GOETHE UNIVERSITÄT FRANKFURT AM MAIN

www.uni-frankfurt.de

Einführung in die Methoden der Künstlichen Intelligenz

Maschinelles Lernen

Dr. David Sabel

WS 2012/13

Stand der Folien: 14. Februar 2013

Einführung

- Direkte Programmierung eines intelligenten Agenten nicht möglich (bisher)
- Daher benötigt: **Maschinelles Lernen**
- Viele Ansichten / Begriffe, was maschinelles Lernen ist
- Erfolgreichste Ansätze verwenden statistische / stochastische Methoden
- Basieren auf Adaption von Werten / Gewichten

Lernen und Agenten

Lernen soll Performanz des Agenten verbessern:

- Verbesserung der internen Repräsentation
- Optimierung bzw. Beschleunigung der Erledigung von Aufgaben.
- Erweiterung des Spektrums oder der Qualität der Aufgaben, die erledigt werden können.

Beispiele

- Anpassung / Erweiterung des Lexikons e. computerlinguistischen Systems
  Inhalt wird angepasst, aber auch gleichzeitig die Semantik
- Bewertungsfunktion im Zweipersonenspiel (Adaption der Gewichte), war für Dame und Backgammon erfolgreich
- Lernen einer Klassifikation durch Trainingsbeispiele

Struktur eines lernenden Systems

- **Agent**: (ausführende Einheit, performance element). Soll verbessert werden anhand von Erfahrung
- **Lerneinheit: (learning element)** Steuerung des Lernvorgangs. Vorgaben was schlecht ist. Bewertungseinheit (critic) und Problemgenerator
- **Umwelt**: Umwelt in der aggiert wird (künstlich oder real)

Lernmethoden

- **Überwachtes Lernen (supervised learning)**
  - Es gibt einen „allwissenden Lehrer“
  - Er sagt dem Agent, nach seiner Aktion, ob diese richtig / falsch wahr
  - unmittelbares Feedback
  - Alternative: Gebe positiv/negative Beispiele am Anfang vor

Lernmethoden

- **Überwachtes Lernen (supervised learning)**
  - Es gibt einen „allwissenden Lehrer“
  - Er sagt dem Agent, nach seiner Aktion, ob diese richtig / falsch wahr
  - unmittelbares Feedback
  - Alternative: Gebe positiv/negative Beispiele am Anfang vor
- **Unüberwachtes Lernen (unsupervised learning)**
  - Agent, weiß nicht, was richtig ist
  - Bewertung der Güte der Aktion
  - z.B. Agent misst den Effekt selbst

Lernmethoden

- **Überwachtes Lernen (supervised learning)**
  - Es gibt einen „allwissenden Lehrer“
  - Er sagt dem Agent, nach seiner Aktion, ob diese richtig / falsch wahr
  - unmittelbares Feedback
  - Alternative: Gebe positiv/negative Beispiele am Anfang vor
- **Unüberwachtes Lernen (unsupervised learning)**
  - Agent, weiß nicht, was richtig ist
  - Bewertung der Güte der Aktion
  - z.B. Agent misst den Effekt selbst
- **Lernen durch Belohnung/Bestrafung (reinforcement learning)**
  - Lernverfahren belohnen gute Aktion, bestrafen schlechte
  - D.h. Aktion ist bewertbar, aber man kennt den richtigen Parameter nicht

Lernmethoden (2)

Mögliche Vorgehensweisen:

- inkrementell,
- alle Beispiele auf einmal.

Mögliche Rahmenbedingungen:

- Beispielwerte: exakt / ungefähr (fehlerhaft)
- nur positive bzw. positive und negative Beispiele

