6. Bayes-Klassifikation

Wahrscheinlichkeitstheorie

- Objekte (Attributwerte, Klasse) werden durch Zufallsprozeß erzeugt
- Optimale Entscheidung durch Schluß von Daten auf Einzelfall

Bayestheorie

1. Hypothesen sind nicht nur wahr oder falsch
2. Hypothesen treffen „weiche" Entscheidungen
3. Jedes Lernbeispiel erhöht/vermindert inkrementell die Hypothesenwahrscheinlichkeiten
4. Vorwissen läßt sich mit den Lerndaten verzahnen
5. Mathematisch abgesicherter Votierungsmechanismus

(Schukat-Talamazzini 2002)

Was sind Bayes-Klassifikatoren?

- Statistische Klassifikatoren
  - Vorhersage der *Class-Membership-Probability* für verschiedene Klassen
  - Beruht auf dem Satz von Bayes
- Verschiedene Verfahren:
  - Naiver Bayes-Klassifikator:
    Relativ einfach zu implementierendes Verfahren, beruhend auf Annahme der Unabhängigkeit zwischen den einzelnen Merkmalen (deshalb naiv)
  - Bayes-Netzwerk (Bayesian Belief Network):
    Mögliche Abhängigkeiten zwischen Merkmalen werden in Form eines Graphen modelliert, der entweder durch den Benutzer vorgegeben wird oder durch das System selbst „gelernt" wird.

- A-Priori-Wahrscheinlichkeiten modellieren Faktenwissen über die Häufigkeit einer Klasse und das Auftreten von Merkmalen, z.B.
  - 20% der Objekte sind Äpfel
  - 30% sind Orangen

  } A-Priori Wahrsch. f. Klassenzugehörigk.
  - 50% der Objekte sind rund
  - 40% haben Farbe orange

  } A-Priori Merkmalshäufigkeit
- Bedingte Wahrscheinlichkeiten („A-Posteriori“) modellieren Zusammenhänge zwischen Klassen und Merkmalen:
  - 100% der Orangen sind rund: $P(\text{rund} \mid \text{Orange}) = 100\%$
  - 100% der Äpfel sind rund: $P(\text{rund} \mid \text{Apfel}) = 100\%$
  - 90% der Orangen sind orange: $P(\text{orange} \mid \text{Orange}) = 90\%$

(Böhm 2003)

Die **bedingte Wahrscheinlichkeit** $P(A|B)$ (A gegeben B, A unter der Bedingung B) wird definiert durch:

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

falls $P(B) \neq 0$ gilt.

$P(A|B)$ ist die Wahrscheinlichkeit für das Eintreten des Ereignisses A, unter der Bedingung, dass das Ereignis B eintritt (eingetreten ist).

Sind A und B unabhängige Ereignisse, gilt $P(A \cap B) = P(A) \cdot P(B)$ und somit $P(A|B) = P(A)$

Beispiel: Zwei Würfel

Σ	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Wahrscheinlichkeit für jeden Eintrag: $1/36$

Σ	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Wahrscheinlichkeit, dass die Summe der beiden Würfel größer als 8 ist, wenn der erste Würfel eine 5 zeigt:

$$P(A|B) = \frac{3/36}{6/36} = 0.5$$

Σ	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Wahrscheinlichkeit, dass die Summe der beiden Würfel größer als 8 ist, wenn beide Würfel die gleiche Zahl zeigen:

$$P(A|B) = \frac{2/36}{6/36} = 1/3$$

(Klawonn 2004)

Der Satz von Bayes:

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

Beweis:

$$\frac{P(E|H) \cdot P(H)}{P(E)} = \frac{\frac{P(E \cap H)}{P(H)} \cdot P(H)}{P(E)} = P(H|E)$$

Interpretation: Die Wahrscheinlichkeit, dass eine Hypothese H zutrifft, wenn ein Ereignis E eintritt, lässt sich aus der Kenntnis der Wahrscheinlichkeiten für

- das Eintreten der Hypothese ganz allgemein,
- das Eintreten des Ereignisses ganz allgemein und
- das Eintreten des Ereignisses, wenn die Hypothese gültig ist,

berechnen.

Beispiel: Seltene Ereignisse

Eine bestimmte Krankheit kommt unter 10000 Menschen im Mittel einmal vor.

Ein Test auf diese Krankheit hat bei bisher 100000 Untersuchungen 1000 mal die falsche Diagnose gestellt.

