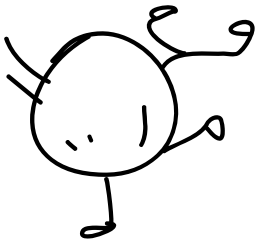


Grundlagen II

Vorlesung 186.844

23.10.2014

TECHNISCHE
UNIVERSITÄT
WIEN
Vienna University of Technology

Überblick

I. Herleitung Bayes-Theorem
II. Statistische Grundlagen für stetige Zufallsvariablen
III. Bayes-Theorem für stetige Merkmale
IV. Parameterschätzung für Bayes-Theorem
V. Statistische Grundlagen für zwei oder mehrere Zufallsvariablen

© Nicole M. Artner

Herleitung Bayes-Theorem

© Nicole M. Artner

Wiederholung

Randverteilung (marginal distribution):

$$\boxed{P(A) = \sum_{B} P(A,B)}$$

— Summenregel (sum rule)

Verbundwahrscheinlichkeit (joint probability):

$$\boxed{\begin{aligned} P(A,B) &= P(B|A)P(A) \\ P(B,A) &= P(A|B)P(B) \end{aligned}}$$

— Produktregel (product rule)

© Nicole M. Artner

Herleitung …

… basierend auf der Produktregel und der Symmetrieeigenschaft $P(A,B) = P(B,A)$, kann man direkt auf folgenden Zusammenhang zwischen den bedingten Wahrscheinlichkeiten schließen:

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

$$\Downarrow$$

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

Bayes-Theorem

© Nicole M. Artner

Statistische Grundlagen für stetige Zufallsvariablen

© Nicole M. Artner

Zufallsvariablen

- in der letzten VO haben wir uns mit diskreten Zufallsvariablen beschäftigt
  - Beispiel: „Hund oder Katze“
  - 4 Gewichtsstufen und 2 Temperaturstufen

- Temperatur und Gewicht sind eigentlich stetige Merkmale …

© Nicole M. Artner

Stetige Zufallsvariablen

Zufallsvariable $x \in \mathbb{R}$ heißt stetig, weil sie unendlich viele verschiedene Werte annehmen kann.

Wichtig: Die Wahrscheinlichkeit, dass x einen bestimmten Wert annimmt, dass das Ereignis $x = a$ eintritt, ist Null:

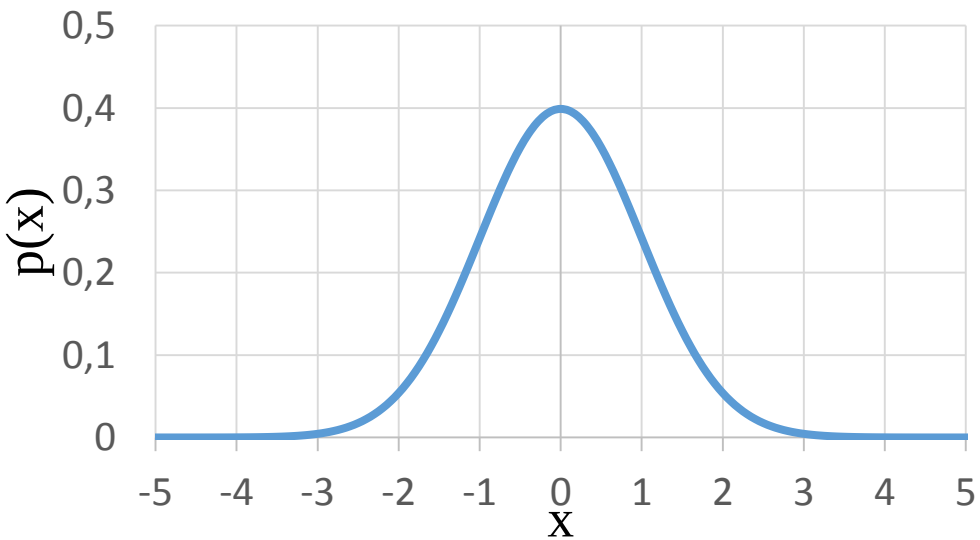
$$P(x = a) = \int_a^a p(x)dx = 0 \ (\forall a \epsilon \mathbb{R})$$

© Nicole M. Artner

Dichtefunktion

Für **stetige Zufallsvariablen** wird eine **Dichtefunktion (DF)** $p(x)$ statt einer Wahrscheinlichkeitsfunktion $p_i = P(x = v_i)$ verwendet.

... Dichtefunktion wird auch als „probability desity function“ oder kurz „pdf“ bezeichnet, z.B.: Normalverteilung/Gauß-Verteilung



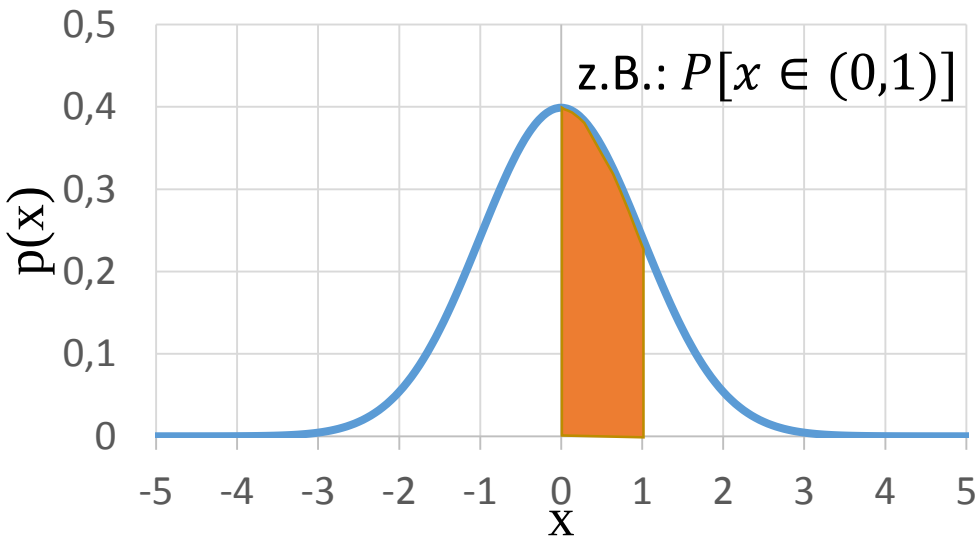
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

© Nicole M. Artner

Wahrscheinlichkeit

Die Wahrscheinlichkeit P das x in das Intervall $[a, b]$ fällt, kann durch Integration ermittelt werden:

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



Folgende Voraussetzungen muss eine DF erfüllen:

$$p(x) \geq 0$$

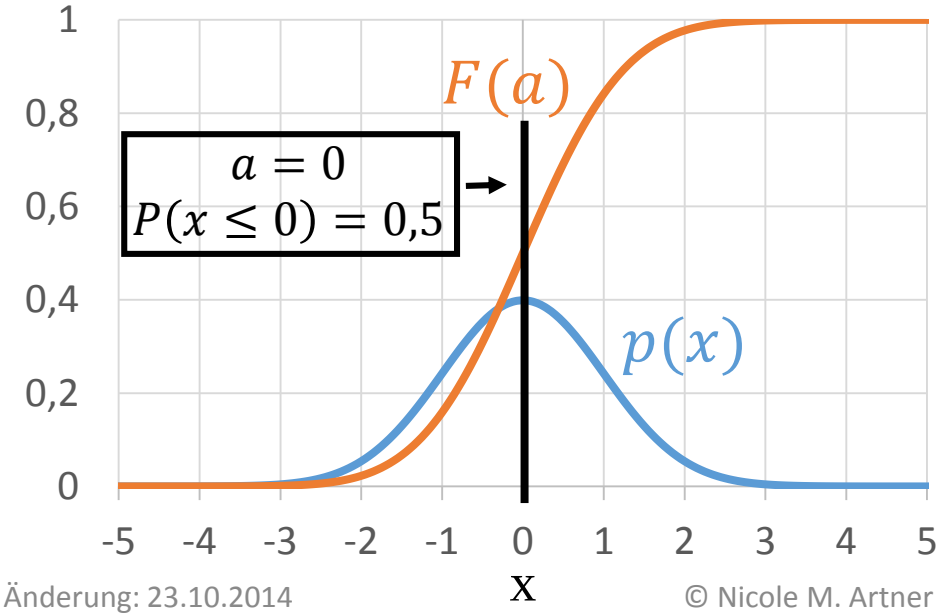
$$\int_{-\infty}^{\infty} p(x)dx = 1$$

© Nicole M. Artner

Verteilungsfunktion

Die **Verteilungsfunktion (VF)** $F(a)$ gibt die Wahrscheinlichkeit für das Ereignis $x \leq a$ an, d.h. die Wahrscheinlichkeit, dass x in das Intervall $(-\infty, a]$ fällt.

- Wahrscheinlichkeit $P(x \leq a)$ kann direkt von $F(a)$ abgelesen werden …



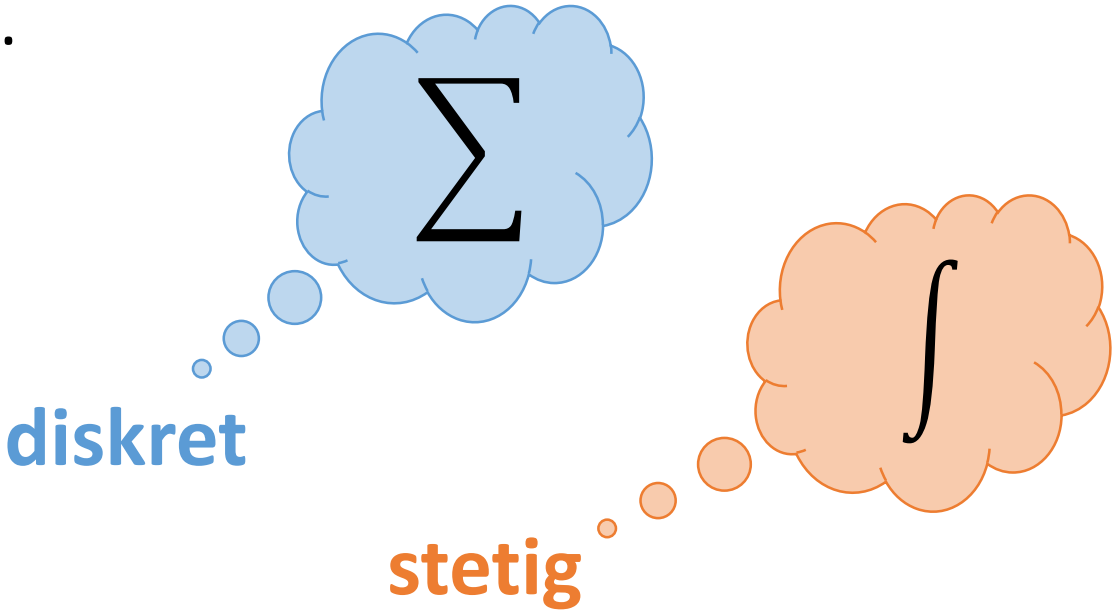
$$F(a) = \int_{-\infty}^{a} p(x)dx$$

© Nicole M. Artner

Kenngrößen von Verteilungen

- Kenngrößen für stetige Verteilungen sind in den meisten Fällen sehr ähnlich zu jenen von diskreten Verteilungen

