Moodle PW: ajmvs

Elements of Statistical Learning

Mehrdimensionale Verteilungen

Def.: $X_i: (\Omega, \mathcal{A}, P) \to \mathbb{R}$ reelle ZV, $i = 1, \dots, p$

Dann ist $X = (X_1, \dots, X_p)^T$ mit $X(\omega) := (X_1(\omega), \dots, X_p(\omega))^T$

eine mehrdim. ZV.

Die Funktion $F_X : \mathbb{R}^p \to [0,1]$ mit $F_X(x_1, \dots, x_p) := P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_p \leq x_p)$

heißt multivar. Verteilungsfkt.

Existiert eine Funktion $f_X: \mathbb{R}^p \to \mathbb{R}^+$ mit

$$F_X(x_1, \dots, x_p) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_3} f(t_1, \dots, t_p)\, dt_p\, dt_{p-1} \dots dt_1$$

dann heißt f Dichte von X.

Die Funktion $F_{X_i} : \mathbb{R} \to [0,1]$ mit $F_{X_i}(x_i) = P(X_i \leq x_i)$

$$= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{x_i} \dots \int_{-\infty}^{\infty} f(t_1, \dots, t_p)\, dt_1 \dots dt_p$$

Def.: $X_1, \dots, X_n$ unabh. $\Leftrightarrow F(x_1, \dots, x_n) = \prod_{i=1}^{n} F(x_i) \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n$

Satz: Sind $X_1, \dots, X_n$ stetig mit Dichten $f_1, \dots, f_n$, dann gilt

$$X_1, \dots, X_n \text{ unabh.} \Leftrightarrow f_X(x_1, \dots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n \quad (\text{mit } X = (X_1, \dots, X_n))$$

Beweis: $F(x_1, \dots, x_n) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_n} f_X(t_1, \dots, t_n)\, dt_n\, dt_{n-1} \dots t_1$

$$= \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} f_{X_1}(t_1) \cdot \dots \cdot f_{X_n}(t_n)\, dt_n \dots dt_1$$

$$= \int_{-\infty}^{x_1} f_{X_1}(t_1) \cdot \dots \cdot \int_{-\infty}^{x_n} f_{X_n}(t_n)\, dt_n \dots dt_1$$

$$= \int_{-\infty}^{x_1} f_{X_1}(t_1) \cdot \ldots \cdot \int_{-\infty}^{x_n} f_{X_n}(t_n)\, dt_n \ldots dt_1$$

$$= F_{X_1}(x_1) \cdot \ldots \cdot F_{X_n}(x_n)$$

Bedingte Verteilungen

Bsp.: Blutzucker $= b_0 + b_1 \cdot HbA1c + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$ und unabh. von X

$$Y = b_0 + b_1 \cdot X + \varepsilon$$
$$= -10 + 3 \cdot X + \varepsilon$$

$$E(Y) = E(b_0 + b_1 \cdot X + \varepsilon) = b_0 + b_1 \cdot E(X) + 0 = -10 + 3 \cdot 60 = 170$$

$E(Y)$ für Patient mit $HbA1c = 70$?

$$\Rightarrow \underbrace{E(Y \mid X=70)}_{ZV?} = -10 + 3 \cdot 70 + 0 = 200$$

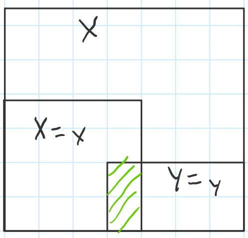
Def. (bedingte Verteilungen):

Diskret: (X,Y) diskret, dann definiert für jedes y mit $P(Y=y) > 0$

$$\boxed{P(X=x \mid Y=y) = \frac{P(X=x, Y=y)}{P(Y=y)}}$$ die bedingte Verteilung von X bedingt auf $Y=y$.

Wir schreiben $X \mid Y=y$ für die durch $P(X=x \mid Y=y)$ definierte ZV.

$$E(X \mid Y=y) = \sum_{k=1}^{\infty} k \cdot P(X=k \mid Y=y)$$



Stetig: (X,Y) stetig mit gem. Dichte $f(x,y)$ und Randdichten $f_X(x)$ und $f_Y(y)$. Dann definiert mit $f_Y(y) > 0$ die bedingte Dichte

$$f(x,y)$$

bedingte Dichte

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

die bedingte Verteilung von X bed. auf $Y=y$.

$$E(X|Y=y) = \int x \cdot f_X(x|y)\,dx = \int x \cdot \frac{f(x,y)}{f_Y(y)}\,dx$$

Rechenregeln für bedingte Erwartungswerte

$$E(X+Y \mid Z=z) = E(X|Z=z) + E(Y|Z=z)$$

$$E(X|X=x) = x$$

$$E(X|Y=y) = E(X), \text{ falls } X, Y \text{ unabh.}$$

Bemerkung: $E(X|Y)$ ist eine ZV ("Erwartungswert von X für jedes beliebige Y")

$$\Rightarrow E(E(X|Y)) = E(X)$$

Satz: X, Y unabh. $\Leftrightarrow$ $f(x|y) = f_X(x)$ bzw. $P(X=x|Y=y)$

$$= P(X=x) \quad \forall x, y$$