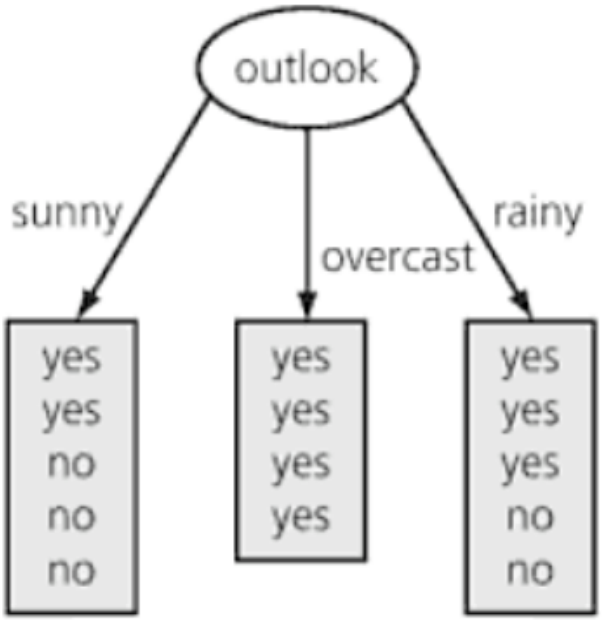
- Schritt 3: Aus den Teilentropien und der Gesamtentropie den Informationgain ermitteln

$$Gain(\mathbf{X}, windy) = H(\mathbf{X}) - \frac{|\mathbf{X}_{false}|}{|\mathbf{X}|} H(\mathbf{X}_{false}) - \frac{|\mathbf{X}_{true}|}{|\mathbf{X}|} H(\mathbf{X}_{true})$$

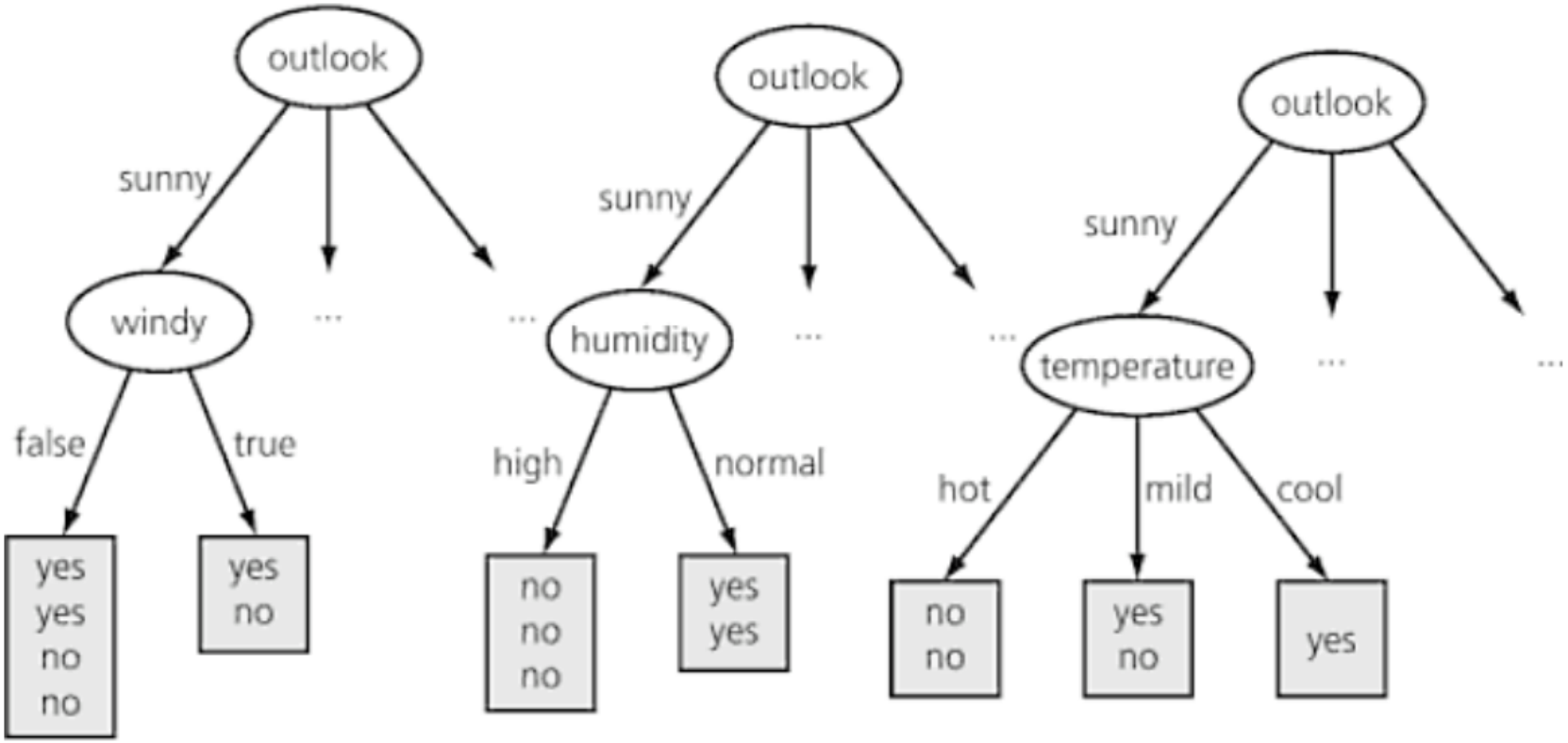
$$Gain(\mathbf{X}, windy) = 0.940286 - \frac{8}{14} * 0.8112781 - \frac{6}{14} * 1 = 0.04812703$$