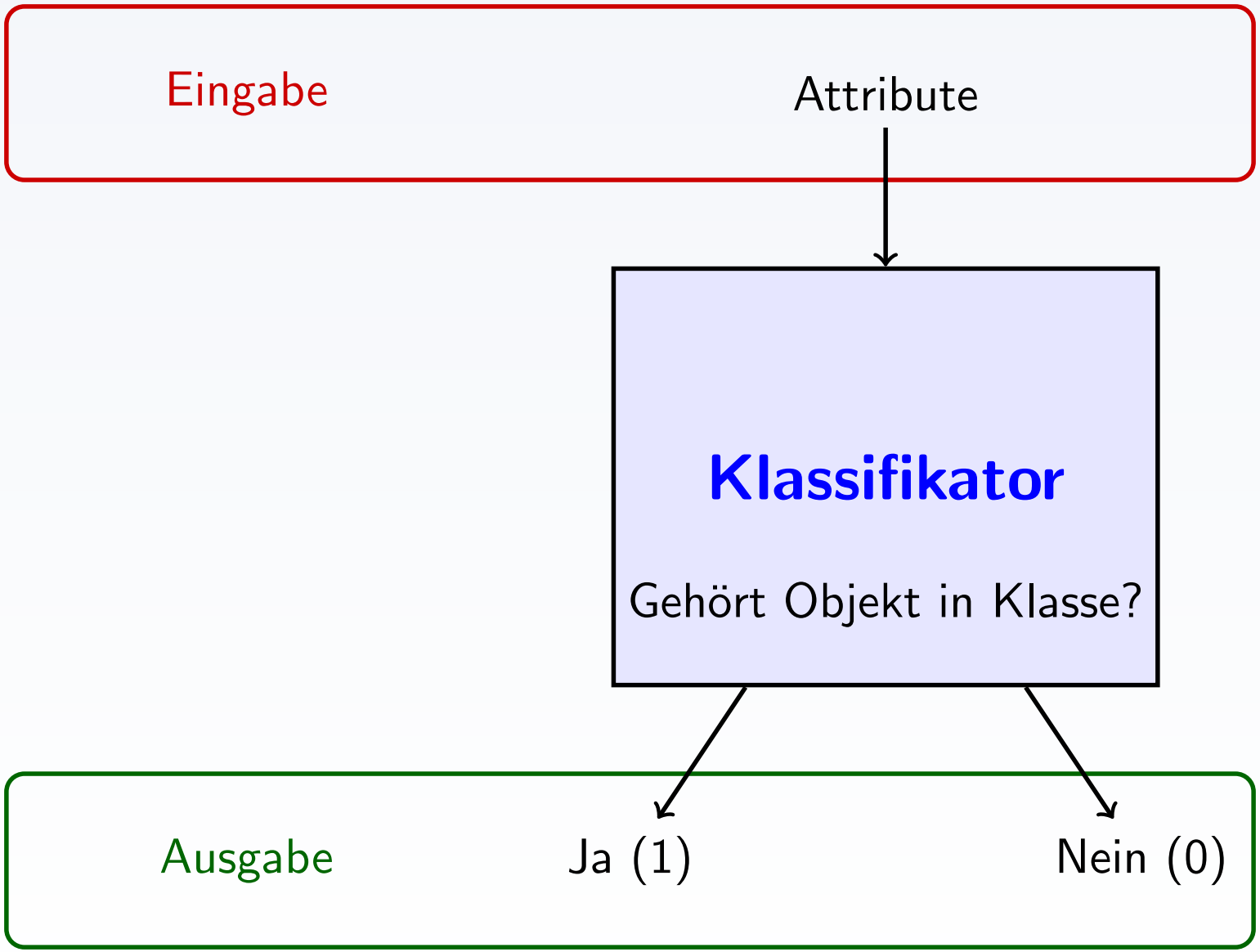
Klassifikationsverfahren

Klassifikation anhand von Eigenschaften (Attributen)

Beispiele:

- Vogel: kann-fliegen, hat-Federn, Farbe,...
- Vorhersage, ob ein Auto im kommenden Jahr einen Defekt hat: Alter, Kilometerstand, letzte Reparatur, Marke, ...
- Medizinischer Test auf Krankheit: Symptome, Blutwerte, ...
- Kreditwürdigkeit e. Bankkunden: Einkommen, Alter, Eigentumsverhältnisse, Adresse

Klassifikator



Abstrakte Situation

- Menge M von Objekten (mit innerer Struktur)
- Programm $P : M \to \{0, 1\}$
- Wahre Klassifikation $K : M \to \{0, 1\}$

Abstrakte Situation

- Menge M von Objekten (mit innerer Struktur)
- Programm $P : M \to \{0, 1\}$
- Wahre Klassifikation $K : M \to \{0, 1\}$

Eingabe: Objekt x

- Wenn $K(x) = P(x)$, dann liegt das Programm richtig
  - **richtig-positiv**: Wenn $P(x) = 1$ und $K(x) = 1$
  - **richtig-negativ**: Wenn $P(x) = 0$ und $K(x) = 0$
- Wenn $K(x) \neq P(x)$, dann liegt das Programm falsch:
  - **falsch-positiv**: Wenn $P(x) = 1$, aber $K(x) = 0$
  - **falsch-negativ**: Wenn $P(x) = 0$, aber $K(x) = 1$

Beispiel: Schwangerschaftstest

Beispieldaten: 200 durchgeführte Tests

Test ergab ...	positiv	negativ
Schwangere Frauen	59	1
Nichtschwangere Frauen	15	125

Beispiel: Schwangerschaftstest

Beispieldaten: 200 durchgeführte Tests

Test ergab ...	positiv	negativ
Schwangere Frauen	59	1
Nichtschwangere Frauen	15	125

- Richtig positiv: Frau schwanger, Test sagt schwanger

Beispiel: Schwangerschaftstest

Beispieldaten: 200 durchgeführte Tests

Test ergab ...	positiv	negativ
Schwangere Frauen	59	1
Nichtschwangere Frauen	15	125

- Richtig positiv: Frau schwanger, Test sagt schwanger
- Falsch negativ: Frau schwanger, Test sagt nicht schwanger

Beispiel: Schwangerschaftstest

Beispieldaten: 200 durchgeführte Tests

Test ergab ...	positiv	negativ
Schwangere Frauen	59	1
Nichtschwangere Frauen	15	125

- Richtig positiv: Frau schwanger, Test sagt schwanger
- Falsch negativ: Frau schwanger, Test sagt nicht schwanger
- Falsch positiv: Frau nicht schwanger, Test sagt schwanger

Beispiel: Schwangerschaftstest

Beispieldaten: 200 durchgeführte Tests

Test ergab ...	positiv	negativ
Schwangere Frauen	59	1
Nichtschwangere Frauen	15	125

- Richtig positiv: Frau schwanger, Test sagt schwanger
- Falsch negativ: Frau schwanger, Test sagt nicht schwanger
- Falsch positiv: Frau nicht schwanger, Test sagt schwanger
- Richtig negativ: Frau nicht schwanger, Test sagt nicht schwanger

Beispiel: Schwangerschaftstest

Beispieldaten: 200 durchgeführte Tests

Test ergab ...	positiv	negativ
Schwangere Frauen	59	1
Nichtschwangere Frauen	15	125

- Richtig positiv: Frau schwanger, Test sagt schwanger
- Falsch negativ: Frau schwanger, Test sagt nicht schwanger
- Falsch positiv: Frau nicht schwanger, Test sagt schwanger
- Richtig negativ: Frau nicht schwanger, Test sagt nicht schwanger

Wie gut ist der Test?

