

Lernen nach Bayes

Prof. Dr.-Ing. J. Marius Zöllner

Prof. Dr.-Ing. Rüdiger Dillmann

Dipl. Inform. Michael Weber



Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft



Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825

- Motivation
- Theorem von Bayes
- MAP- / ML-Hypothesen
- Optimaler Bayes-Klassifikator
- Naiver Bayes-Klassifikator
- Beispiel: Klassifikation von Texten
- Bayes'sche Netze
- Der EM-Algorithmus
- Zusammenfassung

Was ist Lernen nach Bayes?

Statistische Lernverfahren:

- Kombinieren vorhandenes Wissen (a priori Wahrscheinlichkeiten) mit beobachteten Daten
- Hypothesen können mit einer Wahrscheinlichkeit angegeben werden.
- Jedes Beispiel kann die Glaubwürdigkeit einer bestehenden Hypothese erhöhen oder verringern:
→ kein Ausschluss bestehender Hypothesen
- Mehrere mögliche Hypothesen können gemeinsam ausgewertet werden, um genauere Ergebnisse zu erzielen.

Warum Lernen nach Bayes?

Erfolgreiche Lernverfahren:

- Naiver Bayes-Klassifikator
- Bayes'sche Netze

Analyse anderer Lernverfahren:

- „Gold-Standard“ für die Beurteilung von (nicht statistischen) Lernverfahren

Praktische Probleme:

- Initiales Wissen über viele Wahrscheinlichkeiten notwendig
 - Aber: Oft Schätzung basierend auf Hintergrundwissen, vorhandenen Daten, etc. möglich
- Erheblicher Rechenaufwand für optimale Bayes'sche Hypothese im allgemeinen Fall
 - Linear mit Anzahl der möglichen Hypothesen
 - Aber: In speziellen Fällen deutliche Reduzierung des Rechenaufwands möglich

Produktregel: Konjunktion zweier Ereignisse A und B

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

Summenregel: Disjunktion zweier Ereignisse A und B

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

Theorem der totalen Wahrscheinlichkeit: Für sich gegenseitig ausschließende Ereignisse A_1, \dots, A_n mit $\sum_{i=1}^n P(A_i) = 1$ gilt

$$P(B) = \sum_{i=1}^n P(B|A_i)P(A_i)$$

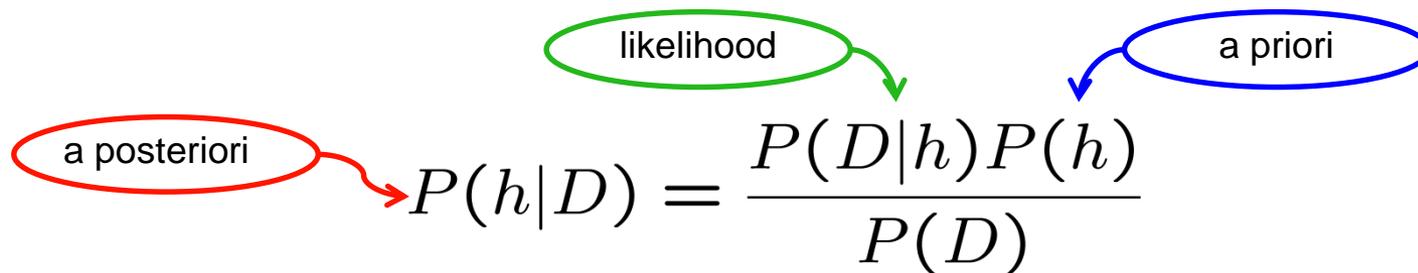
Theorem von Bayes

$P(h)$ Wahrscheinlichkeit, dass h aus H gültig ist (a priori, d.h. vor Beobachtung von D).

$P(D)$ Wahrscheinlichkeit, dass D als Ereignisdatensatz auftritt (ohne Wissen über gültige Hypothese).

$P(D|h)$ Wahrscheinlichkeit des Auftretens von D in einer Welt, in der h gilt.

$P(h|D)$ Wahrscheinlichkeit, dass h wahr ist gegeben die beobachteten Daten D (a posteriori).


$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Bayes-Theorem Herleitung

$$P(A \wedge B) = P(A \wedge B)$$

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

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

Beispiel: Medizinische Diagnose

■ Gegeben:

- 0.8% der Bevölkerung leiden an Krebs.
- Hat man Krebs, fällt der Test in 98% der Fälle positiv aus.
- Hat man keinen Krebs, erzeugt der Test dennoch fälschlicherweise in 3% der Fälle ein positives Ergebnis.

■ Gesucht:

- Wahrscheinlichkeit mit der eine Person deren Test positiv ausgefallen ist tatsächlich Krebs hat?

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

$$P(B) = \sum_{i=1}^n P(B|A_i)P(A_i) \quad P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Beispiel: Medizinische Diagnose

Vorwissen über spezielle Krebserkrankung / Labortest:

$$P(\text{Krebs}) = 0.008 \quad P(\neg\text{Krebs}) = 0.992$$

$$P(\oplus|\text{Krebs}) = 0.98 \quad P(\ominus|\text{Krebs}) = 0.02$$

$$P(\oplus|\neg\text{Krebs}) = 0.03 \quad P(\ominus|\neg\text{Krebs}) = 0.97$$

Gesucht: $P(\text{Krebs}|\oplus)$

$$\begin{aligned} P(\text{Krebs}|\oplus) &= \frac{P(\oplus|\text{Krebs})P(\text{Krebs})}{P(\oplus)} \\ &= \frac{P(\oplus|\text{Krebs})P(\text{Krebs})}{P(\oplus|\text{Krebs})P(\text{Krebs}) + P(\oplus|\neg\text{Krebs})P(\neg\text{Krebs})} \\ &= 0.98 \cdot 0.008 / (0.98 \cdot 0.008 + 0.03 \cdot 0.992) \\ &= 0.21 \end{aligned}$$

