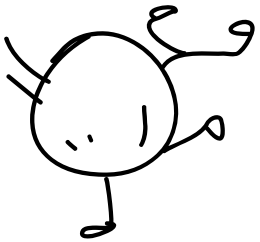


Normalverteilungen

Vorlesung 186.844

6.11.2014

TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

Überblick

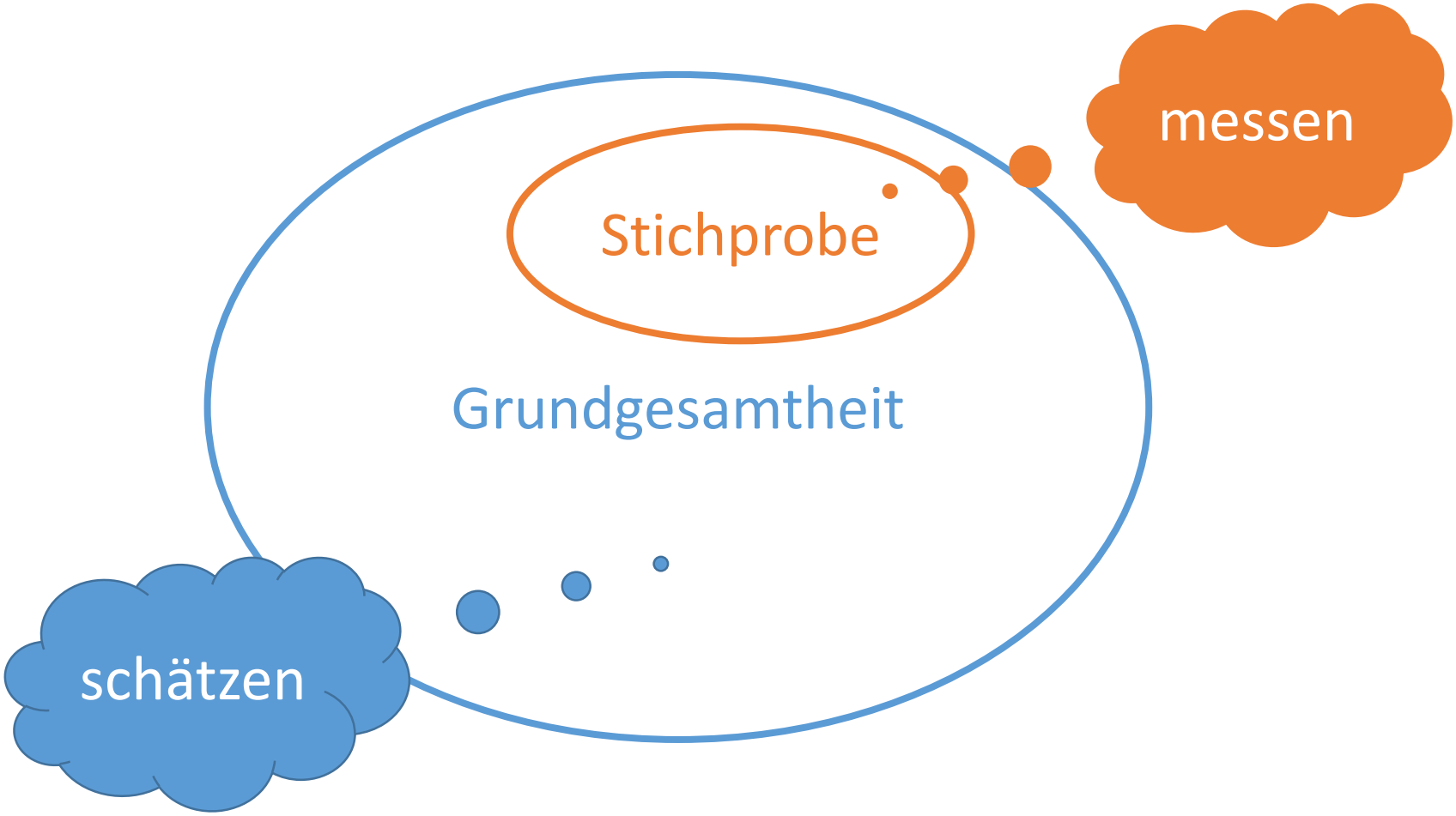
I. Abstecher: Erwartungstreue Schätzer
II. Multivariate Normalverteilungen
III. Bayes-Theorem für multivariate Normalverteilungen
IV. Diskriminantenfunktionen

© Nicole M. Artner

Erwartungstreue Schätzer

© Nicole M. Artner

Grundgesamtheit



© Nicole M. Artner

Eigenschaften von Schätzern

- $\hat{\theta}$ sei der Schätzer des Parameters θ
- wünschenswerte Eigenschaften von $\hat{\theta}$:

Erwartungstreue: $\varepsilon\left[\hat{\theta}\right] = \theta$

Effizienz: je geringer die Varianz des Schätzers $\sigma_{\hat{\theta}}^2$, desto effizienter ist er

Konsistenz: $\hat{\theta} \rightarrow \theta$ für $N \rightarrow \infty$
„Gesetz der großen Zahlen“

© Nicole M. Artner

Schätzung Mittelwert

… ist erwartungstreu, weil

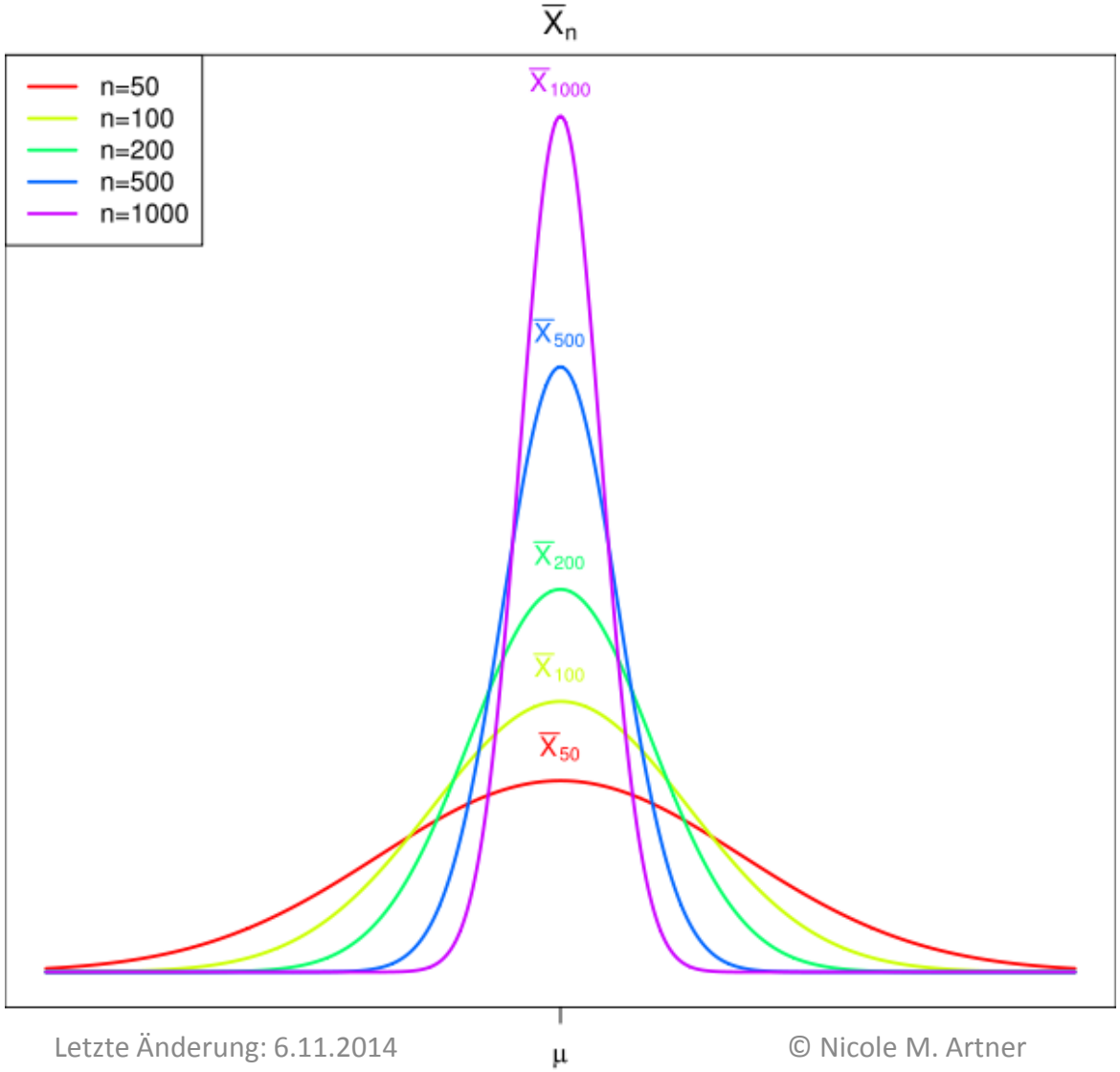
$$\varepsilon[\hat{\mu}] = \varepsilon\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N} \varepsilon[x_i] = \frac{1}{N} \cdot N \cdot \mu = \mu$$

Erklärung:

Wenn man die Stichproben x_i zufällig aus der Grundgesamtheit wählt, dann gilt $\varepsilon[x_i] = \mu$ (siehe Vorlesung 02_Grundlagen_I, Folie 34). Daraus folgt $\mathcal{E}[\hat{\mu}] = \mu$ und das bedeutet, dass der **Mittelwertschätzer erwartungstreu** ist.

© Nicole M. Artner

Effizienz des Mittelwertschätzers



Je geringer die Varianz des Schätzers $\sigma_{\hat{\mu}}^2\ (\bar{X})$:

- desto effizienter ist er
- desto wahrscheinlicher liegt die Schätzung nahe am wahren Wert

© Nicole M. Artner

Schätzung Varianz

… erwartungstreuer Schätzer, wenn wahres Mittel μ bekannt ist:

$$\varepsilon[\hat{\sigma}^2] = \varepsilon\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right) = \frac{1}{N}\sum_{i=1}^{N}\varepsilon[(x_i - \mu)^2] = \frac{1}{N}\cdot N \cdot \sigma^2 = \sigma^2$$

($\varepsilon[(x - \mu)^2] = \sigma^2$ → siehe Vorlesung 02_Grundlagen_I, Folie 37)

ABER:

Der wahre Mittelwert μ ist meist nicht bekannt …

© Nicole M. Artner

Unkorrigierte Stichprobenvarianz

- wird auch als „verzerrte Varianz“ bezeichnet
- verwendet geschätzten Mittelwert $\hat{\mu}$ (aus Stichprobe)

Unkorrigierte Stichprobenvarianz (verzerrte Varianz):

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2 = \frac{(x_1 - \hat{\mu})^2 + \cdots + (x_N - \hat{\mu})^2}{N}$$

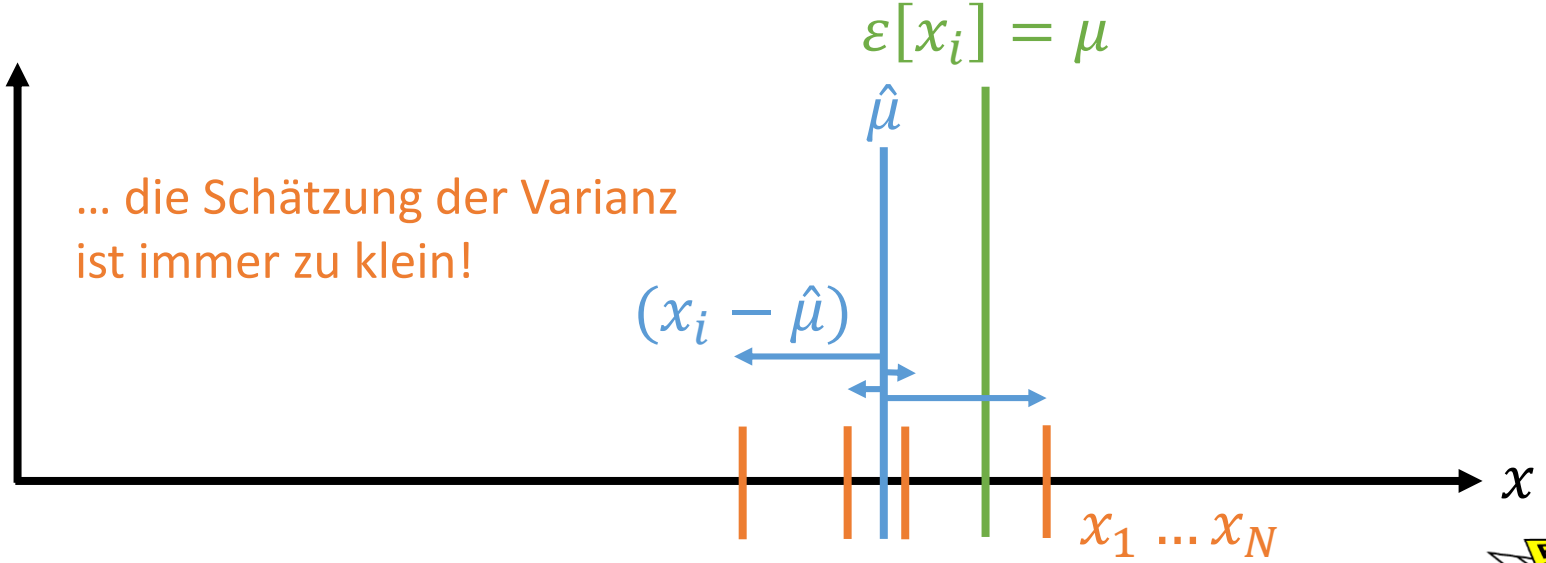
© Nicole M. Artner

Eigenschaften

… der unkorrigierten Stichprobenvarianz

Konsistenz: je mehr Stichproben umso genauer wird der wahre Mittelwert geschätzt und umso besser wird auch die Schätzung der Varianz.

