Machine Learning

Prof. Dr. Fabian Brunner
<fa.brunner@oth-aw.de>

Amberg, 3. November 2020

Übersicht

Wiederholung von Grundbegriffen der Stochastik

- Zufallsexperiment und Wahrscheinlichkeitsraum
- Stochastische Unabhängigkeit von Ereignissen
- Zufallsvariablen
- Erwartungswert und Varianz
- Bedingte Wahrscheinlichkeit und bedingter Erwartungswert

Weiterführendes Material

Vorabbemerkung:

- Auf den folgenden Folien sind einige Definitionen und Aussagen aus der Stochastik zusammengestellt, die im weiteren Verlauf der Vorlesung „Machine Learning“ benötigt werden.
- Die Lektüre dieser Folien kann ein eingehendes Studium der Materie nicht ersetzen.
- Es wird daher dringend empfohlen, als Begleittext ein Stochastik-Buch heranzuziehen.

Literaturhinweise

- K. Bosch: Elementare Einführung in die Wahrscheinlichkeitsrechnung. Vieweg Verlag.
- K. Bosch: Elementare Einführung in die angewandte Statistik. Vieweg-Verlag.
- C. Dietmaier: Mathematik für angewandte Wissenschaften. Springer-Verlag.
- G. Fischer, M. Lehner, A. Puchert: Einführung in die Stochastik. Springer Spektrum.

Grundbegriffe

Grundraum, Ergebnis, Ereignis

Ein **Ergebnis** ω ist ein möglicher Ausgang eines Zufallsexperiments. Der **Grundraum** Ω ist die Menge aller möglichen Ergebnisse eines Zufallsexperiments. Ein Ereignis $A \subset \Omega$ ist eine Teilmenge des Grundraums, d.h. eine Menge gewisser Ergebnisse.

Beispiel: Zweifacher Würfelwurf

- Grundraum: $\Omega = \{(1,1),(1,2),(1,3),\ldots(6,5),(6,6)\}$, $|\Omega| = 36$.
- Elementarereignis „1 und 5“: $\omega = \{(1,5)\}$
- Ereignis „Pasch“: $A = \{(1,1),(2,2),\ldots,(6,6)\}$.

Grundbegriffe

Wahrscheinlichkeitsmaß

Sei Ω ein nichtleerer Grundraum und Σ eine Ereignis-Sigma-Algebra in Ω (d.h. eines bestimmten Systems von Teilmengen von Ω). Eine Abbildung $P : \Sigma \to [0,1]$ heißt **Wahrscheinlichkeitsmaß**, wenn folgende Eigenschaften erfüllt sind:

1. $P(\Omega) = 1$
2. Für paarweise disjunkte Mengen $A_1, A_2, A_3, \ldots \in \Sigma$ gilt stets

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i) \quad \text{(Sigma-Additivität)} .$$

Das Tripel (Ω, Σ, P) heißt Wahrscheinlichkeitsraum. Die Zahl $P(A)$ heißt Wahrscheinlichkeit des Ereignisses A.

Beispiel: Laplace-Verteilung

- Voraussetzung: jedes Ergebnis ist gleich wahrscheinlich.
- Laplace-Wahrscheinlichkeit: $P(A) = \frac{|A|}{|\Omega|}$ für alle $A \in \mathcal{P}(\Omega)$.

Grundbegriffe

Bedingte Wahrscheinlichkeit

Sei (Ω, Σ, P) ein Wahrscheinlichkeitsraum und sei $B \in \Sigma$ ein Ereignis mit $P(B) > 0$. Für jedes Ereignis $A \in \Sigma$ heißt

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

die **bedingte Wahrscheinlichkeit für** A **unter der Bedingung** B. Die Abb.

$$P(\cdot|B) : \Sigma \to [0, 1], \quad A \mapsto P(A|B)$$

heißt **bedingte Verteilung unter der Bedingung** B.