Es wird daher behauptet, dass der Test zu 99% sicher/korrekt ist.

Jemand geht zu einer Vorsorgeuntersuchung, bei der der Test routinemäßig durchgeführt wird.

Mit welcher Wahrscheinlichkeit hat der Betreffende die Krankheit, wenn das Testergebnis positiv ist?

K : Ein Mensch ist an der Krankheit tatsächlich erkrankt.

$\bar{K}$: Ein Mensch hat die Krankheit nicht.

$T = +$: Der Test liefert ein positives Ergebnis.

$T = -$: Der Test liefert ein negatives Ergebnis.

$$P(K) = 1/10000, \qquad P(\bar{K}) = 9999/10000$$

$$P(T = +|K) = 99/100, \qquad P(T = -|K) = 1/100$$

$$P(T = +|\bar{K}) = 1/100, \qquad P(T = -|\bar{K}) = 99/100$$

gesucht: $P(K|T=+)$

Satz von Bayes:

$$P(K|T=+) = \frac{P(T=+|K) \quad P(K)}{P(T=+)}$$

$$\begin{aligned} P(T=+) &= P(T=+|K) \cdot P(K) + \\ &\quad P(T=+|\bar{K}) \cdot P(\bar{K}) \\ &= \frac{10098}{1000000} \end{aligned}$$

$$P(K|T=+) = \frac{\frac{99}{100} \cdot \frac{1}{10000}}{\frac{10098}{1000000}} = \frac{99}{10098} \approx 1\%$$

Plausibilitätserklärung: Bei 10000 Testpersonen ist eine kranke Person zu erwarten.

Bei der einen erkrankten Person wird der Test (sehr wahrscheinlich) ein positives Ergebnis liefern.

Bei den 9999 gesunden Personen wird der Test bei $99.9 \approx 100$ Personen fälschlicherweise ein positives Ergebnis liefern.

Damit wird es bei 10000 Testpersonen ca. 101 positiv getestete Personen geben, von denen eine tatsächlich krank ist.

Klassifikation mittels des Satzes von Bayes

Prinzip des Bayes-Klassifikators:

Aus den Attributen $X_1, \ldots, X_n$ soll das nominale Attribut H vorhergesagt werden.

Wir definieren den Attributvektor $E = (X_1, \ldots, X_n)$.

Ist h einer der Werte, den H annehmen kann, und liegen die Werte $X_1 = x_1, \ldots, X_n = x_n$ vor, so liefert der Satz von Bayes die Wahrscheinlichkeit dafür, dass die Klasse h bei den gegebenen Attributwerten vorliegt:

$$P(H = h | E = (x_1, \ldots, x_n)) =$$

$$\frac{P(\quad = (x_1, \ldots, x_r) | H = h) \cdot P(H = h)}{P(E = (x_1, \ldots, x_n))}$$

Man berechnet diese Wahrscheinlichkeit für alle Klassen h (alle möglichen Werte des nominalen Attributs H) und entscheidet sich für die wahrscheinlichste Klasse.

Da der Nenner bei der Berechnung für alle Klassen h derselbe ist, spielt er für die Entscheidung für die Klasse keine Rolle, so dass man i.A. nur die Werte

$$P(E = (x_1, \ldots, x_n) | H = h) \cdot P(H = h)$$

berechnet.

(Klawonn 2004)

Allgemeine Definition:

Gegeben ist der Hypothesenraum $\mathcal{H}$ und der Datensatz $\omega \subset \Omega$. Wir bezeichnen

$$h_{\mathrm{MAP}} \stackrel{\mathrm{def}}{=} \operatorname*{argmax}_{h \in \mathcal{H}} \mathrm{P}(h|\omega)$$

als **MAP-Hypothese** *(Maximum a posteriori) aus $\mathcal{H}$ für ω.*

Es gilt offensichtlich $h_{\mathrm{MAP}} = \operatorname{argmax}_{h \in \mathcal{H}}(\mathrm{P}(\omega|h) \cdot \mathrm{P}(h))$

Davon zu unterscheiden ist:

Gegeben ist der Hypothesenraum $\mathcal{H}$ und der Datensatz $\omega \subset \Omega$. Wir bezeichnen

$$h_{\mathrm{ML}} \stackrel{\mathrm{def}}{=} \operatorname*{argmax}_{h \in \mathcal{H}} \mathrm{P}(\omega|h)$$

als **ML-Hypothese** *(Maximum-Likelihood) aus $\mathcal{H}$ für ω. Es heißt $\mathrm{P}(\omega|\cdot)$ auch die „Likelihood-funktion" für ω.*

Es gilt $h_{\mathrm{MAP}} = h_{\mathrm{ML}}$ für die uniforme a priori Verteilung $\mathrm{P}(h) \equiv \mathrm{const}$, welche bei fehlendem Vorwissen anzunehmen ist.
LAPLACE: *„principle of insufficient reason"*

(Schukat-Talamazzini 2002)

Achtung: Begriff "Likelihood" wird nicht immer konsequent im Sinne dieser Definition verwendet!