Erwartungstreue: nicht erfüllt → $\varepsilon[\hat{\sigma}^2] \neq \sigma^2$



© Nicole M. Artner

Beispiel

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$$

… wir wollen zeigen, dass $\varepsilon[\hat{\sigma}^2] \neq \sigma^2$

$N = 2$ Messungen

$X = \{x_1, x_2\}$ Beobachtungen (Stichprobe)

$$\varepsilon\left[\frac{\left(x_1 - \frac{x_1 + x_2}{2}\right)^2 + \left(x_2 - \frac{x_1 + x_2}{2}\right)^2}{2}\right] =$$

$$\frac{1}{2}\varepsilon\left[\left(\frac{x_1}{2} - \frac{x_2}{2}\right)^2 + \left(\frac{x_2}{2} - \frac{x_1}{2}\right)^2\right] =$$

© Nicole M. Artner

Beispiel

$$\frac{1}{2}\varepsilon\left[\boxed{\left(\frac{x_1}{2}-\frac{x_2}{2}\right)^2+\left(\frac{x_2}{2}-\frac{x_1}{2}\right)^2}\right] \rightarrow 2\left(\frac{x_1}{2}-\frac{x_2}{2}\right)^2=\frac{1}{2}(x_1-x_2)^2$$

$$=\frac{1}{2}\varepsilon\left[\frac{1}{2}(x_1-x_2)^2\right]=\frac{1}{4}\varepsilon[(x_1-x_2)^2]=\frac{1}{4}\varepsilon[x_1^2-2x_1x_2+x_2^2]$$

$$=\frac{1}{4}(\varepsilon[x_1^2]-2\varepsilon[x_1x_2]+\varepsilon[x_2^2])=\frac{1}{4}(2\varepsilon[x^2]-2\varepsilon[x]^2)$$

$$=\frac{1}{2}(\boxed{\varepsilon[x^2]-\varepsilon[x]^2})$$

$$\sigma^2=\varepsilon[x^2]-\varepsilon[x]^2$$

→ siehe Vorlesung 02_Grundlagen_I, Folie 37

© Nicole M. Artner

Schlussfolgerung

$$= \frac{1}{2}(\varepsilon[x^2] - \varepsilon[x]^2) = \boxed{\frac{\sigma^2}{2} \neq \sigma^2}$$

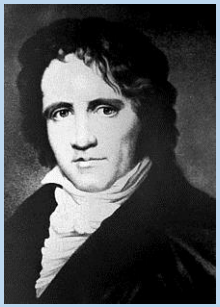
Die unkorrigierte Stichprobenvarianz ist NICHT erwartungstreu!

Was müsste man tun, damit der Schätzer erwartungstreu wird?

- in unserem Beispiel: mit 2 multiplizieren
- Allgemeiner: mit Korrekturfaktor $\frac{N}{N-1}$ multiplizieren

Besselsche Korrektur

$$\frac{N}{N-1}$$



[Bildquelle: http://de.wikipedia.org/wiki/Friedrich_Wilhelm_Bessel]

© Nicole M. Artner

Korrigierte Stichprobenvarianz

… ist erwartungstreu

Korrigierte Stichprobenvarianz (unverzerrte Varianz):

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})^2 = \frac{N}{N-1} \cdot \frac{(x_1 - \hat{\mu})^2 + \cdots + (x_N - \hat{\mu})^2}{N}$$

Umso größer die Stichprobe desto kleiner wird der Korrekturfaktor. Die Verwendung der korrigierten Stichprobenvarianz ist also nur bei kleinen Stichproben (Trainingsdatensätzen) relevant!

© Nicole M. Artner

Multivariate Normalverteilungen

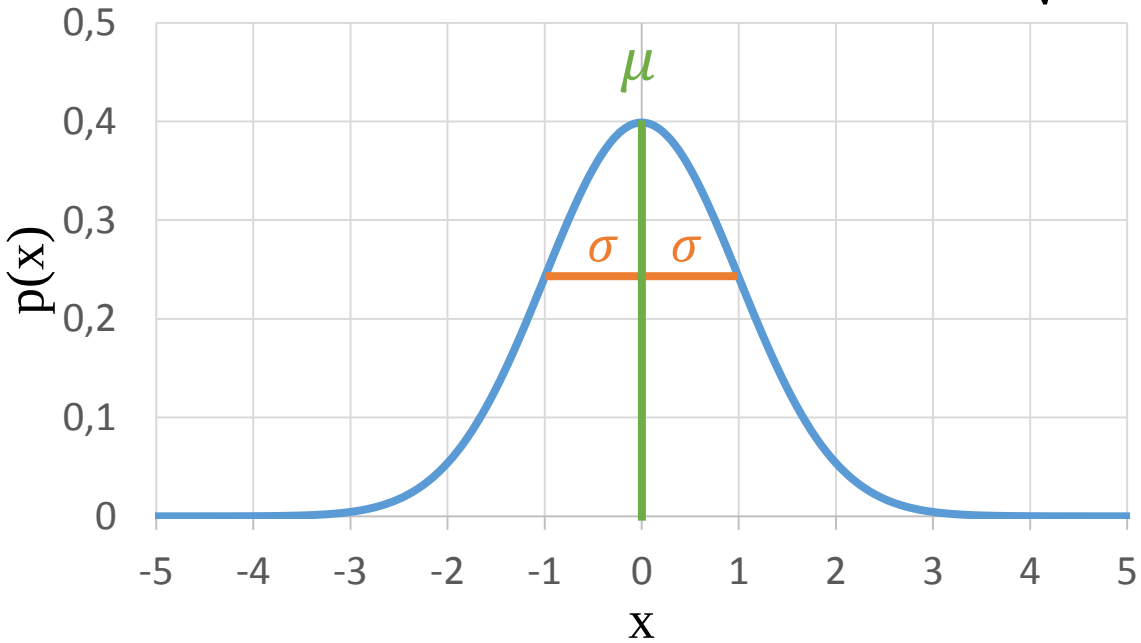
© Nicole M. Artner

WH: Univariate Normalverteilung

- Parameter:
  - Mittelwert μ
  - Varianz σ^2

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

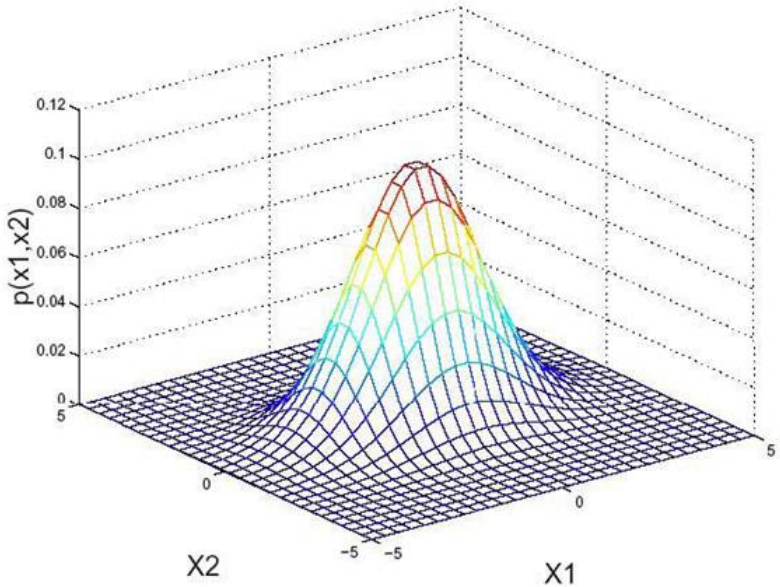
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



© Nicole M. Artner

Multivariate Normalverteilung

- Parameter:
  - Mittelwertvektor $\boldsymbol{\mu}$
  - Kovarianzmatrix $\boldsymbol{\Sigma}$



$$p(\boldsymbol{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right]}$$

© Nicole M. Artner

WH: Mittelwertvektor

$$\boldsymbol{\mu} = \mathcal{E}[\mathbf{x}] = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} \mathcal{E}[x_1] \\ \mathcal{E}[x_2] \\ \vdots \\ \mathcal{E}[x_p] \end{bmatrix}$$

$\mu_i = \mathcal{E}[x_i]$ 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

WH: 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

WH: 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

Eigenschaften

… einer Kovarianzmatrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix} \begin{matrix} \textbf{Varianzen} \\ \textbf{Kovarianzen} \\ \\ \\ \end{matrix}$$

- **symmetrisch**: $\sigma_{ij} = \sigma_{ji}$ für $1 \leq i, j \leq p$ und somit $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T$
- **positiv semi-definitiv**: $\boldsymbol{\Sigma}$ hat nicht-negative Eigenwerte
- diese Eigenschaften gelten auch für $\widehat{\boldsymbol{\Sigma}}$

© Nicole M. Artner

Im Detail …

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]}$$

$\boldsymbol{x}$ … p-dimensionaler Merkmalsvektor (Spaltenvektor)

$\boldsymbol{\mu}$ … p-dimensionaler Mittelwertvektor (Spaltenvektor)

$\boldsymbol{\Sigma}$ … $p \times p$-dimensionale Kovarianzmatrix

$|\boldsymbol{\Sigma}|$ … Determinante von $\boldsymbol{\Sigma}$

$\boldsymbol{\Sigma}^{-1}$ … Inverse von $\boldsymbol{\Sigma}$

$(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$… Mahalanobisdistanz

© Nicole M. Artner

Determinante und Inverse

Determinante von $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ ist $\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$

$|\boldsymbol{\Sigma}| = 0$ … wenn $\boldsymbol{\Sigma}$ singulär ist → hat $\boldsymbol{\Sigma}$ keine Inverse

Inverse $\boldsymbol{\Sigma}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

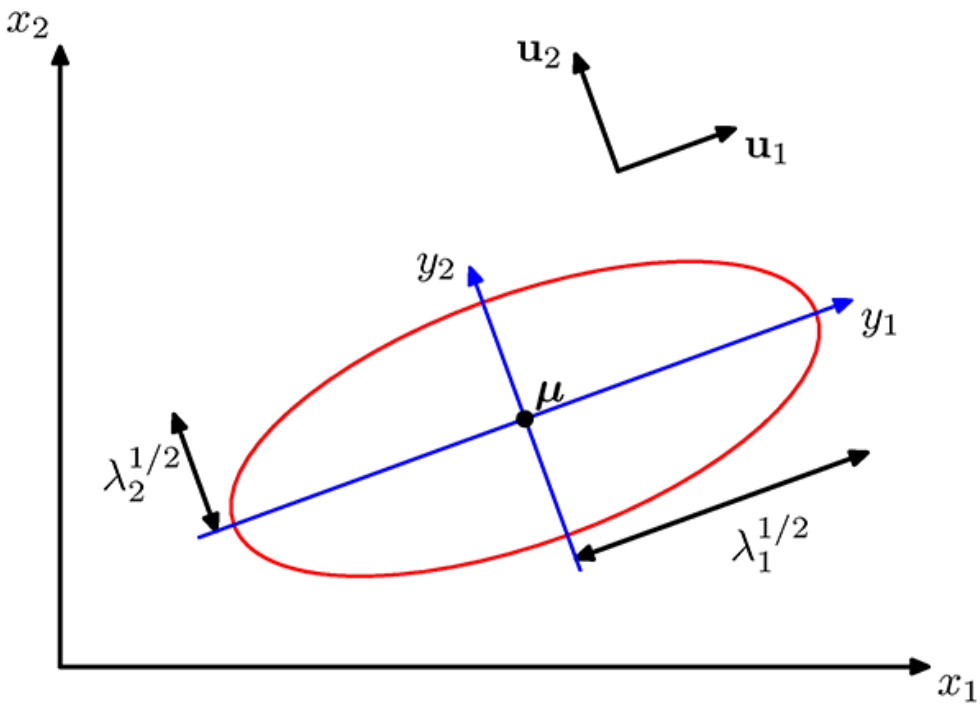
Division durch Null nicht möglich!