Auswahl von Hypothesen

Ziel: Finden der Hypothese h aus H mit der größten Wahrscheinlichkeit gegeben die beobachteten Daten D . Dies ist die Maximum a posteriori (MAP) Hypothese

$$\begin{aligned}h_{MAP} &= \arg \max_{h \in H} P(h|D) \\ &= \arg \max_{h \in H} \frac{P(D|h)P(h)}{P(D)} && \text{Bayes} \\ &= \arg \max_{h \in H} P(D|h)P(h) && P(D) = \text{const.}\end{aligned}$$

Unter der Annahme $P(h_i) = P(h_j)$ lässt sich diese zur Maximum Likelihood (ML) Hypothese vereinfachen:

$$h_{ML} = \arg \max_{h_i \in H} P(D|h_i)$$

Beispiel: Medizinische Diagnose

$$h_{MAP} = \arg \max_{h \in H} P(D|h)P(h)$$

Vorwissen über spezielle Krebserkrankung / Labortest:

$$P(\text{Krebs}) = 0.008 \quad P(\neg\text{Krebs}) = 0.992$$

$$P(\oplus|\text{Krebs}) = 0.98 \quad P(\ominus|\text{Krebs}) = 0.02$$

$$P(\oplus|\neg\text{Krebs}) = 0.03 \quad P(\ominus|\neg\text{Krebs}) = 0.97$$

Beobachtung: neuer Patient, Labortest \oplus

$$P(\oplus|\text{Krebs})P(\text{Krebs}) = 0.98 \cdot 0.008 = 0.0078$$

$$P(\oplus|\neg\text{Krebs})P(\neg\text{Krebs}) = 0.03 \cdot 0.992 = 0.0298$$

Brute Force Lernen von MAP-Hypothesen

1. Berechne für jede Hypothese $h \in H$ die a posteriori Wahrscheinlichkeit:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

2. Gib die Hypothese h_{MAP} mit der größten a posteriori Wahrscheinlichkeit aus:

$$h_{MAP} = \arg \max_{h \in H} P(D|h)P(h)$$

Problemstellung:

- Endlicher Hypothesenraum H über dem Raum der Instanzen X .
- Aufgabe: Lernen eines Zielkonzepts: $c : X \rightarrow \{0, 1\}$
- Feste Sequenz von Instanzen: $\langle x_1, \dots, x_m \rangle$
- Sequenz der Zielwerte: $D = \langle d_1, \dots, d_m \rangle$

Annahmen:

- Trainingsdaten D sind nicht verrauscht (d.h. $d_i = c(x_i)$)
- Zielkonzept c ist in H enthalten
- Kein Grund a priori anzunehmen, dass irgendeine Hypothese wahrscheinlicher ist als eine andere

$$\text{Vorwissen: } P(D|h) = \begin{cases} 1 & \text{falls } h(x_i) = d_i, \forall d_i \in D \\ 0 & \text{sonst} \end{cases}$$
$$P(h) = \frac{1}{|H|}$$

Berechnung der a posteriori Wahrscheinlichkeit:

1. Fall (konsistente Hypothesen)

$$P(h | D) = \frac{1 \cdot \frac{1}{|H|}}{P(D)} = \frac{1 \cdot \frac{1}{|H|}}{\sum_{h\text{-konsistent}} P(D | h)P(h)} = \frac{1 \cdot \frac{1}{|H|}}{\frac{|VS_{H,D}|}{|H|}} = \frac{1}{|VS_{H,D}|}$$

2. Fall (sonst):

$$P(h|D) = \frac{0 \cdot P(h)}{P(D)} = 0$$

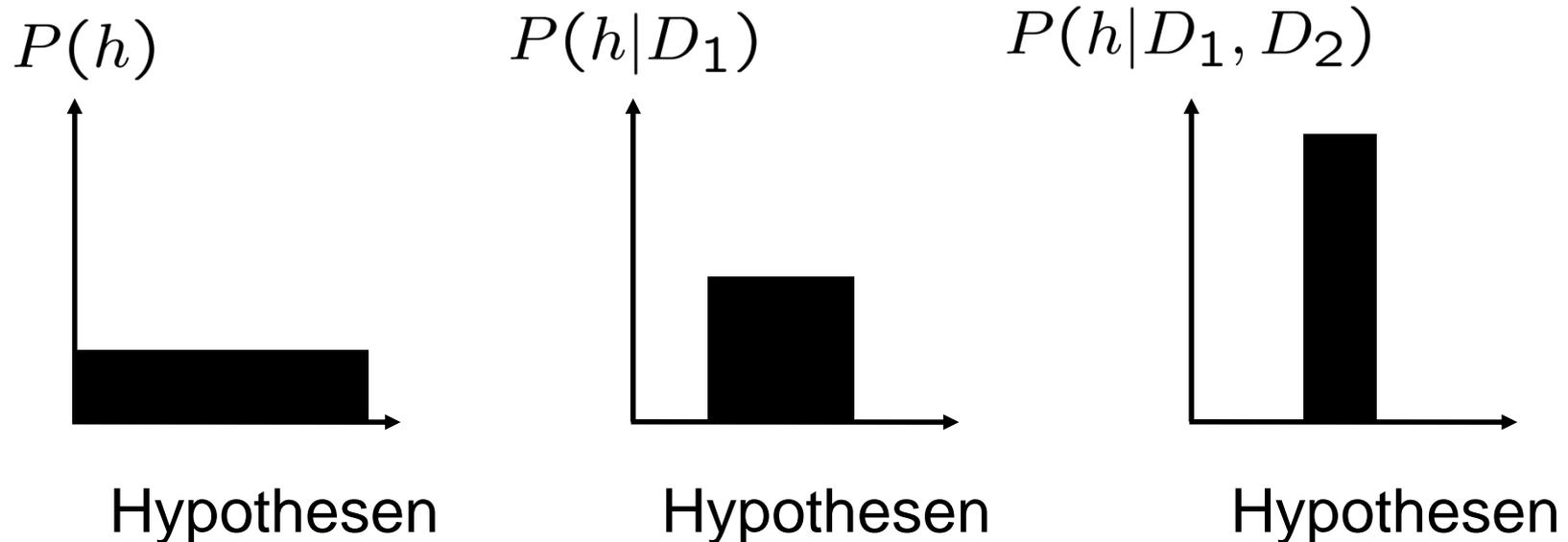
Ergebnis:

$$P(h|D) = \begin{cases} \frac{1}{|VS_{H,D}|} & \text{falls } h \text{ konsistent mit } D \\ 0 & \text{sonst} \end{cases}$$

$VS_{H,D}$ Menge der h aus H , die konsistent mit D sind
(Versionenraum von H)

- Definition: Ein Lernverfahren ist ein konsistenter Lerner, wenn es eine Hypothese liefert, die keine Fehler auf den Trainingsdaten macht.
- Unter obigen Voraussetzungen gibt jeder konsistente Lerner eine MAP-Hypothese aus
- Methode um induktiven Bias auszudrücken.