Grundbegriffe

Unabhängigkeit von Ereignissen

Zwei Ereignisse A und B heißen **unabhängig**, falls

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

Bemerkungen:

- Gilt $P(B) > 0$, so ist die obige Bedingung gleichbedeutend mit

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

- Wenn die Ereignisse nicht unabhängig sind, können wir aus dem einen etwas über das andere lernen.
- Können disjunkte Ereignisse unabhängig sein?

Grundbegriffe

Totale Wahrscheinlichkeit

Sei $\Omega = B_1 \cup \ldots \cup B_n$ eine Zerlegung in paarweise disjunkte Ereignisse und sei $P(B_j) > 0$ für alle $j = 1, \ldots, n$. Dann gilt für jedes Ereignis A:

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i) \ .$$

Formel von Bayes

Seien A und B Ereignisse mit $P(A) > 0$ und $P(B) > 0$. Dann gilt

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

Unter den Voraussetzungen des Satzes von der totalen Wahrscheinlichkeit gilt ferner

$$P(B_i|A) = \frac{P(A|B_i) \cdot P(B_i)}{\sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)} \ .$$

Grundbegriffe

Zufallsvariable

Sei (Ω, Σ, P) ein Wahrscheinlichkeitsraum. Eine **Zufallsvariable** ist eine Abbildung $X : \Omega \to \mathbb{R}$ mit der Eigenschaft $X^{-1}(A) \in \Sigma$ für alle $A \in \mathcal{B}(\mathbb{R})$. Die Zufallsvariable induziert durch

$$P_X(A) := P(X \in A) = P(X^{-1}(A)) = P(\{\omega \in \Omega \,:\, X(\omega) \in A\}) \text{ für alle } A \in \mathcal{B}(\mathbb{R})$$

ein Wahrscheinlichkeitsmaß auf $\mathcal{B}(\mathbb{R})$. Man nennt P_X die **Wahrscheinlichkeitsverteilung von X (unter P)**.

In der obigen Definition bezeichnet $\mathcal{B}(\mathbb{R})$ die Borel-Mengen von $\mathbb{R}$, d.h. ein System von Teilmengen, welches die für uns relevanten Teilmengen (z.B. Intervalle) enthält.

Beispiel

- Beim zweifachen Würfelwurf ist die Augensumme X eine Zufallsvariable.
- Ist $\omega = \{(\omega_0, \omega_1)\} \in \{1, 2, \ldots, 6\} \times \{1, 2, \ldots, 6\}$ ein Elementarereignis, so gilt

$$X(\omega) = \omega_0 + \omega_1 \ .$$

Diskrete und stetige Zufallsvariablen

Diskrete Zufallsvariable

- Eine Zufallsvariable heißt diskret, falls sie eine endliche oder abzählbare Wertemenge $W = \{x_1, x_2, \ldots\}$ besitzt.
- Die Funktion

$$f(x) = \begin{cases} P(X = x) & \text{für } x \in W \\ 0 & \text{sonst} \end{cases}$$

  heißt Wahrscheinlichkeitsfunktion von X.

Stetige Zufallsvariable

Eine Zufallsvariable X heißt stetig, falls es eine Funktion f_X gibt mit:

$$P(a < X \leq b) = \int_a^b f_X(x)dx \qquad \text{für alle } a, b \in \mathbb{R} \text{ mit } a < b \ .$$

Die Funktion f_X heißt **Wahrscheinlichkeitsdichte** (oder nur **Dichte**) von X.

Verteilungsfunktion

Kumulative Verteilungsfunktion

Sei (Ω, Σ, P) ein Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine Zufallsvariable. Die Funktion

$$F_X : \mathbb{R} \to \mathbb{R}, x \mapsto P_X(]-\infty; x]) = P(X \leq x)$$

heißt **(kumulative) Verteilungsfunktion** von X.