- meist reicht es aus die Summen durch Integrale zu ersetzten …



© Nicole M. Artner

Erwartung

Die Erwartung $\varepsilon[]$ einer Funktion $h(x)$ einer stetigen Zufallsvariable x ist definiert als

$$\mathcal{E}[h(x)] = \int_{-\infty}^{\infty} h(x)p(x)dx$$

- zum Vergleich: Erwartung für diskrete Zufallsvariable x

$$\varepsilon[h(x)] = \sum_{x \in X} h(x)P(x)$$

© Nicole M. Artner

Mittelwert

Der **Mittelwert** einer stetigen Zufallsvariable x ist definiert als

$$\mathcal{E}[x] = \mu = \int_{-\infty}^{\infty} x\, p(x)\, dx$$

- zum Vergleich: Mittelwert für diskrete Zufallsvariable x

$$\varepsilon[x] = \mu = \sum_{x \in X} x\, P(x) = \sum_{i=1}^{m} v_i p_i$$

© Nicole M. Artner

Varianz

Die **Varianz** σ^2 ist definiert als

$$\sigma^2 = \mathcal{E}[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx$$

- es gilt natürlich auch: $\sigma^2 = \mathcal{E}[x^2] - (\mathcal{E}[x])^2$
- zum Vergleich: Varianz für diskrete Zufallsvariable x

$$\sigma^2 = \varepsilon[(x-\mu)^2] = \sum_{x \in X} (x-\mu)^2 P(x)$$

© Nicole M. Artner

Bayes-Theorem für stetige Merkmale

© Nicole M. Artner

Diskret ⇔ Stetig

- diskret:

$$P(\omega_j | x = i) = \frac{P(x = i | \omega_j) P(\omega_j)}{P(x = i)}$$

- für eine stetige Zufallsvariable/Merkmal x mit zugehöriger Dichtefunktion (pdf) $p(x)$:

Wahrscheinlichkeiten → P
Dichtefunktionen → p

$$P(\omega_j | x) = \frac{p(x | \omega_j) P(\omega_j)}{p(x)}$$

© Nicole M. Artner

Dichtefunktion einer Klasse

$p(x|\omega_j)$ … wird als „class conditional“ (pdf) bezeichnet

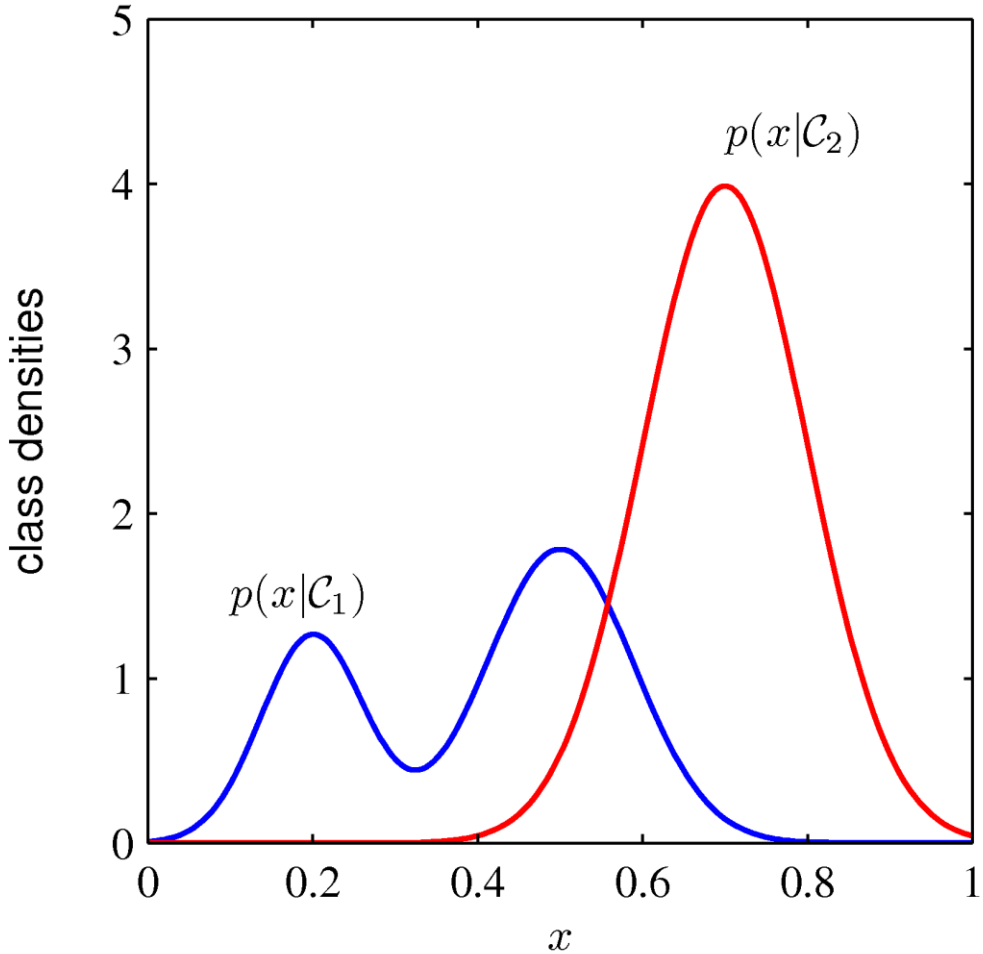
- beschreibt die stetige Verteilung eines Merkmals x für eine gegebene Klasse w_j
- besitzt alle Eigenschaften einer „normalen“ Dichtefunktion

$$p(x|w_j) \geq 0$$

$$\int_{-\infty}^{\infty} p(x|w_j)dx = 1$$

© Nicole M. Artner

Beispiel: class conditional pdfs



© Nicole M. Artner

Likelihood

Falls man $p(x|\omega_j)$ als Funktion der Klasse j für ein festes x betrachtet, spricht man von der „**Likelihood**“ von j bezüglich x.

Likelihood $\neq$ Wahrscheinlichkeit

© Nicole M. Artner

Bayes-Theorem

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$

$$posterior = \frac{likelihood \times prior}{evidence}$$

© Nicole M. Artner

Evidence

- entspricht dem Normalisierungsfaktor (siehe vorherige Vorlesung, Bayes-Theorem für diskrete Merkmale)

- $p(x)$ ist für alle Klasse identisch → keinen Einfluss auf das Verhältnis der posteriors

- für Bestimmung der Klasse mit größter posteriori Wahrscheinlichkeit ist $likelihood \times prior$ ausreichend

- bei identischen priors wird mit likelihoods klassifiziert

© Nicole M. Artner

Likelihood Verhältnis

… im Fall von $c = 2$ (zwei Klassen), kann man die Bayes-Entscheidungsregel auch so formulieren:

Entscheide für w_1, falls

$$P(w_1|x) > P(w_2|x)$$

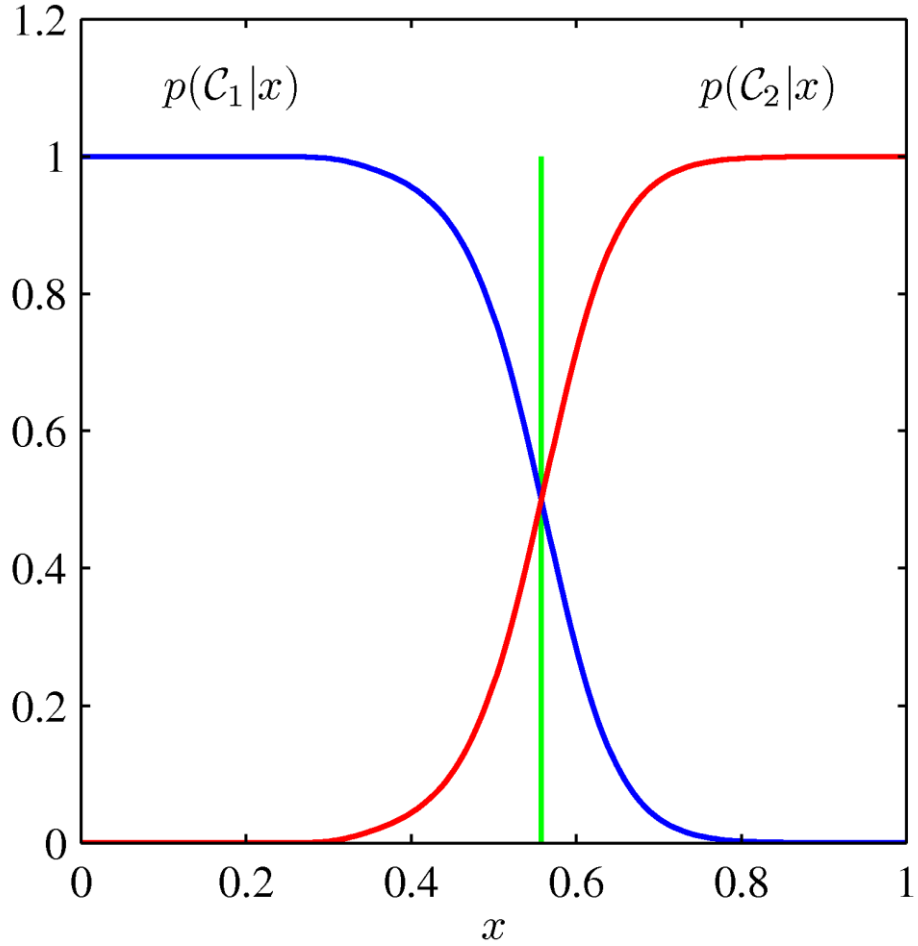
$$p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$$

Likelihood Verhältnis (Ratio) — $\boxed{\dfrac{p(x|\omega_1)}{p(x|\omega_2)}} > \boxed{\dfrac{P(\omega_2)}{P(\omega_1)}}$ — Schwellwert (Threshold)