Entwicklung der a posteriori Wahrscheinlichkeiten mit wachsender Anzahl von Trainingsdaten:



Inkonsistente Hypothesen $P \rightarrow 0$

Lernen einer reell-wertigen Funktion I

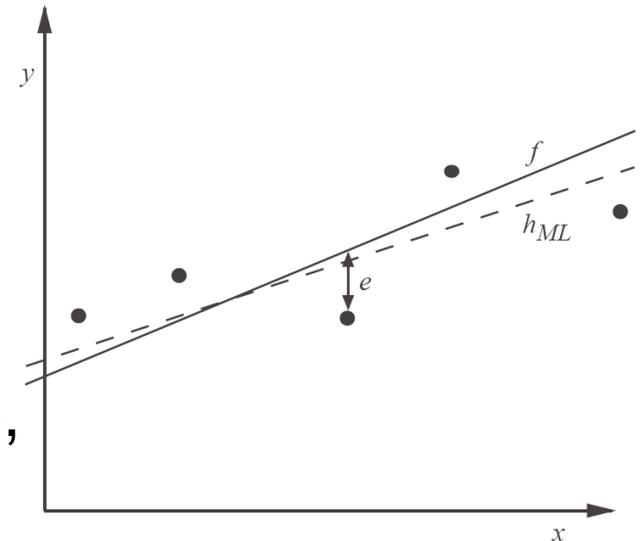
Gesucht: reell-wertige Zielfunktion f

Gegeben: Beispiele $\langle x_i, d_i \rangle$ mit
verrauschten Trainingswerten für d_i :

- $d_i = f(x_i) + e_i$
- e_i ist eine Zufallsvariable (Rauschen),
die unabhängig für alle x_i ent-
sprechend einer Normalverteilung
mit Mittelwert $\mu = 0$ gezogen wird

Die Maximum Likelihood Hypothese h_{ML} ist diejenige,
welche die Summe der Fehlerquadrate minimiert:

$$h_{ML} = \arg \min_{h \in H} \sum_{i=1}^m (d_i - h(x_i))^2$$



Lernen einer reell-wertigen Funktion II

$$h_{ML} = \arg \max_{h \in H} P(D|h)$$

d_i unabhängig

$$= \arg \max_{h \in H} \prod_{i=1}^m P(d_i|h)$$

Rauschen
normalverteilt

$$= \arg \max_{h \in H} \prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2} \left(\frac{d_i - h(x_i)}{\sigma} \right)^2$$

Monotonie

$$= \arg \max_{h \in H} \sum_{i=1}^m \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2} \left(\frac{d_i - h(x_i)}{\sigma} \right)^2$$
$$= \arg \min_{h \in H} \sum_{i=1}^m (d_i - h(x_i))^2$$

Klassifikation neuer Instanzen

Bisher: Suche nach der Hypothese mit der größten Wahrscheinlichkeit gegeben die Daten D .

Jetzt: Welches ist die wahrscheinlichste Klassifikation v_j einer neuen Instanz x ?

Beispiel:

$$P(h_1|D) = 0.4, P(h_2|D) = 0.3, P(h_3|D) = 0.3$$

$$h_1(x) = \oplus, h_2(x) = \ominus, h_3(x) = \ominus$$

→ $h_{MAP}(x)$ ist nicht die wahrscheinlichste Klassifikation!

Optimale Klassifikation nach Bayes:

$$v_{OB} = \arg \max_{v_j \in V} \sum_{h_i \in H} P(v_j | h_i) P(h_i | D)$$

Beispiel:

$$P(h_1 | D) = 0.4, \quad P(\ominus | h_1) = 0, \quad P(\oplus | h_1) = 1$$

$$P(h_2 | D) = 0.3, \quad P(\ominus | h_2) = 1, \quad P(\oplus | h_2) = 0$$

$$P(h_3 | D) = 0.3, \quad P(\ominus | h_3) = 1, \quad P(\oplus | h_3) = 0$$

$$\sum_{h_i \in H} P(\oplus | h_i) P(h_i | D) = 0.4$$

$$\sum_{h_i \in H} P(\ominus | h_i) P(h_i | D) = 0.6$$

Optimaler Bayes-Klassifikator II

Vorteil: Kein anderes Klassifikationsverfahren (bei gleichem Hypothesenraum und Vorwissen) schneidet im Durchschnitt besser ab!

Nachteil: Sehr kostenintensiv bei großer Hypothesenanzahl!

Gibbs Algorithmus:

- Wähle h aus H zufällig gemäß $P(h|D)$.
- Nutze $h(x)$ als Klassifikation von x .
- Bestimme Erwartungswert wie vorher

Eigenschaft: Unter bestimmten Annahmen gilt

$$E[\text{error}_{Gibbs}] \leq 2E[\text{error}_{BayesOptimal}]$$

Ensemble Lernen

Gegeben:

- Instanz x : Konjunktion von Attributen $\langle a_1, a_2 \dots a_n \rangle$
- Endliche Menge von Klassen $V = \{v_1, \dots, v_m\}$
- Menge klassifizierter Beispiele

Gesucht:

- Wahrscheinlichste Klasse für eine neue Instanz

$$\begin{aligned} v_{MAP} &= \arg \max_{v_j \in V} P(v_j | a_1, a_2 \dots a_n) \\ &= \arg \max_{v_j \in V} \frac{P(a_1, a_2 \dots a_n | v_j) P(v_j)}{P(a_1, a_2 \dots a_n)} \\ &= \arg \max_{v_j \in V} P(a_1, a_2 \dots a_n | v_j) P(v_j) \end{aligned}$$

$P(v_i)$ lässt sich leicht aus dem Auftreten der Klasse v_i in der Trainingsmenge berechnen - einfaches Zählen.