Eigenschaften der Verteilungsfunktion:

- $0 \leq F_X(x) \leq 1$ für alle x
- F_X ist monoton wachsend.
- $\lim_{x\to\infty} F_X(x) = 1$ und $\lim_{x\to-\infty} F_X(x) = 0$.
- F_X ist rechtsseitig stetig.
- Für die Verteilungsfunktion F_X einer stetigen Zufallsvariable X gilt

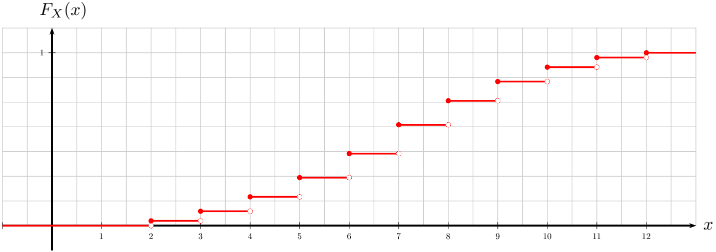
$$F_X(x) = \int_{-\infty}^{x} f_X(u)du \; .$$

Beispiel zur Verteilungsfunktion

Zweifacher Wurf eines Würfels

$$\Omega = \{(1,1),(1,2),(1,3),\ldots,(6,5),(6,6)\}\ ,$$
$$A = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\} \quad \text{(Ereignis: „Pasch“)}$$
$$X((\omega_1,\omega_2)) := \omega_1 + \omega_2 \quad \text{(Zufallsvariable: Augensumme)}\ ,$$

x_i	1	2	3	4	5	6	7	8	9	10	11	12
$P(X = x_i)$	0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
$F_X(x_i)$	0	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	1



Erwartungswert

Erwartungswert

Der Erwartungswert einer Zufallsvariablen ist gegeben durch

$$X \text{ diskret:} \qquad E(X) := \mu_X := \sum_{i \in \mathbb{N}} x_i P(X = x_i) = \sum_{i \in \mathbb{N}} x_i f_X(x_i) \ ,$$

$$X \text{ stetig:} \qquad E(X) := \mu_X := \int_{-\infty}^{\infty} x f_X(x) dx \ .$$

Eigenschaften und Rechenregeln für den Erwartungswert

1. Linearität: seien X und Y Zufallsvariablen und $a, b \in \mathbb{R}$ Konstanten. Dann gilt
$$E(aX + bY) = aE(X) + bE(Y) \ .$$
2. Nichtnegativität: Gilt $X \geq 0$, dann auch $E(X) \geq 0$.
3. Gilt $X \geq 0$ und $E(X) = 0$, dann folgt $P(X = 0) = 1$.

Varianz und Standardabweichung

Varianz einer Zufallsvariable

Die **Varianz** einer Zufallsvariablen X ist gegeben durch

$$X \text{ diskret}: \qquad \mathrm{Var}(X) := \sigma^2 := \sum_{i \in \mathbb{N}} (x_i - \mu)^2 f_X(x_i) \;,$$

$$X \text{ stetig}: \qquad \mathrm{Var}(X) := \sigma^2 := \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx \;.$$

Die Größe $\sigma = \sqrt{\mathrm{Var}(X)}$ heißt **Standardabweichung** von X.

Eigenschaften der Varianz

1. $\mathrm{Var}(X) \geq 0$
2. Für Konstanten $a, b \in \mathbb{R}$ gilt

$$\mathrm{Var}(aX + b) = a^2 \, \mathrm{Var}(X) \;.$$

3. $\mathrm{Var}(X) = E(X^2) - \mu_X^2 \;.$

Gemeinsam verteilte Zufallsvariablen

Gemeinsame Verteilung, diskreter Fall

Seien X und Y diskrete Zufallsvariablen auf Ω mit Wertevorräten $W_X = \{x_1, x_2, \ldots\}$ und $W_Y = \{y_1, y_2, \ldots\}$. Dann ist die **gemeinsame Wahrscheinlichkeitsfunktion** von X und Y gegeben durch

$$f_{X,Y}(x_i, y_j) = \begin{cases} P(X = x_i, Y = y_j) & \text{falls } (x_i, y_j) \in W_X \times W_Y \ , \\ 0 & \text{sonst} \ . \end{cases}$$