Maßzahlen

M Gesamtmenge aller zu untersuchenden Objekte:

Recall (Richtig-Positiv-Rate, hit rate)

$$\frac{|\{x \in M \mid P(x) = 1 \wedge K(x) = 1\}|}{|\{x \in M \mid K(x) = 1\}|}$$

D.h. Anteil richtig klassifizierter, positiver Objekte

Maßzahlen

M Gesamtmenge aller zu untersuchenden Objekte:

Recall (Richtig-Positiv-Rate, hit rate)

$$\frac{|\{x \in M \mid P(x) = 1 \wedge K(x) = 1\}|}{|\{x \in M \mid K(x) = 1\}|}$$

D.h. Anteil richtig klassifizierter, positiver Objekte

Beispiel (60 Schwangere, 59 mal positiv)

$$\frac{59}{60} \approx 98,3\%$$

Maßzahlen (2)

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Richtig-Negativ-Rate, correct rejection rate

$$\frac{|\{x \in M \mid P(x) = 0 \wedge K(x) = 0\}|}{|\{x \in M \mid K(x) = 0\}|}$$

D.h. Anteil richtig klassifizierter, negativer Objekte

Maßzahlen (2)

Richtig-Negativ-Rate, correct rejection rate

$$\frac{|\{x \in M \mid P(x) = 0 \wedge K(x) = 0\}|}{|\{x \in M \mid K(x) = 0\}|}$$

D.h. Anteil richtig klassifizierter, negativer Objekte

Beispiel (140 Nicht-Schwangere, 125 mal negativ)

$$\frac{125}{140} \approx 89,3\%$$

Maßzahlen (3)

Precision (Präzision, positiver Vorhersagewert)

$$\frac{|\{x \in M \mid P(x) = 1 \wedge K(x) = 1\}|}{|\{x \in M \mid P(x) = 1\}|}$$

D.h. Anteil der richtigen unten den als scheinbar richtig erkannten

Maßzahlen (3)

Precision (Präzision, positiver Vorhersagewert)

$$\frac{|\{x \in M \mid P(x) = 1 \wedge K(x) = 1\}|}{|\{x \in M \mid P(x) = 1\}|}$$

D.h. Anteil der richtigen unten den als scheinbar richtig erkannten

Beispiel (59 Schwangere richtig erkannt, Test positiv: $59 + 15 = 74$)

$$\frac{59}{74} \approx 79,8\%$$

Maßzahlen (4)

Negative-Vorhersage Rate

$$\frac{|\{x \in M \mid P(x) = 0 \wedge K(x) = 0\}|}{|\{x \in M \mid P(x) = 0\}|}$$

D.h. Anteil der richtig als falsch klassifizierten unter allen als falsch klassifizierten

Maßzahlen (4)

Negative-Vorhersage Rate

$$\frac{|\{x \in M \mid P(x) = 0 \wedge K(x) = 0\}|}{|\{x \in M \mid P(x) = 0\}|}$$

D.h. Anteil der richtig als falsch klassifizierten unter allen als falsch klassifizierten

Beispiel (125 Nicht-Schwangere richtig erkannt, Test negativ: 125 + 1 = 126)

$$\frac{125}{126} \approx 99,2\%$$

Maßzahlen (5)

F-Maß: Harmonisches Mittel aus Precision und Recall:

$$F = 2 \cdot \frac{(\text{precision} \cdot \text{recall})}{(\text{precision} + \text{recall})}$$

Maßzahlen (5)

F-Maß: Harmonisches Mittel aus Precision und Recall:

$$F = 2 \cdot \frac{(\text{precision} \cdot \text{recall})}{(\text{precision} + \text{recall})}$$

Beispiel (Precision = 79,8 % und Recall = 98,3 %)

$$F = 2 \cdot \frac{0,798 \cdot 0,983}{0,798 + 0,983} \approx 2 \cdot \frac{0,784}{1,781} \approx 0,88$$

Weitere Beispiele

Bei seltenen Krankheiten möglich:

- Guter Recall (Anteil der Kranken, die erkannt wurden),
- aber schlechte Präzision (viele false-positives)

Bsp: Klassifikator: Körpertemperatur über 38,5 C $\Longrightarrow$ Gelbfieber.

- In Deutschland haben 10.000 Menschen Fieber mit 38,5 C aber nur 1 Mensch davon hat Gelbfieber
- Recall $= \frac{1}{1} = 1$
- Precision $= \frac{1}{10.000} = 0,0001$
- F-Wert ≈ 0

Fazit: Man muss immer beide Maße betrachten!