$P(a_1, a_2 \dots a_n | v_j)$ ist schwerer zu berechnen: Auszählen aller Kombinationen über Attributwerte \rightarrow dazu ist eine riesige Trainingsmenge notwendig.

Vereinfachende Annahme (a_i bedingt unabhängig):

$$P(a_1, a_2 \dots a_n | v_j) = \prod_i P(a_i | v_j)$$

Naiver Bayes-Klassifikator:

$$v_{NB} = \arg \max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$

Definition:

X ist bedingt unabhängig von Y gegeben Z , wenn die Wahrscheinlichkeitsverteilung von X bei gegebenem Wert von Z unabhängig vom Wert von Y ist:

$$(\forall x_i, y_j, z_k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

Oder kompakter: $P(X|Y, Z) = P(X|Z)$

Beispiel:

Donner ist bedingt unabhängig von Regen gegeben Blitz

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

Zusammenfassung:

- $P(v_j)$ und $P(a_i|v_j)$ werden basierend auf den Häufigkeiten in den Trainingsdaten geschätzt.
- Wahrscheinlichkeiten für Klassifikation entspricht gelernter Hypothese.
- Neue Instanzen werden klassifiziert unter Anwendung obiger MAP Regel.
- Wenn Annahme (bedingte Unabhängigkeit der Attribute) erfüllt ist, ist v_{NB} äquivalent zu einer MAP-Klassifikation.

→ Keine explizite Suche im Hypothesenraum!

Beispiel I

Gesucht: $P(v_j)$ und $P(a_i|v_j)$

Vorhersage	Temperatur	Luftfeuchtigkeit	Wind	Tennis?
sonnig	heiß	hoch	schwach	nein
sonnig	heiß	hoch	stark	nein
bedeckt	heiß	hoch	schwach	ja
regnerisch	warm	hoch	schwach	ja
regnerisch	kalt	normal	schwach	ja
regnerisch	kalt	normal	stark	nein
bedeckt	kalt	normal	stark	ja
sonnig	warm	hoch	schwach	nein
sonnig	kalt	normal	schwach	ja
regnerisch	warm	normal	schwach	ja
sonnig	warm	normal	stark	ja
bedeckt	warm	hoch	stark	ja
bedeckt	heiß	normal	schwach	ja
regnerisch	warm	hoch	stark	nein

Beispiel II

Neue Instanz: $\langle \text{sonnig, kalt, hoch, stark} \rangle$

$$v_{NB} = \arg \max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$

$$P(\text{Tennis}=\text{ja}) = \frac{9}{14} = 0.64$$

$$P(\text{Wind}=\text{stark} | \text{Tennis}=\text{ja}) = \frac{3}{9} = 0.33$$

$$P(\text{Tennis}=\text{nein}) = \frac{5}{14} = 0.36$$

$$P(\text{Wind}=\text{stark} | \text{Tennis}=\text{nein}) = \frac{3}{5} = 0.60$$

⋮

$$P(\text{ja})P(\text{sonnig}|\text{ja})P(\text{kalt}|\text{ja})P(\text{hoch}|\text{ja})P(\text{stark}|\text{ja}) = 0.0053$$

$$P(\text{nein})P(\text{sonnig}|\text{nein})P(\text{kalt}|\text{nein})P(\text{hoch}|\text{nein})P(\text{stark}|\text{nein}) = 0.0206$$

→ Klassifikation: Tennis = nein

Normierte Wahrscheinlichkeit: $\frac{0.0206}{0.0206 + 0.0053} = 0.795$

Problem: Was, wenn für eine Klasse v_j ein Attribut a_i einen bestimmten Wert in den Daten gar nicht annimmt?

$$\hat{P}(a_i|v_j) = 0 \rightarrow \hat{P}(v_j) \prod_i \hat{P}(a_i|v_j) = 0$$

Lösung: $\hat{P}(a_i|v_j) \leftarrow \frac{n_c + mp}{n + m}$ (m- Laplace Schätzer)

n Anzahl der Beispiele mit $v = v_j$

n_c Anzahl der Beispiele mit $v = v_j$ und $a = a_j$

p A priori Wahrscheinlichkeit für $\hat{P}(a_i|v_j)$ z.B: $p = \frac{1}{|\text{Werte}(a_i)|}$

m Anzahl der „virtuellen Beispiele“ gewichtet mit a priori Wahrscheinlichkeit p

Anwendungen:

- Lernen, welche Nachrichten interessant sind
- Lernen der Themenzugehörigkeit von Webseiten

Motivation:

- Statistische Verfahren sind sehr erfolgreich bei der Klassifikation von Texten.

Frage:

- Welche Attribute sind geeignet, um Textdokumente zu repräsentieren?

Gesucht Zielfunktion: Interessant? : Dokument $\rightarrow \{\oplus, \ominus\}$

1. Repräsentation jedes Textes als Vektor aus Wörtern:
Ein Attribut pro Wortposition im Dokument.
2. Lernphase: Verwende die Trainingsbeispiele zum
Schätzen von $P(\oplus)$, $P(\ominus)$, $P(doc|\oplus)$, $P(doc|\ominus)$

$$\text{Es gilt } P(doc|v_j) = \prod_{i=1}^{\text{length}(doc)} P(a_i = w_k|v_j)$$

wobei $P(a_i = w_k|v_j)$ die Wahrscheinlichkeit ist, das Wort w_k an der Position a_i zu finden, gegeben v_j .

