Wahrscheinlichkeitsrechnung

Irene Winkler & Nico Görnitz

Arbeitsgruppe Maschinelles Lernen

Bsp: Linear Regression mal anders

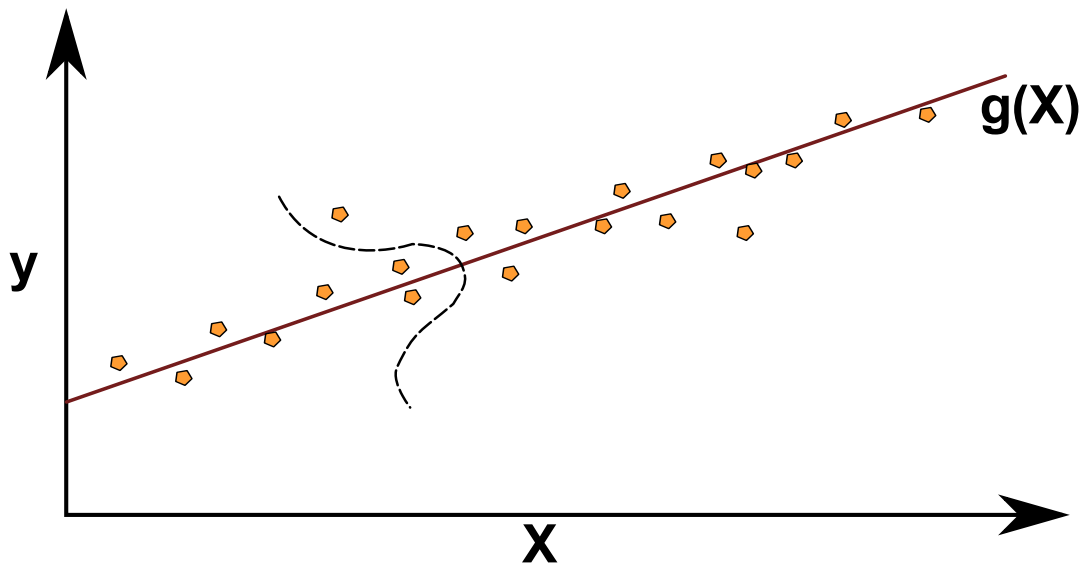
Unser Model beinhaltet nun eine explizite Rausch-Annahme:

$$Y = g_w(X) + \epsilon \quad \text{und} \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2)$$

dann folgt $Y - g_w(X) = \epsilon$ und $(Y - g_w(X)) \sim \mathcal{N}(0, \sigma_n^2)$. Das heisst

$$p_w(Y|X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(Y - g_w(X))^2}{2\sigma_n^2}\}$$

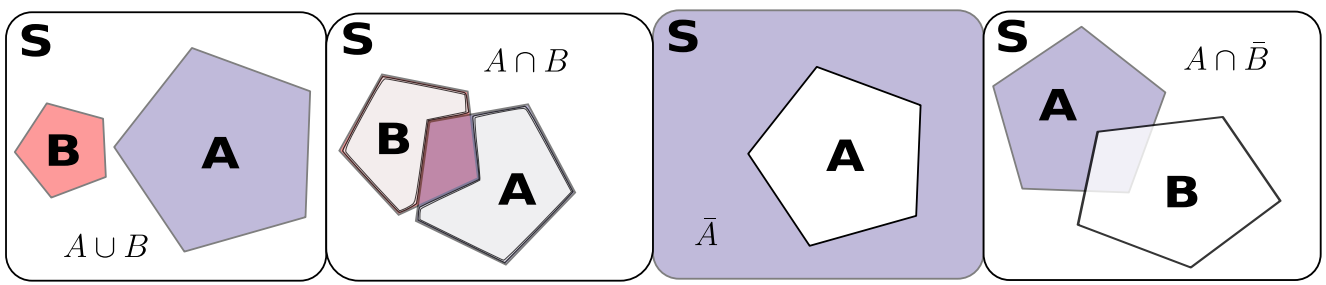
Und wir wollen $\mathbf{w}^* = \text{argmax}_{\mathbf{w}}\, p_w(Y|X) = \prod_{i=1}^n p_{\mathbf{w}}(y_i|x_i)$ unter der Annahme, dass alle Datenpunkte unabhaengig sind. Die negative log-likelihood ist $\mathbf{w}^* = \text{argmin}_{\mathbf{w}}\, \frac{1}{2} \sum_i (y_i - g_{\mathbf{w}}(x_i))^2$. Wenn wir unsere Modelfunktion nun wieder $g_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$ setzen...