© Nicole M. Artner

Bayes-Entscheidungsregel

$$P(w_1|x) > P(w_2|x)$$

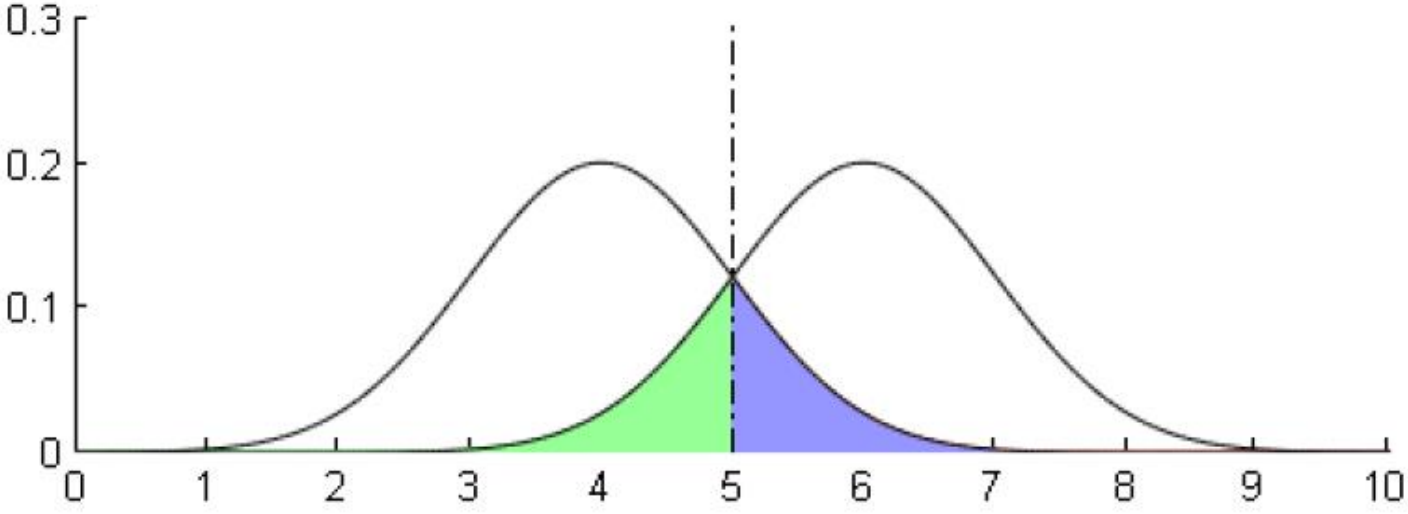


© Nicole M. Artner

Bayes-Entscheidungsregel

$$p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$$

für $P(\omega_1) = P(\omega_2) = 0{,}5$

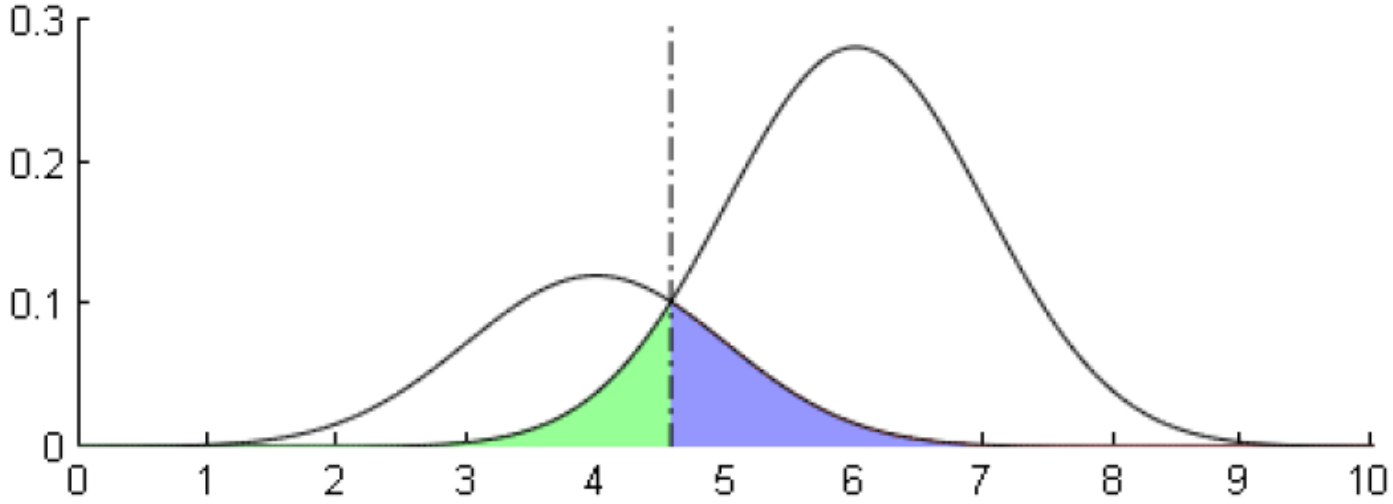


© Nicole M. Artner

Bayes-Entscheidungsregel

$$p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$$

- für $P(\omega_1) = 0{,}3$ und $P(\omega_2) = 0{,}7$



© Nicole M. Artner

Fehlerwahrscheinlichkeit

… die (bedingte) **Wahrscheinlichkeit der Fehlklassifikation** (conditional error) $P(error|x)$ ist

- $P(w_1|x)$, falls man für w_2 entscheidet und
- $P(w_2|x)$, falls man für w_1 entscheidet.

Die **Fehlerrate** (oder der mittlere Fehler) $P(error)$ berechnet sich als

$$P(error) = \int_{-\infty}^{\infty} P(error|x)\, p(x)\, dx$$

Parameterschätzung für Bayes-Theorem

© Nicole M. Artner

Parameterschätzung …

… man spricht auch vom Training oder Supervised Learning

Mit Hilfe von Trainingsdaten, bestehend aus N Stichproben die c Klassen zugeordnet sind, schätzt man:

- priors $P(\omega_j)$
- evidence $p(x)$
- class conditional pdfs $p(x|\omega_j)$

Ziel: Klassifikator soll Muster die nicht in den Trainingsdaten waren richtig klassifizieren.

© Nicole M. Artner

Schätzung der Priors

- meist einfach zu schätzen
- Schätzungen basieren auf
  - Klassenverteilung in den Trainingsdaten
  - zusätzliches Wissen über die Verteilung
- oft werden einfach gleiche Priors für alle Klassen angenommen

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$

© Nicole M. Artner

Schätzung der Evidence

- Evidence (Normalisierungsfaktor) erhält man als:

$$p(x) = \sum_{j=1}^{c} p(x|\omega_j)P(\omega_j)$$

- man braucht unter anderem die Dichtefunktionen der Klassen (class conditional pdfs)

$$P(\omega_j|x) = \frac{\overset{?}{p(x|\omega_j)}P(\omega_j)}{p(x)}$$

© Nicole M. Artner

Schätzung der Dichtefunktionen

… die class conditional pdfs sind am schwierigsten zu schätzen

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$

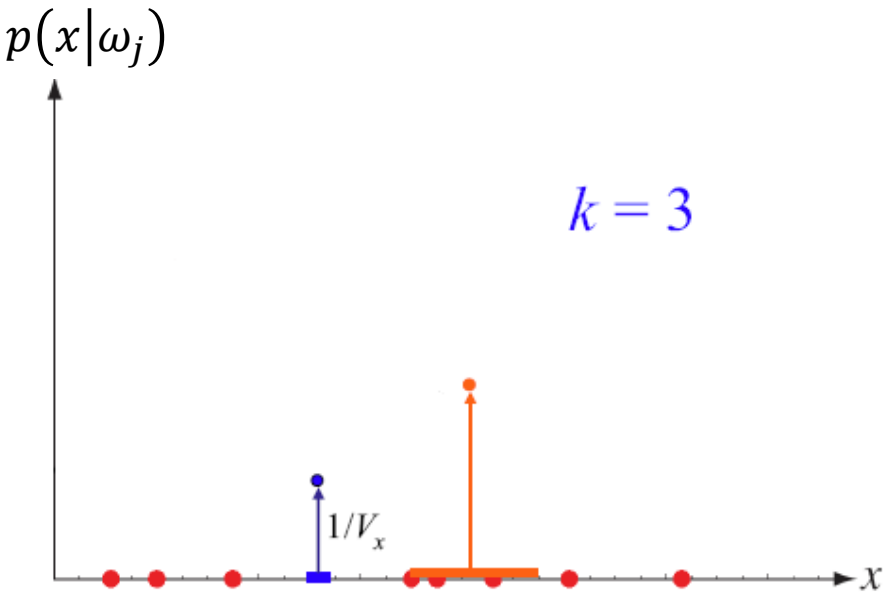
nicht-parametrische Verfahren	parametrische Verfahren
• Anzahl der Parameter nicht fix • keine Annahme über die Verteilung	• Anzahl der Parameter ist fix • man macht am Beginn eine Annahme über die Verteilung (z.B. Normalverteilung)

© Nicole M. Artner

k-NN Schätzung

- k-Nearest Neighbor Estimation: einfaches **nicht-parametrisches** Verfahren zur Dichteschätzung aus n Stichproben von Klasse j

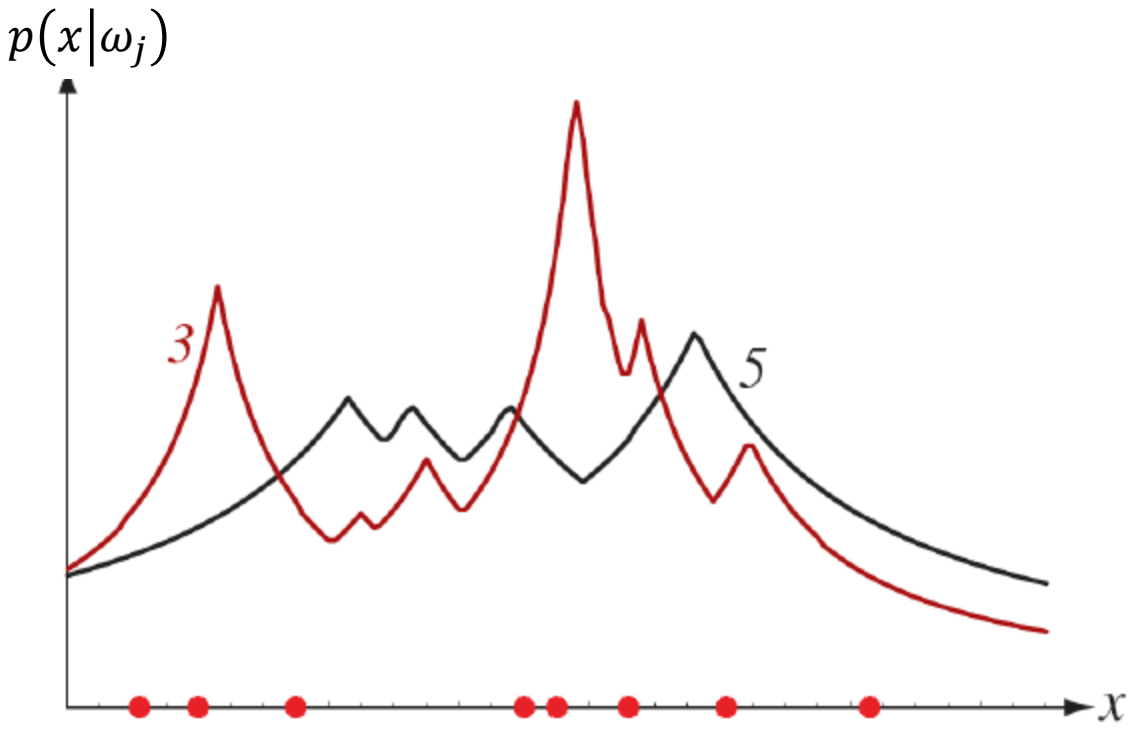
- zentriere „Zelle“ über x
- lass Zelle wachsen bis k Stichproben enthalten sind
- V_x → Größe der Zelle (z.B. Länge der Linie)
- $\frac{1}{V_x}$ → Schätzung der Dichte