Für diese VO reicht es aus, wenn ihr die Determinante und Inverse einer 2×2 Matrix bestimmten könnt.

© Nicole M. Artner

Mahalanobisdistanz

$$d^2(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$$



- alle Punkte $\{\boldsymbol{x} : d^2(\boldsymbol{x}) = c\}$ liegen auf einem Hyperellipsoid im $\mathbb{R}^p$ mit Mittelpunkt $\boldsymbol{\mu}$
- für alle diese $\boldsymbol{x}$ liefert die Dichtefunktion $p(\boldsymbol{x})$ denselben Wert
- Hauptachsen der Ellipse sind definiert durch **Eigenvektoren** $\boldsymbol{u}_1$ und $\boldsymbol{u}_2$ von $\boldsymbol{\Sigma}$
- **Eigenwerte** λ_1 und λ_2 bestimmen die Länge der Achsen

© Nicole M. Artner

Bayes-Theorem für multivariate Normalverteilungen

© Nicole M. Artner

Vorgehensweise

… anhand eines Beispiels mit 2 Klassen

Iris virginica

Iris versicolor

- jeweils 4 Merkmale
  - Sepalumlänge, -breite, Petalumlänge, -breite

© Nicole M. Artner

Hypothese

- die Dichtefunktion $p(\boldsymbol{x}|\omega_j)$ jeder Klasse j kann durch eine multivariate Normalverteilung $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ beschrieben werden

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

© Nicole M. Artner

Parameterschätzung

- aus dem Trainingsdatensatz schätzt man Mittelwertvektor $\hat{\boldsymbol{\mu}}_j$ und Kovarianzmatrix $\hat{\boldsymbol{\Sigma}}_j$ je Klasse

Iris virginica

Trainingsdaten

$$\hat{\boldsymbol{\mu}}_1 = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \qquad \hat{\boldsymbol{\Sigma}}_1 = \begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

Iris versicolor

Trainingsdaten

$$\hat{\boldsymbol{\mu}}_2 = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \qquad \hat{\boldsymbol{\Sigma}}_2 = \begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

© Nicole M. Artner

Parameterschätzung

… mit 4 Merkmalen

$$\hat{\boldsymbol{\mu}}_j = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \qquad \hat{\boldsymbol{\Sigma}}_j = \begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

- 4 Mittelwerte
- 4 Varianzen
- 6 Kovarianzen

Insgesamt 4 + (4 + 6) = 14 Parameter je Klasse *j*

Kann man die Anzahl der zu schätzenden Parameter reduzieren?

© Nicole M. Artner

Naive Bayes-Klassifikator

- um Anzahl der Parameter zu reduzieren, werden Annahmen über die Form der Kovarianzmatrizen gemacht

- **Naive Bayes-Klassifikator** nimmt an, dass es **keine Korrelation zwischen den Merkmalen** gibt → $\sigma_{xy} = 0$

- je Klasse 8 Parameter
  - 4 Mittelwerte + 4 Varianzen
  - statt 14 Parameter

$$\widehat{\boldsymbol{\Sigma}}_j = \begin{bmatrix} .. & 0 & 0 & 0 \\ 0 & .. & 0 & 0 \\ 0 & 0 & .. & 0 \\ 0 & 0 & 0 & .. \end{bmatrix}$$

© Nicole M. Artner

weitere Annahmen …

- Alle Klassen haben die **gleiche Kovarianzmatrix**
  - im Beispiel: $\widehat{\boldsymbol{\Sigma}}_1 = \widehat{\boldsymbol{\Sigma}}_2$
  - 4 Mittelwerte je Klasse
  - 10 Parameter (1 Kovarianzmatrix) für alle Klassen
  - nur 2 x 4 + 10 = 18 Parameter → statt 14 x 2 = 28

- Alle Klassen haben die **gleiche Kovarianzmatrix** und sie ist **diagonal** → $\sigma_{xy} = 0$
  - 4 Mittelwerte je Klasse
  - 4 Varianzen (1 diagonale Kovarianzmatrix) für alle Klassen
  - nur 2 x 4 + 4 = 12 Parameter → statt 28 Parameter im allgemeinen Fall

© Nicole M. Artner

Dichtefunktionen

… mit den geschätzten Parametern

Iris virginica

$$p(\boldsymbol{x}|\omega_1) = \frac{1}{(2\pi)^{\frac{p}{2}}\left|\widehat{\boldsymbol{\Sigma}}_1\right|^{\frac{1}{2}}} e^{\left[-\frac{1}{2}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_1)^T \widehat{\boldsymbol{\Sigma}}_1^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_1)\right]}$$

Iris versicolor

$$p(\boldsymbol{x}|\omega_2) = \frac{1}{(2\pi)^{\frac{p}{2}}\left|\widehat{\boldsymbol{\Sigma}}_2\right|^{\frac{1}{2}}} e^{\left[-\frac{1}{2}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_2)^T \widehat{\boldsymbol{\Sigma}}_2^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_2)\right]}$$

© Nicole M. Artner

Bayes-Klassifikator

- man kann anhand der Mahalanobisdistanz klassifizieren, wenn:
  - keine normalisierten/genauen Klassenwahrscheinlichkeiten $P(w_j|\boldsymbol{x})$ notwendig sind
  - die Priors gleich sind, im Beispiel: $P(w_1) = P(w_2)$

$$p(\boldsymbol{x}|\omega_1) = \cancel{\frac{1}{(2\pi)^{\frac{p}{2}}|\widehat{\boldsymbol{\Sigma}}_1|^{\frac{1}{2}}}} e^{\left[-\frac{1}{2}\boxed{(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_1)^T\widehat{\boldsymbol{\Sigma}}_1^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_1)}\right]} \rightarrow$$

Mahalanobisdistanz Klasse 1

$$p(\boldsymbol{x}|\omega_2) = \cancel{\frac{1}{(2\pi)^{\frac{p}{2}}|\widehat{\boldsymbol{\Sigma}}_2|^{\frac{1}{2}}}} e^{\left[-\frac{1}{2}\boxed{(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_2)^T\widehat{\boldsymbol{\Sigma}}_2^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_2)}\right]} \rightarrow$$

Mahalanobisdistanz Klasse 2

© Nicole M. Artner

Bayes-Klassifikator

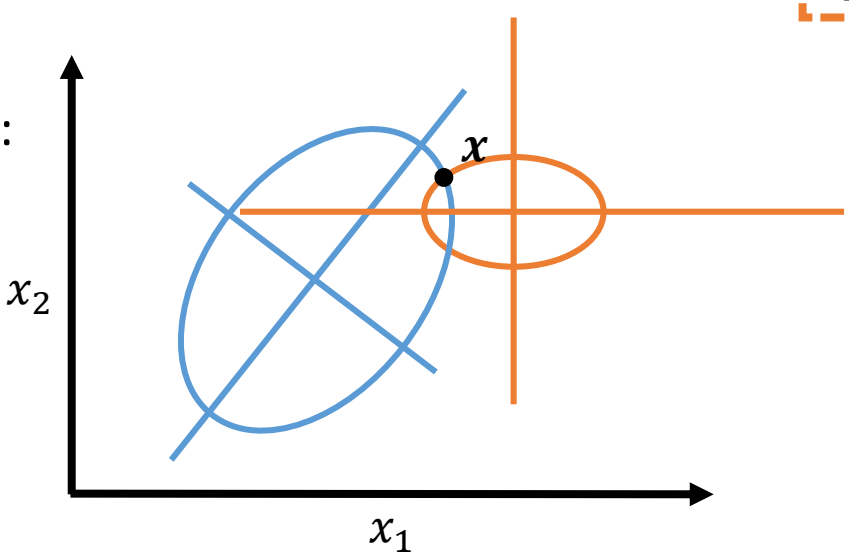
$$d_{w_1}^2(\boldsymbol{x}) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_1)^T \hat{\boldsymbol{\Sigma}}_1^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_1) \quad d_{w_2}^2(\boldsymbol{x}) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_2)^T \hat{\boldsymbol{\Sigma}}_2^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_2)$$

Iris virginica

$\boldsymbol{x}$ wird der Klasse mit der **kleinsten** Mahalanobisdistanz zugewiesen

Iris versicolor

Vereinfachte Darstellung:



© Nicole M. Artner

Diskriminanten-funktionen

© Nicole M. Artner

Was ist das?

Ein Klassifikator kann als eine **Gruppe von Diskriminanten-funktionen** $g_i(\boldsymbol{x}), i = 1, \dots, c$ dargestellt werden. Der Klassifikator weist den Merkmalsvektor $\boldsymbol{x}$ der Klasse w_i zu, wenn gilt: $g_i(\boldsymbol{x}) > g_j(\boldsymbol{x})$ für alle $j \neq i$.

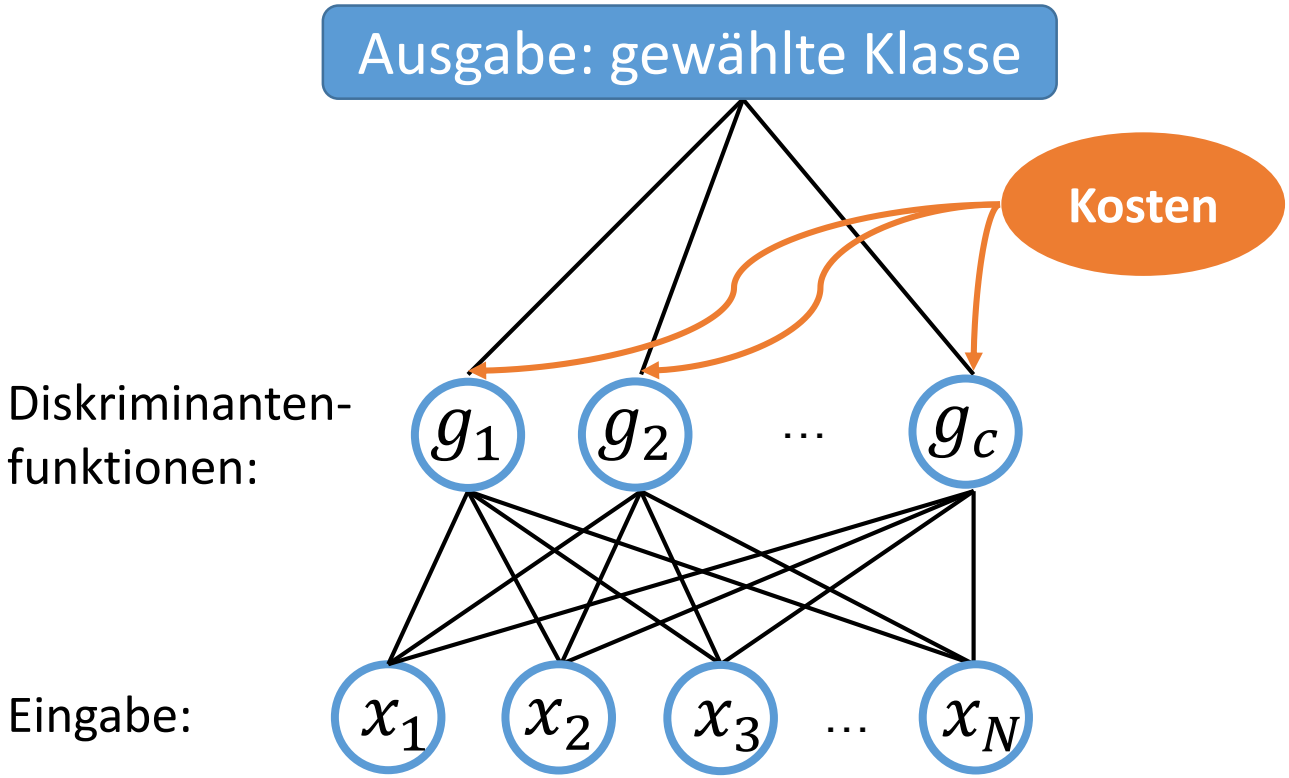
Beschreibung eines Bayes-Klassifikators mit Diskriminanten-funktionen:

$$g_i(\boldsymbol{x}) = P(w_i|\boldsymbol{x})$$

© Nicole M. Artner

Was ist das?