Axiome der Wahrscheinlichkeit

"nothing but common sense reduced to calculation." (Laplace)

- Eine Experiment, welches unter identischen Bedingungen verschiedene Ergebnisse hat, heisst *Zufallsexperiment*
- Ein einzelnes Ergebnis eines Zufallsexperiments heisst *Elementarereignis*
- Eine Menge S die aus allen möglichen Elementarereignissen eines Zufallsexperiments besteht, heisst *Ereignisraum*.



Axiome der Wahrscheinlichkeit

Axiom 1 Für jedes Ereignis A gilt: $P(A) \geq 0$

Axiom 2 Für das sichere Ereignis S gilt: $P(S) = 1$

Axiom 3 Für alle disjunkten Ereignisse A_i gilt:
$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$

Daraus lässt sich nun folgern:

- Sei $A_1 \subset A_2$, dann $P(A_1) \leq P(A_2)$ oder $P(A_2 - A_1) = P(A_2) - P(A_1)$
- Für jedes Ereignis A gilt: $0 \leq P(A) \leq 1$
- Das unmögliche Ereignis hat $P(\emptyset) = 0$
- Für das komplementäre Ereignis, gilt: $P(\bar{A}) = 1 - P(A)$
- Sei $A = A_1 \cup A_2 \cup \ldots$ mit disjunkten Teilmengen A_i, dann gilt:
  $P(A) = P(A_1) + P(A_2) + \ldots$, wenn $A = S$ ist, dann gilt $P(A) = 1$
- Für allgemeine Ereignisse A und B gilt:
  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Bedingte Wahrscheinlichkeit & Unabhängigkeit

- Die Wahrscheinlichkeit von B, wenn A eingetreten ist:
  $P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \Leftrightarrow \quad P(A \cap B) = P(A)P(B|A)$
- Es gilt $P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B)$
- Sollte $P(B|A)$ gar nicht von A abhängen, dann gilt
  $P(B|A) = P(B)$ und somit $P(A \cap B) = P(A)P(B)$

Summen-, Produkt- & Bayesregel

- Summenregel:

$$P(A) = \sum_j P(A|A_j)P(A_j)$$

- Produktregel:

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

- Bayesregel (2 x Anwendung Produktregel):

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

Kombinatorik

...ist die intelligente Art zu zählen.

- *Permutation*: Wenn man n Objekte hat, von denen man r angeordnet haben möchte, so gibt es $P_{n,r} = n(n-1)(n-2)\dots(n-r+1) = \frac{n!}{(n-r)!}$ Möglichkeiten, dies zu tun. Insbesondere ist $P_{n,n} = n!$.
- Wenn die Reihenfolge keine Rolle spielt (also wenn abc das Gleiche sein soll, wie acb oder bca), dann gibt es $C_{n,r} = \binom{n}{r}$ Möglichkeiten.