© Nicole M. Artner

k-NN Schätzung

Ergebnis für Beispiel mit 8 Stichproben mit $k = 3$ und k =5 aus Duda et al.:

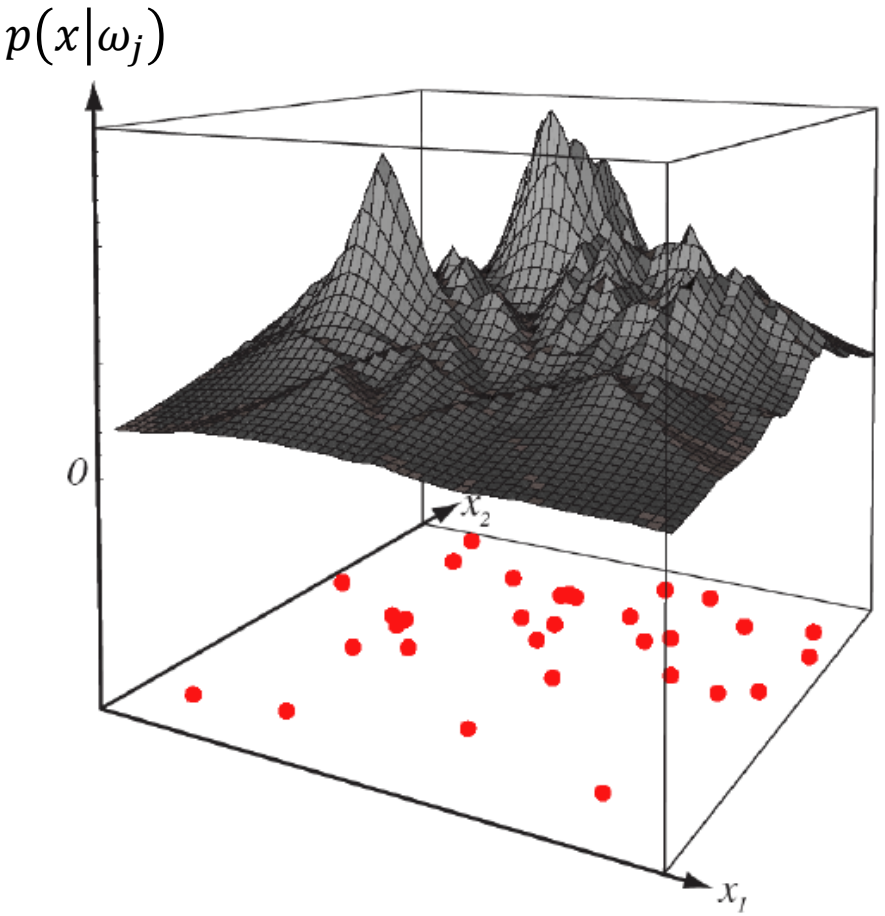


[Quelle: Duda et al., 2001]

© Nicole M. Artner

Dichteschätzung für nD

- k-NN Schätzverfahren ist auch für Dichteschätzungen mit zwei oder mehr Dimensionen anwendbar
- zum Beispiel:
  - 2D-Zelle Kreis oder Rechteck
  - V_x Flächeninhalt der Form



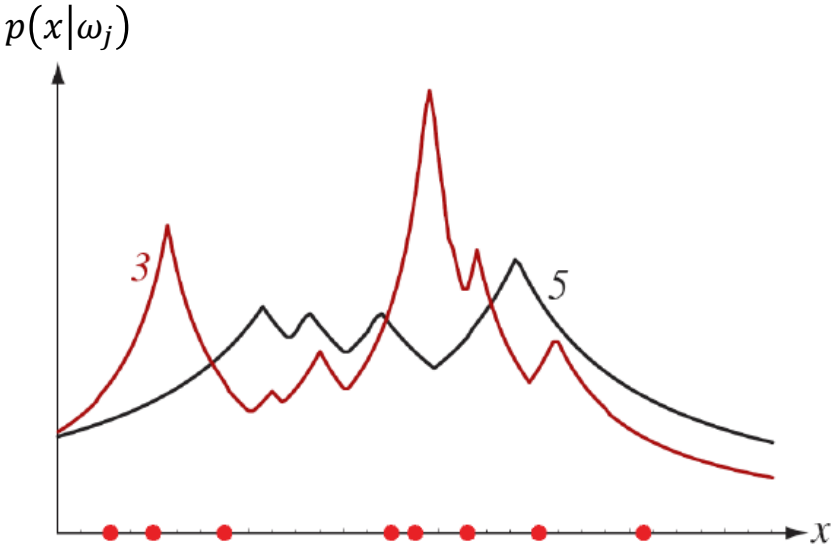
[Quelle: Duda et al., 2001]

Anmerkungen Dichteschätzung

- Form der geschätzten Dichtefunktion hängt stark vom Wert für k ab

- Geschätzte Dichtefunktion ist üblicherweise nicht glatt

- Schätzung nähert sich der tatsächlichen Dichtefunktion an, wenn Anzahl der Stichproben $\mathrm{N} \rightarrow \infty$



[Quelle: Duda et al., 2001]

Link: k-NN Klassifikator

- basiert auf der Idee des k-NN Schätzers
- ABER: statt die likelihood zu schätzen, wird direkt die posteriori Wahrscheinlichkeit geschätzt
  - aus Trainingsdaten werden k „Datenpunkte“ mit der kleinsten Distanz zu x gefunden
  - k_j gehören zur Klasse j
  - posteriori Wahrscheinlichkeit für Klasse j wird folgendermaßen geschätzt:

$$\mathrm{P}(\omega_j | x) = \frac{k_j}{k}$$

© Nicole M. Artner

Parametrisches Verfahren

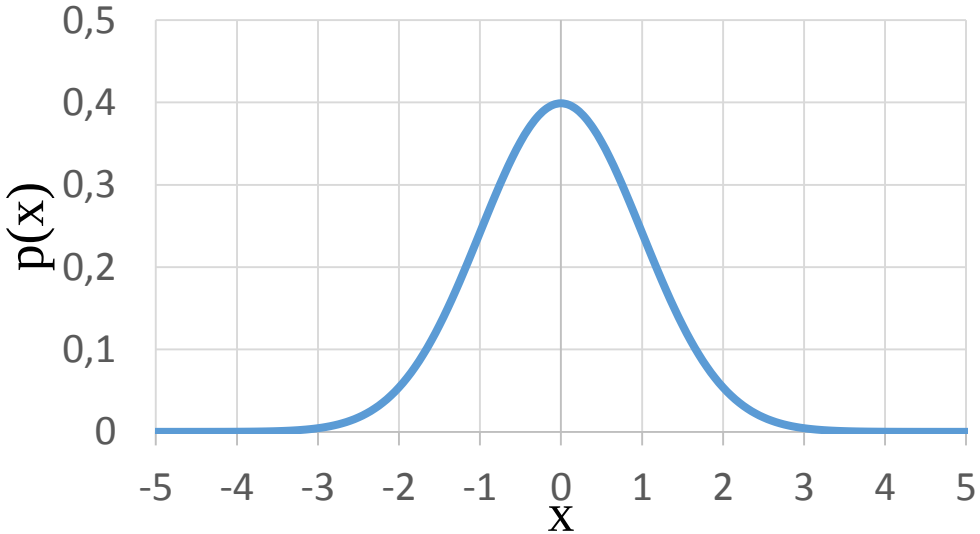
1. Annahme über die Verteilung der Dichtefunktion wird getroffen
2. Schätzen der Parameter der Verteilung

Vorteile	Nachteile
• weniger Parameter zu schätzen • kleinerer Trainings-datensatz notwendig	• nicht so flexibel wie nicht-parametrische Verfahren

© Nicole M. Artner

Normalverteilung

- häufig wird eine Normalverteilung angenommen
- Gründe
  - eine der meist studierten Verteilungen
  - gute analytische Eigenschaften
  - gutes Modell für die Verteilung eines Merkmals



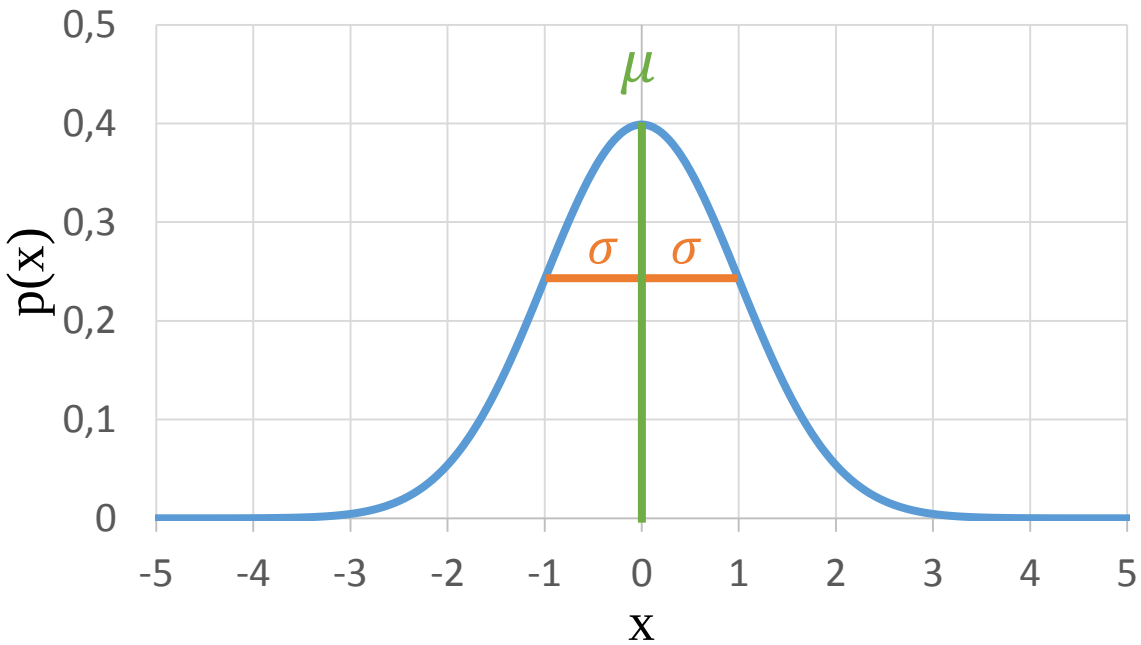
© Nicole M. Artner

Normalverteilung

- Parameter:
  - Mittelwert μ
  - Varianz σ^2

$$p(x) \sim N(\mu, \sigma^2)$$

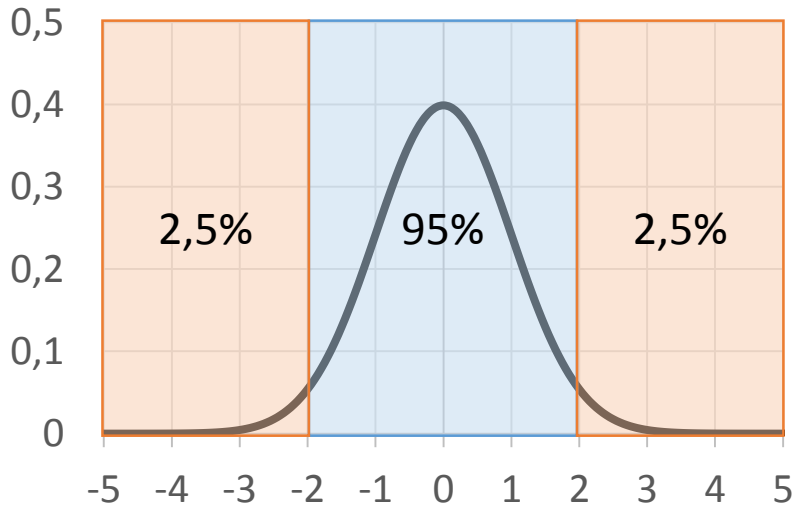
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