Weiteres Vorgehen

Ziel: Finde effizientes Klassifikatorprogramm

Vorher: Kurzer Exkurs zu Wahrscheinlichkeiten und zur Entropie

Exkurs: Wahrscheinlichkeiten, Entropie

Sei X ein Orakel (n-wertige Zufallsvariable)

- X liefert Wert a_i aus $\{a_1, \ldots, a_n\}$
- $p_i =$ Wahrscheinlichkeit, dass X den Wert a_i liefert
- Folge von Orakelausgaben: $b_1, \ldots, b_m$
  Je länger die Folge: Anteil der a_i in der Folge nähert sich p_i
- $\{p_1, \ldots, p_n\}$ nennt man auch
  diskrete Wahrscheinlichkeitsverteilung
  der Menge $\{a_1, \ldots, a_n\}$ bzw. des Orakels X
- Es gilt stets $\sum_i p_i = 1$
- Sind a_i Zahlen, dann ist der
  Erwartungswert $E(X) = \sum_i p_i \cdot a_i$

Exkurs: Wahrscheinlichkeiten, Entropie (2)

Urnenmodell:

X benutzt einen Eimer mit Kugeln beschriftet mit $a_1, \ldots, a_n$ und zieht bei jeder Anfrage zufällig eine Kugel (und legt sie zurück)

Dann gilt:

$$p_i = \textbf{relative Häufigkeit} \text{ von } a_i\text{-Kugeln in der Urne}$$

$$= \frac{a_i\text{-Kugeln in der Urne}}{\text{Anzahl alle Kugeln in der Urne}}$$

Exkurs: Wahrscheinlichkeiten, Entropie (3)

Gegeben: Wahrscheinlichkeitsverteilung p_i, $i = 1, \ldots, n$

Informationsgehalt des Zeichens a_k

$$I(a_k) = \log_2(\frac{1}{p_k}) = -\log_2(p_k) \geq 0$$

Interpretation:

- „Grad der Überraschung beim Ziehen des Symbols a_i“, oder auch:
- „Wie oft muss man das Orakel im Mittel fragen, um a_i zu erhalten“

D.h.

- Selten auftretenden Zeichen: haben hohe Überraschung
- Bei nur einem Zeichen:$p_1 = 1$, $I(p_1) = 0$

Exkurs: Wahrscheinlichkeiten, Entropie (4)

Entropie (Mittlerer Informationsgehalt)

$$I(X) = \sum_{i=1}^{n} p_i * I(a_i) = \sum_{i=1}^{n} p_i * \log_2(\frac{1}{p_i}) = -\sum_{i=1}^{n} p_i * \log_2(p_i) \geq 0$$

entspricht in etwa, der „mittleren Überraschung“

Beispiele

8 Objekte mit gleicher Wahrscheinlichkeit ($p_i = \frac{1}{8}$)

- Informationsgehalt jedes a_i: $\log_2(\frac{1}{\frac{1}{8}}) = \log_2 8 = 3$
- Entropie $\sum_{i=1}^{8} p_i * 3 = \sum_{i=1}^{8} \frac{1}{8} * 3 = 3$

Beispiele

8 Objekte mit gleicher Wahrscheinlichkeit ($p_i = \frac{1}{8}$)

- Informationsgehalt jedes a_i: $\log_2(\frac{1}{\frac{1}{8}}) = \log_2 8 = 3$
- Entropie $\sum_{i=1}^{8} p_i * 3 = \sum_{i=1}^{8} \frac{1}{8} * 3 = 3$

1000 Objekte mit gleicher Wahrscheinlichkeit ($p_i = \frac{1}{1000}$)

- Informationsgehalt jedes a_i: $-\log_2(1/1000) = 9.966$
- Entropie $= 9.996$

Beispiele

8 Objekte mit gleicher Wahrscheinlichkeit ($p_i = \frac{1}{8}$)

- Informationsgehalt jedes a_i: $\log_2(\frac{1}{\frac{1}{8}}) = \log_2 8 = 3$
- Entropie $\sum_{i=1}^{8} p_i * 3 = \sum_{i=1}^{8} \frac{1}{8} * 3 = 3$

1000 Objekte mit gleicher Wahrscheinlichkeit ($p_i = \frac{1}{1000}$)

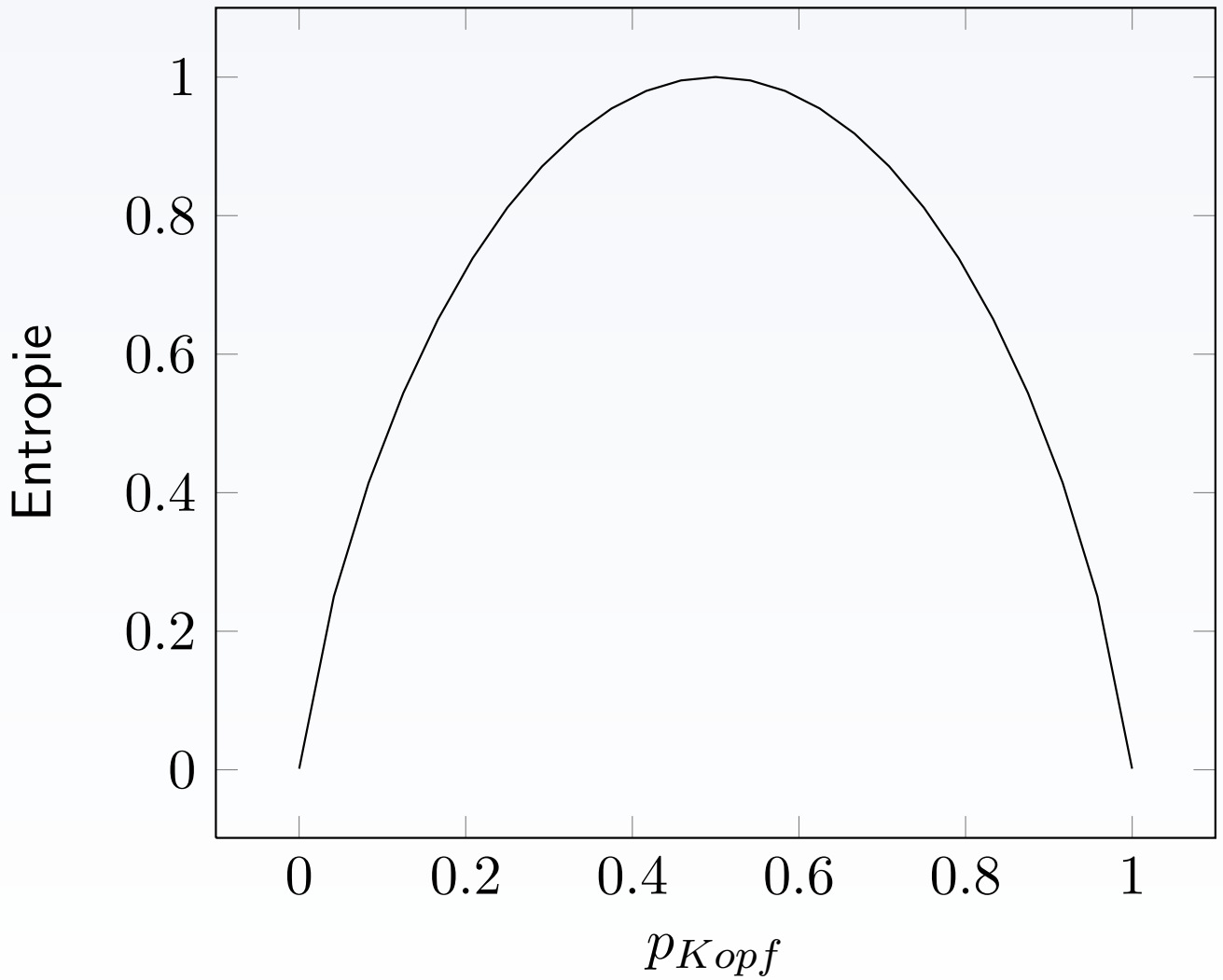
- Informationsgehalt jedes a_i: $-\log_2(1/1000) = 9.966$
- Entropie $= 9.996$