Zusätzliche „weichere“ Annahme (bag of words):

$$P(a_i = w_k|v_j) = P(a_m = w_k|v_j), \forall i, m$$

1. Sammle Vokabular:

Vokabular \leftarrow Alle Wörter und Token aus den Beispielen

2. Berechne $P(v_j)$ und $P(w_k|v_j)$ für alle v_j :

$\text{docs}_j \leftarrow$ Untermenge der Beispiele mit v_j

$$P(v_j) \leftarrow \frac{|\text{docs}_j|}{|\text{Beispiele}|}$$

$\text{Text}_j \leftarrow$ Konkatenation aller Elemente von docs_j

$n \leftarrow$ Gesamtanzahl Wortpositionen in Text_j

$n_k \leftarrow$ Anzahl Vorkommen von w_k in Text_j

$$P(w_k|v_j) \leftarrow \frac{n_k + 1}{n + |\text{Vokabular}|} \quad (\text{Laplace Schätzer})$$

Klassifikation von Texten: Klassifikationsphase

1. Positionen \leftarrow Alle Positionen, die ein Token enthalten, das in Vokabular vorkommt.
2. Berechne v_{NB}

$$v_{NB} = \arg \max_{v_j \in V} P(v_j) \prod_{i \in \text{Positionen}} P(a_i | v_j)$$



(Token-Positionen die nicht im Beispiel vorhanden sind werden ignoriert)

Klassifikation von Texten: Anwendung (s. Mitchel)

Anwendung:

- Gegeben: 20 Newsgroups mit ca. je 1000 Beiträgen
- Gesucht: Zuordnung neuer Beiträge zu den Newsgroups

Klassifikator:

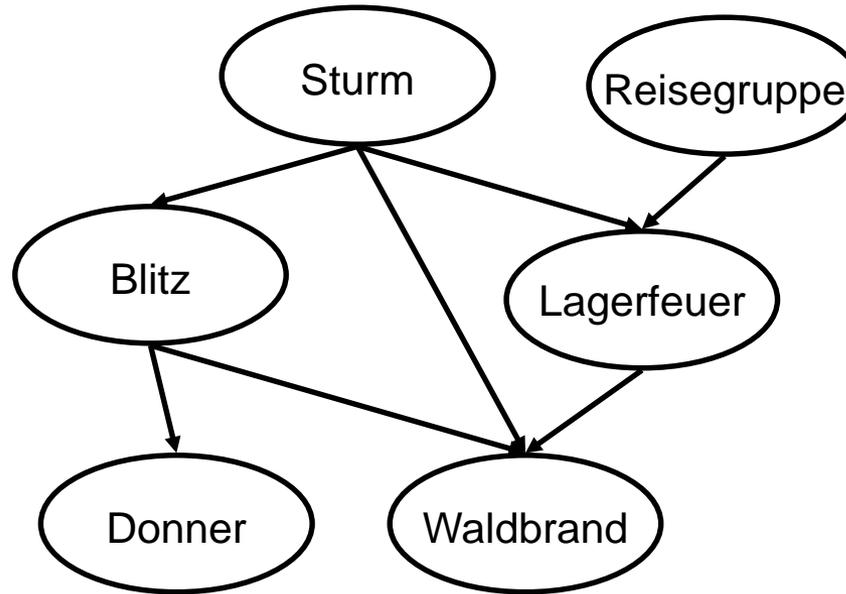
- Naiver Bayes-Klassifikator wie oben, aber
 - 100 häufigsten Wörter wurden aus Vokabular entfernt
 - Wörter mit $w_k < 3$ wurden aus Vokabular entfernt

Ergebnis:

- Klassifikationsgüte von 89% (vgl. zufälliges Raten: 5%)

Motivation:

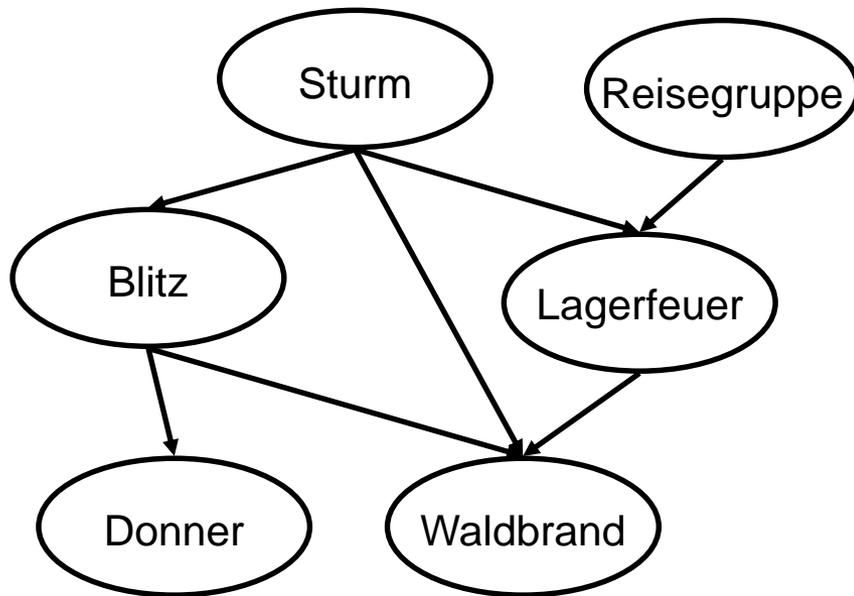
- Naive Bayes-Annahme der bedingten Unabhängigkeit oft zu restriktiv.
- Ohne solche, vereinfachenden Annahmen ist Lernen nach Bayes jedoch oft nicht möglich.
- Bayes'sche Netze beschreiben bedingte Abhängigkeiten/Unabhängigkeiten bzgl. Untermengen von Variablen.
 - Erlauben somit die Kombination von a priori Wissen über bedingte (Un-)Abhängigkeiten von Variablen mit den beobachteten Trainingsdaten.



- Bayes'sche Netze beschreiben bedingte Abhängigkeiten/Unabhängigkeiten bzgl. Untermengen von Variablen.
 - Erlauben somit die Kombination von a priori Wissen über bedingte (Un-)Abhängigkeiten von Variablen mit den beobachteten Trainingsdaten.