© Nicole M. Artner

Dichte in Normalverteilung

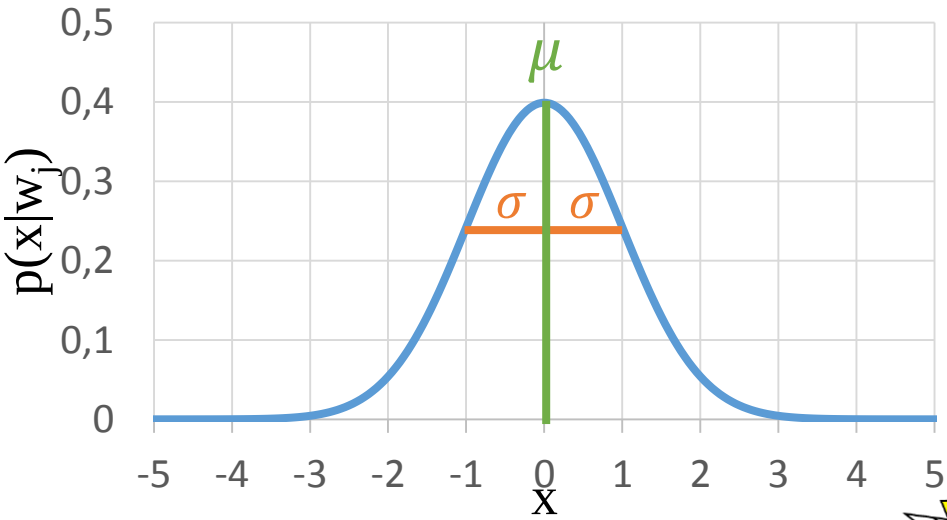
- am höchsten in der Nähe des Mittelwerts
- Werteverteilung / Dichte:
  - $P[|x-\mu| \leq \sigma] \approx 0{,}68$ → 68% der Daten
  - $P[|x-\mu| \leq 2\sigma] \approx 0{,}95$ → 95% der Daten
  - $P[|x-\mu| \leq 3\sigma] \approx 0{,}997$ → mehr als 99% der Daten



© Nicole M. Artner

Schätzung der Parameter

1. Annahme der Verteilung: Normalverteilung
2. Trainingsdaten:
   - $N = \{x_1, \dots, x_N\}$ Stichproben mit Klassenlabel w_j
   - $j = 1, 2, \dots, c$
   - N_j sind die Stichproben der Klasse w_j



© Nicole M. Artner

Schätzung Mittelwert

Die Schätzung des Mittelwerts $\hat{\mu}_j$ für Klasse j aus den Trainingsdaten x_i der Klasse j

$$\hat{\mu}_j = \frac{1}{N_j} \sum_{x_i \in \omega_j} x_i$$

- $\hat{\mu}_j$ ist eine Schätzung des tatsächlichen Mittelwerts μ_j

© Nicole M. Artner

Schätzung Varianz

Falls tatsächlicher Mittelwert μ_j bekannt ist, wird Varianz folgendermaßen geschätzt:

$$\hat{\sigma}_j^2 = \frac{1}{N_j} \sum_{x_i \in \omega_j} \left(x_i - \mu_j\right)^2$$

Bei geschätztem Mittelwert $\hat{\mu}_j$, wird folgender Schätzer verwendet:

$$\hat{\sigma}_j^2 = \frac{1}{N_j - 1} \sum_{x_i \in \omega_j} \left(x_i - \hat{\mu}_j\right)^2 = \frac{1}{N_j - 1} \left(\sum_{x_i \in \omega_j} x_i^2 - N_j \hat{\mu}_j^2 \right)$$

© Nicole M. Artner

Beispiel aus der Botanik

- Fisher's Iris Datensatz
- 50 Stichproben (Beobachtungen) von drei Arten von Schwertlilien
- vier Merkmale der Blüten
  - Länge und die Breite des Sepalum (Kelchblatt) und des Petalum (Kornblatt)

Iris setosa

Iris virginica

Iris versicolor

© Nicole M. Artner

Auszug aus den Daten

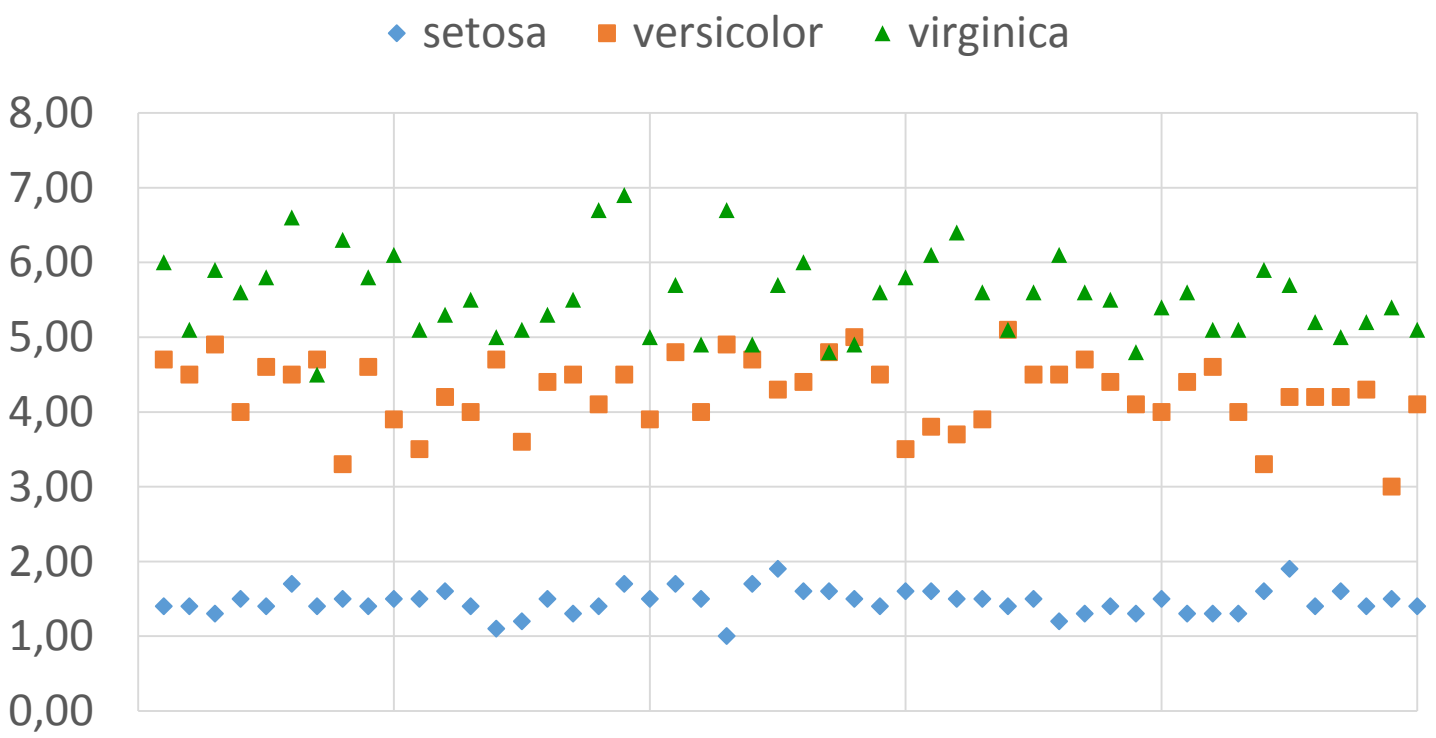
Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	I. setosa
4.9	3.0	1.4	0.2	I. setosa
4.7	3.2	1.3	0.2	I. setosa
4.6	3.1	1.5	0.2	I. setosa
5.0	3.6	1.4	0.2	I. setosa
5.4	3.9	1.7	0.4	I. setosa
4.6	3.4	1.4	0.3	I. setosa
5.0	3.4	1.5	0.2	I. setosa
…	…	…	…	…

Quelle: http://en.wikipedia.org/wiki/Iris_flower_data_set#Data_set

© Nicole M. Artner

Trainingsdaten für Beispiel

- 1 Merkmal: Petalumlänge



© Nicole M. Artner

Schätzung der Priors

- wir nehmen an, dass die a priori Wahrscheinlichkeiten für alle 3 Klassen gleich sind:

$$P(\omega_j|x) = \frac{p(x|\omega_j)\boxed{P(\omega_j)}}{p(x)}$$

$$P(\omega_1) = P(\omega_2) = P(\omega_3) = \frac{1}{3}$$

© Nicole M. Artner

Schätzung der Parameter

- für class conditional pdfs (Dichtefunktionen der Klassen)
- geschätzte Mittelwerte und Standardabweichungen (Varianz) von den Trainingsdaten:

Petalumlänge	setosa	versicolor	virginica
Mittelwert	1,46	4,26	5,55
Standardabweichung	0,17	0,47	0,55