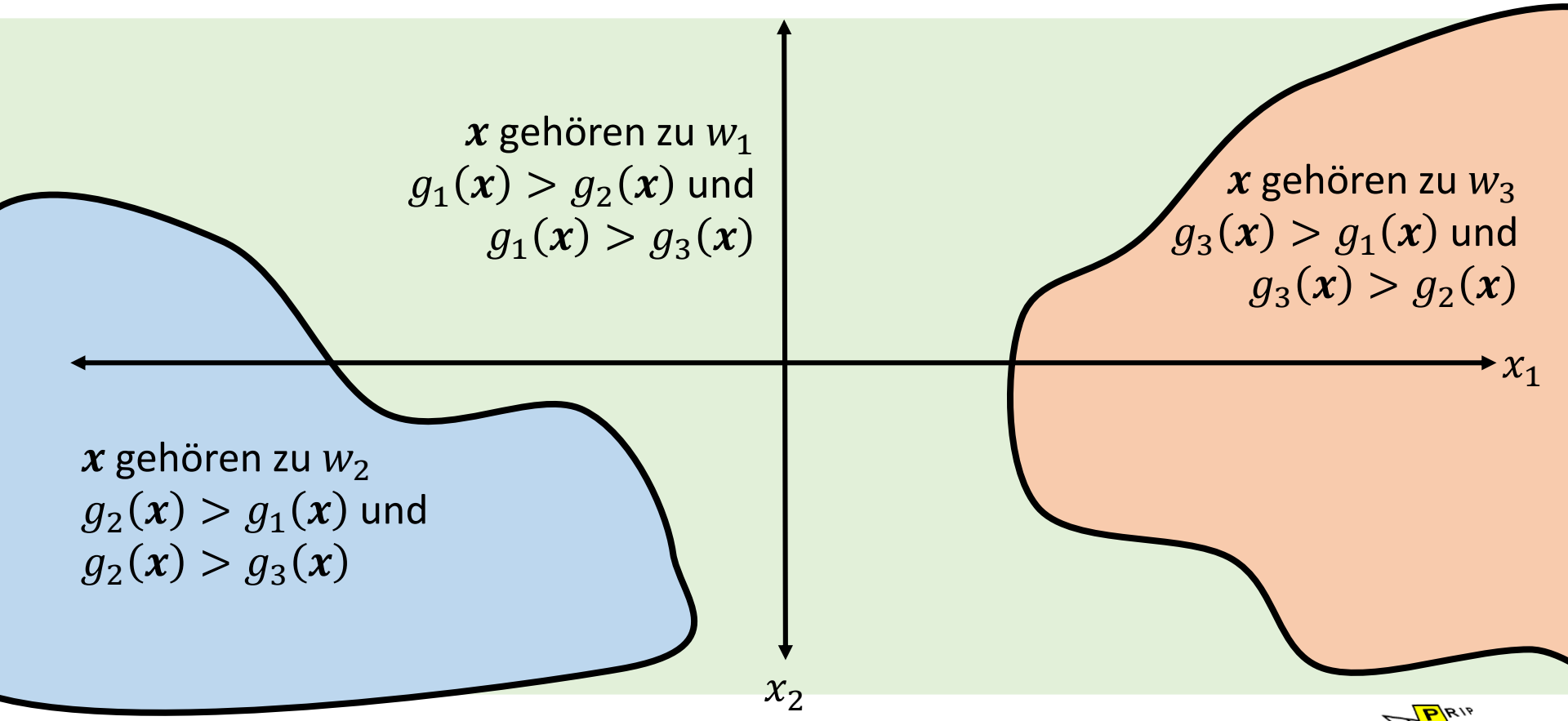
Man kann sich einen Klassifikator als ein Netzwerk vorstellen, dass c Diskriminantenfunktionen berechnet und die Klasse mit dem größten Diskriminanten auswählt.



© Nicole M. Artner

Beispiel

- 2 Merkmale $\boldsymbol{x} = \{x_1, x_2\}$
- 3 Klassen w_1, w_2, w_3 → 3 Diskriminatenfunk. $g_1(\boldsymbol{x}), g_2(\boldsymbol{x}), g_3(\boldsymbol{x})$

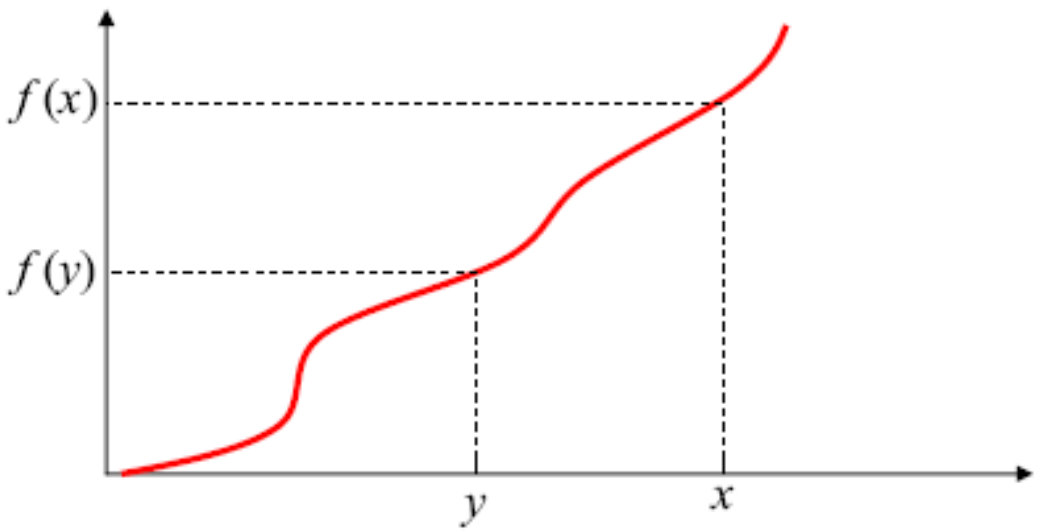


© Nicole M. Artner

Wahl

… der Diskriminantenfunktion ist nicht eindeutig.

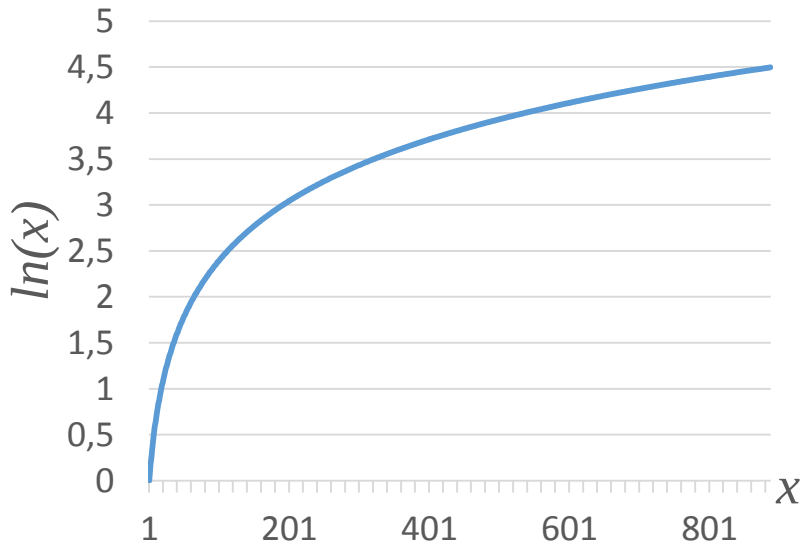
Wenn man alle $g_i(\boldsymbol{x})$ durch $f(g_i(\boldsymbol{x}))$ ersetzt, wobei $f(\cdot)$ eine streng monoton wachsende Funktion ist, dann wird das Ergebnis der Klassifikation dadurch nicht beeinflusst.



© Nicole M. Artner

Optimierung durch kluge Wahl

- Folgende Diskriminantenfunktionen führen alle zum gleichen Klassifikationsergebnis:
  - $g_i(\boldsymbol{x}) = P(\omega_i|\boldsymbol{x})$
  - $g_i(\boldsymbol{x}) = p(\boldsymbol{x}|\omega_i)P(\omega_i)$
  - $g_i(\boldsymbol{x}) = \ln p(\boldsymbol{x}|\omega_i) + \ln P(\omega_i)$



Obwohl alle obigen Diskriminantenfunktionen zum gleichen Ergebnis führen, könnte eine davon besonders **einfach zu verstehen** oder **effizient zu berechnen** sein.

© Nicole M. Artner

Spezialfall: 2 Klassen

Im Fall von 2 Klassen ist es üblich **nur eine Diskriminanten-funktion** zu definieren $g(\boldsymbol{x}) \equiv g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$. Diese Funktion entscheidet für Klasse w_1 wenn $g(\boldsymbol{x}) > 0$; ansonsten für Klasse w_2.

- z.B.: Minimierung der Fehlerrate („minimum-error-rate discriminant function“) beim Klassifizieren mit dem Bayes-Theorem → $g(\boldsymbol{x}) \equiv P(\omega_1|\boldsymbol{x}) - P(\omega_2|\boldsymbol{x})$

Klassifikatoren die nur zwischen zwei Klassen unterscheiden nennt man „Dichotomizer“. Klassifikatoren für $c \geq 3$ bezeich-net man als „Polychotomizer“.

© Nicole M. Artner

Entscheidungsregionen und -grenzen

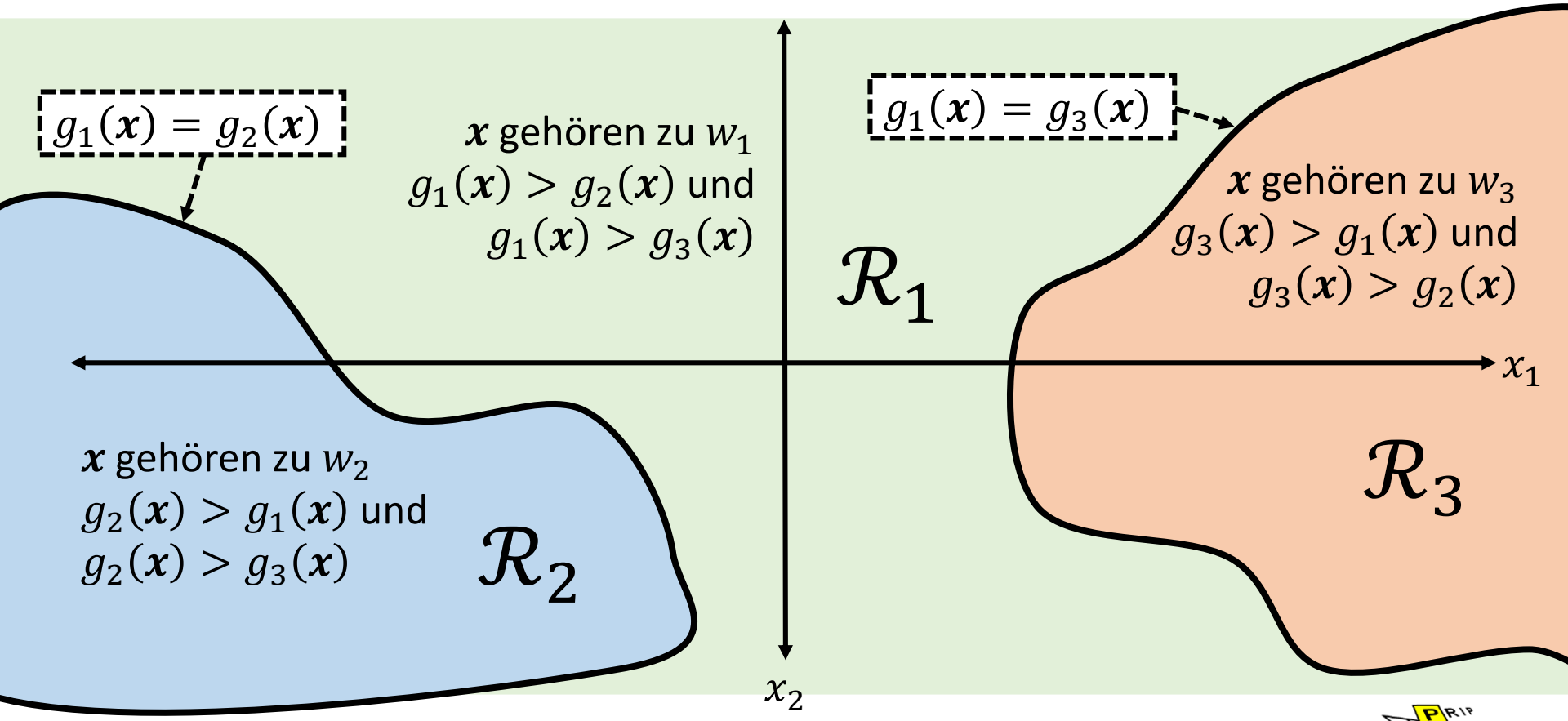
- Diskriminantenfunktionen unterteilen den Merkmalsraum in c Entscheidungsregionen $\mathcal{R}_1, \dots, \mathcal{R}_c$
- wenn $g_i(\boldsymbol{x}) > g_j(\boldsymbol{x})$ für alle $j \neq i$, dann liegt x in $\mathcal{R}_i$
  - $\boldsymbol{x}$ wird der Klasse w_i zugewiesen