Aus einer gemeinsamen Wahrscheinlichkeitsfunktion kann man auch die Wahrscheinlichkeitsfunktionen für X und Y ableiten (Randverteilungen):

$$f_X(x) = P(X = x) = \sum_{i \in \mathbb{N}} f_{X,Y}(x, y_i) \quad \text{für alle } x \in W_X \ ,$$

$$f_Y(y) = P(Y = y) = \sum_{i \in \mathbb{N}} f_{X,Y}(x_i, y) \quad \text{für alle } y \in W_Y \ .$$

Gemeinsam verteilte Zufallsvariablen

Gemeinsame Verteilung, stetiger Fall

Zwei stetige Zufallsvariablen X und Y heißen gemeinsam verteilt, falls es eine Dichtefunktion $f_{X,Y}$ gibt, sodass

$$P(a < X \leq b, c < Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x,y)\, dy\, dx \quad \text{für } a < b, c < d \;.$$

Die Randverteilungen besitzen die Dichtefunktionen

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy \;,$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx \;.$$

Stochastische Unabhängigkeit

Stochastische Unabhängigkeit

Die Zufallsvariablen X und Y heißen **stochastisch unabhängig**, wenn die gemeinsame Wahrscheinlichkeits- bzw. Dichtefunktion gerade gleich dem Produkt der beiden Randverteilungen ist:

$$
\begin{aligned}
X, Y \text{ diskret}: &\quad f_{X,Y}(x_i, y_j) = f_X(x_i) f_Y(y_j) \quad \text{für alle } (x_i, y_j) \in W_X \times W_Y \ , \\
X, Y \text{ stetig}: &\quad f_{X,Y}(x, y) = f_X(x) f_Y(y) \quad \text{für alle } x, y \in \mathbb{R} \ .
\end{aligned}
$$

Bemerkung: Sind X und Y unabhängig, dann gilt

$$E(XY) = E(X)E(Y) \ .$$

Beispiel zur stochastischen Unabhängigkeit

Die gemeinsame Dichtefunktion der Zufallsvariablen X und Y laute

$$f_{X,Y}(x,y) = \begin{cases} 6xy^2 & \text{falls } 0 \leq x, y \leq 1 \ , \\ 0 & \text{sonst} \ . \end{cases}$$

Sind X und Y stochastisch unabhängig?

Lösung:

- Für $0 \leq x \leq 1$ gilt

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_0^1 6xy^2 dy = 6x \int_0^1 y^2 dy = 6x \left[\frac{y^3}{3}\right]_0^1 = 2x \ .$$

Beispiel zur stochastischen Unabhängigkeit

Die gemeinsame Dichtefunktion der Zufallsvariablen X und Y laute

$$f_{X,Y}(x,y) = \begin{cases} 6xy^2 & \text{falls } 0 \le x, y \le 1 \ , \\ 0 & \text{sonst} \ . \end{cases}$$

Sind X und Y stochastisch unabhängig?

Lösung:

- Für $0 \le x \le 1$ gilt

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_0^1 6xy^2 dy = 6x \int_0^1 y^2 dy = 6x \left[\frac{y^3}{3}\right]_0^1 = 2x \ .$$

- Für $0 \le y \le 1$ gilt

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = \int_0^1 6xy^2 dx = 6y^2 \int_0^1 x dx = 6y^2 \left[\frac{x^2}{2}\right]_0^1 = 3y^2 \ .$$

Beispiel zur stochastischen Unabhängigkeit

Die gemeinsame Dichtefunktion der Zufallsvariablen X und Y laute

$$f_{X,Y}(x,y) = \begin{cases} 6xy^2 & \text{falls } 0 \leq x, y \leq 1 \ , \\ 0 & \text{sonst} \ . \end{cases}$$

Sind X und Y stochastisch unabhängig?