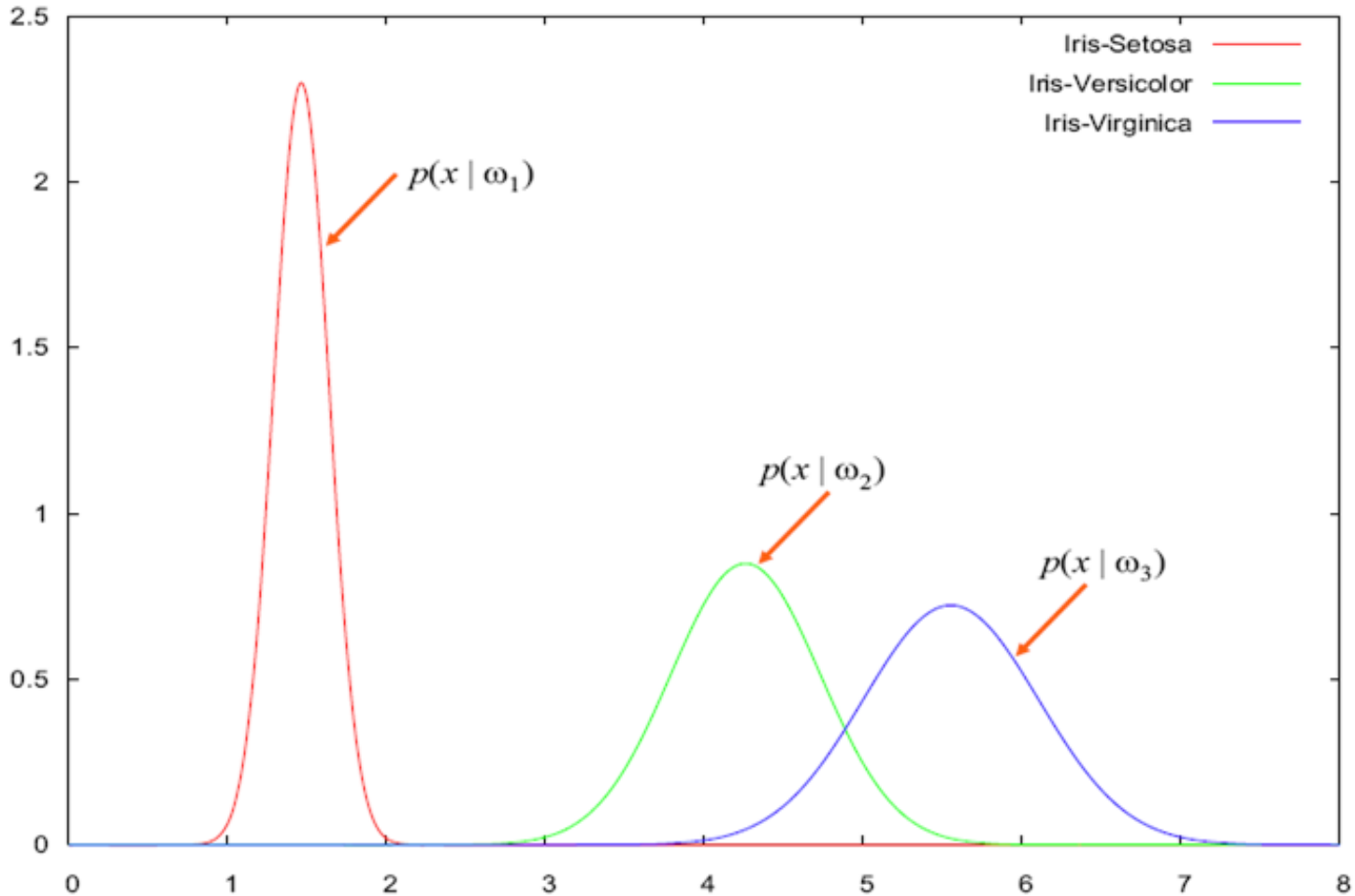
$$p(x|\omega_1) = \frac{1}{0{,}17\sqrt{2\pi}} e^{-\left[\frac{(x-1{,}46)^2}{0{,}058}\right]}$$