- Entscheidungsregionen sind durch Entscheidungs-grenzen („decision boundaries“) getrennt
- Entscheidungsgrenze zwischen $\mathcal{R}_i$ und $\mathcal{R}_j$ ist durch die Gleichung $g_i(\boldsymbol{x}) = g_j(\boldsymbol{x})$ gegeben

© Nicole M. Artner

Beispiel

- Entscheidungsregionen: $\mathcal{R}_1$, $\mathcal{R}_2$ und $\mathcal{R}_3$
- Entscheidungsgrenzen: $g_1(\boldsymbol{x}) = g_2(\boldsymbol{x})$ und $g_1(\boldsymbol{x}) = g_3(\boldsymbol{x})$

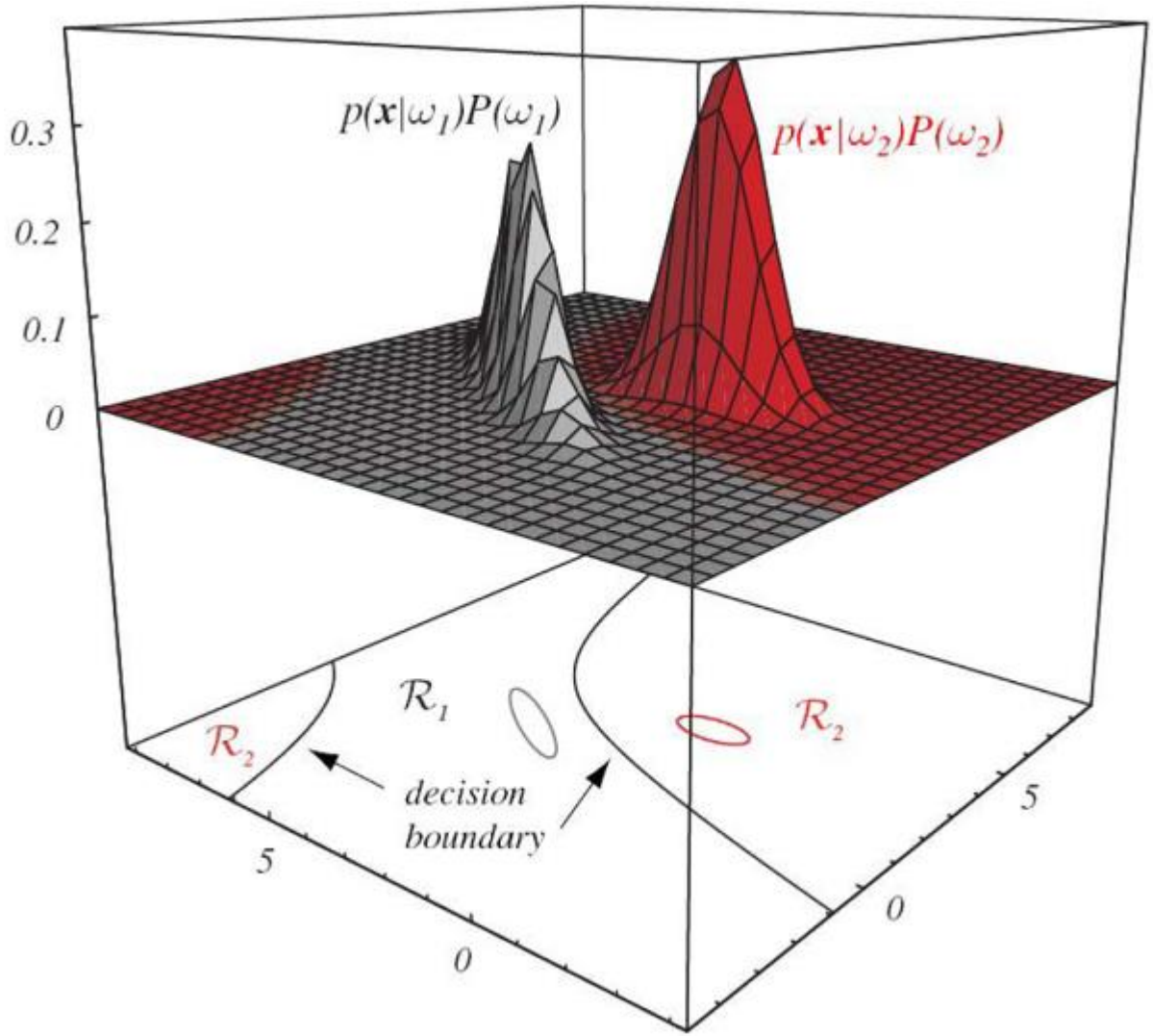


© Nicole M. Artner

Beispiel

[Quelle: Duda et al., 2001]

- bivariate Normal-verteilung je Klasse
- Entscheidungsgrenze besteht aus 2 Hyperbeln
- Entscheidungsregion $\mathcal{R}_2$ ist nicht zusammenhängend



Diskriminatenfunktionen

… für multivariate Normalverteilungen unter Verwendung eines **Bayes-Klassifikators**

- durch Logarithmieren der Posteriors

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

- erhält man die folgenden (optimalen) Diskriminanten-funktionen

$$g_j(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T \Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j) - \frac{1}{2}\ln|\Sigma_j| + \ln P(\omega_j) - \frac{p}{2}\ln 2\pi - \ln p(\boldsymbol{x})$$

Vereinfachung

$$g_j(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T \Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j) - \frac{1}{2}\ln|\Sigma_j| + \ln P(\omega_j) - \cancel{\frac{p}{2}\ln 2\pi - \ln p(\boldsymbol{x})}$$

- die letzten beiden Terme können weggelassen werden, weil sie unabhängig von den Klassen ω_j sind

- Diskriminantenfunktionen vereinfachen sich zu

$$g_j(\boldsymbol{x}) = -\frac{1}{2}\boxed{d_j^2(\boldsymbol{x})} + \left[-\frac{1}{2}\ln\boxed{|\boldsymbol{\Sigma}_j|} + \ln\boxed{P(\omega_j)}\right]$$

Mahalanobisdistanz

Determinante von $\boldsymbol{\Sigma}$

Prior

© Nicole M. Artner

Beispiel

- Berechnung der Entscheidungsgrenze mit Hilfe der Diskriminatenfunktionen
- Annahme: $p(\boldsymbol{x}|\omega_j)$ → multivariate Normalverteilungen

Klasse 1

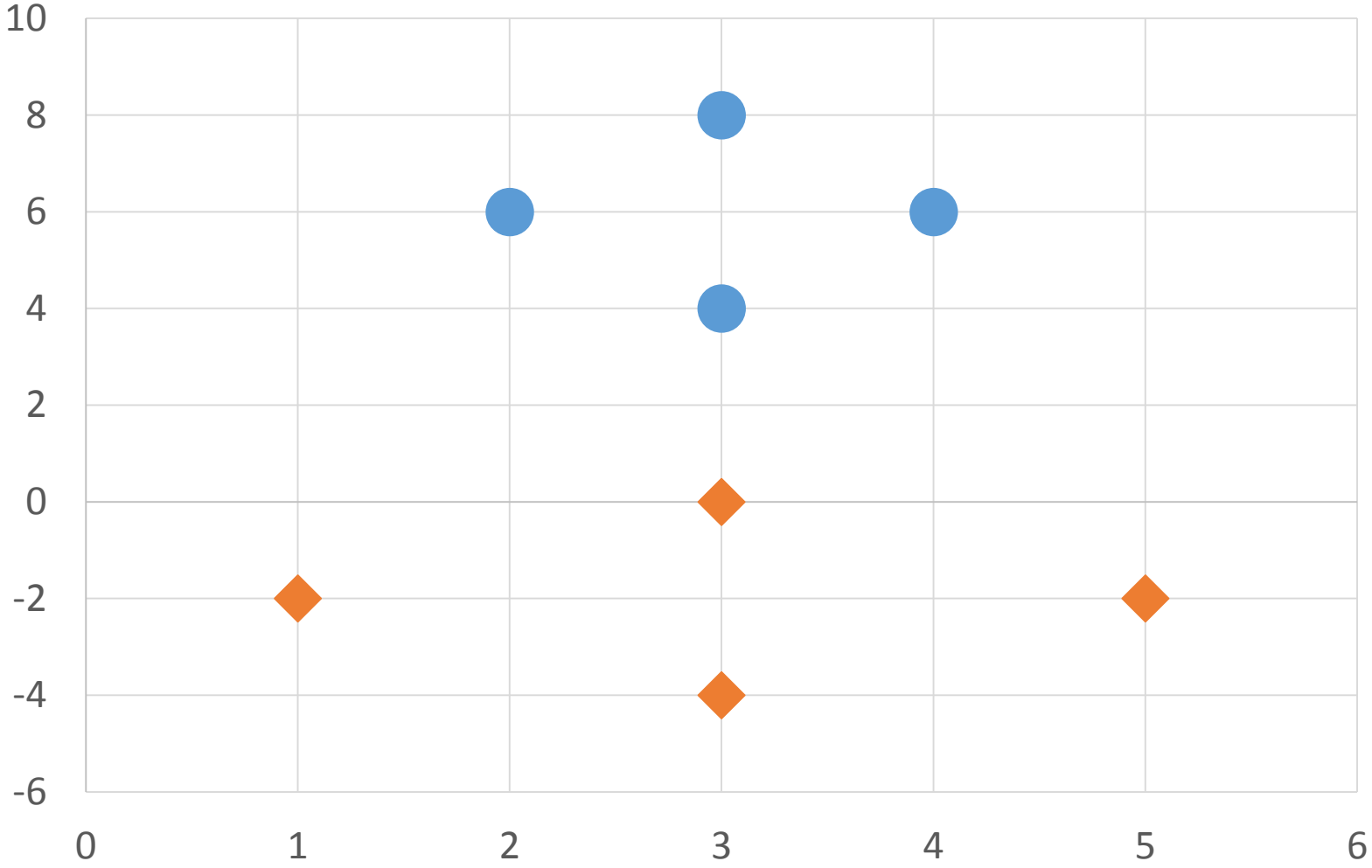
$$\begin{bmatrix}3\\4\end{bmatrix}, \begin{bmatrix}3\\8\end{bmatrix}, \begin{bmatrix}2\\6\end{bmatrix}, \begin{bmatrix}4\\6\end{bmatrix}$$

Klasse 2

$$\begin{bmatrix}3\\0\end{bmatrix}, \begin{bmatrix}3\\-4\end{bmatrix}, \begin{bmatrix}1\\-2\end{bmatrix}, \begin{bmatrix}5\\-2\end{bmatrix}$$

© Nicole M. Artner

Beispieldaten



© Nicole M. Artner

Beispiel

- angenommen die wahren Mittelwerte sind bekannt

$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix} \qquad \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$$

- die Kovarianzmatrizen sehen dann wie folgt aus

$$\widehat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 2 \end{pmatrix} \quad \widehat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

© Nicole M. Artner

Beispiel

- Determinanten

$$|\boldsymbol{\Sigma}| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

$$|\boldsymbol{\Sigma}_1| = 1 \qquad |\boldsymbol{\Sigma}_2| = 4$$

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

- Inverse

$$\widehat{\boldsymbol{\Sigma}}_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \qquad \widehat{\boldsymbol{\Sigma}}_2^{-1} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$$

© Nicole M. Artner

Beispiel

1. Ermittlung von $g_1(\boldsymbol{x})$ und $g_2(\boldsymbol{x})$

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

Einsetzen in …

$$g_j(\boldsymbol{x}) = -\frac{1}{2}\boxed{d_j^2(\boldsymbol{x})} + \left[-\frac{1}{2}\ln\left|\Sigma_j\right| + \ln P\left(\omega_j\right)\right]$$

$$d_j^2(\boldsymbol{x}) = \left(\boldsymbol{x} - \boldsymbol{\mu}_j\right)^T \Sigma_j^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}_j\right)$$

© Nicole M. Artner

Beispiel

- Zwischenergebnisse: Diskriminatenfunktionen der beiden Klassen

$$g_1(\boldsymbol{x}) = -x_1^2 + 6x_1 - \frac{1}{4}x_2^2 + 3x_2 - 18 + \ln 0{,}5$$

