Stochastik für Informatiker

Dr. rer. nat. Johannes Riesterer

Diskrete Wahrscheinlichkeitsverteilung

Es sei Ω eine (nicht leere) albzählbare Menge. Eine Abbildung $P : \mathcal{P}(\Omega) \to [0, 1]$ heißt diskrete Wahrscheinlichkeitsverteilung (Wahrscheinlichkeismaß), falls gilt:

$$P(\Omega) = 1$$

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n), \text{ mit } A_i \cap A_j = \emptyset \text{ für } i \neq j$$

Lemma

Für $A \subset \mathcal{P}(\Omega)$ ist $P(A) = \sum_{\omega \in A} P(\{\omega\})$.

Beispiel: Laplace Wahrscheinlichkeit

Ω endlich und $P(A) = \frac{\#A}{\#\Omega}$.

Diskrete Wahrscheinlichkeitsverteilung

Beispiel: Hash Kollision

Beim Hashing werden zufällig $k \leq n$ Daten auf n Speicherplätze verteilt. Bezeichnen wir mit $A_{k,n}$ die Möglichkeiten der Mehrfachbelegungen von Feldern, so ist das Komplementäre Ereignis $A^c_{k,n} = Perm^n_k(\Omega, o.W.)$, wobei Ω die Menge der Verfügbaren Speicherplätze Darstellt.

Beispiel: Hash Kollision

$$P(A^c_{k,n}) = \frac{\#Perm^n_k(\Omega, o.W.)}{\#Perm^n_k(\Omega, m.W.)} = \frac{n_k}{n^k} = \prod_{i=0}^{k-1}(1 - \frac{i}{n})$$

$$= \exp(\sum_{i=0}^{k-1} \ln(1 - \frac{i}{n})) \leq \exp(\sum_{i=0}^{k-1}(-\frac{i}{n})$$

$$(ln(1 - x) \leq -x \text{ für } x < 1)$$

$$= \exp(-\frac{(k-1)k}{2n})$$

Beispiel: Hash Kollision (Geburtstags-Paradoxon)

Für $n = 365$ und $k = 23$ ist damit $P(A_{k,n}) > \frac{1}{2}$. Die Wahrscheinlichkeit, dass bei einer Gruppe von mehr als 23 Leuten zwei Leute am gleichen Tag Geburtstag haben, ist also größer als $\frac{1}{2}$.

Bedingte Wahrscheinlichkeit

Für $A, B \subset \mathcal{P}(\Omega)$ und $P(B) > 0$ heißt

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

die bedingte Wahrscheinlichkeit (von A unter B).

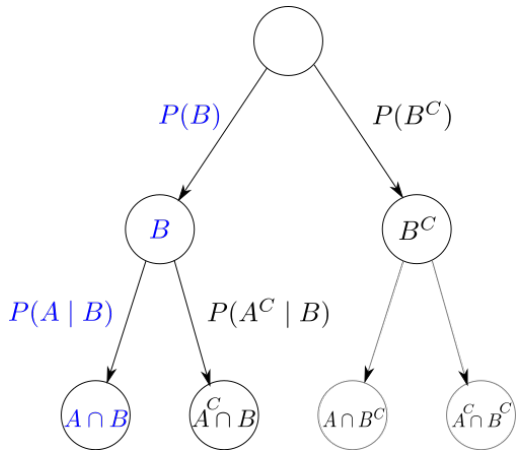

Figure: Quelle: Wikipedia

Satz der totalen Wahrscheinlichkeit

Für eine Zerlegung $\Omega = \bigcup_{j=1}^{n} B_j$, mit $B_i \cap B_k = \emptyset$ für $i \neq k$

$$P(A) = \sum_{j=1}^{n} P(A \mid B_j) \cdot P(B_j)$$

Satz von Bayes

Für $A, B \subset \mathcal{P}(\Omega)$ mit $P(B) > 0$ gilt

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

Beweis

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

Stochastische Unabhängigkeit

Zwei Ereignisse A, B heißen stochastisch Unabhängig, falls

$$P(A \cap B) = P(A) \cdot P(B)$$

gilt. Somit ist auch $P(A|B) = P(A)$ und $P(B|A) = P(B)$.

Naiver Bayes'scher Spam Filter

Gegeben ist eine E-Mail E. Wir möchten anhand des Vorkommens bestimmter Wörter $A_1, \ldots A_n$ in der Mail entscheiden, ob es sich um eine erwünschte Mail H oder eine unerwünschte Mail S (Ham or Spam) handelt. (Typische Wörter wären zum Beispiel "reichwerden", "onlinecasino" ...) Aus einer Datenbank kann man das Vorkommen dieser Wörter in Spam und Ham Mails zählen und damit empirisch die Wahrscheinlichkeiten $P(A_i|S)$ und $P(A_i|H)$ des Vorkommens dieser Wörter in Spam und Ham Mails ermitteln. Wir gehen davon aus, dass es sich bei der Mail prinzipiell mit Wahrscheinlichkeit $P(E = S) = P(E = H) = \frac{1}{2}$ um eine erwünschte Mail H oder eine unerwünschte Mail S handeln kann.

Naiver Bayes'scher Spam Filter

Wir machen zudem die (naive) Annahme, dass das Vorkommen der Wörter stochastisch unabhängig ist, also

$$P(A_1 \cap \cdots \cap A_n|S) = P(A_1|S) \cdots P(A_n|S)$$
$$P(A_1 \cap \cdots \cap A_n|H) = P(A_1|H) \cdots P(A_n|H)$$

gilt.

Naiver Bayes'scher Spam Filter

Mit der Formel von Bayes und der totalen Wahrscheinlichkeit können wir somit berechnen

$$P(E = S|A_1 \cap \cdots \cap A_n) = \frac{P(A_1 \cap \cdots \cap A_n|S) \cdot P(S)}{P(A_1 \cap \cdots \cap A_n)}$$
$$= \frac{P(A_1|S) \cdots P(A_n|S) \cdot P(S)}{P(A_1 \cap \cdots \cap A_n|H) + P(A_1 \cap \cdots \cap A_n|S)}$$
$$= \frac{P(A_1|S) \cdots P(A_n|S) \cdot P(S)}{P(A_1|H) \cdots P(A_n|H) + P(A_1|S) \cdots P(A_n|S)}$$

Bemerkung: $P(E = H|A_1 \cap \cdots \cap A_n) = 1 - P(E = S|A_1 \cap \cdots \cap A_n)$