- Diese Schritte werden für alle Attribute wiederholt:
  $Gain(\mathbf{X}, outlook) = 0.247$
  $Gain(\mathbf{X}, temperature) = 0.029$
  $Gain(\mathbf{X}, humidity) = 0.152$

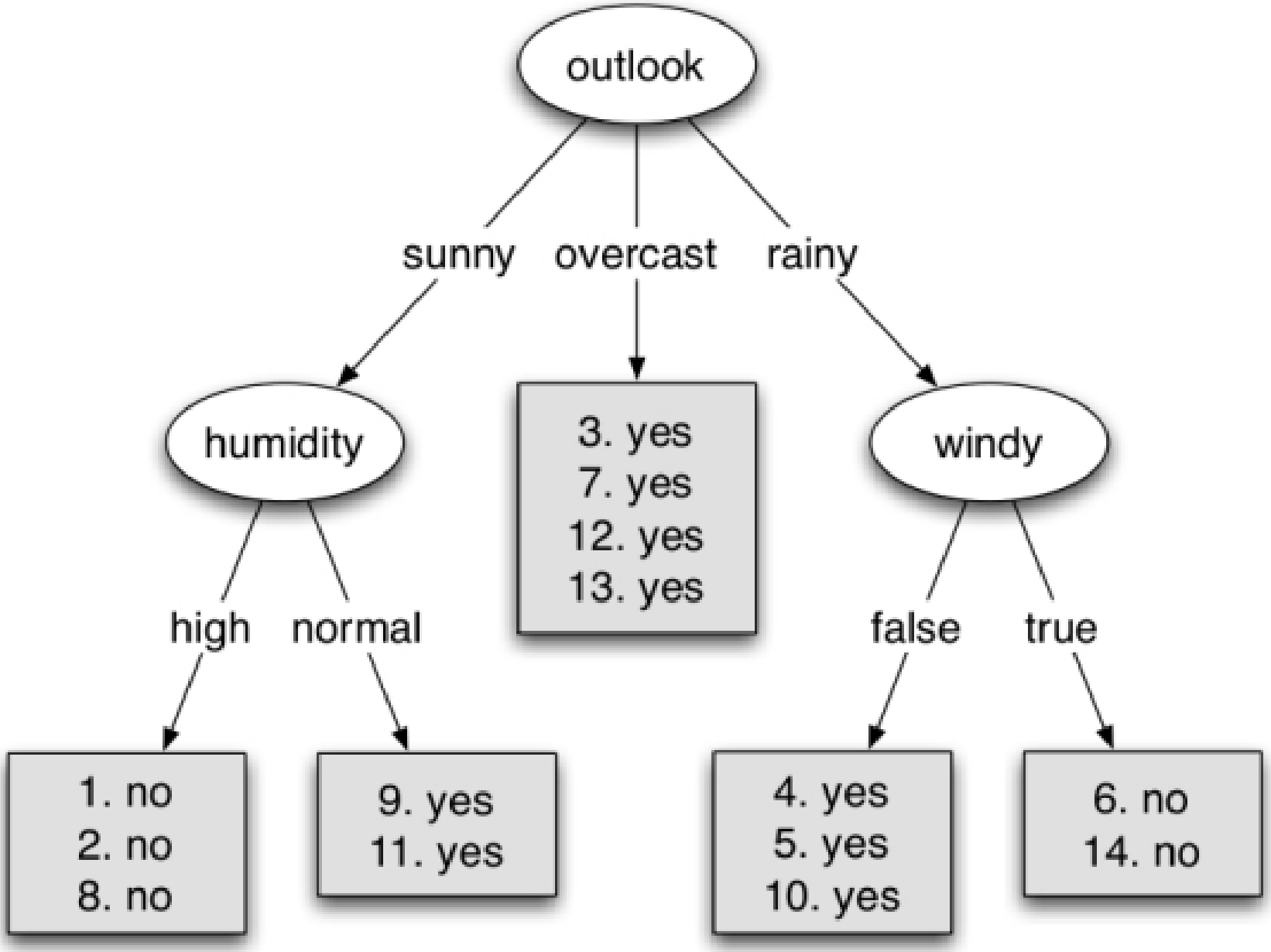
- Outlook als Wurzel setzen



- Rekursiv für jeden Wert des Attributs wiederholen



$$Gain(\mathbf{X}, windy) = 0.020 \qquad Gain(\mathbf{X}, temperature) = 0.571$$
$$Gain(\mathbf{X}, humidity) = 0.971$$


outlook
sunny
overcast
rainy
humidity
3. yes
7. yes
12. yes
13. yes
windy
high
normal
false
true
1. no
2. no
8. no
9. yes
11. yes
4. yes
5. yes
10. yes
6. no
14. no

Eigenschaften des ID3 Algorithmus

- Auswahl der Knoten nach höchstem Informationsgewinn über den gesamten Datensatz
- Robust gegenüber Unregelmäßigkeiten
- Keine vollständige Suche (nicht zwingend der optimale Baum)
- Flache Bäume werden bevorzugt
- Neigt zu Overfitting

Ausblick

weiterführende Recherche:

- NB mit kontinuierlichen Attributen
  - Diskretisierung der Attribute
  - Dichtefunktionen
- Entscheidungsbäume optimieren
  - Pruning
- C4.5 Algorithmus

Vielen dank für Eure Aufmerksamkeit


Wolfgang Ertel.
Grundkurs Künstliche Intelligenz: Eine praxisorientierte Einführung.
Springer Fachmedien Wiesbaden, 3. aufl. 2013 edition, 2013.
ISBN 9783834816771.


Russell and Peter Norvig.
Artificial intelligence: A modern approach.
2010.


Serlo.
Bedingte warscheinlichkeit, March 2018.
URL `https://de.serlo.org/mathe/stochastik/bedingte-wahrscheinlichkeit-unabhaengigkeit/bedingte-wahrscheinlichkeit/bedingte-wahrscheinlichkeit`.