Repräsentieren eine gemeinsame Wahrscheinlichkeitsverteilung von Zufallsvariablen:

- Gerichteter, azyklischer Graph
- Jede Zufallsvariable wird durch einen Knoten im Graph repräsentiert
- Definition: X ist Nachfolger von Y , wenn ein gerichteter Pfad von Y nach X existiert.
- Die Kanten repräsentieren die Zusicherung, dass eine Variable von ihren Nicht-Nachfolgern bedingt unabhängig ist, gegeben ihre direkten Vorgänger.
- Lokale Tabellen mit bedingten Wahrscheinlichkeiten für jede Variable gegeben ihre direkten Vorgänger.



$P(L|S, R)$

	S, R	$S, \neg R$	$\neg S, R$	$\neg S, \neg R$
L	0.4	0.1	0.8	0.2
$\neg L$	0.6	0.9	0.2	0.8



- Es gilt: $P(y_1, \dots, y_n) = \prod_{i=1}^n P(y_i | \text{Vorgänger}(Y_i))$

wobei $\text{Vorgänger}(Y_i)$ die Menge der direkten Vorgänger von Y_i ist.

Wie lassen sich die Werte einer oder mehrerer Netzvariablen bestimmen, gegeben die beobachteten Werte von anderen?

- Netz enthält alle benötigten Informationen
- Ableitung einer einzigen Variable ist einfach
- Aber: Der allgemeine Fall ist NP-vollständig

In der Praxis:

- Einige Netztopologien erlauben exakte Inferenz
- Verwendung von Monte Carlo Methoden zur Zufallssimulation von Netzen:
Berechnung von approximierten Lösungen

Aufgabenstellungen:

- Struktur des Netzes bekannt oder unbekannt
- Alle Variablen direkt beobachtbar oder nur teilweise

Struktur bekannt, alle Variablen beobachtbar:

- Lernen wie für Naiven Bayes-Klassifikator

Struktur bekannt, nur einige Variablen beobachtbar:

- Gradientenanstieg, EM

Struktur unbekannt:

- Heuristische Verfahren

Der EM-Algorithmus

EM = Expectation Maximization

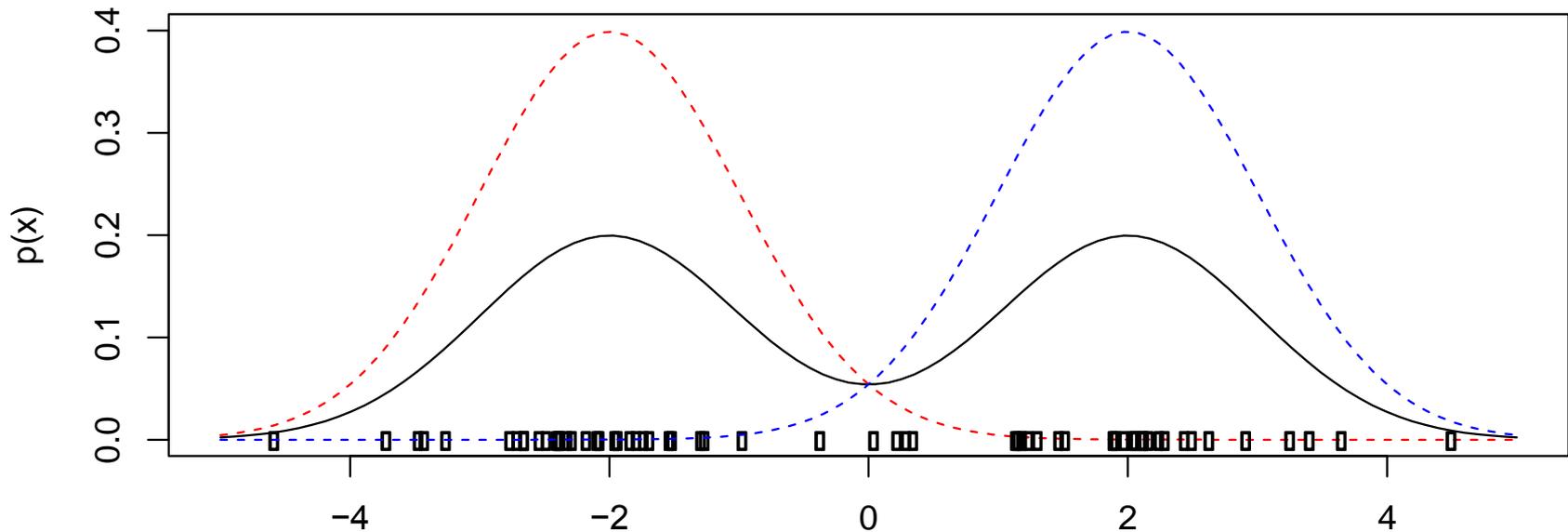
Problemstellungen:

- Daten sind nur partiell beobachtbar
- Unüberwachtes Clustering (Zielwert ist nicht beobachtbar)
- Überwachtes Lernen (einige Attribute der Instanzen sind nicht beobachtbar)

Anwendungen:

- Trainieren von Bayes'schen Netzen
- Lernen von Hidden Markov Modellen

EM - Mixtur aus k Gaußverteilungen



Generierung jeder Instanz x wie folgt:

- Wähle eine der k Gaußverteilungen mit gleichmäßiger Wahrscheinlichkeit
- Generiere eine Instanz zufällig entsprechend der gewählten Gaußverteilung

EM für die Bestimmung der k Mittelwerte I

Gegeben:

- Instanzen aus X generiert entsprechend einer Mixtur aus k Gaußverteilungen
- Unbekannte Mittelwerte $\langle \mu_1, \dots, \mu_k \rangle$
- Es ist unbekannt welche Instanz x_i entsprechend welcher Gaußverteilung generiert wurde

