Kapitel ML: V

V. Statistische Lernverfahren

- Wahrscheinlichkeitsbegriff
- Bayes-Classifier
- Maximum-a-Posteriori-Hypothesen

Wahrscheinlichkeitsbegriff

Definition 1 (Zufallsexperiment, Zufallsbeobachtung)
Ein Zufallsexperiment ist ein – im Prinzip – beliebig oft wiederholbarer Vorgang, der sich wie folgt beschreiben lässt:

1. Anordnung.
   Eine durch eine Beschreibung festgelegte reproduzierbare Gegebenheit.
2. Prozedur.
   Vorgehen zur Durchführung des Experiments mit der Anordnung.
3. Unvorhersagbarkeit des Resultats.

Werden Anordnung und Prozedur nicht künstlich geschaffen, so spricht man von natürlichen Zufallsexperimenten bzw. von Zufallsbeobachtungen.

Bemerkungen:

- Der Vorgang kann mehrmalig mit demselben System oder einmalig mit gleichartigen Systemen durchgeführt werden.
- Auch Zufallsexperimente laufen kausal im Sinne von Ursache und Wirkung ab. Die Zufälligkeit des Ergebnisses beruht nur auf dem Fehlen von Informationen über die Ursachenkette. Somit kann ein Zufallsexperiment durch neue Erkenntnisse sein ganze Zufälligkeit verlieren.

Wahrscheinlichkeitsbegriff

Definition 2 (Ergebnisraum, Ereignisraum)
Eine Menge $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ heißt Ergebnisraum eines Zufallsexperiments, wenn jedem Versuchsausgang höchstens ein Element $\omega \in \Omega$ zugeordnet ist. Die Elemente von Ω heißen Ergebnisse.

Jede Teilmenge A eines Ergebnisraums Ω heißt Ereignis. Ein Ereignis A tritt genau dann ein, wenn ein Ergebnis ω vorliegt mit $\omega \in A$. Die Menge aller Ereignisse $\mathcal{P}(\Omega)$ heißt Ereignisraum.

Wahrscheinlichkeitsbegriff

Definition 2 (Ergebnisraum, Ereignisraum)

Eine Menge $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ heißt Ergebnisraum eines Zufallsexperiments, wenn jedem Versuchsausgang höchstens ein Element $\omega \in \Omega$ zugeordnet ist. Die Elemente von Ω heißen Ergebnisse.

Jede Teilmenge A eines Ergebnisraums Ω heißt Ereignis. Ein Ereignis A tritt genau dann ein, wenn ein Ergebnis ω vorliegt mit $\omega \in A$. Die Menge aller Ereignisse $\mathcal{P}(\Omega)$ heißt Ereignisraum.

Definition 3 (wichtige Ereignistypen)

Sei Ω ein Ergebnisraums, $A \subseteq \Omega$ und $B \subseteq \Omega$ zwei Ereignisse. Dann sei vereinbart:

$\emptyset$		unmögliches Ereignis
Ω		sicheres Ereignis
$\overline{A} := \Omega \setminus A$		Gegenereignis zu A
$\lvert A \rvert = 1$		Elementarereignis
$A \subseteq B$	$\Leftrightarrow$	A ist Teilereignis von B bzw. „A zieht B nach sich“
$A = B$	$\Leftrightarrow$	$A \subseteq B \ \wedge \ B \subseteq A$
$A \cap B$	$\Leftrightarrow$	A und B sind unvereinbar (ansonsten vereinbar)

Wahrscheinlichkeitsbegriff

Klassische Begriffsbildung

Empirisches Gesetz der großen Zahlen:

Es gibt Ereignisse, deren relative Häufigkeit nach einer hinreichend großen Anzahl von Versuchen ungefähr gleich einem festen Zahlenwert ist.

Wahrscheinlichkeitsbegriff

Klassische Begriffsbildung

Empirisches Gesetz der großen Zahlen:

Es gibt Ereignisse, deren relative Häufigkeit nach einer hinreichend großen Anzahl von Versuchen ungefähr gleich einem festen Zahlenwert ist.