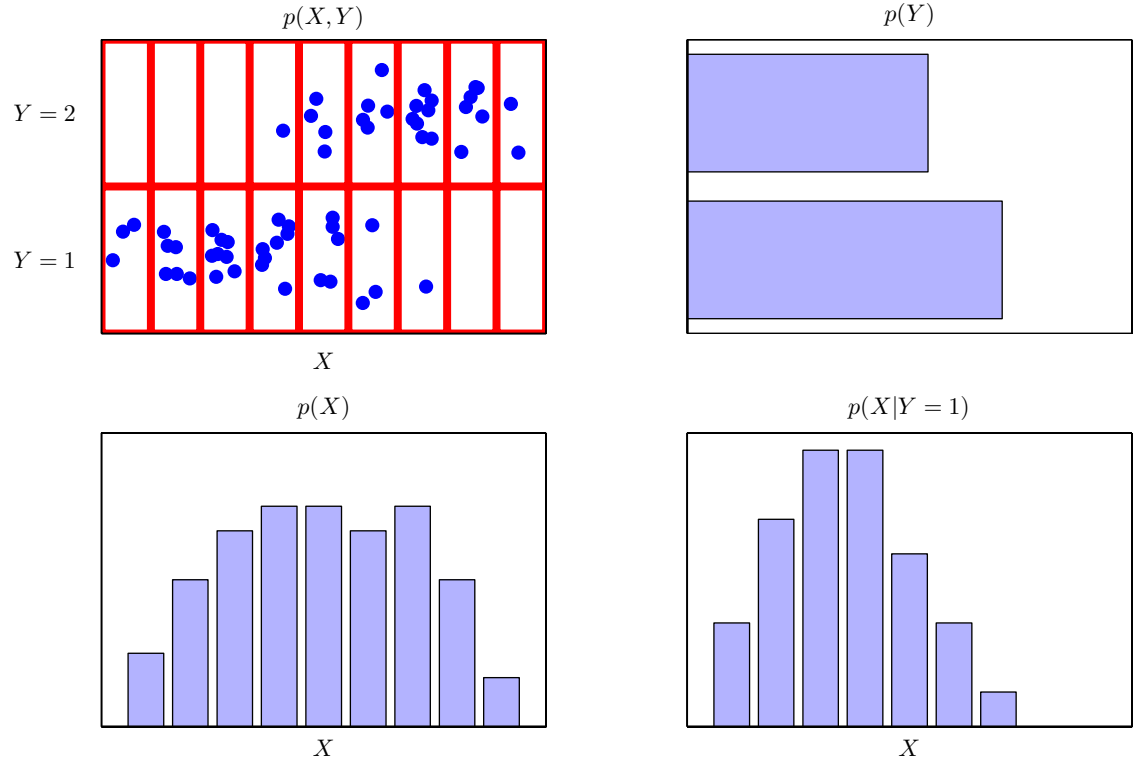
Zufallsvariablen

Wenn man jedem Punkt im Ereignisraum eine Zahl zuweisen würde, so hätte man eine Funktion definiert. Diese Funktion heisst *Zufallsvariable* (zutreffender wäre Zufallsfunktion).

Diskrete Zufallsvariablen

$X \in \{x_1, \ldots, x_n\}$, $Y \in \{y_1, \ldots, y_m\}$ diskrete Zufallsvariablen

- Randverteilung $P(X)$, $P(Y)$
- Gemeinsame Verteilung $P(X, Y)$
- Bedingte bzw. Posteriorverteilung $P(X|Y), P(Y|X)$



Diskrete Zufallsvariablen

- Summenregel:

$$P(X = x) = \sum_{y \in Y} P(X = x, Y = y)$$

- Produktregel:

$$P(X = x, Y = y) = P(Y = y | X = x) P(X = x)$$

- Bayesregel (2 x Anwendung Produktregel):

$$P(Y = y | X = x) = \frac{P(X = x | Y = y) P(Y = y)}{P(X = x)}$$

Kontinuierliche Zufallsvariablen

Sei $X \in \mathbb{R}$ eine reellwertige Zufallsvariable. Eine Funktion $p : \mathbb{R} \mapsto \mathbb{R}$ heisst *Dichte* von X falls für $a < b$

$$P(X \in [a, b]) = \int_a^b p(x)dx$$

Für jede Dichte gilt

$$\int_{-\infty}^{\infty} p(x)dx = 1 \qquad \text{und} \qquad p(x) \geq 0, \forall x \in \mathbb{R}$$

Die (kumulative) *Verteilungsfunktion* $F_X : \mathbb{R} \mapsto \mathbb{R}$ ist definiert durch

$$F_X(c) = P(X \leq c) = \int_{-\infty}^{c} p(x)dx$$

Es gilt

$$0 \leq F_X(c) \leq 1$$

$$P(X \in [a, b]) = F(b) - F(a)$$

Kontinuierliche Zufallsvariablen

Veallgemeinerung für mehrdimensionale Zufallsvariable $X \in \mathbb{R}^d$:

- $p : \mathbb{R}^d \mapsto \mathbb{R}$ Dichte von X falls

$$P(X \in [a_1, b_1] \times \cdots \times [a_d, b_d]) = \int_{a_1}^{b_1} \cdots \int_{a_d}^{b_d} p(\mathbf{x}) d\mathbf{x}$$

Für jede Dichte gilt

$$\int_{\mathbb{R}^d} p(\mathbf{x}) d\mathbf{x} = 1 \qquad \text{und} \qquad p(\mathbf{x}) \geq 0, \forall \mathbf{x} \in \mathbb{R}^d$$

Kontinuierliche Zufallsvariablen

Seien $X \in \mathbb{R}^d, Y \in \mathbb{R}^l$ zwei Zufallsvariablen mit gemeinsamer Dichte p, und den Randverteilungen p_X und p_Y.

- Summenregel:

$$p_X(\mathbf{x}) = \int_{\mathbb{R}^l} p(\mathbf{x}, \mathbf{y}) d\mathbf{y}$$

- Produktregel:

$$p(\mathbf{x}, \mathbf{y}) = p_{Y|X=\mathbf{x}}(\mathbf{y}) p(\mathbf{x})$$

- Bayesregel:

$$p_{Y|X=\mathbf{x}}(\mathbf{y}) = \frac{p_{X|Y=\mathbf{y}}(\mathbf{x}) p_Y(\mathbf{y})}{p_X(\mathbf{x})}$$

Unabhängigkeit

Zwei Zufallsvariablen X und Y sind *unabhängig* wenn

- (Im diskreten:)

$$P(X = \mathbf{x}, Y = \mathbf{y}) = P(X = \mathbf{x})P(Y = \mathbf{y})$$

- (Im kontinuierlichen:)

$$p(\mathbf{x}, \mathbf{y}) = p_X(\mathbf{x}) \cdot p_Y(\mathbf{y})$$

Mode & Median

- *Mode*: wird der wahrscheinlichste Wert/Belegung einer Zufallsvariablen genannt. Sollten es 2/3/viele Werte sein, die die höchsten Wahrscheinlichkeiten haben, so heisst die Verteilung bi-/tri-/multimodal.
- *Median*: der Wert x für den gilt, dass $P(X < x) \leq 0.5$ und $P(X > x) \leq 0.5$

Erwartungswert

= mittlerer Wert der Funktion einer Zufallsvariablen $f(X)$

- Ist X eine diskrete Zufallsvariable so ist der Erwartungswert

$$\mathbb{E}(f(X)) := \sum_x f(\mathbf{x}) P(X = \mathbf{x})$$

- Ist $X \in \mathbb{R}^d$ eine kontinuierliche Zufallsvariable so ist der Erwartungswert

$$\mathbb{E}(f(X)) := \int_{\mathbb{R}^d} f(\mathbf{x}) p(\mathbf{x}) dx$$

- Der Erwartungswert ist linear:

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$$

Varianz & Standartabweichung

Seien X und Y zwei eindimensionale Zufallsvariablen.