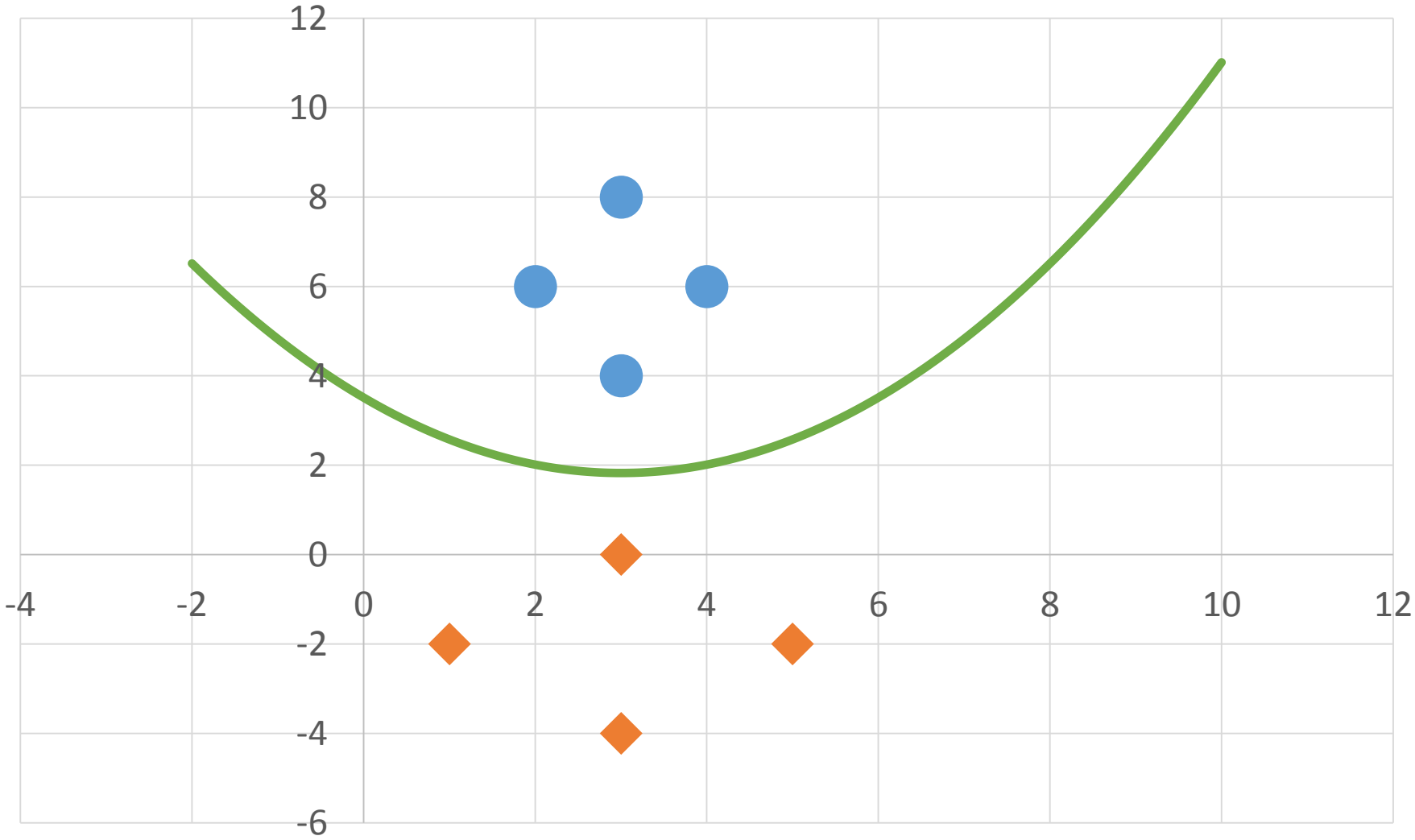
$$g_2(\boldsymbol{x}) = -\frac{1}{4}x_1^2 + \frac{3}{2}x_1 - \frac{1}{4}x_2^2 - x_2 - \frac{13}{4} + \ln 2 + \ln 0{,}5$$

2. Finden der Entscheidungsgrenze durch $g_1(\boldsymbol{x}) = g_2(\boldsymbol{x})$

Ergebnis? Rechne selbst …

© Nicole M. Artner

Ergebnis des Beispiels



© Nicole M. Artner

Spezialfall 1

$$\boldsymbol{\Sigma_i} = \boldsymbol{\sigma^2 I}$$

- Merkmale sind unabhängig $\sigma_{ij} = 0$ für alle $i \neq j$
- alle Merkmale haben dieselbe Varianz $\sigma_{ii} = \sigma^2$

$$g_j(\boldsymbol{x}) = -\frac{1}{2} d_j^2(\boldsymbol{x}) + \left[-\frac{1}{2}\ln|\Sigma_j| + \ln P(\omega_j)\right]$$

$$\Downarrow$$

$$g_j(\boldsymbol{x}) = \left(-\frac{1}{2}\right)\left(\frac{1}{\sigma^2}\right)(\boldsymbol{x} - \boldsymbol{\mu}_j)^T(\boldsymbol{x} - \boldsymbol{\mu}_j) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_j| + \ln P(\omega_j)$$

$$\Downarrow$$

$$= \frac{1}{2\sigma^2}\left(\cancel{\boldsymbol{x}^T\boldsymbol{x}} - 2\boldsymbol{\mu}_j^T\boldsymbol{x} + \boldsymbol{\mu}_j^T\boldsymbol{\mu}_j\right) - \cancel{\frac{1}{2}\ln|\boldsymbol{\Sigma}_j|} + \ln P(\omega_j)$$

… für alle Klassen gleich

© Nicole M. Artner

Spezialfall 1

… man erhält die äquivalente lineare Diskriminaten-funktion

$$g_j(\boldsymbol{x}) = \frac{1}{2\sigma^2}\left(-2\boldsymbol{\mu}_j^T\boldsymbol{x} + \boldsymbol{\mu}_j^T\boldsymbol{\mu}_j\right) + \ln P(\omega_j)$$

$$\Downarrow$$

$$g_j(\boldsymbol{x}) = \boldsymbol{w}_j^T\boldsymbol{x} + w_{j0}$$

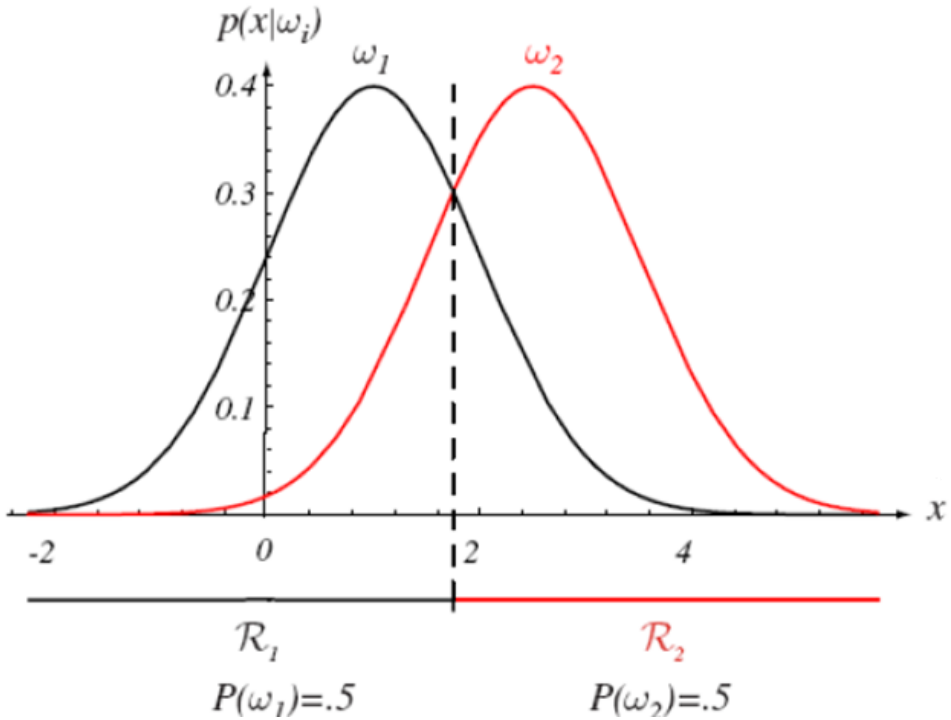
$$\boldsymbol{w}_j = \frac{1}{\sigma^2}\boldsymbol{\mu}_j$$

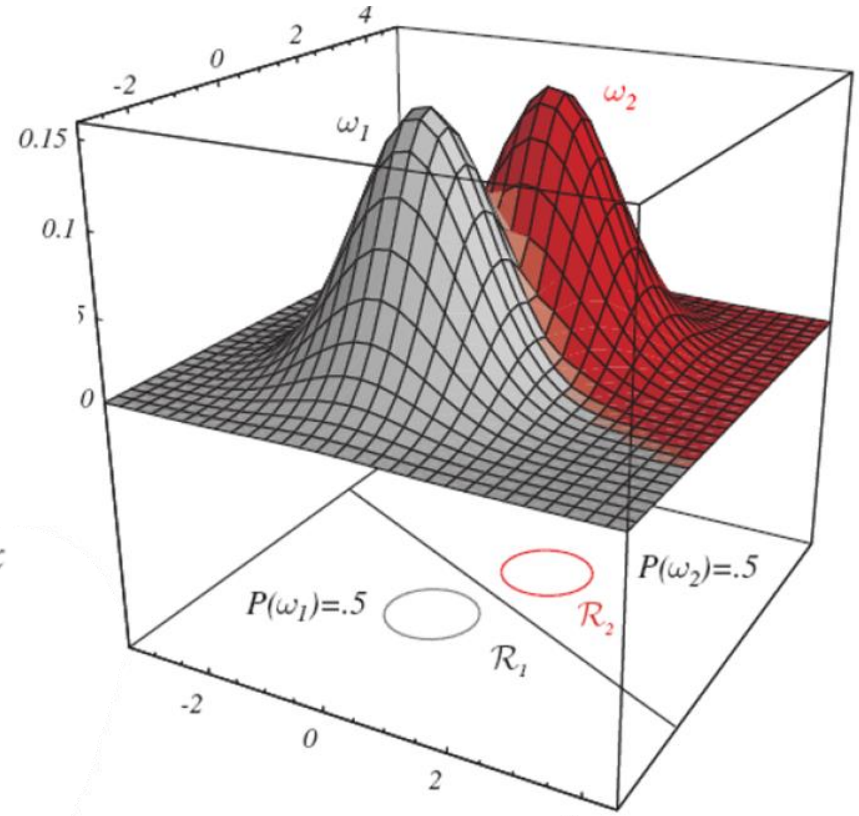
$$w_{j0} = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j + \ln P(\omega_j)$$

© Nicole M. Artner

[Quelle: Duda et al., 2001]