Definition 4 (klassischer / Laplace'scher Wahrscheinlichkeitsbegriff)
Wird jedem Elementarereignis aus Ω die gleiche Wahrscheinlichkeit zugeordnet, so gilt für die Wahrscheinlichkeit $P(A)$ eines Ereignisses A:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{Anzahl der für } A \text{ günstigen Elementarereignisse}}{\text{Anzahl aller möglichen Elementarereignisse}}$$

Bemerkungen:

- Ein Zufallsexperiment, bei dem auf Basis einer Analyse der Anordnung und Prozedur angenommen werden kann, dass alle Ergebnisse gleich wahrscheinlich sind, heißt Laplace-Experiment. Die Wahrscheinlichkeiten der Ergebnisse heißen Laplace-Wahrscheinlichkeiten.
- Laplace-Annahme: Annahme, dass es sich bei einem Experiment um ein Laplace-Experiment handelt. Trifft die Laplace-Annahme nicht zu, so können die Wahrscheinlichkeiten nur als relative Häufigkeiten bei der Durchführung vieler Versuche geschätzt werden.
- Strenggenommen handelt es sich bei der Definition des Laplace'scher Wahrscheinlichkeitsbegriffs nicht um eine Definition.
- Motiviert durch das empirische Gesetz der großen Zahlen wurde versucht, den Wahrscheinlichkeitsbegriff *frequentistisch*, über den (fiktiven) Grenzwert der relativen Häufigkeit zu definieren (hier insbesondere v. Mises, 1951). Dieser Versuch scheiterte, weil dieser Grenzwert – im Gegensatz zu Grenzwerten in der Infinitesimalrechnung – nicht nachweisbar ist.

Wahrscheinlichkeitsbegriff

Axiomatische Begriffsbildung

Prinzip einer axiomatischen Begriffsbildung:

1. Postulierung einer Funktion, die jedem Element des Ereignisraums eine Wahrscheinlichkeit zuordnet.
2. Formulierung grundlegender Eigenschaften dieser Funktion in Form von Axiomen.

Definition 5 (Wahrscheinlichkeitsmaß [Kolmogorow 1933]**)**
Sei Ω eine Menge, genannt Ergebnisraum, und sei $\mathcal{P}(\Omega)$ die Menge aller Ereignisse, genannt Ereignisraum. Dann heißt eine Funktion $P : \mathcal{P}(\Omega) \to \mathbf{R}$, die jedem Ereignis $A \in \mathcal{P}(\Omega)$ eine relle Zahl $P(A)$ zuordnet, Wahrscheinlichkeitsmaß, wenn sie folgende Eigenschaften besitzt:

1. $P(A) \geq 0$ (Axiom I)
2. $P(\Omega) = 1$ (Axiom II)
3. $A \cap B = \emptyset$ impliziert $P(A \cup B) = P(A) + P(B)$ (Axiom III)

Wahrscheinlichkeitsbegriff

Axiomatische Begriffsbildung

Definition 6 (Wahrscheinlichkeitsraum)

Sei Ω ein Ergebnisraum, $\mathcal{P}(\Omega)$ ein Ereignisraum und $P : \mathcal{P}(\Omega) \rightarrow \mathbf{R}$ ein Wahrscheinlichkeitsmaß. Dann heißt das Paar (Ω, P) bzw. das Tripel $(\Omega, \mathcal{P}(\Omega), P)$ Wahrscheinlichkeitsraum.

Satz 1 (Folgerungen aus den Axiomen von Kolmogorov)

1. $P(A) + P(\overline{A}) = 1$ (aus Axiomen II und III)
2. $P(\emptyset) = 0$ (aus 1. mit $A = \Omega$)
3. Monotoniegesetz des Wahrscheinlichkeitsmaßes:
   $A \subseteq B \;\Rightarrow\; P(A) \leq P(B)$ (aus Axiomen I und II)
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (aus Axiom III)
5. Seien $A_1, A_2 \ldots, A_p$ paarweise unvereinbar, dann gilt:
   $P(A_1 \cup A_2 \cup \ldots \cup A_p) = P(A_1) + P(A_2) + \ldots + P(A_p)$

Bemerkungen:

- Die drei Axiome heißen auch das Axiomensystem von Kolmogorow.
- $P(A)$ wird als „Wahrscheinlichkeit für das Eintreffen von A“ bezeichnet.
- Es ist nicht definiert, wie die Wahrscheinlichkeiten P verteilt sind.
- Allgemein nennt man eine Funktion, welche die drei Bedingungen der Wahrscheinlichkeitsmaß-Definition erfüllt, ein nicht-negatives, normiertes und additives Maß.