8 Objekte: $p_1 = 0.4994, p_2 = 0.4994, p_i = 0.001$ für $i = 3, \ldots, 8$

- Informationsgehalt a_1, a_2: $-\log_2(0.04994) \approx 1.002$
  a_i: $-\log_2(0.001) \approx 9.966$
- Entropie: $2 * 0.4994 * 1.002 + 6 * 0.001 * 9.996 \approx 1.061$

Beispiele

Bernoulli-Experiment: p_{Kopf} und $p_{Zahl} = 1 - p_{Kopf}$



Entscheidungsbaumlernen

Lernen von Entscheidungsbäumen (1)

Objekt mit Attributen

- Es gibt eine endliche Menge A von Attributen.
- zu jedem Attribut $a \in A$: Menge von möglichen Werten W_a. Wertebereich endlich, oder $\mathbb{R}$.
- Ein Objekt wird beschrieben durch eine Funktion $A \to \times_{a \in A} W_a$.
  Alternativ: Tupel mit $|A|$ Einträgen
- Ein Konzept K ist repräsentiert durch ein Prädikat P_K auf der Menge der Objekte. $P_K \subseteq$ Alle Objekte

Beispiel:

- Alle Objekte: Bücher
- Attribute: (Autor, Titel, Seitenzahl, Preis, Erscheinungsjahr).
- Konzepte „billiges Buch“ (Preis ≤ 10); „umfangreiches Buch“ (Seitenzahl ≥ 500), „altes Buch“ (Erscheinungjahr < 1950)

Entscheidungsbaum

Entscheidungsbaum zu einem Konzept K:

- endlicher Baum
- innere Knoten: Abfragen eines Attributwerts
  - Bei reellwertigen Attributen $a \leq v$ für $v \in \mathbb{R}$. 2 Kinder: Für Ja und Nein
  - Bei diskreten Attributen a mit Werten $w_1, \ldots, w_n$: n Kinder: Für jeden Wert eines
- Blätter: Markiert mit „Ja“ oder „Nein“ (manchmal auch mit „Ja oder Nein“)
- Pro Pfad: Jedes Attribut (außer rellwertige) nur einmal