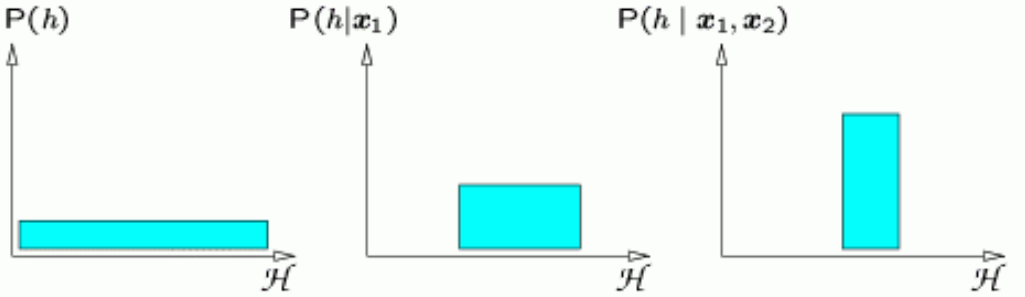
Bedeutung der MAP-Hypothese:

MAP-Hypothese und konsistente Lernverfahren

Ein Lernverfahren heißt *konsistent*, wenn es stets eine Hypothese als Output liefert, die auf den Trainingsdaten *keine* Fehler macht.

Satz:

Es seien Ω, $\mathcal{H}$ gegeben, die Lerndaten ω seien unverrauscht und die a priori Verteilung $P(h)$ der Hypothesen sei uniform. Dann gilt:
Jedes konsistente Lernverfahren produziert eine MAP-Hypothese als Ergebnis.



Entwicklung der a-posteriori-W'keiten P(*h* | *D*) beim Durchlaufen der Trainingsdaten.
Zuerst haben alle Hypothesen die gleiche W'keit; es scheiden diejenigen aus, die mit den neuen Trainingsdaten nicht konsistent sind.

(Schukat-Talamazzini 2002, Mitchell 1997 S. 162)

Prinzip der minimalen Beschreibungslänge:

Durch Logarithmieren erhält man den folgenden Formelausdruck, der interpretationsfähig ist:

Minimum Description Length (MDL)

$$h_{MAP} = \operatorname*{argmax}_{h \in \mathcal{H}} P(\omega|h) \cdot P(h)$$
$$= \operatorname*{argmin}_{h \in \mathcal{H}} \{-\log_2 P(\omega|h) - \log_2 P(h)\}$$

Interpretation:

Satz (Shannon 1949)

Ein Zufallsprozeß erzeuge Zeichenfolgen über dem Alphabet $\{s_1, \ldots, s_L\}$ *mit den Wahrscheinlichkeiten* $q_1, \ldots, q_L$.

Die **optimale** *Codierung dieser Quelle verwendet für jedes Zeichen* s_l *ein Codewort der Länge* $-\log_2 q_l$ *Bit.*

Ihre mittlere Codewortlänge beträgt $\mathcal{H}(q_1, \ldots, q_L)$ *Bit.*

Somit:

$-\log_2 P(h)$ ist die Codewortlänge von *h* unter der optimalen Codierung des Hypothesenraums $\mathcal{H}$

$-\log_2 P(\omega \mid h)$ ist die Codewortlänge der Trainingsdaten ω, wenn die Hypothese h vorausgesetzt wird (d.h. Sender und Empfänger kennen *h*) in optimaler Codierung.

$\Rightarrow$ *MDL-Prinzip:*

h_{MAP} ist diejenige Hypothese *h*, die die Summe aus der Codierungslänge der Hypothesen und der Codierungslänge der Trainingsdaten unter der Voraussetzung der entsprechenden Hypothese minimiert.

$h_{MAP} = h_{MDL} = \text{argmin}_h \{ \text{Länge}_C(h) + \text{Länge}_{C'}(\omega \mid h) \}$

= "kürzeste Erklärung" der Trainingsdaten

(vgl. Mitchell 1997 S. 172 ff.)