Wahrscheinlichkeitsbegriff

Bedingte Wahrscheinlichkeit

Definition 7 (bedingte Wahrscheinlichkeit)
Seien $(\Omega, \mathcal{P}(\Omega), P)$ ein Wahrscheinlichkeitsraum und $A, B \in \mathcal{P}(\Omega)$ zwei Ereignisse. Die Wahrscheinlichkeit, dass A eintritt, wenn man bereits weiß, dass B eingetreten ist, ist wie folgt definiert:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{falls } P(B) > 0$$

$P(A \mid B)$ heißt bedingte Wahrscheinlichkeit für A unter der Bedingung B.

Wahrscheinlichkeitsbegriff

Bedingte Wahrscheinlichkeit

Satz 2 (totale Wahrscheinlichkeit)

Seien $(\Omega, \mathcal{P}(\Omega), P)$ ein Wahrscheinlichkeitsraum, $B_1, \ldots, B_k$ disjunkte Ereignisse mit $\Omega = B_1 \cup \ldots \cup B_k$, $P(B_i) > 0$, $i = 1, \ldots, k$, und $A \in \mathcal{P}(\Omega)$. Dann gilt:

$$P(A) = \sum_{i=1}^{k} P(B_i) \cdot P(A \mid B_i)$$

Wahrscheinlichkeitsbegriff

Bedingte Wahrscheinlichkeit

Satz 2 (totale Wahrscheinlichkeit)

Seien $(\Omega, \mathcal{P}(\Omega), P)$ ein Wahrscheinlichkeitsraum, $B_1, \ldots, B_k$ disjunkte Ereignisse mit $\Omega = B_1 \cup \ldots \cup B_k$, $P(B_i) > 0$, $i = 1, \ldots, k$, und $A \in \mathcal{P}(\Omega)$. Dann gilt:

$$P(A) = \sum_{i=1}^{k} P(B_i) \cdot P(A \mid B_i)$$

Beweis 1

$$\begin{aligned} P(A) &= P(\Omega \cap A) \\ &= P((B_1 \cup \ldots \cup B_k) \cap A) \\ &= P((B_1 \cap A) \cup \ldots \cup (B_k \cap A)) \\ &= \sum_{i=1}^{k} P(B_i \cap A) \\ &= \sum_{i=1}^{k} P(A \cap B_i) \;=\; \sum_{i=1}^{k} P(B_i) \cdot P(A \mid B_i) \end{aligned}$$

Bemerkungen:

- Die bedingte Wahrscheinlichkeit $P(A \mid B)$ erfüllt für variables A und festes B die Axiome von Kolmogorov und stellt ein Wahrscheinlichkeitsmaß, P_B, dar.

- Wichtige Folgerungen aus der Definition der bedingten Wahrscheinlichkeit:

  1. $P(A \cap B) = P(B) \cdot P(A \mid B)$ (Multiplikationsregel)
  2. $P(A \cap B) = P(B \cap A) = P(A) \cdot P(B \mid A)$
  3. $P(B) \cdot P(A \mid B) = P(A) \cdot P(B \mid A) \Leftrightarrow P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(B) \cdot P(A \mid B)}{P(A)}$
  4. $P(\overline{A} \mid B) = 1 - P(A \mid B)$

- Im Allgemeinen gilt: $P(A \mid \overline{B}) \neq 1 - P(A \mid B)$.