© Nicole M. Artner

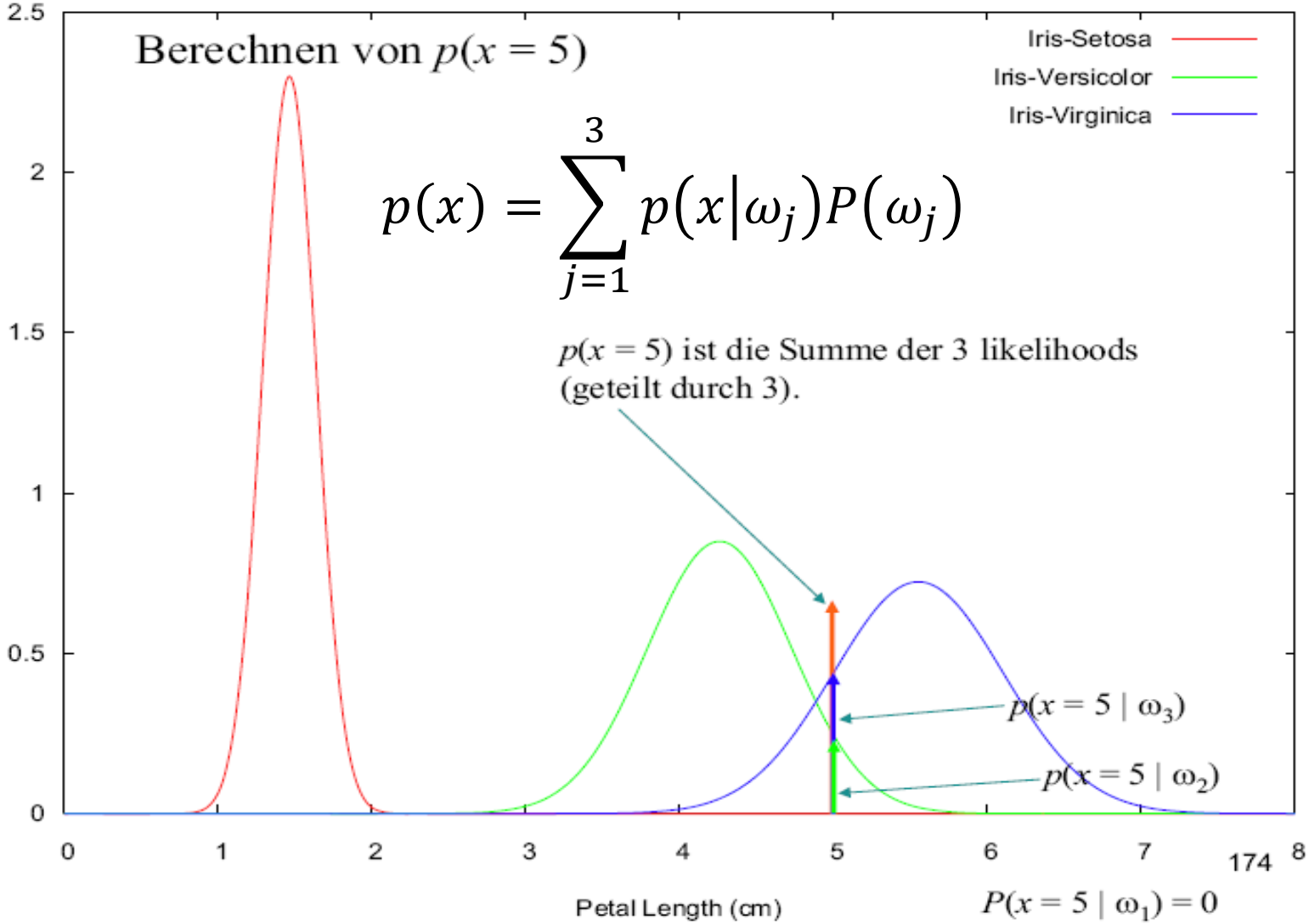
Geschätzte Dichtefunktionen

… für jede Klasse/Lilienart

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$

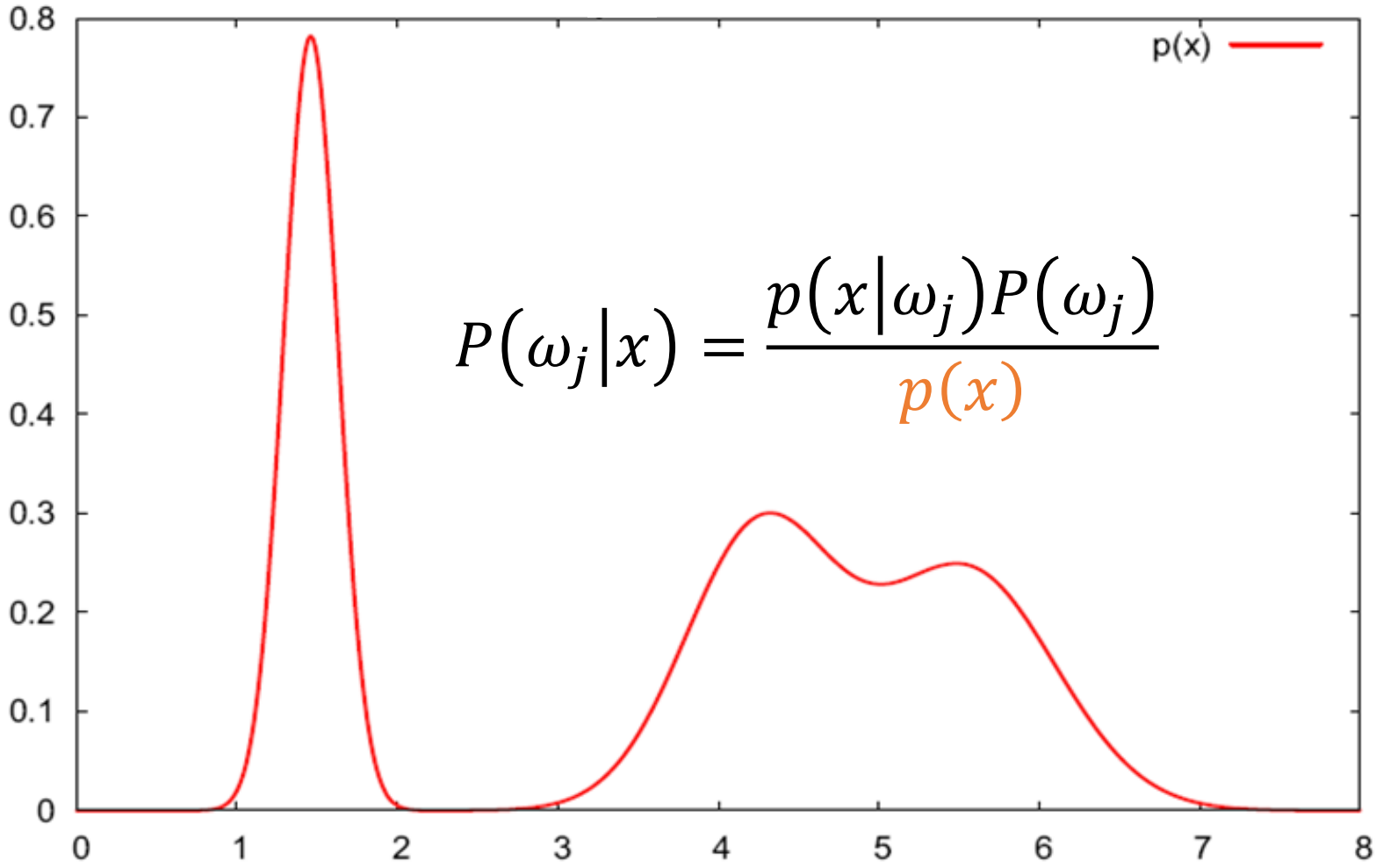


Schätzung Evidence



$$p(x) = \sum_{j=1}^{3} p(x|\omega_j)P(\omega_j)$$

Schätzung Evidence

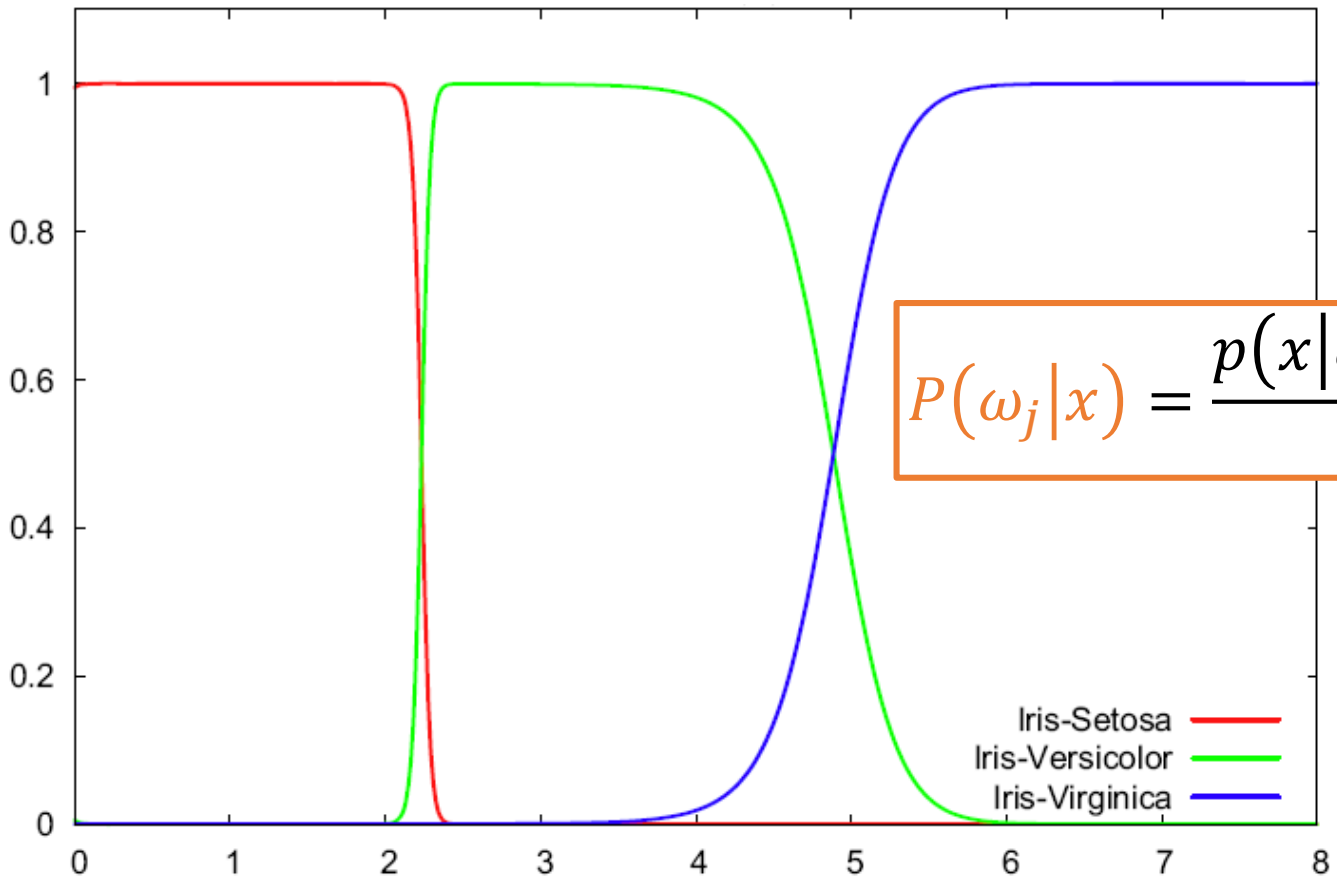


$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$

© Nicole M. Artner

Klassifikation

… mit Bayes-Entscheidungsregel



$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$

© Nicole M. Artner

Resümee

- ✓ diskrete und stetige Zufallsvariablen
- ✓ Bayes-Theorem für diskrete und stetige Merkmale
- ✓ Parameterschätzung
- ✓ Schätzung von Dichtefunktionen

Was fehlt noch?

Mehr Merkmale!

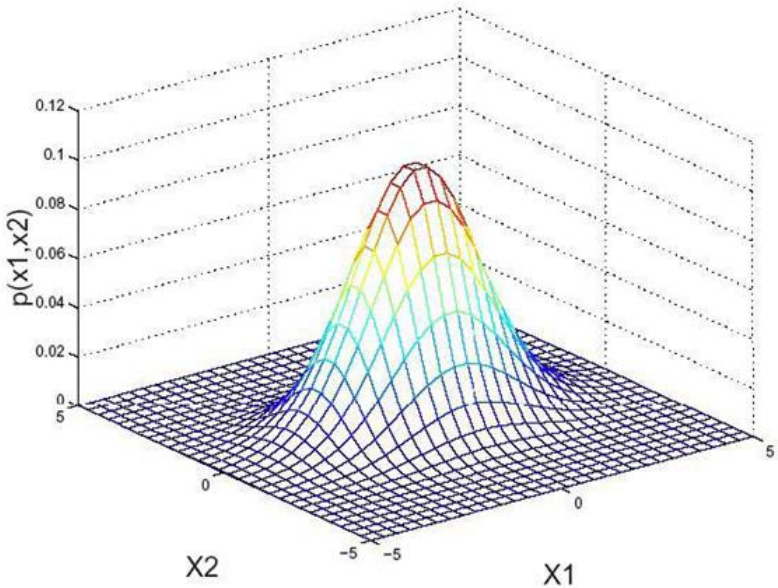
Statistische Grundlagen für zwei oder mehrere Zufallsvariablen

© Nicole M. Artner

Erwartung

Die Erwartung $\varepsilon[]$ einer bivariaten Funktion $h(x, y)$ von zwei diskreten Zufallsvariablen ist:

$$\mathcal{E}[h(x,y)] = \sum_{x \epsilon X} \sum_{y \in Y} h(x,y) P(x,y)$$



Mittelwerte und Varianzen

- Mittelwerte beider Zufallsvariablen

$$\mu_x = \mathcal{E}[x] = \sum_{x \epsilon X} \sum_{y \in Y} x P(x, y)$$

$$\mu_y = \mathcal{E}[y] = \sum_{x \epsilon X} \sum_{y \in Y} y P(x, y)$$

- Varianzen beider Zufallsvariablen

$$\sigma_x^2 = \mathcal{E}[(x - \mu_x)^2] = \sum_{x \epsilon X} \sum_{y \in Y} (x - \mu_x)^2 P(x, y)$$

$$\sigma_y^2 = \mathcal{E}\left[(y - \mu_y)^2\right] = \sum_{x \epsilon X} \sum_{y \in Y} (y - \mu_y)^2 P(x, y)$$

© Nicole M. Artner

Beispiel: Würfel

$P(x) = \frac{1}{6}, x = \{1, 2, 3, 4, 5, 6\}$ und $P(y) = \frac{1}{6}, y = \{1, 2, 3, 4, 5, 6\}$

$P(x, y) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$ (x und y unabhängig)

- Berechnung des Mittelwerts μ_x:

$$\mu_x = 1 \sum_{y=1}^{6} P(x, y) + 2 \sum_{y=1}^{6} P(x, y) + 3 \sum_{y=1}^{6} P(x, y) + \cdots +$$

$$+6 \sum_{y=1}^{6} P(x, y) = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3{,}5$$

Kovarianz

Die Kovarianz ist ein Maß für die statistische Abhängigkeit zwischen Zufallsvariablen:

$$\sigma_{xy} = \mathcal{E}\left[(x-\mu_x)(y-\mu_y)\right] = \sum_{x\in X}\sum_{y\in Y}(x-\mu_x)(y-\mu_y)P(x,y)$$

Schätzung der Kovarianz:

$$\hat{\sigma}_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\hat{\mu}_x)(y_i-\hat{\mu}_y)$$

… falls die wahren Mittelwerte bekannt sind, wird mit $\frac{1}{N}$ normalisiert.

© Nicole M. Artner

Eigenschaften der Kovarianz

x und y unabhängig:

$$P(x|y) = P(x) \Rightarrow \sigma_{xy} = 0$$

Achtung: Kovarianz $\sigma_{xy} = 0$ kann zwei Bedeutungen haben.
(1) x und y unabhängig oder (2) x und y unkorreliert

Positive Kovarianz σ_{xy}:
x und y vergrößern und verkleinern sich gemeinsam

Negative Kovarianz σ_{xy}:
x wird größer wenn y kleiner wird

© Nicole M. Artner

Bivariat → Multivariat

- von zwei Zufallsvariablen (Merkmalen) zu „beliebig“ vielen

In der Praxis erfolgt die Klassifikation meist anhand von multivariaten Merkmalsvariablen:

p-dimensionaler Merkmalsvektor — $\boxed{\mathbf{x}} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \in \mathbb{R}^p$

© Nicole M. Artner

Mittelwertvektor

$$\boldsymbol{\mu} = \mathcal{E}[\mathbf{x}] = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} \mathcal{E}[x_1] \\ \mathcal{E}[x_1] \\ \vdots \\ \mathcal{E}[x_1] \end{bmatrix}$$

$\mu_1 = \mathcal{E}[x_1]$ ist der Mittelwert der $i$ten Komponente ($i$tes Merkmal) des Merkmals-vektors $\mathbf{x}$.

Der Schätzer des Mittelwertvektors ergibt sich, analog zum univariaten Fall, als

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^{\boxed{N}} \boldsymbol{x}_i$$

N p-dimensionale Merkmalsvektoren

© Nicole M. Artner

Kovarianzmatrix

Eine Kovarianzmatrix $\boldsymbol{\Sigma}$ setzt sich aus zentralen Momenten 2. Ordnung zusammen. Die Elemente σ_{ij} bezeichnet man als Varianz wenn $(i = j)$ und als Kovarianz wenn $(i \neq j)$.