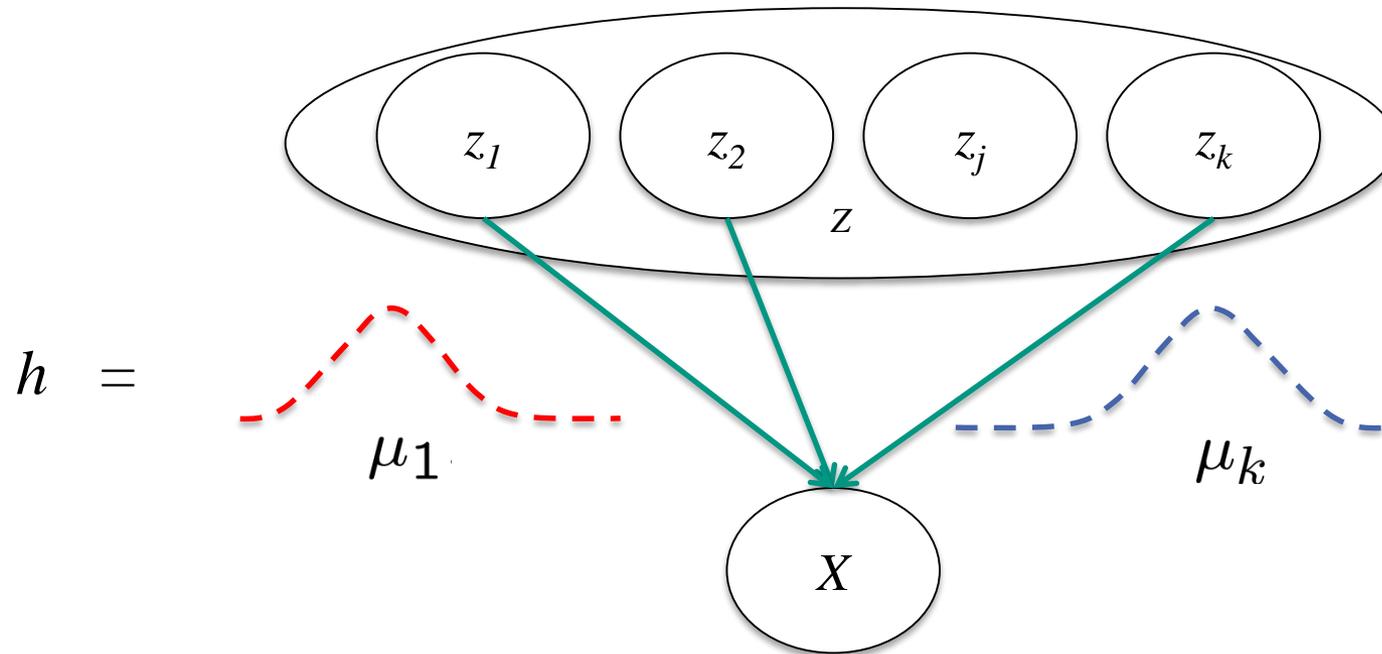
Gesucht: Maximum Likelihood Schätzung für

$$h = \langle \mu_1, \dots, \mu_k \rangle$$

- Erweiterte Sicht:
Beschreibung der Instanzen durch: $y_i = \langle x_i, z_{i1}, z_{i2} \rangle$
 z_{ij} ist 1, falls x_i entsprechend der j -ten Gaußverteilung gezogen wurde, sonst 0

→ x_i beobachtbar, z_{ij} nicht beobachtbar

EM – Modell – „anschaulich“



Wenn x und h bekannt $\rightarrow z$

Wenn z und x bekannt $\rightarrow h$



EM für die Bestimmung der k ($=2$) Mittelwerte II

Initialisierung: Wähle $h = \langle \mu_1, \mu_2 \rangle$ zufällig.

E-Schritt: Berechne den Erwartungswert $E[z_{ij}]$ für jede versteckte Variable z_{ij} , unter der Annahme, dass die aktuelle Hypothese $h = \langle \mu_1, \mu_2 \rangle$ gültig ist.

M-Schritt: Berechne eine neue Maximum Likelihood Hypothese $h' = \langle \mu'_1, \mu'_2 \rangle$, unter der Annahme, dass die Werte der versteckten Variablen die im E-Schritt berechneten Erwartungswerte annehmen. Ersetze $h = \langle \mu_1, \mu_2 \rangle$ durch die neue Hypothese $h' = \langle \mu'_1, \mu'_2 \rangle$ und iteriere.

EM für die Bestimmung der k (=2) Mittelwerte III

Beispiel: EM für obige Mixtur aus 2 Gaußverteilungen.

$$\begin{aligned} \text{E-Schritt: } E[z_{ij}] &= \frac{p(x = x_i | \mu = \mu_j)}{\sum_{n=1}^2 p(x = x_i | \mu = \mu_n)} \\ &= \frac{\exp\left(-\frac{1}{2\sigma^2} (x_i - \mu_j)^2\right)}{\sum_{n=1}^2 \exp\left(-\frac{1}{2\sigma^2} (x_i - \mu_n)^2\right)} \end{aligned}$$

$$\text{M-Schritt: } \mu_j \leftarrow \frac{1}{m} \sum_{i=1}^m E[z_{ij}] x_i$$

- Konvergiert gegen eine lokale Maximum Likelihood Hypothese und liefert Schätzungen für die versteckten Variablen z_{ij} .
- Sucht die Maximum Likelihood Hypothese h' , welche $E [\ln P (Y | h')]$ maximiert, wobei
 - Y die vollständigen Daten sind (beobachtbare plus versteckte Variablen)
 - Der Erwartungswert über die möglichen Werte der versteckten Variablen berechnet wird

Allgemeines EM-Problem

Gegeben:

- Beobachtbare Daten $X = \{x_1, \dots, x_m\}$
- Nicht beobachtbare Daten $Z = \{z_1, \dots, z_m\}$
- Parametrisierte Wahrscheinlichkeitsverteilung $P(Y|h)$, wobei
 - $Y = \{y_1, \dots, y_m\}$ die vollständigen Daten $Y = X \cup Z$ sind
 - h die Parameter sind

Gesucht:

- Hypothese h , welche $E[\ln P(Y|h)]$ (lokal) maximiert

Definiere Likelihood Funktion Q , welche $Y = X \cup Z$ unter Verwendung der beobachtbaren Daten X und der aktuellen Parameter h berechnet, um Z zu schätzen.

$$Q \leftarrow E \left[\ln P \left(Y | h^* \right) | h, X \right]$$

E-Schritt: Berechne $P(Z | X, h)$ unter Verwendung der aktuellen Hypothese h und der beobachtbaren Daten X .