Bedeutung: Tradeoff zwischen Komplexität der Hypothese und Zahl der Fehler bei ihrer Anwendung auf die Trainingsdaten.

optimale Bayes-Klassifikation

unterscheide die Fragen:

- was ist die wahrscheinlichste Hypothese aus $\mathcal{H}$, die die Trainingsdaten erklärt?
- was ist die wahrscheinlichste Klassifikation einer neuen Instanz, wenn die Trainingsdaten bekannt sind?

auf eine neue Instanz stets die MAP-Hypothese anzuwenden, muss nicht immer die richtige Antwort auf die 2. Frage sein!

Beispiel:

3 Hypothesen *g*, *h*, *i*
$P(g \mid \omega) = 0{,}4$
$P(h \mid \omega) = 0{,}3$
$P(i \mid \omega) = 0{,}3$
$\Rightarrow$ *g* ist die MAP-Hypothese

x sei neue Instanz:
g klassiert *x* als positiv
h und *i* klassieren *x* als negativ
$\Rightarrow P(x \text{ positiv}) = 0{,}4$
$P(x \text{ negativ}) = 0{,}6$
$\Rightarrow$ wahrscheinlichste Klassifikation ("negativ") entspricht nicht dem Ergebnis der MAP-Hypothese!

wahrscheinlichste Klassifikation:
kombiniere die Vorhersagen aller Hypothesen, gewichtet mit den a-posteriori-Wahrscheinlichkeiten der Hypothesen!

Optimaler Bayes-Klassifikator:

Für ein K-Klassenproblem über Ω mit Hypothesenraum $\mathcal{H}$ mit den **a posteriori** *Klassenwahrscheinlichkeiten*

$$\mathrm{P}(x \in \Omega_\kappa \mid \omega) = \sum_{h \in \mathcal{H}} \mathrm{P}(x \in \Omega_\kappa \mid h) \cdot \mathrm{P}(h|\omega)$$

von x bei Vorliegen der Lerndaten ω heißt

$$\delta_{\text{Bayes}}(x) = \operatorname*{argmax}_{\kappa} \sum_{h \in \mathcal{H}} \mathrm{P}(x \in \Omega_\kappa \mid h) \cdot \mathrm{P}(h|\omega)$$

die **Bayes-Entscheidungsregel** *für das Datum x.*

Satz:

Sind Datenraum Ω, Hypothesenraum $\mathcal{H}$, Lernmaterial ω und a priori Verteilung $\mathrm{P} : \mathcal{H} \to \mathbb{R}$ fest vorgegeben, so realisiert die Bayes-Entscheidungsregel denjenigen Klassifikator mit minimaler asymptotischer Fehlerrate.

Nachteil: hoher Aufwand (a-posteriori-W'keiten für alle Hypothesen müssen berechnet werden!)

deshalb oft Vereinfachung mit einem Monte-Carlo-Ansatz:

Gibbs' Auswürfelverfahren

GEGEBEN: $\omega, x \in \Omega$

1. *Wähle zufällig ein $h^* \in \mathcal{H}$ nach der a posteriori Verteilung $\mathrm{P}(\cdot|\omega)$*
2. *Setze $\delta_{\text{Gibbs}}(x) = h^*(x)$*

(Schukat-Talamazzini 2002)

Man kann beweisen: Der Erwartungswert des Fehlers ist beim Gibbs-Verfahren im ungünstigsten Fall nur doppelt so groß wie beim optimalen Bayes-Klassifikator.

Dennoch gibt es auch hier Nachteile (alle Hypothesen müssen verwaltet werden; Nichtdeterminismus).
Deshalb oft doch Verwendung nur der MAP-Hypothese ("brute-force Bayes-Lernen").

Welche Werte nimmt man für die Wahrscheinlichkeiten in der Formel?

Die Wahrscheinlichkeit $P(H = h)$ lässt sich aus gegebenen Daten sehr einfach schätzen:

$$P(H = h) = \frac{\text{Anzahl der Daten aus Klasse } h}{\text{Anzahl der Daten}}$$

Theoretisch ließe sich die Wahrscheinlichkeit $P(E = (x_1, \ldots, x_n)|H = h)$ analog berechnen:

$$P(E = (x_1, \ldots, x_n)|H = h) =$$

$$\frac{\text{Anzahl der Daten aus Klasse } h \text{ mit Werten } (x_1, \ldots, x_n)}{\text{Anzahl der Daten aus Klasse } h}$$

Dies würde aber bedeuten, dass bei $n = 10$ nominalen Attributen $X_1, \ldots, X_{10}$ mit jeweils drei Werten mindestens $3^{10} = 59049$ Daten vorhanden sein müssten, um allein alle möglichen Attributwertkombinationen abzudecken.