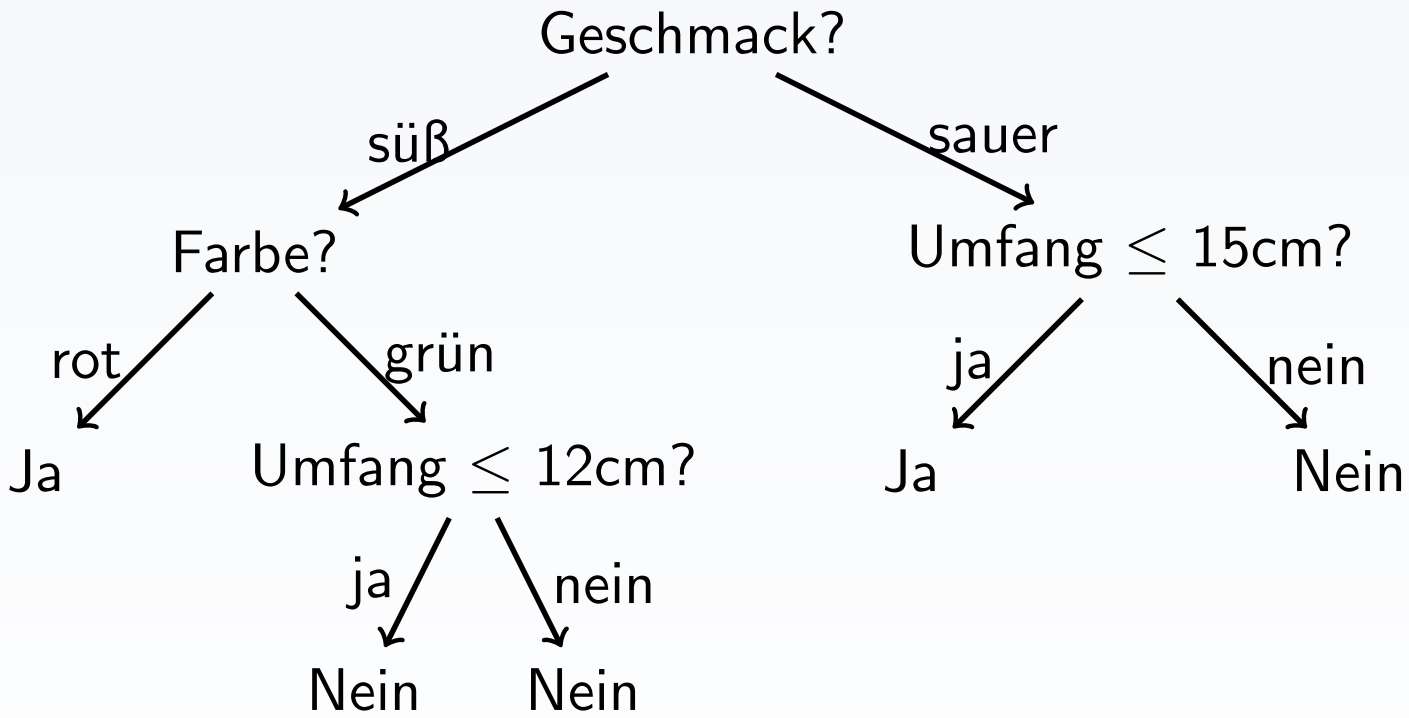
Der Entscheidungsbaum gibt gerade an, ob ein Objekt zum Konzept gehört.

Beispiel

Objekte: Äpfel mit Attributen:

Geschmack (süß/sauer), Farbe (rot/grün), Umfang (in cm)

Konzept: „guter Apfel“



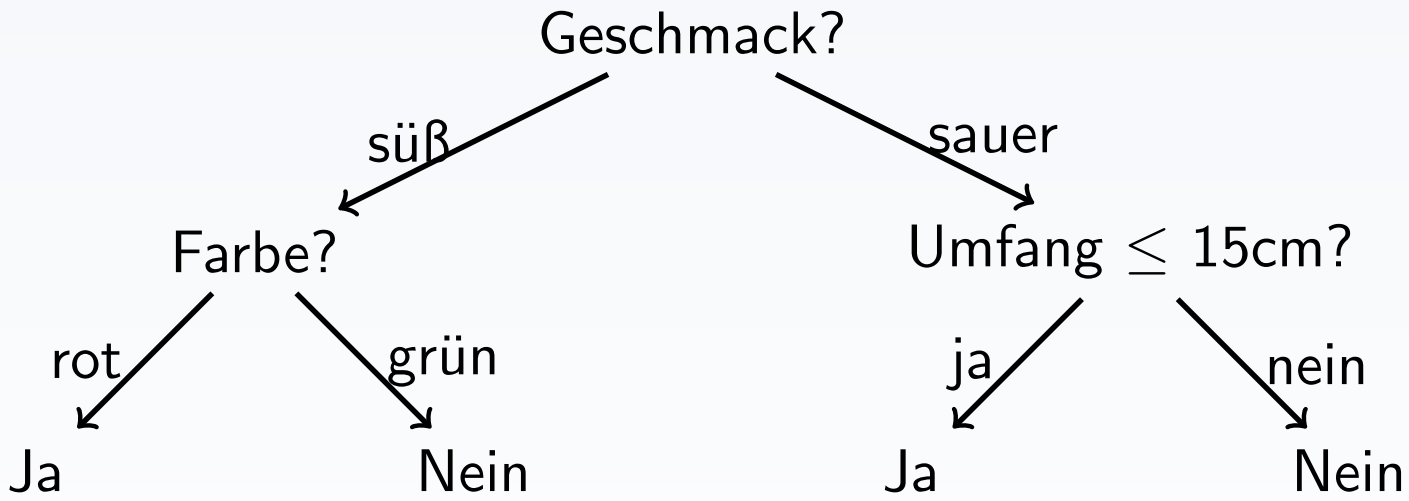
Ist der Baum „optimal“?

Beispiel

Objekte: Äpfel mit Attributen:

Geschmack (süß/sauer), Farbe (rot/grün), Umfang (in cm)

Konzept: „guter Apfel“



Ist der Baum „optimal“?Nein

Entscheidungsbäume (2)

Wofür werden sie benutzt?

→ als Klassifikator für Konzepte

Woher kommt der Baum?

→ Durch Lernen einer Trainingsmenge

Was ist zu beachten?

→ Der Baum sollte möglichst **kurze Pfade** haben

→ Trainingsmenge **muss** positive **und** negative Beispiele beinhalten

Gute Entscheidungsbäume

Ein **guter Entscheidungsbaum** ist ein möglichst kleiner, d.h. der eine **möglichst kleine mittlere Anzahl** von Anfragen bis zur Entscheidung benötigt.