Wahrscheinlichkeitsbegriff

Unabhängigkeit von Ereignissen

Definition 8 (stochastische Unabhängigkeit von zwei Ereignissen)
Seien $(\Omega, \mathcal{P}(\Omega), P)$ ein Wahrscheinlichkeitsraum und $A, B \in \mathcal{P}(\Omega)$ zwei Ereignisse. A und B heißen stochastisch unabhängig (bei P) genau dann, wenn gilt:

$$P(A \cap B) = P(A) \cdot P(B) \qquad \text{„Multiplikationsregel“}$$

Bei Vorliegen stochastischer Unabhängigkeit und $0 < P(B) < 1$ gelten folgende Äquivalenzen:

$$\begin{aligned} & P(A \cap B) && = P(A) \cdot P(B) \\ \Leftrightarrow \quad & P(A \mid B) && = P(A \mid \overline{B}) \\ \Leftrightarrow \quad & P(A \mid B) && = P(A) \end{aligned}$$

Wahrscheinlichkeitsbegriff

Unabhängigkeit von Ereignissen

Definition 9 (stochastische Unabhängigkeit von m Ereignissen)
Seien $(\Omega, \mathcal{P}(\Omega), P)$ ein Wahrscheinlichkeitsraum und $A_1, \ldots, A_p \in \mathcal{P}(\Omega)$ Ereignisse. $A_1, \ldots, A_p$ heißen gemeinsam stochastisch unabhängig (bei P) genau dann, wenn für alle Teilmengen $\{A_{i_1}, \ldots, A_{i_l}\}$ aus $\{A_1, \ldots, A_p\}$ die Multiplikationsregel gilt:

$$P(A_{i_1} \cap \ldots \cap A_{i_l}) = P(A_{i_1}) \cdot \ldots \cdot P(A_{i_l}),$$

mit $i_1 < i_2 < \ldots < i_l$ und $2 \leq l \leq m$.

Kapitel ML: V (Fortsetzung)

V. Statistische Lernverfahren

- Wahrscheinlichkeitsbegriff
- Bayes-Klassifikation
- Maximum-a-Posteriori-Hypothesen

Bayes-Klassifikation

Satz 3 (Bayes)

Seien $(\Omega, \mathcal{P}(\Omega), P)$ ein Wahrscheinlichkeitsraum, $B_1, \ldots, B_k$ disjunkte Ereignisse mit $\Omega = B_1 \cup \ldots \cup B_k$, $P(B_i) > 0$, $i = 1, \ldots, k$, und $A \in \mathcal{P}(\Omega)$ mit $P(A) > 0$. Dann gilt:

$$P(B_i \mid A) = \frac{P(B_i) \cdot P(A \mid B_i)}{\sum_{i=1}^{k} P(B_i) \cdot P(A \mid B_i)}$$

$P(B_i)$ heißt *a-Priori-Wahrscheinlichkeit* von B_i.

$P(B_i \mid A)$ heißt *a-Posteriori-Wahrscheinlichkeit* von B_i.

Bayes-Klassifikation

Satz 3 (Bayes)

Seien $(\Omega, \mathcal{P}(\Omega), P)$ ein Wahrscheinlichkeitsraum, $B_1, \ldots, B_k$ disjunkte Ereignisse mit $\Omega = B_1 \cup \ldots \cup B_k$, $P(B_i) > 0$, $i = 1, \ldots, k$, und $A \in \mathcal{P}(\Omega)$ mit $P(A) > 0$. Dann gilt:

$$P(B_i \mid A) = \frac{P(B_i) \cdot P(A \mid B_i)}{\sum_{i=1}^{k} P(B_i) \cdot P(A \mid B_i)}$$

$P(B_i)$ heißt *a-Priori-Wahrscheinlichkeit* von B_i.

$P(B_i \mid A)$ heißt *a-Posteriori-Wahrscheinlichkeit* von B_i.

Beweis 2

Aus den bedingten Wahrscheinlichkeiten für $P(A \mid B_i)$ und $P(B_i \mid A)$ folgt:

$$P(B_i \mid A) = \frac{P(A \cap B_i)}{P(A)} = \frac{P(B_i) \cdot P(A \mid B_i)}{P(A)}$$

Anwendung des Satzes der totalen Wahrscheinlichkeit für $P(A)$ im Nenner des Bruches gibt die Behauptung.