Man geht daher von der (naiven, unrealistischen) Annahme aus, dass die Attribute $X_1, \ldots, X_n$ (bei vorgegebener Klasse jeweils) unabhängig sind, d.h.

$$P(E = (x_1, \ldots, x_n)|H = h) =$$

$$P(X_1 = x_1|H = h) \cdot \ldots \cdot P(X_n = x_n|H = h)$$

$P(X_i = x_i | H = h)$ lässt sich wiederum einfach berechnen/schätzen:

$$P(X_i = x_i | H = h) =$$

$$\frac{\text{Anzahl der Daten aus Klasse } h \text{ mit } X_i = x_i}{\text{Anzahl der Daten aus Klasse } h}$$

(Klawonn 2004)

Somit:

Der "naive Bayes-Klassifikator"

Motivation

Bei hochdimensionalen Merkmalsvektoren schwierige Schätzung der bedingten Wahrscheinlichkeiten $P(M \mid C)$ und damit $P(C \mid M)$:

- M besteht aus vielen einzelnen Komponenten, die UND-verknüpft sind:

$$P(C \mid M_1 \wedge M_2 \wedge \ldots) = \frac{P(M_1 \wedge M_2 \wedge \ldots \mid C) \cdot P(C)}{P(M_1 \wedge M_2 \wedge \ldots)}$$

- Bei d verschiedenen Merkmalen und jeweils r verschiedenen Werten ergeben sich r^d verschiedene Merkmalskombinationen

Probleme:

- Die Wahrscheinlichkeiten lassen sich nicht mehr abspeichern
- Man bräuchte $>> r^d$ Trainingsdatensätze, um die Wahrscheinlichkeit der einzelnen Merkmalskombinationen überhaupt ermitteln zu können

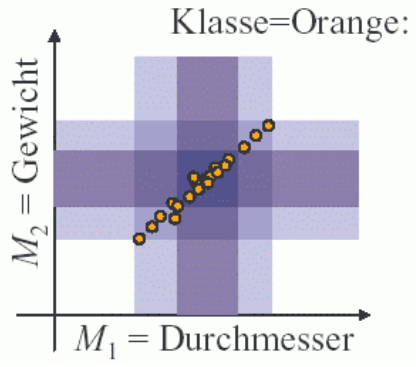
Lösung dieses Problems beim naiven Bayes-Klassifikator:

Annahme der Bedingten Unabhängigkeit

d.h. bei jeder einzelnen Klasse werden die Merkmale so behandelt als wären sie voneinander statistisch unabhängig:

$$P(M_1 \wedge M_2 \mid C) = P(M_1 \mid C) \cdot P(M_2 \mid C)$$

Was bedeutet dies?



- Annahme kann falsch sein
- Dies führt *nicht* unbedingt dazu, dass die Klassifikation versagt
- Aber schlechte Leistung, wenn…
  - alle Merkmale bei mehreren Klassen etwa gleich verteilt sind
  - Unterschiede nur in „Relationen" der Merkmale zueinander

(Böhm 2003)

„Naiver" Bayesklassifikator

… entsteht aus der MAP-Regel

$$\delta_{\text{MAP}}(\boldsymbol{x}) = \underset{\lambda}{\operatorname{argmax}} \{P(\boldsymbol{x}|\lambda) \cdot P(\lambda)\}$$

durch Postulation bedingter Unabhängigkeiten über Ω

formale Def.:

Es sei $\Omega = \mathcal{X}_1 \times \ldots \times \mathcal{X}_N$ und ein K-Klassenproblem gegeben. Die Entscheidungsregel

$$\delta_{\text{NB}}(\boldsymbol{x}) = \underset{\lambda}{\operatorname{argmax}} \left\{ P(\lambda) \cdot \prod_{n=1}^{N} P(x_n|\lambda) \right\}$$

heißt **naiver Bayesklassifikator**.

(Schukat-Talamazzini 2002)

In Worten:

Der Datensatz enthalte ausschließlich diskrete Attribute.
Aus den Werten x_1, ..., x_n der Attribute X_1, ..., X_n soll der Wert des Zielattributs H vorhergesagt werden.