- Die *Varianz* von X ist definiert als

$$\mathrm{Var}(X) := \mathbb{E}[(X - \mathbb{E}(X))^2]$$

- Die *Standardabweichung* von X ist definiert als

$$Std(X) := \sqrt{\mathrm{Var}(X)}$$

Wichtiges Bsp: Die Normalverteilung

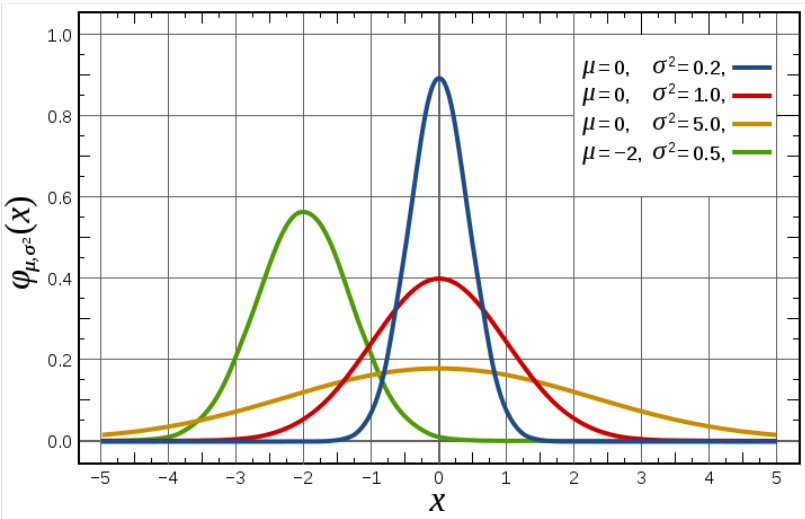
Eine Zufallsvariable $X \in \mathbb{R}$ ist normalverteilt, $X \sim \mathcal{N}(\mu, \sigma^2)$, wenn

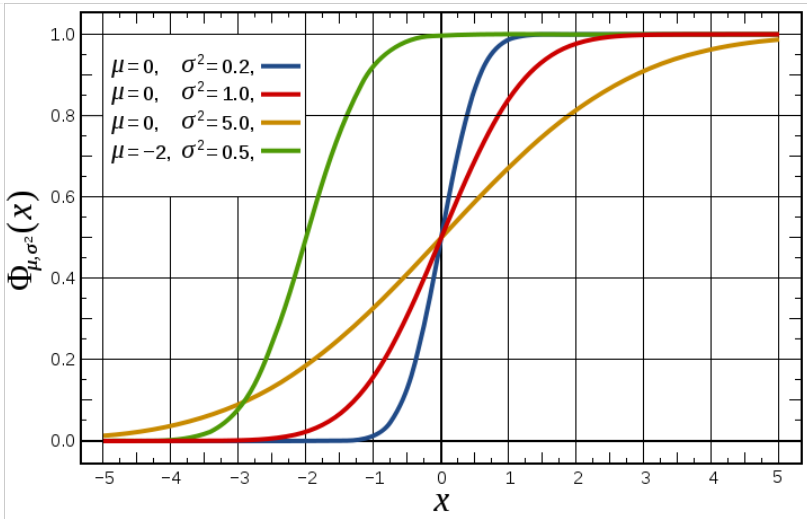
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

ihre Dichtefunktion ist.

$$\mathbb{E}(X) = \mu$$

$$\mathrm{Var}(X) = \sigma^2$$





Moments & Moment Generating Function

Das r-te Moment einer Zufallsvariablen X mit Mittelwert μ ist definiert als:

$$\mu_r := E[(X - \mu)^r]$$

Die Moment Generating Function von X ist definiert als (wenn die Reihe konvergiert):

$$M_X(t) := E[e^{tX}]$$

Die r-te Ableitung der Funktion an der Stelle $t = 0$ liefert das r-te Moment (deswegen der Name):

$$\mu_r = \frac{d^r}{dt^r} M_X(t)|_{t=0}$$

Skewness & Kurtosis

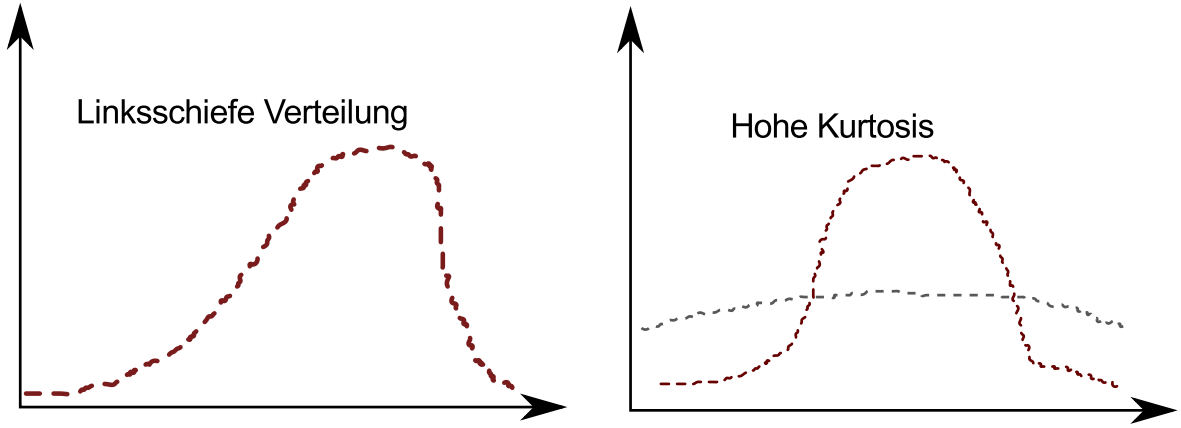
- *Skewness*: Wenn eine Verteilung nicht symmetrisch ist, sagt diese Zahl was über darüber aus, ob sie zur linken oder rechten Seite hin sich neigt:

$$\alpha_3 = \frac{\mathbb{E}((X-\mu)^3)}{\sigma^3}$$

- *Kurtosis*: Ein Mass für die Stärke des Peaks in der Verteilung

$$\alpha_4 = \frac{\mathbb{E}((X-\mu)^4)}{\sigma^4}$$

Für die Normalverteilung gilt $\alpha_4 = 3$.



Kovarianz

Seien X und Y zwei eindimensionale Zufallsvariablen.

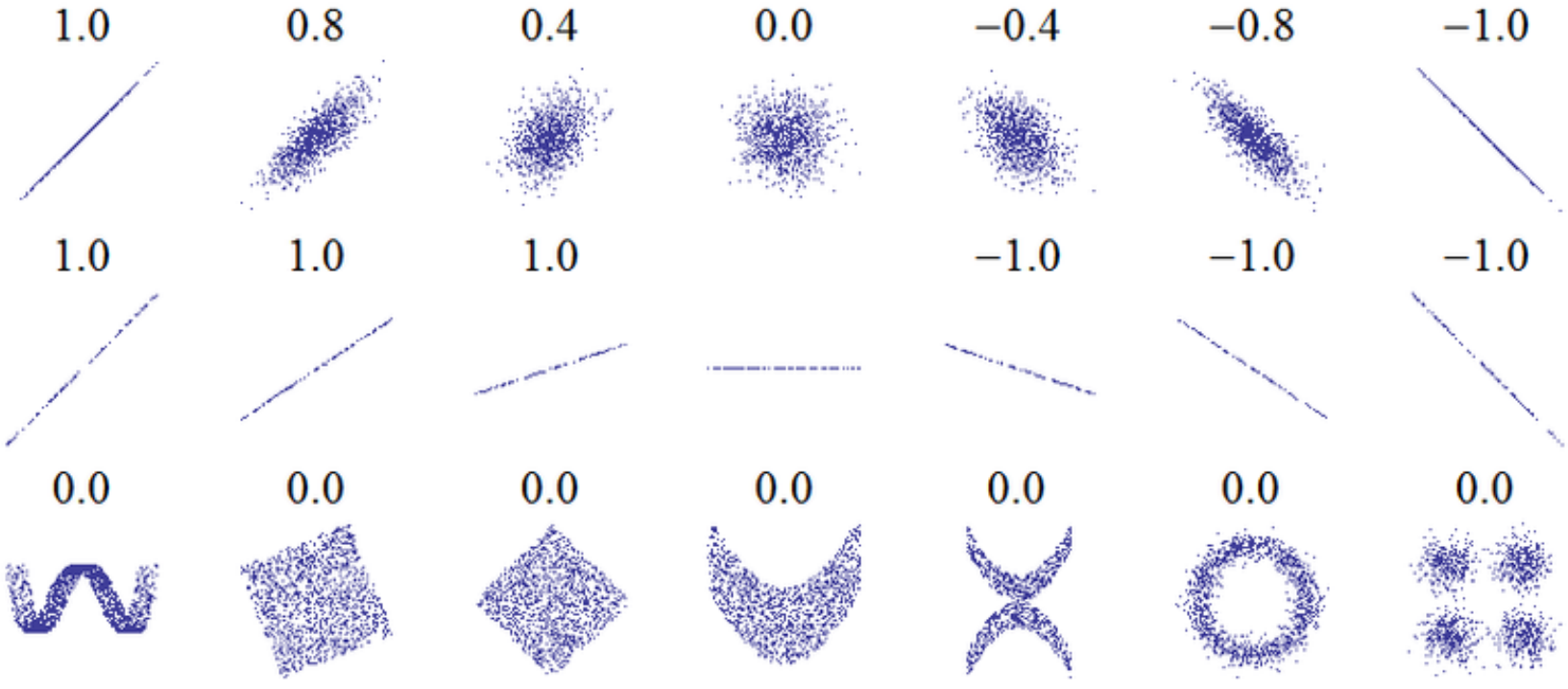
- Die *Kovarianz* von X und Y ist definiert als