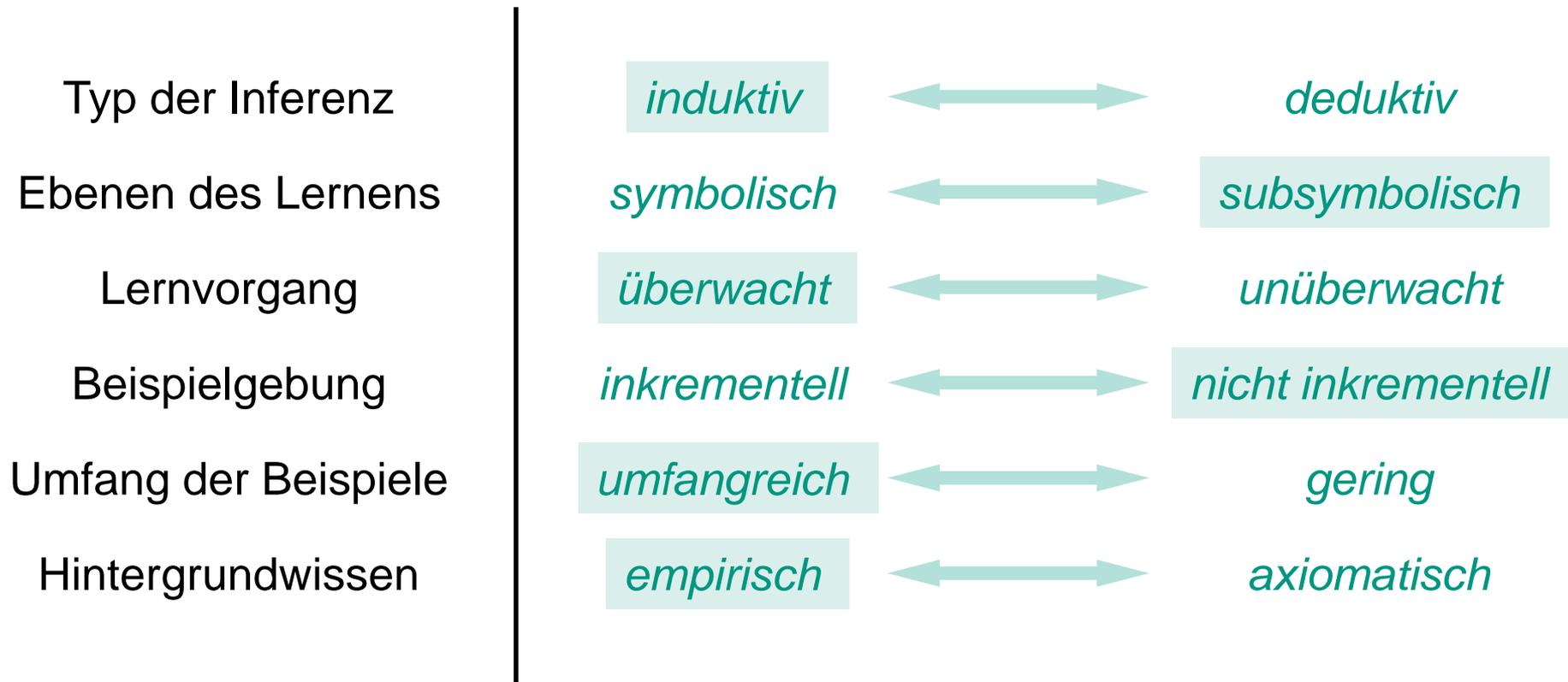
M-Schritt: Ersetze Hypothese h mit Hypothese h' , welche die Q Funktion maximiert.

$$h' \leftarrow \arg \max_{h'} Q$$

- Bayes-Methoden ermitteln a posteriori Wahrscheinlichkeiten für Hypothesen basierend auf angenommenen a priori Wahrscheinlichkeiten und beobachteten Daten.
- Mit Bayes-Methoden kann wahrscheinlichste Hypothese (MAP-Hypothese) bestimmt werden („optimale“ Hypothese).
- Der Optimale Bayes-Klassifikator bestimmt die wahrscheinlichste Klassifikation einer neuen Instanz aus den gewichteten Vorhersagen aller Hypothesen.
- Der Naive Bayes-Klassifikator ist ein erfolgreiches Lernverfahren. Annahme: bedingte Unabhängigkeit der Attributwerte.

- Bayes-Methoden erlauben die Analyse anderer Lernalgorithmen, die nicht direkt das Bayes-Theorem anwenden.
- Bayes'sche Netze beschreiben gemeinsame Wahrscheinlichkeitsverteilungen mit Hilfe eines gerichteten Graphen und lokalen Wahrscheinlichkeitstabellen.
- Bayes'sche Netze modellieren bedingte Unabhängigkeiten in Untermengen von Variablen. Weniger restriktiv als der Naive Bayes-Klassifikator.
- Der EM-Algorithmus erlaubt den Umgang mit nicht beobachtbaren Zufallsvariablen.

Einordnung



- [1] *Tom Mitchell: **Machine Learning***. McGraw-Hill, New York, 1997. Kap 6.
- [2] *Tom Mitchell: **Homepage***.
 - <http://www-2.cs.cmu.edu/~tom/>
 - Programm und Daten zum Naiven Bayes-Klassifikator
- [3] *Andrew W. Moore: **Data Mining Tutorials***.
<http://www-2.cs.cmu.edu/~awm/tutorials/>
- [4] *S. Russel, P. Norvig: **Artificial Intelligence: A Modern Approach***. Prentice Hall, 2nd Edition, 2003.
- [5] *Christopher M. Bishop: **Pattern Recognition and Machine Learning***. Springer, 2006.