Bemerkungen:

- Anwendung zur Klassifikation: Aus den a-Priori-Wahrscheinlichkeiten für die Klassen, $P(\text{Klasse}=c_i)$, und den Wahrscheinlichkeiten für die unmittelbar beobachtbaren Zusammenhänge, $P(\text{Attribut=Wert} \mid \text{Klasse}=c_i)$, lässt sich in einer *Umkehrschlusssituation* mittels Bayes die Wahrscheinlichkeit $P(\text{Klasse}=c_i \mid \text{Attribut=Wert})$ berechnen.
- Oft, aber nicht notwendigerweise, handelt es sich bei den $P(\text{Attribut=Wert} \mid \text{Klasse}=c_i)$ um Wahrscheinlichkeiten für *Kausalzusammenhänge*: das Ereignis „Klasse=c_i“ verursacht das Ereignis „Attribut=Wert“.
- Klassen und Attribut-Wert-Paare werden als Ereignisse aufgefasst; Ereignisse sind Teilmengen eines Ergebnisraums $\Omega = \{\omega_1, \ldots, \omega_n\}$.

Bayes-Klassifikation

Satz von Bayes für kombinierte Ereignisse

Bezeichne $P(B_i \mid A_1, \ldots, A_p)$ die Wahrscheinlichkeit für das Eintreten von Ereignis B_i unter der Voraussetzung, dass die Ereignisse $A_1, \ldots, A_p$ eingetreten sind.

Anwendung zur Klassifikation:

- die A_j, $j = 1, \ldots, p$, entsprechen den Ereignissen „Attribut_j=Wert_j“,
  die B_i, $i = 1, \ldots, k$, entsprechen den Ereignissen „Klasse=c_i“
- beobachtbarer Zusammenhang: $A_1, \ldots, A_p \mid B_i$
- Umkehrschlusssituation: $B_i \mid A_1, \ldots, A_p$

Bayes-Klassifikation

Satz von Bayes für kombinierte Ereignisse

Bezeichne $P(B_i \mid A_1, \ldots, A_p)$ die Wahrscheinlichkeit für das Eintreten von Ereignis B_i unter der Voraussetzung, dass die Ereignisse $A_1, \ldots, A_p$ eingetreten sind.

Anwendung zur Klassifikation:

- die A_j, $j = 1, \ldots, p$, entsprechen den Ereignissen „Attribut_j=Wert_j“, die B_i, $i = 1, \ldots, k$, entsprechen den Ereignissen „Klasse=c_i“
- beobachtbarer Zusammenhang: $A_1, \ldots, A_p \mid B_i$
- Umkehrschlusssituation: $B_i \mid A_1, \ldots, A_p$

➜ Wenn sich Daten zur Berechnung von $P(B_i)$ und $P(A_1, \ldots, A_p \mid B_i)$ akquirieren lassen, lässt sich $P(B_i \mid A_1, \ldots, A_p)$ mit dem Satz von Bayes berechnen:

$$P(B_i \mid A_1, \ldots, A_p) = \frac{P(B_i) \cdot P(A_1, \ldots, A_p \mid B_i)}{P(A_1, \ldots, A_p)} \qquad (\star)$$

Bemerkungen:

- $P(B_i \mid A_1, \ldots, A_p)$ heißt bedingte Wahrscheinlichkeit für B_i unter den Bedingungen $A_1, \ldots, A_p$.
- Andere Schreibweisen für $P(B_i \mid A_1, \ldots, A_p)$: $P(B_i \mid A_1 \wedge \ldots \wedge A_p)$, $P(B_i \mid A_1 \cap \ldots \cap A_p)$
- Beobachtbarer, evtl. kausaler Zusammenhang: Es wird (wurde in der Vergangenheit) immer wieder festgestellt, dass in der Situation B_i die Symptome $A_1, \ldots, A_p$ auftreten.
- Umkehrschlusssituation: Es treten die Symptome $A_1, \ldots, A_p$ auf; gesucht ist die Wahrscheinlichkeit für das Vorliegen von B_i.