Spezialfall 1

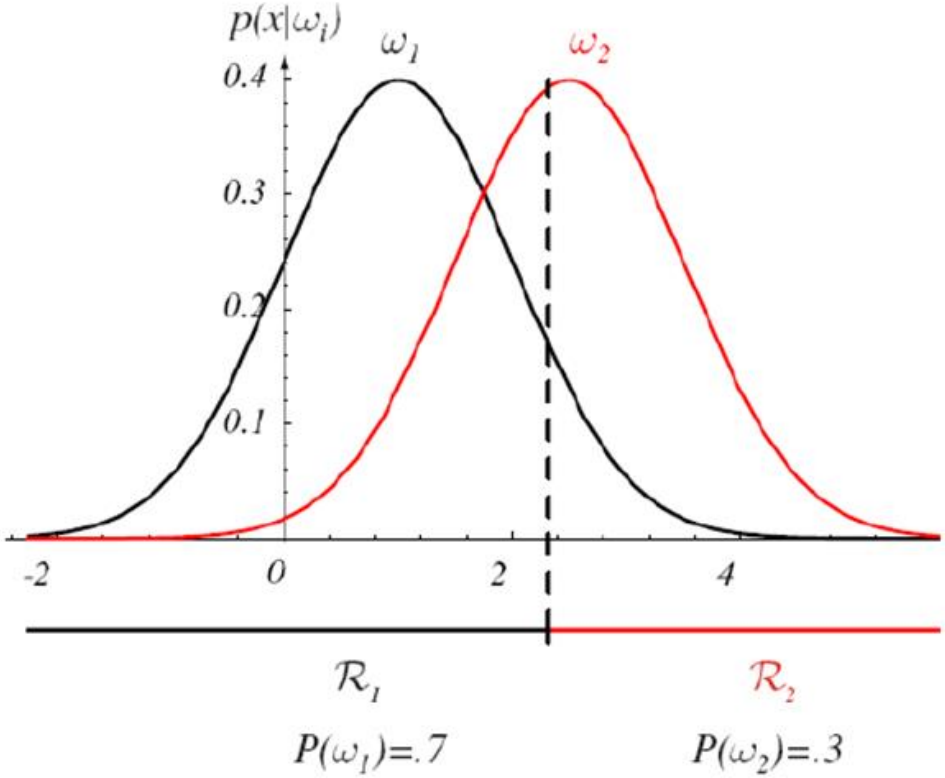


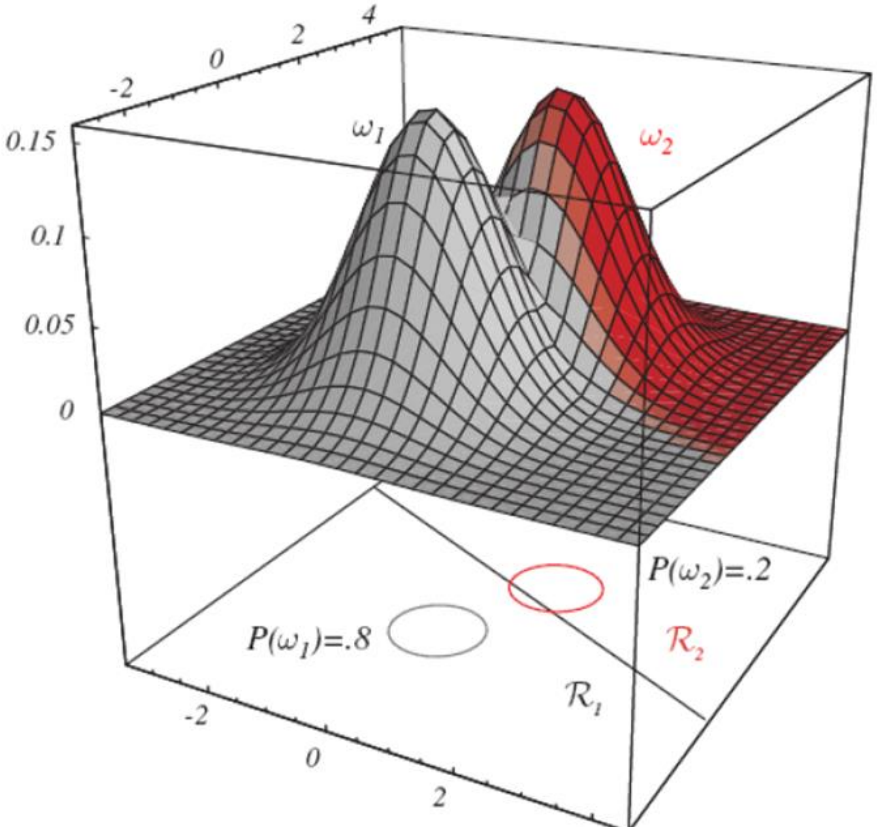


- univariate (links) bzw. bivariate (rechts) Normalverteilungen mit $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = I\sigma^2$
- Entscheidungsgrenzen sind linear und normal zur Verbindungsstrecke zwischen den beiden Klassenmitteln
- bei gleichen Priors verläuft Entscheidungsgrenze durch $(\mu_1 + \mu_2)/2$
- ansonsten wird sie von Prior der wahrscheinlicheren Klasse wegverschoben.

[Quelle: Duda et al., 2001]

Spezialfall 1

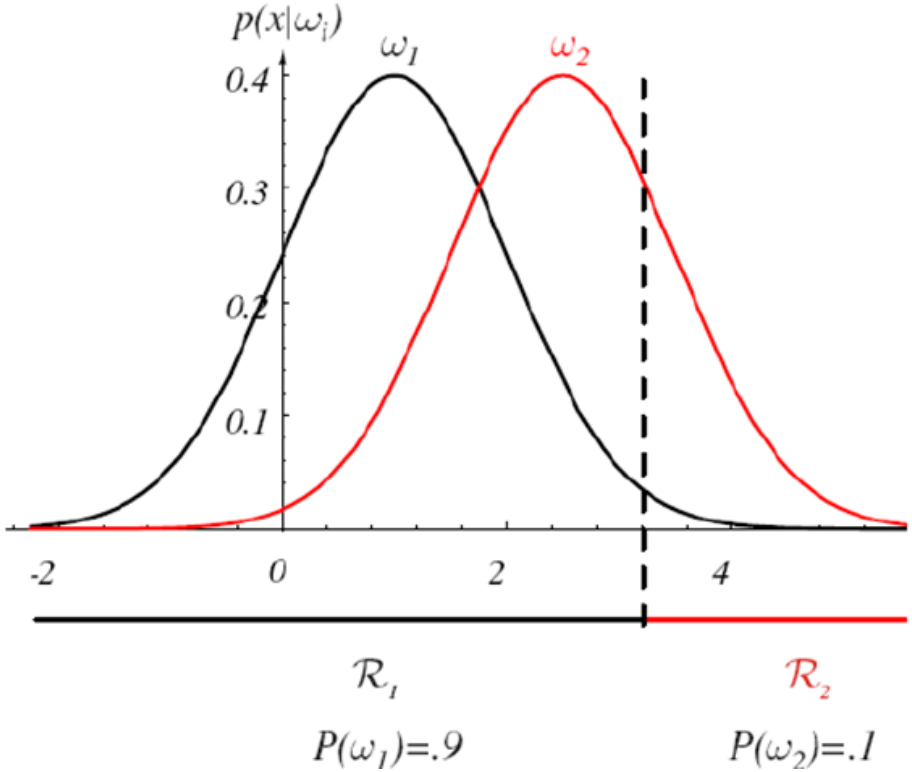


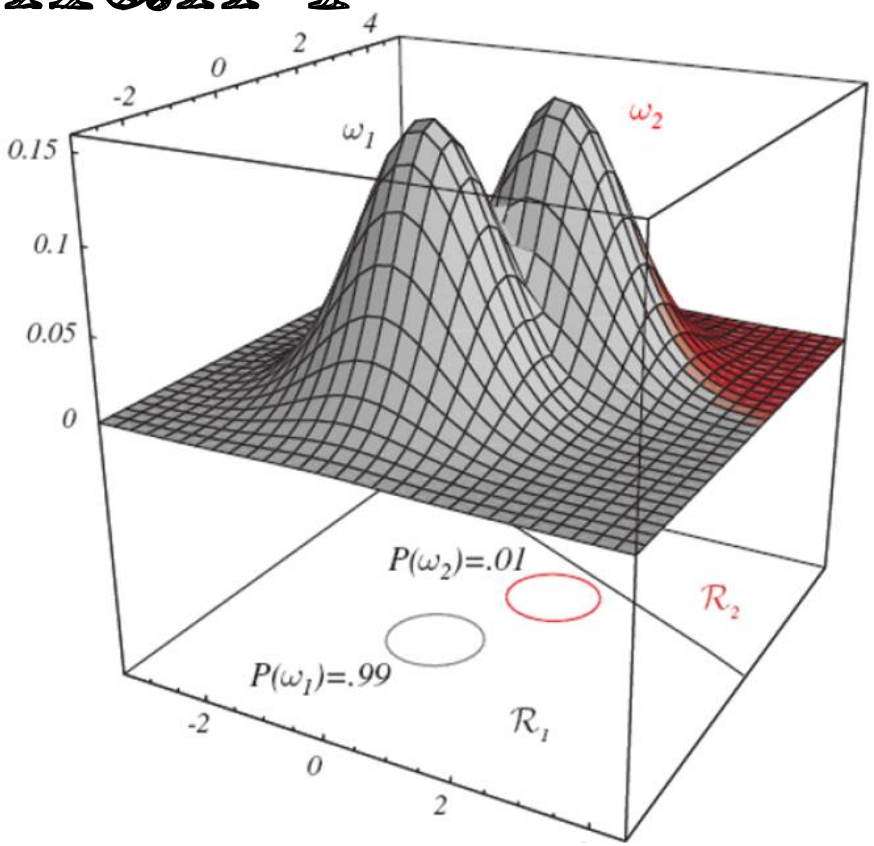


- univariate (links) bzw. bivariate (rechts) Normalverteilungen mit $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = I\sigma^2$
- Entscheidungsgrenzen sind linear und normal zur Verbindungsstrecke zwischen den beiden Klassenmitteln
- bei gleichen Priors verläuft Entscheidungsgrenze durch $(\mu_1 + \mu_2)/2$
- ansonsten wird sie von Prior der wahrscheinlicheren Klasse wegverschoben.

[Quelle: Duda et al., 2001]

Spezialfall 1



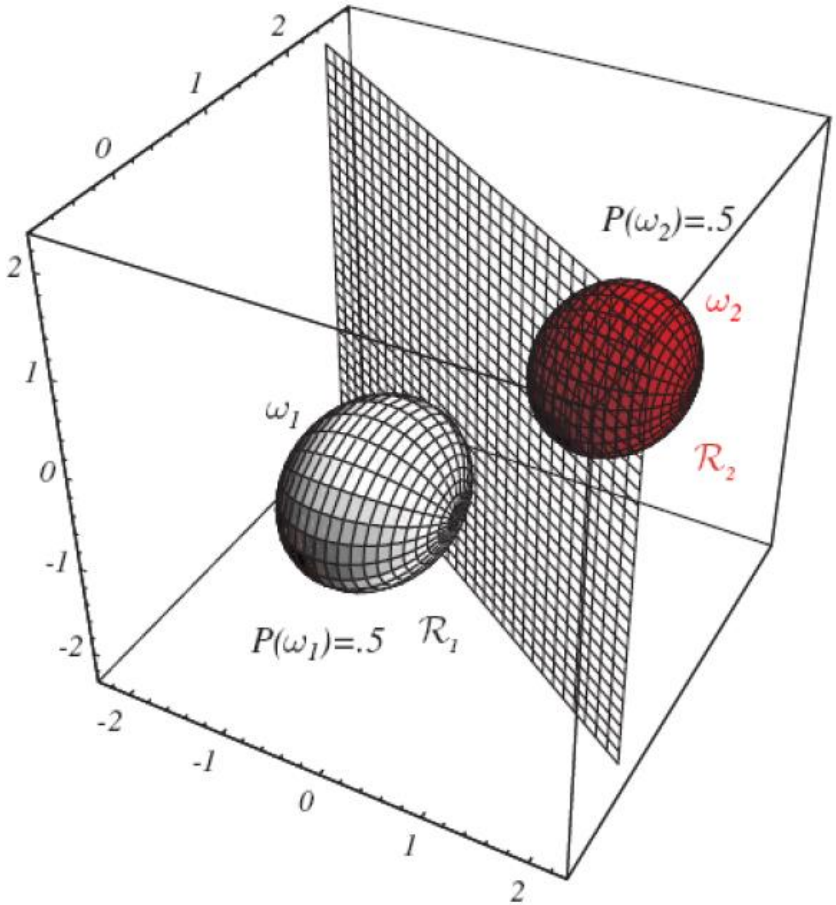


- univariate (links) bzw. bivariate (rechts) Normalverteilungen mit $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = I\sigma^2$
- Entscheidungsgrenzen sind linear und normal zur Verbindungsstrecke zwischen den beiden Klassenmitteln
- bei gleichen Priors verläuft Entscheidungsgrenze durch $(\mu_1 + \mu_2)/2$
- ansonsten wird sie von Prior der wahrscheinlicheren Klasse wegverschoben.

[Quelle: Duda et al., 2001]

Spezialfall 1

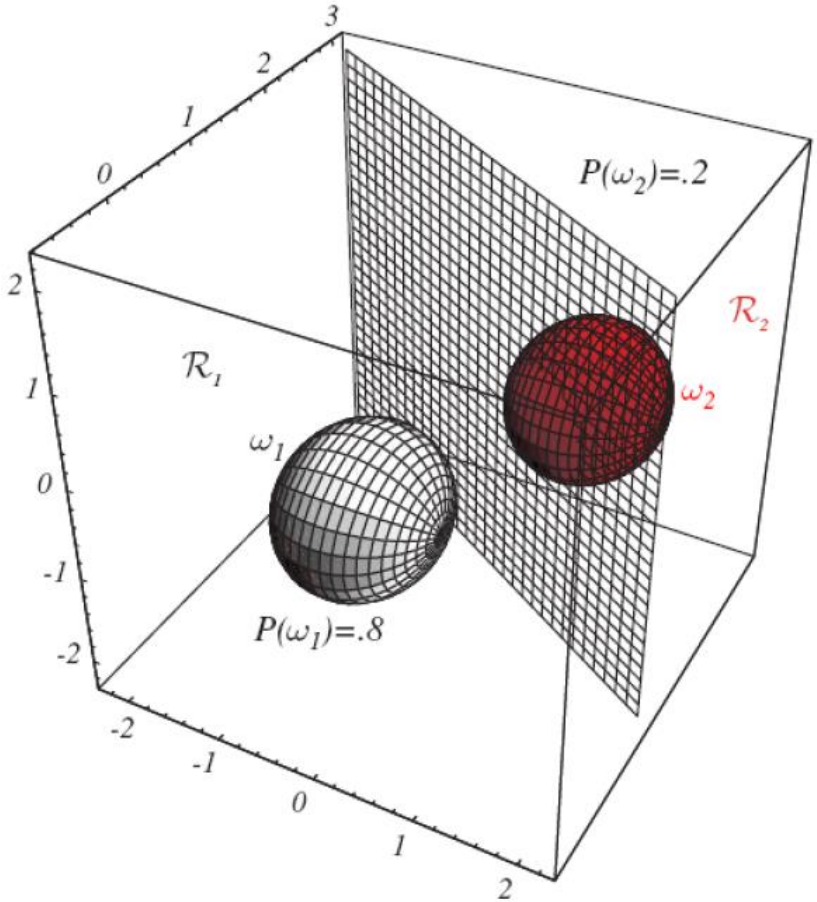
- Entscheidungsgrenze für zwei trivariate Normalverteilungen mit $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathrm{I}\sigma^2$.



