

Wahrscheinlichkeitstheorie

Florian Fink

(basierend auf Folien von Benjamin Roth, Helmut Schmid und Michaela Geierhos)

Centrum für Informations- und Sprachverarbeitung
Ludwig-Maximilian-Universität München
`beroth@cis.uni-muenchen.de`

Zufallsexperiment

In der Statistik geht es um die Wahrscheinlichkeit von Ereignissen:

Beispiel 1: Wie wahrscheinlich ist es, dass die Summe zweier geworfener Würfel den Wert 7 ergibt?

Beispiel 2: Wie wahrscheinlich ist es, dass eine Email Spam ist?

Zufallsexperiment: Experiment (Versuch) mit mehreren möglichen Ausgängen

Beispiel 3: Wurf mit zwei Würfeln

Ergebnis: Resultat eines Experimentes

Beispiel 4: 3 Augen auf Würfel 1 und 4 Augen auf Würfel 2

Stichprobe: Folge von Ergebnissen bei einem wiederholten Experiment

Ereignisraum

Ω Ergebnisraum: Menge aller möglichen Ergebnisse

$$\Omega = \{\text{nom}, \text{gen}, \text{dat}, \text{acc}\}$$

ω Elementarereignisse: Elemente des Ereignisraums

$$\omega_i = \text{nom}|\text{gen}|\text{dat}|\text{acc} \text{ und } \omega_i \in \Omega$$

A Ereignis: Teilmenge von Ω ; d.h. $A_i \subseteq \Omega$

$$A_i = \{\text{nom}, \text{dat}, \text{acc}\}$$

\mathcal{F} Ereignisraum: Menge aller möglichen Ereignismenge
(Potenzmenge von Ω)

Wahrscheinlichkeitsraum I

Ein wohlgeformter Wahrscheinlichkeitsraum besteht aus:

- ▶ einem Ergebnisraum Ω
- ▶ einem Ereignisraum F und
- ▶ einer Wahrscheinlichkeitsfunktion P , wobei
 - ▶ \emptyset als unmögliches Ereignis und
 - ▶ Ω als sicheres Ereignis bezeichnet werden

Wahrscheinlichkeitsraum II

Es gilt das Axiomensystem nach Kolmogoroff:

- ▶ **K1 Nichtnegativität**

$$P(A) \geq 0 \text{ für alle } A \in \mathcal{F}$$

- ▶ **K2 Additivität**

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i) \text{ falls für alle } A_i, A_j \text{ gilt } A_i \cap A_j = \emptyset$$

- ▶ **K3 Normierung**

$$P(\Omega) = 1$$

Elementare Wahrscheinlichkeitsrechnung

Klassische Wahrscheinlichkeitsdefinition:

$$P(A) = \frac{|A|}{|\Omega|} \sim \frac{\text{Anzahl der günstigen Fälle}}{\text{Anzahl der möglichen Fälle}}$$

Rechenregeln:

1. $0 \leq P(A) \leq 1$
2. $P(A \text{ oder } B) = P(A) + P(B)$
3. $P(\neg A) = 1 - P(A)$
4. $P(A \text{ und } B) = P(A) \times P(B)$

Bedingte und A-priori Wahrscheinlichkeit

Bedingte Wahrscheinlichkeit:

$$P(A|B)$$

- ▶ die Wahrscheinlichkeit, dass Ereignis A eintritt, wenn Ereignis B eingetreten ist, oder
- ▶ die Wahrscheinlichkeit, dass A zutrifft, wenn B wahr ist

A priori-Wahrscheinlichkeit: Wahrscheinlichkeit eines Ereignisses vor der Betrachtung zusätzlichen Wissens $P(A)$

A posteriori-Wahrscheinlichkeit: Neue Wahrscheinlichkeit, die aus der Betrachtung zusätzlichen Wissens resultiert $P(A|B)$

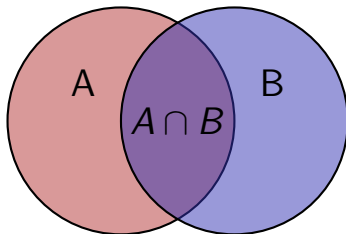
Bedingte Wahrscheinlichkeit

Definition:

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

andere Schreibweise:

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



Illustration

Bedingte Wahrscheinlichkeit: Beispiel

Beispielkorpus

Das	<i>dete</i>
Der	<i>dete</i>
Haus	<i>nomn</i>
Haus	<i>nomn</i>
Baum	<i>nomn</i>

Beispielanfrage

- ▶ Ereignis $A = \{\text{Wort mit Lexem} = \text{"Haus"}\}$
- ▶ Ereignis $B = \{\text{Wort mit categ} = \text{nomn}\}$

Wie groß ist die Wahrscheinlichkeit von A wenn B gegeben ist?