$$\sigma_{ij} = \mathcal{E}\left[(x_i - \mu_i)(x_j - \mu_j)\right]$$

$$Cov(\boldsymbol{x}) = \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \boxed{\sigma_{p1}} & \cdots & \boxed{\sigma_{pp}} \end{bmatrix} = \mathcal{E}[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T]$$

linearer Zusammenhang zwischen Komponente p und 1 → Kovarianz

Dispersion (Energie) → Varianz $\sigma_{pp} = \sigma_p^2$

© Nicole M. Artner

Schätzung der Kovarianzmatrix

- Ein Schätzer der Kovarianz-Matrix ist

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^T$$

© Nicole M. Artner

Beispiel

- Gegeben folgende Trainingsdaten:
  - 2 Merkmale und 2 Klasse

Klasse 1

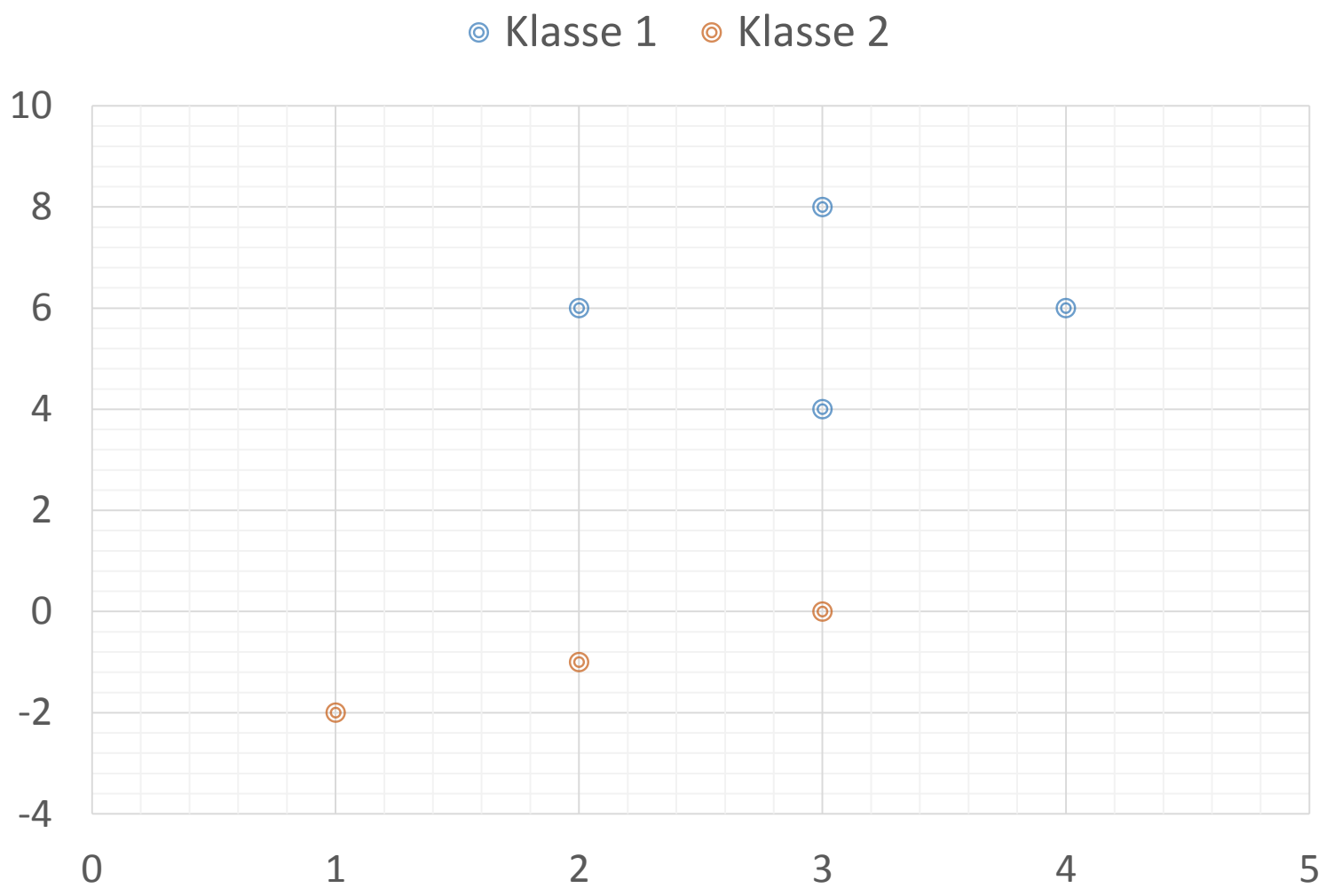
$$\begin{bmatrix}3\\4\end{bmatrix}, \begin{bmatrix}3\\8\end{bmatrix}, \begin{bmatrix}2\\6\end{bmatrix}, \begin{bmatrix}4\\6\end{bmatrix}$$

Klasse 2

$$\begin{bmatrix}3\\0\end{bmatrix}, \begin{bmatrix}1\\-2\end{bmatrix}, \begin{bmatrix}2\\-1\end{bmatrix}$$

© Nicole M. Artner

Trainingsdaten



Schätzung Mittelwertvektor

- für Klasse 1

$$\hat{\boldsymbol{\mu}}_1 = \frac{1}{4}\left\{\begin{bmatrix}3\\4\end{bmatrix} + \begin{bmatrix}3\\8\end{bmatrix} + \begin{bmatrix}2\\6\end{bmatrix} + \begin{bmatrix}4\\6\end{bmatrix}\right\}$$

$$= \frac{1}{4}\left\{\begin{bmatrix}12\\24\end{bmatrix}\right\}$$

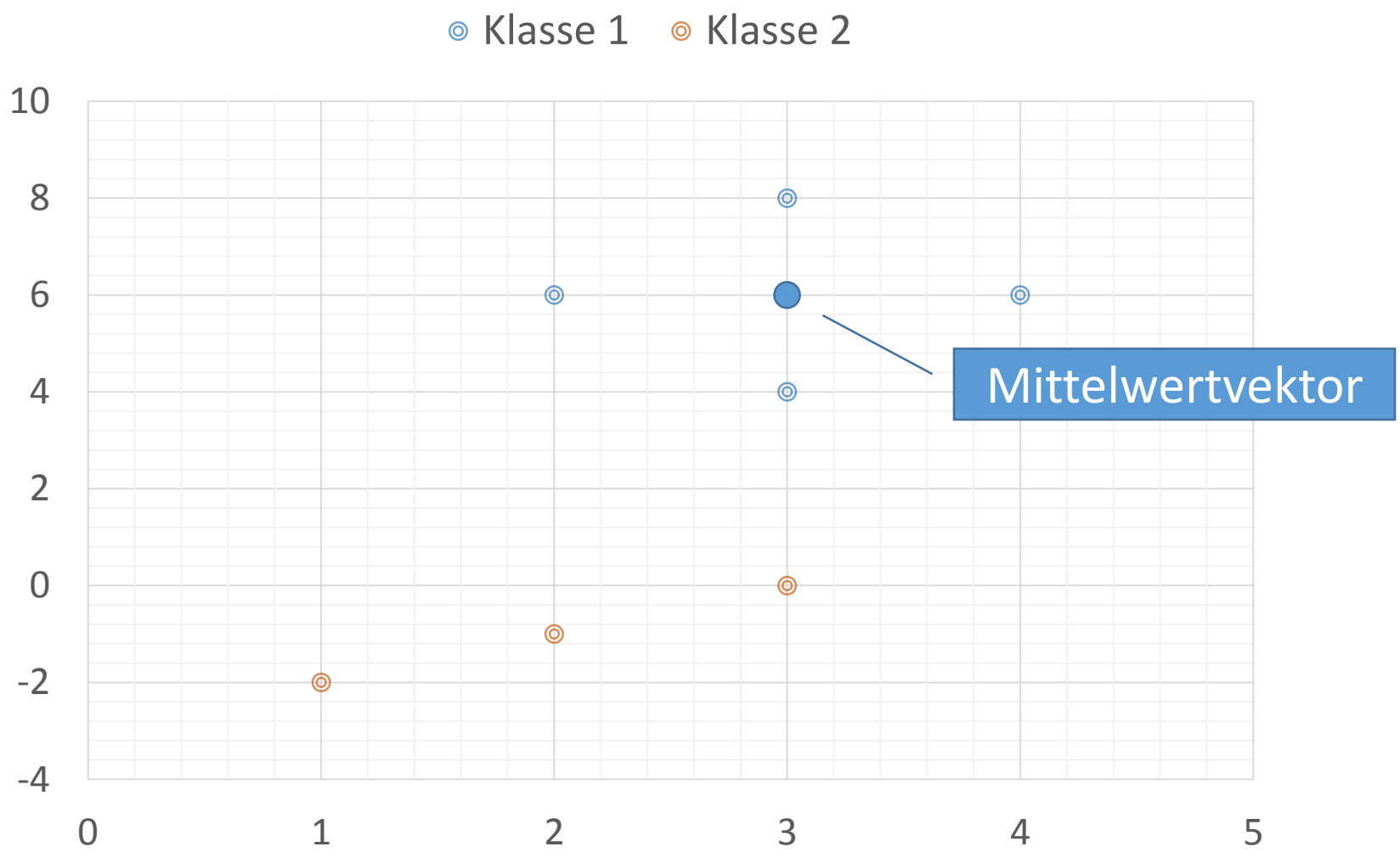
$$= \begin{bmatrix}3\\6\end{bmatrix}$$

Klasse 1

$$\begin{bmatrix}3\\4\end{bmatrix}, \begin{bmatrix}3\\8\end{bmatrix}, \begin{bmatrix}2\\6\end{bmatrix}, \begin{bmatrix}4\\6\end{bmatrix}$$

© Nicole M. Artner

Schätzung Mittelwertvektor



© Nicole M. Artner

Schätzung Kovarianzmatrix

Klasse 1

$\begin{bmatrix}3\\4\end{bmatrix}, \begin{bmatrix}3\\8\end{bmatrix}, \begin{bmatrix}2\\6\end{bmatrix}, \begin{bmatrix}4\\6\end{bmatrix}$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^T$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{3}\left\{\left(\left[\begin{bmatrix}3\\4\end{bmatrix} - \begin{bmatrix}3\\6\end{bmatrix}\right]\left[\begin{bmatrix}3\\4\end{bmatrix} - \begin{bmatrix}3\\6\end{bmatrix}\right]^T\right) + \left(\left[\begin{bmatrix}3\\8\end{bmatrix} - \begin{bmatrix}3\\6\end{bmatrix}\right]\left[\begin{bmatrix}3\\8\end{bmatrix} - \begin{bmatrix}3\\6\end{bmatrix}\right]^T\right) + \right.$$

$$\left.\left(\left[\begin{bmatrix}2\\6\end{bmatrix} - \begin{bmatrix}3\\6\end{bmatrix}\right]\left[\begin{bmatrix}2\\6\end{bmatrix} - \begin{bmatrix}3\\6\end{bmatrix}\right]^T\right) + \left(\left[\begin{bmatrix}4\\6\end{bmatrix} - \begin{bmatrix}3\\6\end{bmatrix}\right]\left[\begin{bmatrix}4\\6\end{bmatrix} - \begin{bmatrix}3\\6\end{bmatrix}\right]^T\right)\right\} =$$

$$\frac{1}{3}\left\{\left(\begin{bmatrix}0\\-2\end{bmatrix}\begin{bmatrix}0 & -2\end{bmatrix}\right) + \left(\begin{bmatrix}0\\2\end{bmatrix}\begin{bmatrix}0 & 2\end{bmatrix}\right) + \left(\begin{bmatrix}-1\\0\end{bmatrix}\begin{bmatrix}-1 & 0\end{bmatrix}\right) + \left(\begin{bmatrix}1\\0\end{bmatrix}\begin{bmatrix}1 & 0\end{bmatrix}\right)\right\} =$$

$$\frac{1}{3}\left\{\begin{pmatrix}0 & 0\\0 & 4\end{pmatrix} + \begin{pmatrix}0 & 0\\0 & 4\end{pmatrix} + \begin{pmatrix}1 & 0\\0 & 0\end{pmatrix} + \begin{pmatrix}1 & 0\\0 & 0\end{pmatrix}\right\} = \begin{pmatrix}2/3 & 0\\0 & 8/3\end{pmatrix}$$

© Nicole M. Artner

Übung

… berechne selbst die Schätzung von Mittelwertvektor und Kovarianzmatrix für Klasse 2.

Klasse 2

$$\begin{bmatrix} 3 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ -2 \end{bmatrix}, \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

© Nicole M. Artner