Für jede Klasse (jeden Wert von *H*) wird die *a-posteriori*-Wahrscheinlichkeit

$P(H=h \mid X_1 = x_1, ..., X_n = x_n) =$
$P(X_1 = x_1 \mid H = h) \cdot ... \cdot P(X_n = x_n \mid H = h) \cdot P(H = h)$

unter der Annahme der stochastischen Unabhängigkeit von X_1, ..., X_n geschätzt.

$(x_1, \ldots, x_n)$ wird der Klasse h mit der größten Wahrscheinlichkeit $P(H = h | X_1 = x_1, \ldots, X_n = x_n)$ zugeordnet.

Dieser Klassifikator wird als *naiv* bezeichnet, da die Unabhängigkeit der Attribute $X_1, \ldots, X_n$ angenommen wird.

Auch wenn diese Annahme in der Realität selten zutrifft, so liefert der naive Bayes-Klassifikator in der Praxis häufig trotzdem recht gute Ergebnisse, wenn nicht zu viele Attribute stark korreliert sind.

Beispiel:

Es sei der folgende Datensatz mit den Attributen Größe (klein (k), mittel (m), groß (g)), Gewicht (leicht (l), normal (n), schwer (s)), lange Haare (ja (j), nein (n)) und Geschlecht (weiblich (w), männlich (m)) gegeben.

Das Geschlecht soll anhand der Attribute Größe, Gewicht und lange Haare vorhergesagt werden.

Nr.	Größe	Gewicht	lange Haare	Geschlecht
1	m	n	n	m
2	k	l	j	w
3	g	s	n	m
4	k	n	j	w
5	g	n	j	w
6	k	l	n	w
7	k	s	n	m
8	m	n	n	w
9	m	l	j	w
10	g	n	n	m

Wie klassifiziert ein naiver Bayes-Klassifikator das Tupel (g, l, j)?

Dazu müssen wir

$$P(\text{Geschlecht} = m \mid \text{Größe} = g, \text{Gewicht} = l, \text{lange\_Haare} = j)$$

$$= P(\text{Größe} = g \mid \text{Geschlecht} = m) \cdot P(\text{Gewicht} = l \mid \text{Geschlecht} = m) \cdot P(\text{lange\_Haare} = j \mid \text{Geschlecht} = m) \cdot P(\text{Geschlecht} = m)$$

und

$$P(\text{Geschlecht} = w \mid \text{Größe} = g, \text{Gewicht} = l, \text{lange\_Haare} = j)$$

$$= P(\text{Größe} = g \mid \text{Geschlecht} = w) \cdot P(\text{Gewicht} = l \mid \text{Geschlecht} = w) \cdot P(\text{lange\_Haare} = j \mid \text{Geschlecht} = w) \cdot P(\text{Geschlecht} = w)$$

berechnen.

$P(\text{Größe} = g \mid \text{Geschlecht} = m) = 2/4 = 1/2$

Nr.	Größe	Gewicht	lange Haare	Geschlecht
1	m	n	n	m
2	k	l	j	w
3	g	s	n	m
4	k	n	j	w
5	g	n	j	w
6	k	l	n	w
7	k	s	n	m
8	m	n	n	w
9	m	l	j	w
10	g	n	n	m

(... usw. ...)

$\Rightarrow$

$$P(\text{Geschlecht} = w \mid \text{Größe} = g, \text{Gewicht} = l, \text{lange\_Haare} = j)$$

$$= \frac{1}{6} \cdot \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{5} = 1/30$$

$$> 0 =$$

$$P(\text{Geschlecht} = w \mid \text{Größe} = g, \text{Gewicht} = l, \text{lange\_Haare} = j)$$

Klassifikation von (g, l, j): weiblich (w)

Das (Eingabe-)Tupel (g, l, j) wurde durch den naiven Bayes-Klassifikator klassifiziert.

Im Datensatz kommt kein Objekt mit diesen Attributwerten vor.

Ein nicht-naiver Bayes-Klassifikator würde für dieses Tupel keine Klassifikation liefern.

Problem der Null-Wahrscheinlichkeiten:

Im Beispiel war zu sehen, dass der gesamte a-posteriori-Wahrscheinlichkeitswert für eine Klasse Null wird, wenn ein Attributwert in Verbindung mit der entspr. Klasse nie auftritt.

Beispiel: *Gewicht*= s für die Klasse w oder *lange_Haare*= j für die Klasse m.