Berechnung

$$\begin{aligned} P(A = \text{'Haus'} | B = \text{nomn}) &= \\ P(A = \text{'Haus'}) &\times \frac{P(B = \text{nomn} \cap A = \text{'Haus'})}{P(B = \text{nomn})} = \\ \frac{2}{5} &\times \frac{1}{3} = \frac{2}{3} \end{aligned}$$

Gemeinsame Verteilungen und Randverteilungen

Die **gemeinsame Verteilung** zweier Zufallsvariablen X und Y :

$$p(x, y) = P(X = x, Y = y) = P(A_x \cap A_y)$$

Die **Randverteilung** von zwei Zufallsvariablen X und Y :

$$p_X(x) = \sum_y p(x, y) \quad p_Y(y) = \sum_x p(x, y)$$

Unabhängigkeit: Die Zufallsvariablen X und Y sind statistisch unabhängig, falls gilt:

$$p(x, y) = p_X(x)p_Y(y)$$

Beispiel: Beim Wurf zweier Würfel sind deren Augenzahlen statistisch voneinander unabhängig.

Kettenregel

Eine gemeinsame Wahrscheinlichkeit kann in ein Produkt bedingter Wahrscheinlichkeiten umgewandelt werden.

$$\begin{aligned} P(A_1 \cap A_2 \cap \dots \cap A_n) &= P(A_1)P(A_2|A_1)\dots P(A_n|A_1 \cap \dots \cap A_{n-1}) \\ &= \prod_{i=1}^n P(A_i|A_1 \cap \dots \cap A_{i-1}) \end{aligned}$$

Theorem von Bayes

Ermöglicht es, $P(B|A)$ aus $P(A|B)$ zu berechnen,
d.h. die Betrachtung der Abhängigkeitsverhältnisse umzukehren

Bedingte Wahrscheinlichkeit

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

Regel von Bayes (vgl. Kettenregel)

$$P(A \cap B) = P(A) \times P(B|A) = P(B) \times P(A|B)$$

Theorem von Bayes

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

Wahrscheinlichkeitsschätzung

$$\tilde{p}(x) = \frac{n(x)}{N}$$

Die **Relative Häufigkeit** $n(x)/N$ ist die Zahl der Vorkommen (*counts*) $n(x)$ eines Ereignisses x geteilt durch die Stichprobengröße n .

Für zunehmende Stichprobengröße n , konvergiert die relative Häufigkeit zu der tatsächlichen Wahrscheinlichkeit eines Ereignisses.

genauer: Die Wahrscheinlichkeit, dass die relative Häufigkeit um mehr als ϵ von der tatsächlichen Wahrscheinlichkeit abweicht, konvergiert für zunehmende Stichprobengröße gegen 0.

Wahrscheinlichkeitsschätzung durch relative Häufigkeit

Beispiel:

- ▶ Zufallsereignis: Wortvorkommen ist ein bestimmtes Wort
- ▶ $n(x)$: Anzahl der Vorkommen (*counts*) des Wortes in einem Corpus
- ▶ N : Anzahl aller Wortvorkommen im Corpus.

Wort	$n(\text{Wort})$	$\tilde{p}(\text{Wort})$
meet		
deadline		
single		
...		

hot
stock
tip

reminder
deadline
meet
thanks

meet
hot
single

thanks
for
tip

deadline
approaching

Wahrscheinlichkeitsschätzung durch relative Häufigkeit

Beispiel:

- ▶ Zufallsereignis: Wortvorkommen ist ein bestimmtes Wort
- ▶ $n(x)$: Anzahl der Vorkommen (*counts*) des Wortes in einem Corpus
- ▶ N : Anzahl aller Wortvorkommen im Corpus.

Wort	$n(\text{Wort})$	$\tilde{p}(\text{Wort})$
meet	2	$\frac{2}{15} \approx 0.133$
deadline	2	$\frac{2}{15} \approx 0.133$
single	1	$\frac{1}{15} \approx 0.067$
...		

hot
stock
tip

reminder
deadline
meet
thanks

meet
hot
single

thanks
for
tip

deadline
approaching

Relative Häufigkeit für bedingte Wahrscheinlichkeiten

$$\tilde{p}(x|y) = \frac{n(x, y)}{n_y}$$

Auch bedingte Wahrscheinlichkeiten können anhand von relativen Häufigkeiten geschätzt werden.