$$\mathrm{Cov}(X,Y) := \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

- Die *Korrelation* von X und Y ist definiert als

$$\mathrm{Corr}(X,Y) := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

Es gilt $-1 \leq \mathrm{Corr}(X,Y) \leq 1$.
Maß für den linearen Zusammenhang


1.0
0.8
0.4
0.0
−0.4
−0.8
−1.0
1.0
1.0
1.0
−1.0
−1.0
−1.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0

Rechenregeln

- Verschiebungssatz

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \\ \mathrm{Var}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 \end{aligned}$$

- Die Kovarianz ist bilinear

$$\begin{aligned} \mathrm{Cov}(aX+b, cY+d) &= ac\,\mathrm{Cov}(X,Y) \\ \mathrm{Var}(aX+b) &= a^2\,\mathrm{Var}(X) \\ \mathrm{Cov}(X+Z,Y) &= \mathrm{Cov}(X,Y) + \mathrm{Cov}(Z,Y) \end{aligned}$$

- Ausserdem gilt

$$\mathrm{Cov}(X,Y)^2 \leq \mathrm{Var}(X)\,\mathrm{Var}(Y)$$

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$$

- Sind X, Y unabhängig, so gilt:

$$\begin{aligned} \mathbb{E}(XY) &= \mathbb{E}(X)\mathbb{E}(Y) \\ \mathrm{Cov}(X,Y) &= 0 \end{aligned}$$

Kovarianzmatrix

Sei $X = (X_1, \dots X_n)^T$ eine mehrdimensionale Zufallsvariable.
Dann ist ihre Kovarianzmatrix gegeben durch

$$\begin{aligned} \mathrm{Cov}(X) &= \mathbb{E}((X - \mathbb{E}(X)) \cdot (X - \mathbb{E}(X))^T) \\ &= \mathbb{E}(XX^T) - \mathbb{E}(X)\mathbb{E}(X)^T \\ &= \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \cdots & \mathrm{Cov}(X_n, X_n) \end{pmatrix} \end{aligned}$$

Die Kovarianzmatrix ist

- symmetrisch
- positiv semi-definit

Bedingte Erwartungswerte

- Sind X, Y diskrete Zufallsvariablen so ist der bedingte Erwartungswert

$$\mathbb{E}(f(Y)|X = x) := \sum_{y} f(\mathbf{y}) P(Y = \mathbf{y}|X = x)$$

- Sind $X, Y \in \mathbb{R}^d$ kontinuierliche Zufallsvariablen so ist der bedingte Erwartungswert

$$\mathbb{E}(f(Y)|X = x) := \int_{\mathbb{R}^d} f(\mathbf{y}) p(\mathbf{y}|X = x) dy$$

Chebyshevs Ungleichung & Law of Large Numbers

Sei X eine Zufallsvariable mit Mittelwert μ und Varianz σ^2, dann ist für $\epsilon > 0$

$$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

Seien $X_1, X_2, \ldots, X_n$ unabhängige Zufallsvariablen mit Mittelwert μ und Varianz σ^2. Sei $S_n = \sum_i X_i$, dann ist

$$\lim_{n \to \infty} P(|\frac{S_n}{n} - \mu| \geq \epsilon) = 0$$