Bayes-Klassifikation

Naive-Bayes

Der Aufbau einer Datenbasis, um verlässliche Werte für $P(A_1, \ldots, A_p \mid B_i)$ schätzen zu können, ist in der Praxis unmöglich. Vorgehensweise:

(a) Naive-Bayes-Assumption:

„Unter dem Ereignis (der Bedingung) B_i sind die Ereignisse $A_1, \ldots, A_p$ stochastisch unabhängig.“

in Zeichen: $$P(A_1, \ldots, A_p \mid B_i) \overset{NB}{=} \prod_{j=1}^{p} P(A_j \mid B_i)$$

Bayes-Klassifikation

Naive-Bayes

Der Aufbau einer Datenbasis, um verlässliche Werte für $P(A_1, \ldots, A_p \mid B_i)$ schätzen zu können, ist in der Praxis unmöglich. Vorgehensweise:

(a) Naive-Bayes-Assumption:

„Unter dem Ereignis (der Bedingung) B_i sind die Ereignisse $A_1, \ldots, A_p$ stochastisch unabhängig."

in Zeichen: $$P(A_1, \ldots, A_p \mid B_i) \overset{NB}{=} \prod_{j=1}^{p} P(A_j \mid B_i)$$

(b) $P(A_1, \ldots, A_p)$ ist eine Konstante und braucht nicht bekannt zu sein, um das unter der Naive-Bayes-Assumption wahrscheinlichste Ereignis $B_{NB} \in \{B_1, \ldots, B_k\}$ mit Hilfe von Bayes zu bestimmen:

$$\underset{B_i \in \{B_1,\ldots,B_k\}}{\text{argmax}} \frac{P(B_i) \cdot P(A_1, \ldots, A_p \mid B_i)}{P(A_1, \ldots, A_p)} \overset{NB}{=} \underset{B_i \in \{B_1,\ldots,B_k\}}{\text{argmax}} P(B_i) \cdot \prod_{j=1}^{p} P(A_j \mid B_i) = B_{NB}$$

Bemerkungen:

- Trifft die Naive-Bayes-Assumption zu, so maximiert B_{NB} auch die a-Posteriori-Wahrscheinlichkeit $P(B_{NB} \mid A_1, \ldots, A_p)$, die mit dem Satz von Bayes $(\star)$ berechnet wird.
- Das „Lernen“ bei einem Naive-Bayes-Classifier besteht in der Schätzung der a-Priori-Wahrscheinlichkeiten $P(B_i)$ und der Wahrscheinlichkeiten $P(A_j \mid B_i)$ für die beobachtbaren Zusammenhänge auf Basis von Traningsdaten D. Die geschätzten Wahrscheinlichkeiten entsprechen der gelernten Hypothese, die dazu verwendet wird, neue Beispiele gemäß der Optimierungsformel für B_{NB} zu klassifizieren.
- Der Hypothesenraum H besteht aus der Menge aller Werte, die $P(B_i)$ und $P(A_j \mid B_i)$ annehmen können. Beachte, dass bei der Konstruktion eines Naive-Bayes-Classifiers der Hypothesenraum H nicht durchsucht wird, sondern dass sich die gesuchte Hypothese direkt aus einer Häufigkeitsanalyse von D berechnet.

Bayes-Klassifikation

Naive-Bayes

Sei neben der Naive-Bayes-Assumption weiterhin vorausgesetzt:

(c) die Menge der B_i ist vollständig: $\sum_{i=1}^{k} P(B_i) = 1$

(d) die B_i schließen sich gegenseitig aus: $P(B_i, B_\iota) = 0, \quad 1 \leq i, \iota \leq k, \; i \neq \iota$

Bayes-Klassifikation

Naive-Bayes

Sei neben der Naive-Bayes-Assumption weiterhin vorausgesetzt:

(c) die Menge der B_i ist vollständig: $\sum_{i=1}^{k} P(B_i) = 1$

(d) die B_i schließen sich gegenseitig aus: $P(B_i, B_\iota) = 0, \quad 1 \leq i, \iota \leq k, \; i \neq \iota$

dann gilt:

$$P(A_1, \ldots, A_p) = \sum_{i=1}^{k} P(B_i) \cdot P(A_1, \ldots, A_p \mid B_i) \quad (\texttt{totale Wahrscheinlichkeit})$$

$$\overset{NB}{=} \sum_{i=1}^{k} P(B_i) \cdot \prod_{j=1}^{p} P(A_j \mid B_i) \qquad \text{(Naive-Bayes-Assumption)}$$

Bayes-Klassifikation

Naive-Bayes

Sei neben der Naive-Bayes-Assumption weiterhin vorausgesetzt:

(c) die Menge der B_i ist vollständig: $\sum_{i=1}^{k} P(B_i) = 1$

(d) die B_i schließen sich gegenseitig aus: $P(B_i, B_\iota) = 0, \quad 1 \leq i, \iota \leq k, \; i \neq \iota$

dann gilt:

$$P(A_1, \ldots, A_p) = \sum_{i=1}^{k} P(B_i) \cdot P(A_1, \ldots, A_p \mid B_i) \quad (\texttt{totale Wahrscheinlichkeit})$$

$$\overset{NB}{=} \sum_{i=1}^{k} P(B_i) \cdot \prod_{j=1}^{p} P(A_j \mid B_i) \qquad \text{(Naive-Bayes-Assumption)}$$

und es folgt für die bedingten Wahrscheinlichkeiten:

$$P(B_i \mid A_1, \ldots, A_p) = \frac{P(B_i) \cdot P(A_1, \ldots, A_p \mid B_i)}{P(A_1, \ldots, A_p)} \overset{NB,c,d}{=} \frac{P(B_i) \cdot \prod_{j=1}^{p} P(A_j \mid B_i)}{\sum_{i=1}^{k} P(B_i) \cdot \prod_{j=1}^{p} P(A_j \mid B_i)}$$

Bemerkungen:

- Auf Basis von $\operatorname{argmax}_{B_i \in \{B_1,\ldots,B_k\}} P(B_i) \cdot \prod_{j=1}^{p} P(A_j \mid B_i)$ lässt sich eine Rangordnung zwischen den Ereignissen B_i berechnen.
- Unter weiteren bestimmten Voraussetzungen lassen sich die Rangordnungswerte der B_i in die a-Posteriori-Wahrscheinlichkeiten $P(B_i \mid A_1, \ldots, A_p)$ umrechnen. Kann nämlich (c) Vollständigkeit und (d) Exklusivität der B_i angenommen werden, so gilt, dass die Summe aller a-Posteriori-Wahrscheinlichkeiten Eins sein muss, in Zeichen: $\sum_{i=1}^{k} P(B_i \mid A_1, \ldots, A_p) = 1$.

  Unter diesen Voraussetzungen dürfen die Rangordnungswerte nach der Normalisierung als Wahrscheinlichkeiten interpretiert werden. Die Normalisierung geschieht mittels Division durch die Summe aller Rangordnungswerte, $\sum_{i=1}^{k} P(B_i) \cdot \prod_{j=1}^{p} P(A_j \mid B_i)$.

Bayes-Klassifikation

Naive-Bayes: Zusammenfassung

Sei D eine Menge von Trainingsbeispielen der Form $(\mathbf{x}, c(\mathbf{x}))$; die k Ausprägungen (Klassen) des Zielkonzeptes c entsprechen den Ereignissen $B_1, \ldots, B_k$, die Ausprägungen (Werte) der $|\mathbf{x}| = p$ Attribute den Ereignissen $A_1, \ldots, A_p$.

Bayes-Klassifikation

Naive-Bayes: Zusammenfassung

Sei D eine Menge von Trainingsbeispielen der Form $(\mathbf{x}, c(\mathbf{x}))$; die k Ausprägungen (Klassen) des Zielkonzeptes c entsprechen den Ereignissen $B_1, \ldots, B_k$, die Ausprägungen (Werte) der $|\mathbf{x}| = p$ Attribute den Ereignissen $A_1, \ldots, A_p$.

Konstruktion und Anwendung eines Naive-Bayes-Classifiers:

1. Schätzung von $P(B_i)$ auf Basis der relativen Häufigkeiten in D.
2. Schätzung von $P(A_j \mid B_i)$ auf Basis der relativen Häufigkeiten in D.
3. Klassifikation eines neuen Beispiels $\mathbf{x}$ als B_{NB}, mit

$$B_{NB} = \underset{B_i \in \{B_1,\ldots,B_k\}}{\text{argmax}} \; p(B_i) \cdot \prod_{\substack{A_j \in \mathbf{x} \\ j=1,\ldots,p}} p(A_j \mid B_i)$$