$n(x, y)$ ist hier die Zahl der gemeinsamen Vorkommen der Ereignisse x und y .

n_y ist die Anzahl aller Vorkommen des Ereignisses y .

Es gilt: $n_y = \sum_{x'} n(x', y)$

Relative Häufigkeit für bedingte Wahrscheinlichkeiten

- ▶ Zufallsereignis x : Wortvorkommen ist ein bestimmtes Wort
- ▶ Zufallsereignis y : Wortvorkommen ist in Email einer bestimmten Kategorie, z.B. HAM oder SPAM (HAM= "kein Spam")
- ▶ $n(x, y)$: Anzahl der Wortvorkommen in Emails einer Kategorie im Corpus

Wort	$n(\text{Wort, HAM})$	$\tilde{p}(\text{Wort} \text{HAM})$	$n(\text{Wort, SPAM})$	$\tilde{p}(\text{Wort} \text{SPAM})$
meet				
deadline				
single				
...				

reminder
deadline
meet
thanks

thanks
for
tip

deadline
approaching

hot
stock
tip

meet
hot
single

Relative Häufigkeit für bedingte Wahrscheinlichkeiten

- ▶ Zufallsereignis x : Wortvorkommen ist ein bestimmtes Wort
- ▶ Zufallsereignis y : Wortvorkommen ist in Email einer bestimmten Kategorie, z.B. HAM oder SPAM (HAM= "kein Spam")
- ▶ $n(x, y)$: Anzahl der Wortvorkommen in Emails einer Kategorie im Corpus

Wort	$n(\text{Wort, HAM})$	$\tilde{p}(\text{Wort} \text{HAM})$	$n(\text{Wort, SPAM})$	$\tilde{p}(\text{Wort} \text{SPAM})$
meet	1	$\frac{1}{9} \approx 0.111$	1	$\frac{1}{6} \approx 0.167$
deadline	2	$\frac{2}{9} \approx 0.222$	0	0
single	0	0	1	$\frac{1}{6} \approx 0.167$
...				

reminder
deadline
meet
thanks

thanks
for
tip

deadline
approaching

hot
stock
tip

meet
hot
single

Wahrscheinlichkeit für Wortsequenz

- ▶ Soweit haben wir nur Wahrscheinlichkeiten von Einzelwörtern ausgedrückt und diese geschätzt.
- ▶ Wie können wir die Wahrscheinlichkeiten von ganzen Texten (z.B. Emails) berechnen?
- ▶ Anwendung der bedingten Wahrscheinlichkeit:

$$P(w_1, w_2, \dots, w_n)$$

$$= P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \dots P(w_n|w_1 \dots w_{n-1})$$

- ▶ \Rightarrow löst das Problem nicht wirklich, denn $P(w_n|w_1 \dots w_{n-1})$ kann nicht gut geschätzt werden

Unabhängigkeitsannahme: Bag of Words

- ▶ Eine Lösung: Wir machen die statistische Annahme, dass jedes Wort unabhängig vom Vorkommen anderer Wörter ist.
- ▶ Dies nennt man auch Bag-of-words (BOW) Annahme, weil die Reihenfolge der Wörter irrelevant wird.

$$\begin{aligned} & P(w_1, w_2, \dots, w_n) \\ &= P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \dots P(w_n|w_1 \dots w_{n-1}) \\ & \quad \underset{\text{Unabh.}}{=} P(w_1)P(w_2)P(w_3) \dots P(w_n) \end{aligned}$$

Bedingte Unabhängigkeit

- ▶ Für viele Machine-Learning Algorithmen ist **bedingte Unabhängigkeit** das zentrale Konzept:
Wenn der Wert einer Zufallsvariable y bekannt ist, sind Zufallsvariablen x_1, \dots, x_n unabhängig
- ▶ Mittelweg zwischen:
 - ▶ Keine Unabhängigkeit
 - ▶ Unabhängigkeit aller Zufallsvariablen
- ▶ In unserem Fall:

$$\begin{aligned} &P(w_1, w_2, \dots, w_n | \text{SPAM}) \\ \text{bed. Unabh.} &= P(w_1 | \text{SPAM}) P(w_2 | \text{SPAM}) \dots P(w_n | \text{SPAM}) \end{aligned}$$