Lösung:

- Für $0 \leq x \leq 1$ gilt

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_0^1 6xy^2 dy = 6x \int_0^1 y^2 dy = 6x \left[\frac{y^3}{3}\right]_0^1 = 2x \ .$$

- Für $0 \leq y \leq 1$ gilt

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = \int_0^1 6xy^2 dx = 6y^2 \int_0^1 xdx = 6y^2 \left[\frac{x^2}{2}\right]_0^1 = 3y^2 \ .$$

- Man erhält also

$$f_X(x) = \begin{cases} 2x & \text{falls } 0 \leq x \leq 1 \\ 0 & \text{sonst} \end{cases} \qquad \text{und} \qquad f_Y(y) = \begin{cases} 3y^2 & \text{falls } 0 \leq y \leq 1 \\ 0 & \text{sonst} \ . \end{cases}$$

- Daraus folgt, dass X und Y unabhängig sind, denn:

$$f_X(x)f_Y(y) = f_{X,Y}(x,y) = \begin{cases} 6xy^2 & \text{falls } 0 \leq x, y \leq 1 \\ 0 & \text{sonst} \ . \end{cases}$$

Erwartungswerte bei mehreren Zufallsvariablen

Sind X, Y zwei Zufallsvariablen und sei $g : \mathbb{R}^2 \to \mathbb{R}$ eine reelle Funktion. Dann ist der Erwartungswert von $g(X, Y)$ gegeben durch

$$X \text{ diskret:} \qquad E(g(X,Y)) = \sum_{i,j \in \mathbb{N}} g(x_i, y_j) f_{X,Y}(x_i, y_j) \ ,$$

$$X \text{ stetig:} \qquad E(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\, dy\, dx \ .$$

Beispiel:
Sei X die Augensumme und Y der Betrag der Differenz der Augen beim zweifachen Würfeln. Dann gilt

$$\begin{aligned} E(X \cdot Y) = {} & 3 \cdot \frac{1}{18} + 5 \cdot \frac{1}{18} + 7 \cdot \frac{1}{18} + 9 \cdot \frac{1}{18} + 11 \cdot \frac{1}{18} \\ & + 8 \cdot \frac{1}{18} + 12 \cdot \frac{1}{18} + 16 \cdot \frac{1}{18} + 20 \cdot \frac{1}{18} + \\ & + 15 \cdot \frac{1}{18} + 21 \cdot \frac{1}{18} + 27 \cdot \frac{1}{18} \\ & + 24 \cdot \frac{1}{18} + 32 \cdot \frac{1}{18} + \frac{35}{18} = \frac{245}{18} \ . \end{aligned}$$

Bedingter Erwartungswert

Bedingter Erwartungswert, stetiger Fall

Sei (Ω, Σ, P) ein Wahrscheinlichkeitsraum und seien $X, Y : \Omega \to \mathbb{R}$ zwei stetige Zufallsvariablen mit gemeinsamer Dichte $f(x, y)$ sodass $\int_{-\infty}^{\infty} |y| f_Y(y) dy < \infty$. Sei ferner $f_X(x) = \int_{\mathbb{R}} f(x, y)\, dy$ die Randdichte von X und

$$f(y|x) = \frac{f(x, y)}{f_X(x)}$$

die bedingte Dichte von Y gegeben $X = x$.
Dann definiert der Ausdruck

$$E(Y|X = x) := \int_{\mathbb{R}} y f(y|x)\, dy$$

den **bedingten Erwartungswert** von Y gegeben $X = x$.

Für den diskreten Fall erfolgt die Definition analog mit der Wahrscheinlichkeitsfunktion.

Beispiel zum bedingten Erwartungswert

Es wird das (unabhängige) Werfen zweier idealer Würfel betrachtet. Sei X die Augenzahl des ersten Würfels und Y die Augensumme beider Würfel.

a) Man bestimme $E(Y)$.

b) Man bestimme $E(Y|X = 1)$.

Lösung:

a) $E(Y) = 7$.

b) $E(Y|X = 1) = 4.5$.