4. Gegebenenfalls Berechnung der a-Posteriori-Wahrscheinlichkeit für B_{NB} durch Normalisierung von $p(B_{NB}) \cdot \prod_{\substack{A_j \in \mathbf{x} \\ j=1,\ldots,p}} p(A_j \mid B_{NB})$

Bemerkungen:

- Die „echten“ Wahrscheinlichkeiten P sind unbekannt und werden durch die relativen Häufigkeiten, in Zeichen p, geschätzt.
- Naive-Bayes ist angesagt für Trainingsmengen D, die mittelgroß bis sehr groß sind.
- Naive-Bayes fordert, dass die Attributausprägungen der Beispiele stochastisch unabhängig von den Ausprägungen des Zielkonzeptes sind.

  Die Praxis im Bereich der Textklassifikation zeigt, dass auch bei Verletzung der Naive-Bayes-Assumption hervorragende Klassifikationsergebnisse mit einem Naive-Bayes-Classifier erzielbar sind.
- Möchte man nicht nur Rangordnungswerte sondern auch a-Posteriori-Wahrscheinlichkeiten berechnen, wird noch die Closed-World-Assumption (Vollständigkeit der B_i) und die Single-Fault-Assumption (gegenseitiger Ausschluss der B_i) gefordert.
- Insgesamt muss gelten, dass sich die betrachtete Welt nur langsam ändert: die Daten müssen über einen längeren Zeitraum relativ konstant bleiben.

Bayes-Klassifikation

Naive-Bayes: Beispiel

	Outlook	Temperature	Humidity	Wind	EnjoySport
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rain	mild	high	weak	yes
5	rain	cold	normal	weak	yes
6	rain	cold	normal	strong	no
7	overcast	cold	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cold	normal	weak	yes
10	rain	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rain	mild	nigh	strong	no

Das Zielkonzept der Instanz $\mathbf{x} = (\textit{sunny}, \textit{cool}, \textit{high}, \textit{strong})$ sei unbekannt.

Bayes-Klassifikation

Naive-Bayes: Beispiel (Fortsetzung)

$$B_{NB} = \underset{B_i \in \{\textit{yes},\textit{no}\}}{\text{argmax}}\ p(B_i) \cdot \prod_{A_j \in \mathbf{x}} p(A_j \mid B_i)$$

$$= \underset{B_i \in \{\textit{yes},\textit{no}\}}{\text{argmax}}\ p(B_i) \cdot p(\textit{Outlook=sunny} \mid B_i) \cdot p(\textit{Temperature=cool} \mid B_i) \cdot \ \ldots$$

„$A_j \in \mathbf{x}$“ bezeichnet die im Merkmalsvektor $\mathbf{x}$ codierten Ereignisse. Zum Beispiel werden durch $\mathbf{x} = (\textit{sunny}, \textit{cool}, \textit{high}, \textit{strong})$ die folgenden Ereignisse definiert:

A_1 : *Outlook=sunny*

A_2 : *Temperature=cool*

A_3 : *Humidity=high*

A_4 : *Wind=strong*

Bayes-Klassifikation

Naive-Bayes: Beispiel (Fortsetzung)

Zur Klassifikation von $\mathbf{x}$ sind $2 + 4 * 2$ Wahrscheinlichkeiten zu schätzen:

- $p(\textit{EnjoySport=yes}) = \frac{9}{14} = 0.64$
- $p(\textit{EnjoySport=no}) = \frac{5}{14} = 0.36$
- $p(\textit{Wind=strong} \mid \textit{EnjoySport=yes}) = \frac{3}{9} = 0.33$
- ...

➜ Rangordnung:

1. $p(\textit{EnjoySport=no}) \cdot \prod_{A_j \in \mathbf{x}} p(A_j \mid \textit{EnjoySport=no}) = 0.0206$
2. $p(\textit{EnjoySport=yes}) \cdot \prod_{A_j \in \mathbf{x}} p(A_j \mid \textit{EnjoySport=yes}) = 0.0053$

➜ Normalisierung:

1. $P(\textit{EnjoySport=no} \mid \mathbf{x}) = \frac{0.0206}{0.0053+0.0206} = 0.795$
2. $P(\textit{EnjoySport=yes} \mid \mathbf{x}) = \frac{0.0053}{0.0053+0.0206} = 0.205$