Eine übliche Heuristik, dieses Problem zu umgehen, besteht in der **Laplace-Schätzung** oder einer Verallgemeinerung davon.

Bei der Laplace-Schätzung fängt man bei 1 statt bei 0 an zu zählen.

Laplace-Schätzung für
$P(\text{Größe} = \ldots | \text{Geschlecht} = m)$

Größe	Anzahl	Laplace	P	P_{Laplace}
k	1	2	1/4	2/7
m	1	2	1/4	2/7
g	2	3	2/4	3/7

Bei der verallgemeinerten Laplace-Schätzung fängt man nicht bei 1 sondern bei einem beliebigen Wert $\varepsilon \in \mathbb{R}^+$ an zu zählen.

weiteres Problem unseres Ansatzes:
Die Schätzung der a-posteriori-Wahrscheinlichkeiten beim naiven Bayes-Klassifikator basiert auf dem Abzählen der Häufigkeit von Werten.
Voraussetzung dafür: alle Attribute können nur endlich viele Werte annehmen.

Möglichkeiten bei Attributen mit kontinuierlichem Wertebereich:

Eine Möglichkeit besteht darin, das kontinuierliche Attribut zu diskretisieren.

Dazu wird der Wertebereich des kontinuierlichen Attributs in k disjunkte Intervalle aufgeteilt und so zu einem diskreten (ordinalen) Attribut mit k Werten umgewandelt.

Ein kontinuierlicher Wert wird durch die jeweilige Nummer des Intervalls ersetzt, in dem er sich befindet.

Meistens werden die Intervalle so gewählt, dass sie jeweils ungefähr gleich viele Daten haben.

Alternativ kann man auch eine Verteilungsannahme für die kontinuierlichen Werte treffen, dass sie beispielsweise einer Normalverteilung mit der Dichtefunktion

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

folgen.

Die Parameter der Verteilungs- bzw. Dichtefunktion werden aus den Daten geschätzt

Im Falle einer Normalverteilungsannahme ist μ der Mittelwert des entsprechenden Attributs bei gegebener Klasse und σ^2 die empirische Varianz.

Bei der Schätzung der a-posteriori-Wahrscheinlichkeiten werden die relativen Häufigkeiten dann durch die entsprechenden Werte der Dichtefunktion ersetzt.

Statt

$$P(X = x|H = h)$$

verwendet man den Wert

$$f(x)$$

der Dichtefunktion des Attributs X bei gegebener Klasse $H = h$.

Missing Values

Fehlende Werte einzelner Attribute stellen für den naiven Bayes-Klassifikator kein Problem dar.

Fehlen einzelne Werte im Lerndatensatz, werden sie bei den entsprechenden Häufigkeiten nicht mitgezählt

Fehlen einzelne Werte bei einem neu zu klassifizierenden Objekt, werden die entsprechenden Attribute bei der Berechnung einfach weggelassen.

weitere Ergänzung des Verfahrens:

Gibt man bei der Klassifikation die bedingten Wahrscheinlichkeiten für *alle* Klassen mit an, anstatt nur die Klasse mit der höchsten a-posteriori-Wahrscheinlichkeit auszugeben, so erhält man eine Zusatzinformation über die Sicherheit der Klassifikation:

- ist die W'keit für die Klasse mit der größten W'keit deutlich größer als alle anderen $\Rightarrow$ Klassifikation ziemlich eindeutig
- haben mehrere Klassen relativ hohe W'keiten, so weiß man, welche alternativen Klassen für das Objekt ebenfalls in Frage kommen.

Stark korrelierte Attribute

Bei vielen stark korrelierten Attributen versagt der naive Bayes-Klassifikator im Allgemeinen.

Mittels Dimensionsreduktionsverfahren (z.B. Hauptkomponentenanalyse, Variablenselektion) lässt sich dieses Problem verringern.

Alternativ können auch stark korrelierte Attribute zu einem komplexeren Attribut zusammengefasst werden.

Beispielsweise könnten das Attribut *Tageszeit* mit den Werten

Tag (t) und *Nacht* (n)

und das Attribut *Licht* mit den Werten

an (+) und *aus* (-)

zu einem Attribut mit den vier Werten

(t,+), (t,-), (n,+), (n,-)

zusammengefasst werden.