Wir betrachten: Algorithmen zur Konstruktion von guten Entscheidungsbäumen

Ansatz: Verwende die Entropie
(verwandt zur Konstruktion von Huffman-Bäumen)

Vorgehen: Auswahl des nächsten Attributs

Menge M von Objekten mit Attributen geteilt in positive und negative Beispiele. Annahme: Objekte sind gleichverteilt und spiegeln die Wirklichkeit wider.
Sei

- $p =$ Anzahl positiver Beispiele in M
- $n =$ Anzahl negativer Beispiele in M

Entropie der Menge M:

$$I(M) = \frac{p}{p+n} * \log_2(\frac{p+n}{p}) + \frac{n}{p+n} * \log_2(\frac{p+n}{n})$$

Vorgehen: Auswahl des nächsten Attributs (2)

Wie verändert sie die Entropie nach Wahl von Attribut a?

- Sei $m(a)$ der Wert des Attributs a des Objekts $m \in M$.
- Sei a ein Attribut mit Werten $w_1, \ldots, w_k$
- Dann zerlegt das Attribut a die Menge M in Mengen

$$M_i := \{m \in M \mid m(a) = w_i\}$$

- Seien p_i, n_i die Anzahl positiver/negativer Beispiele in M_i.
- Gewichtete Mittelwert des entstandenen Informationsgehalt nach Auswahl des Attributs a

$$I(M|a) = \sum_{i=1}^{k} P(a = w_i) * I(M_i)$$

Vorgehen: Auswahl des nächsten Attributs (2)

Wie verändert sie die Entropie nach Wahl von Attribut a?

- Sei $m(a)$ der Wert des Attributs a des Objekts $m \in M$.
- Sei a ein Attribut mit Werten $w_1, \ldots, w_k$
- Dann zerlegt das Attribut a die Menge M in Mengen

$$M_i := \{m \in M \mid m(a) = w_i\}$$

- Seien p_i, n_i die Anzahl positiver/negativer Beispiele in M_i.
- Gewichtete Mittelwert des entstandenen Informationsgehalt nach Auswahl des Attributs a

$$I(M|a) = \sum_{i=1}^{k} \underbrace{P(a = w_i)}_{\frac{p_i+n_i}{p+n}} * \underbrace{I(M_i)}_{\frac{p_i}{p_i+n_i} * \log_2(\frac{p_i+n_i}{p_i}) + \frac{n_i}{p_i+n_i} * \log_2(\frac{p_i+n_i}{n_i})}$$

Vorgehen: Auswahl des nächsten Attributs (3)

$$I(M|a) = \sum_{i=1}^{k} \frac{p_i + n_i}{p + n} * \left(\frac{p_i}{p_i + n_i} * \log_2(\frac{p_i + n_i}{p_i}) + \frac{n_i}{p_i + n_i} * \log_2(\frac{p_i + n_i}{n_i}) \right)$$

Wähle das Attribut a mit bestem **Informationsgewinn**:
$I(M) - I(M|a)$

Zur Wohldefiniertheit, setzen wir: $\frac{0}{a} * \log_2(\frac{a}{0}) := 0$

Das Verfahren ID3 (Iterative Dichotomiser 3)

Algorithmus **ID3-Verfahren**

Eingabe: Menge M von Objekten mit Attributen

Algorithmus:

Erzeuge Wurzel als offenen Knoten mit Menge M

while es gibt offene Knoten **do**
  wähle offenen Knoten K mit Menge M
    if M enthält nur positive Beispiele **then**
      schließe K mit Markierung „Ja“
    else-if M enthält nur negative Beispiele **then**
      schließe K mit Markierung „Nein“
    else-if $M = \emptyset$ **then**
      schließe K mit Markierung „Ja“ oder „Nein“
    else
      finde Attribut a mit maximalem Informationsgewinn: $I(M) - I(M|a)$
      markiere K mit a und schließe K
      erzeuge Kinder von K:
        Ein Kind pro Attributausprägung w_i von a mit Menge M_i
      Füge Kinder zu den offenen Knoten hinzu
    end-if
end-while

Bemerkungen

- Praktische Verbesserung: Stoppe auch, wenn der Informationsgewinn zu klein
- Jedes diskrete Attribut wird nur einmal pro Pfad abgefragt, da beim zweiten Mal der Informationsgewinn 0 ist
- Wenn man eine Beispielmenge hat, die den ganzen Tupelraum abdeckt, dann wird genau das Konzept gelernt.
- Reellwertige Attribute: Leichte Anpassung möglich.

Beispiel

Äpfel: Geschmack $\in \{\text{süß}, \text{sauer}\}$ und Farbe $\in \{\text{rot}, \text{grün}\}$.

Menge $M = \{(\text{süß}, \text{rot}), (\text{süß}, \text{grün}), (\text{sauer}, \text{rot}), (\text{sauer}, \text{grün})\}$.

Konzept: „guter Apfel“

Positiv: $\{(\text{süß}, \text{rot}), (\text{süß}, \text{grün})\}$
Negativ: $\{(\text{sauer}, \text{rot}), (\text{sauer}, \text{grün})\}$

$p = 2, n = 2 \Rightarrow I(M) = 0.5 \log_2 2 + 0.5 \log_2 2 = 1$

Beispiel (Forts)

Attribut Geschmack:

- $p_{\text{süß}} = 2, n_{\text{süß}} = 0$
- $p_{\text{sauer}} = 0, n_{\text{sauer}} = 2$
- $I(M|\text{Geschmack}) = \frac{2}{4} * (\frac{2}{2} \log \frac{2}{2} + \frac{0}{2} \log \frac{2}{0}) + \frac{2}{4} * (\frac{0}{2} \log \frac{2}{0} + \frac{2}{2} \log \frac{2}{2}) = 0$
- $I(M) - I(M|\text{Geschmack}) = 1$

Attribut Farbe:

- $p_{\text{rot}} = 1, n_{\text{rot}} = 1$
- $p_{\text{grün}} = 1, n_{\text{grün}} = 1$
- $I(M|\text{Farbe}) = \frac{2}{4} * (\frac{1}{2} \log \frac{2}{1} + \frac{1}{2} \log \frac{2}{1}) + \frac{2}{4} * (\frac{1}{2} \log \frac{2}{1} + \frac{1}{2} \log \frac{2}{1}) = 1$
- $I(M) - I(M|\text{Farbe}) = 0$

Beispiel

süß,rot	süß,grün	sauer,rot	sauer,grün
10	20	4	6

Ergibt:

- $I(M) = 0.8112$
- $I(M|\text{Geschmack}) = 0$
- $I(M) - I(M|\text{Geschmack}) = 0.8112$

Attribut Farbe:

- $I(M|\text{Farbe}) = 0.8086$
- $I(M) - I(M|\text{Farbe}) = 0.0026$

Grund: Die Farben sind in positiv / negativ nicht relativ gleich

Beispiel

süß,rot	süß,grün	sauer,rot	sauer,grün
10	20	3	6

- $I(M) = 0.7793$
- $I(M|\text{Geschmack}) = 0$
- $I(M) - I(M|\text{Geschmack}) = 0.7793$

Attribut Farbe:

- $I(M|\text{Farbe}) = 0.7793$
- $I(M) - I(M|\text{Farbe}) = 0$

Beispiel

Äpfel: Geschmack $\in \{\text{süß}, \text{sauer}\}$ und Farbe $\in \{\text{rot}, \text{grün}\}$, Nr $\in \{1, \ldots, 4\}$

Menge $M = \{(\text{süß}, \text{rot}, 1), (\text{süß}, \text{grün}, 2), (\text{sauer}, \text{rot}, 3), (\text{sauer}, \text{grün}, 4)\}$.

Dann:

- $I(M) = 1$
- $I(M|\text{Geschmack}) = 1$
- $I(M|\text{Farbe}) = 0$
- $I(M|\text{Nr}) = 1$

Unfair: Apfelnr ist eindeutig, und stellt implizit mehr Ja/Nein Fragen dar.
Abhilfe: Weglassen der Apfelnr
Allgemein: ID3 bevorzugt Attribute mit vielen Werten
Daher: C4.5 als Anpassung von ID3

Beispiel: Konzept Apfel schmeckt wie er aussieht

Äpfel: Geschmack $\in \{$süß, sauer$\}$ und Farbe $\in \{$rot, grün, gelb$\}$

Menge $M =$ einmal jede Kombination
positiv: (rot,süß), (grün,sauer), (gelb,süß), (gelb,sauer)

- $I(M) = 0.9183$
- $I(M|\text{Farbe} = 0.6666$ und $I(M) - I(M|\text{Farbe}) = 0.2516$
- $I(M|\text{Geschmack}) = 0.9183$ und $I(M) - I(M|\text{Geschmack}) = 0$

Wähle Farbe und dann Geschmack.

Die Variante C4.5

- ID3 bevorzugt Attribute mit vielen Ausprägungen
- C4.5 ändert dies, und normiert daher den Informationsgewinn
- Algorithmus wie ID3 mit einem Unterschied:

$$\text{normierter Informationsgewinn} = (I(M) - I(M|a)) * \text{Normierungsfaktor}$$

Normierungsfaktor für Attribut a Werten $w_i, i = 1, \ldots, k$:

$$\frac{1}{\sum_{i=1}^{k} P(a = w_i) * \log_2(\frac{1}{P(a = w_i)})}$$

Beispiel

Äpfel: Geschmack $\in \{\text{süß}, \text{sauer}\}$ und Farbe $\in \{\text{rot}, \text{grün}\}$, Nr $\in \{1, \ldots, 4\}$

Menge $M = \{(\text{süß}, \text{rot}, 1), (\text{süß}, \text{grün}, 2), (\text{sauer}, \text{rot}, 3), (\text{sauer}, \text{grün}, 4)\}$.

Dann:

- $I(M) = 1$
- $I(M|\text{Geschmack}) = 1$
- $I(M|\text{Farbe}) = 0$
- $I(M|\text{Nr}) = 1$

Normierungsfaktoren:

- Geschmack: $\frac{1}{2/4*\log_2(4/2)+2/4*\log_2(4/2)} = \frac{1}{1} = 1$
- Farbe: $\frac{1}{2/4*\log_2(4/2)+2/4*\log_2(4/2)} = \frac{1}{1} = 1$
- Nr: $\frac{1}{1/4*\log_2(4/1)+1/4*\log_2(4/1)+1/4*\log_2(4/1)+1/4*\log_2(4/1)} = \frac{1}{2}$

Übergeneralisierung

- **Effekt**: Beispiele werden eingeordnet, aber der Entscheidungsbaum unterscheidet zu fein
- **Grund**: Beispiele nicht repräsentativ bzw. ausreichend.

Beispiel: Krankheitsdiagnose: Alle positiven Beispiele sind jünger als 25 oder älter als 30
Übergeneralisierung: Alter zwischen 25 und 30 $\Longrightarrow$ nicht krank.

Übergeneralisierung (2)

Lösung: **Pruning** des Entscheidungbaums

- Stoppe Aufbau des Baums ab einer gewissen Schranke, da alle weiteren Attribute vermutlich irrelevant.
- Blatt-Markierung: Jenachdem welche Beispiele signifikant sind bisher
- Stoppen kann durch statistische Tests gesteuert werden

Verrauschte Daten: Gleiches Verfahren, d.h. Pruning