(Klawonn 2004)

Beispiel für Anwendung des naiven Bayes-Klassifikators: Klassifikation von Texten

(Böhm 2003)

- Anwendungen (z.B. [Craven et al. 1999], [Chakrabarti, Dom & Indyk 1998])
  - Filterung von Emails
  - Klassifikation von Webseiten
- Vokabular $T = \{t_1, \ldots, t_d\}$ von relevanten Termen
- Repräsentation eines Textdokuments $o = (o_1, \ldots, o_d)$
- o_i: Häufigkeit des Auftretens von t_i in o
- Methode
  - Auswahl der relevanten Terme
  - Berechnung der Termhäufigkeiten
  - Konstruktion des Modells
  - Anwendung des Modells zur Klassifikation neuer Dokumente

Auswahl der Terme

- Reduktion der auftretenden Worte auf Grundformen
  - Stemming
  - Abhängigkeit von der Sprache der Texte
- Einwort- oder Mehrwort-Terme?
- Elimination von Stoppwörtern
- weitere Reduktion der Anzahl der Terme

bis zu 100 000 Terme

Reduktion der Anzahl der Terme

- optimaler Ansatz

  $O(2^{\text{AnzahlTerme}})$ Teilmengen

  optimale Teilmenge läßt sich nicht effizient bestimmen

- Greedy-Ansatz

  bewerte jeden Terms einzeln

  welchen „Informationsgewinn“ liefert er in Bezug auf die Separation der gegebenen Klassen?

  sortiere die Terme nach dieser Maßzahl absteigend

  wähle die ersten d Terme als Attribute aus

Konstruktion des Klassifikators

- Anwendung des naiven Bayes-Klassifikators

aber: Häufigkeiten der verschiedenen Terme typischerweise korreliert

- wichtigste Aufgabe: Schätzung der $P(o_i|\, c)$ aus den Trainingsdokumenten

vermeide $P(o_i|\, c) = 0$

Glättung der beobachteten Häufigkeiten

In typischen Anwendungsbeispielen (z.B. Klassifikation von Webseiten von Informatik-Instituten) können mit dieser Methode Klassifikationsgenauigkeiten von 70-80 % für die meisten Klassen erreicht werden.

Nachteil des naiven Bayes-Klassifikators:
Stochastische Unabhängigkeitsvoraussetzung ist oft für spezielle Attribute verletzt

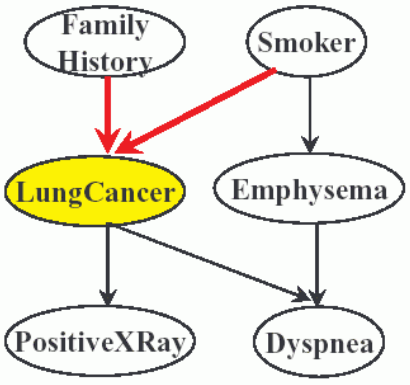
⇒ Berücksichtigung von einzelnen Abhängigkeiten in einem Graphen

Bayes-Netzwerke (*Bayesian belief networks*)

Grundbegriffe

- Graph mit Knoten = *Zufallsvariable* und Kante = *bedingte Abhängigkeit*
- Jede Zufallsvariable ist bei gegebenen Werten für die Vorgänger-Variablen bedingt unabhängig von allen Zufallsvariablen, die keine Nachfolger sind.
- Für jeden Knoten (Zufallsvariable): Tabelle der bedingten Wahrscheinlichkeiten
- Trainieren eines Bayes-Netzwerkes
  - bei gegebener Netzwerk-Struktur und allen bekannten Zufallsvariablen
  - bei gegebener Netzwerk-Struktur und teilweise unbekannten Zufallsvariablen
  - bei apriori unbekannter Netzwerk-Struktur

Beispiel



	FH,S	FH, ¬S	¬FH,S	¬FH, ¬S
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

bedingte Wahrscheinlichkeiten für LungCancer

bei gegebenen Werten für FamilyHistory und Smoker liefert der Wert für Emhysema keine zusätzliche Information über LungCancer

(Böhm 2003)

Bayes-Netzwerke sind aktives Forschungsgebiet

Bayes-Klassifikationsverfahren:

Diskussion

+ hohe Klassifikationsgenauigkeit in vielen Anwendungen
+ Inkrementalität
  Klassifikator kann einfach an neue Trainingsobjekte adaptiert werden
+ Einbezug von Anwendungswissen

- Anwendbarkeit
  die erforderlichen bedingten Wahrscheinlichkeiten sind oft unbekannt
- Ineffizienz
  bei sehr vielen Attributen
  insbesondere Bayes-Netzwerke

(Böhm 2003)