[Quelle: Duda et al., 2001]

Spezialfall 1

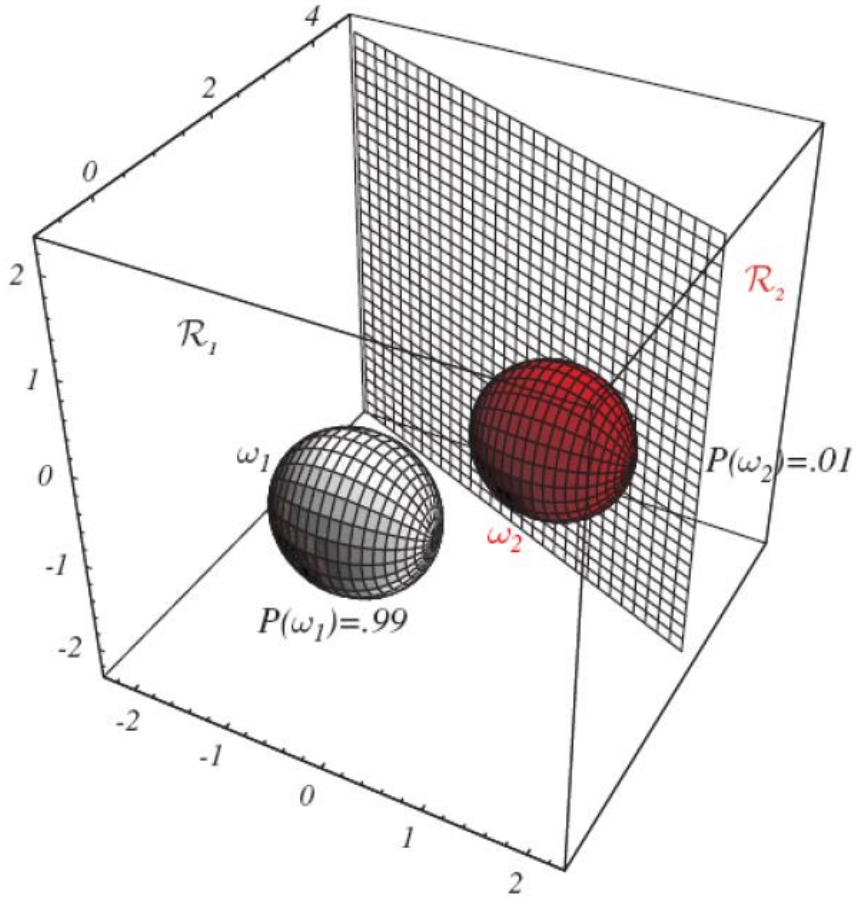
- Entscheidungsgrenze für zwei trivariate Normalverteilungen mit $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathrm{I}\sigma^2$.



[Quelle: Duda et al., 2001]

Spezialfall 1

- Entscheidungsgrenze für zwei trivariate Normalverteilungen mit $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathrm{I}\sigma^2$.



Spezialfall 2

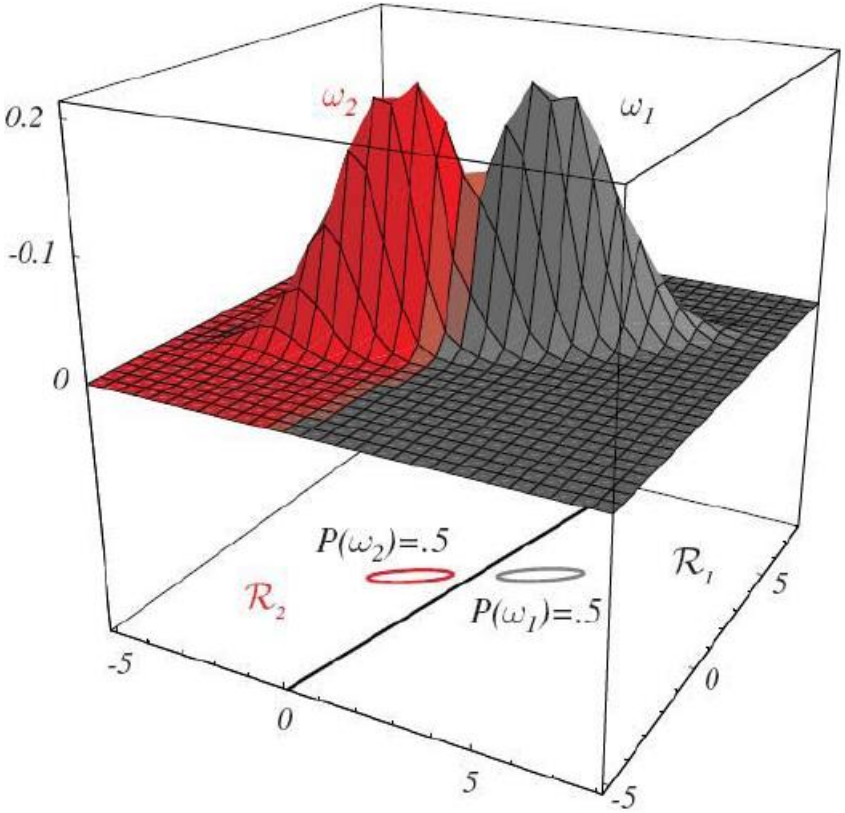
$$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$$

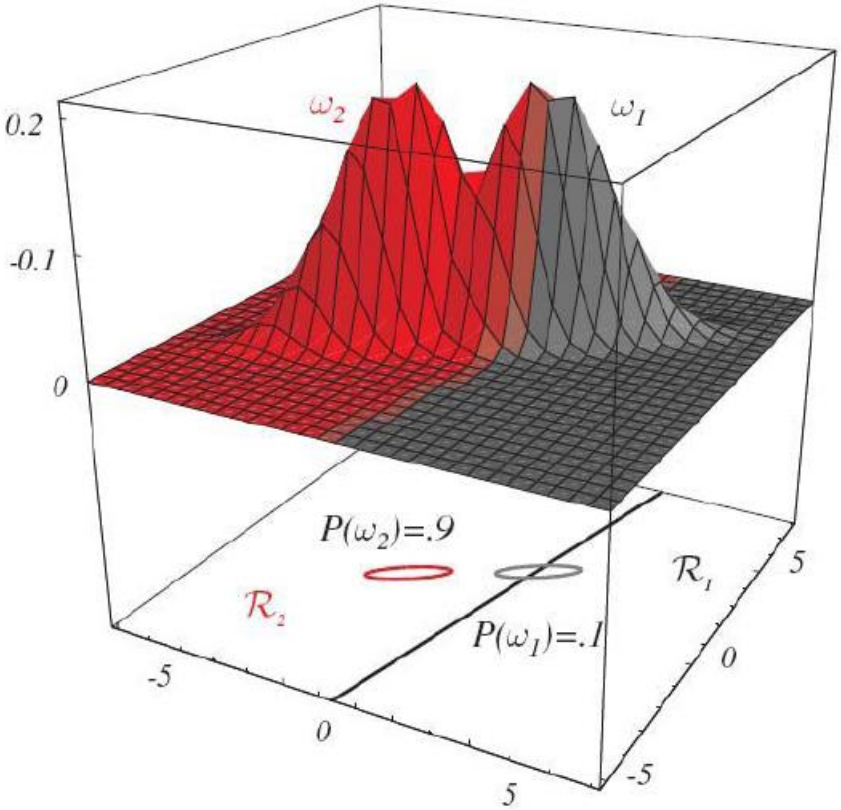
- alle Klassen haben dieselbe Kovarianzmatrix
- Form der Verteilungen ist durch Hyperellipsoide in $\mathcal{R}^p$ gegeben
- Entscheidungsgrenzen sind wieder linear
- jedoch nicht normal zur Verbindungsstrecke zwischen den beiden Klassenmitteln
- Details: siehe Kapitel 2, Duda et al., 2001

© Nicole M. Artner

[Quelle: Duda et al., 2001]

Spezialfall 2





© Nicole M. Artner

Allgemeiner Fall

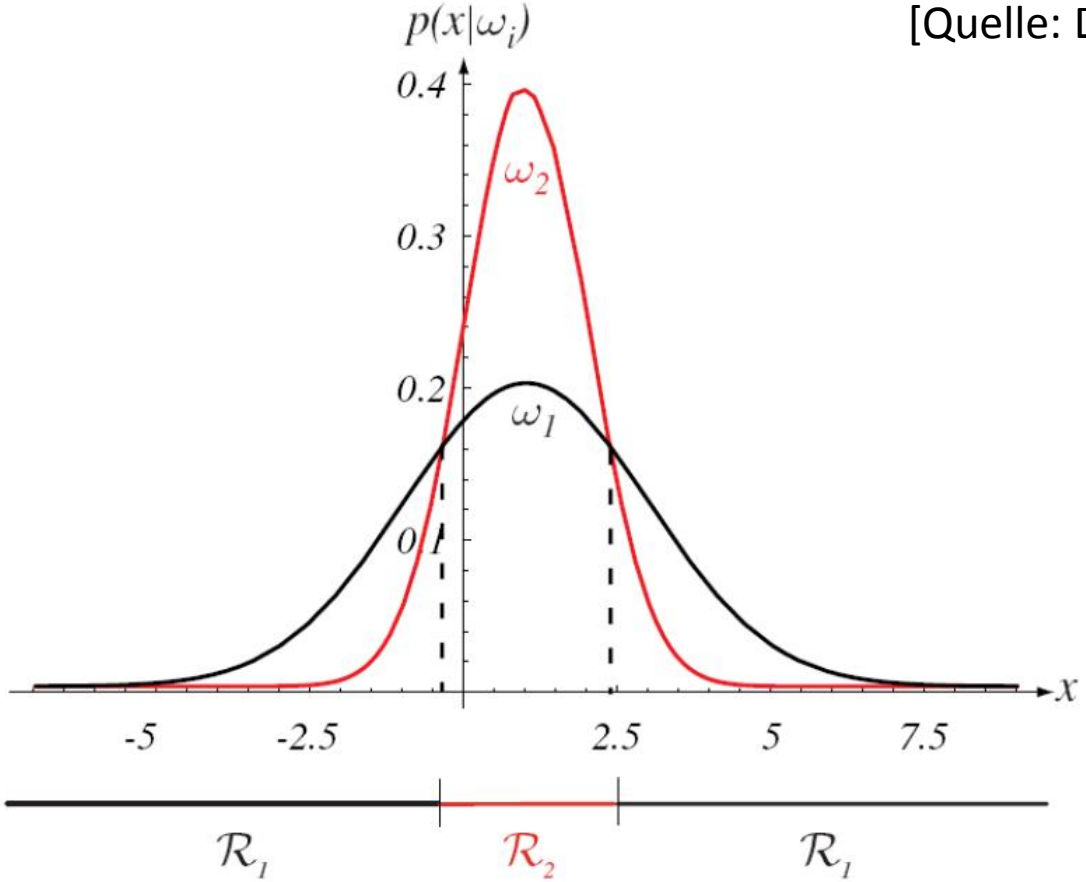
- bliebige Kovarianzmatrix $\boldsymbol{\Sigma_i}$
- Entscheidungsgrenzen sind durch so genannte Hyperquadrics gegeben
- Hyperquadrics sind unter anderem: Hyperebenen, Hyperkugeln, Hyperellipsoide, Hyperparaboloide, etc.
- korrespondierenden Entscheidungsregionen müssen nicht einfach zusammenhängend sein

© Nicole M. Artner

Nicht zusammenhängende Entscheidungsregionen

… entstehen im 1D-Fall z.B. durch Verteilungen mit gleichem Mittelwert aber unterschiedlichen Varianzen

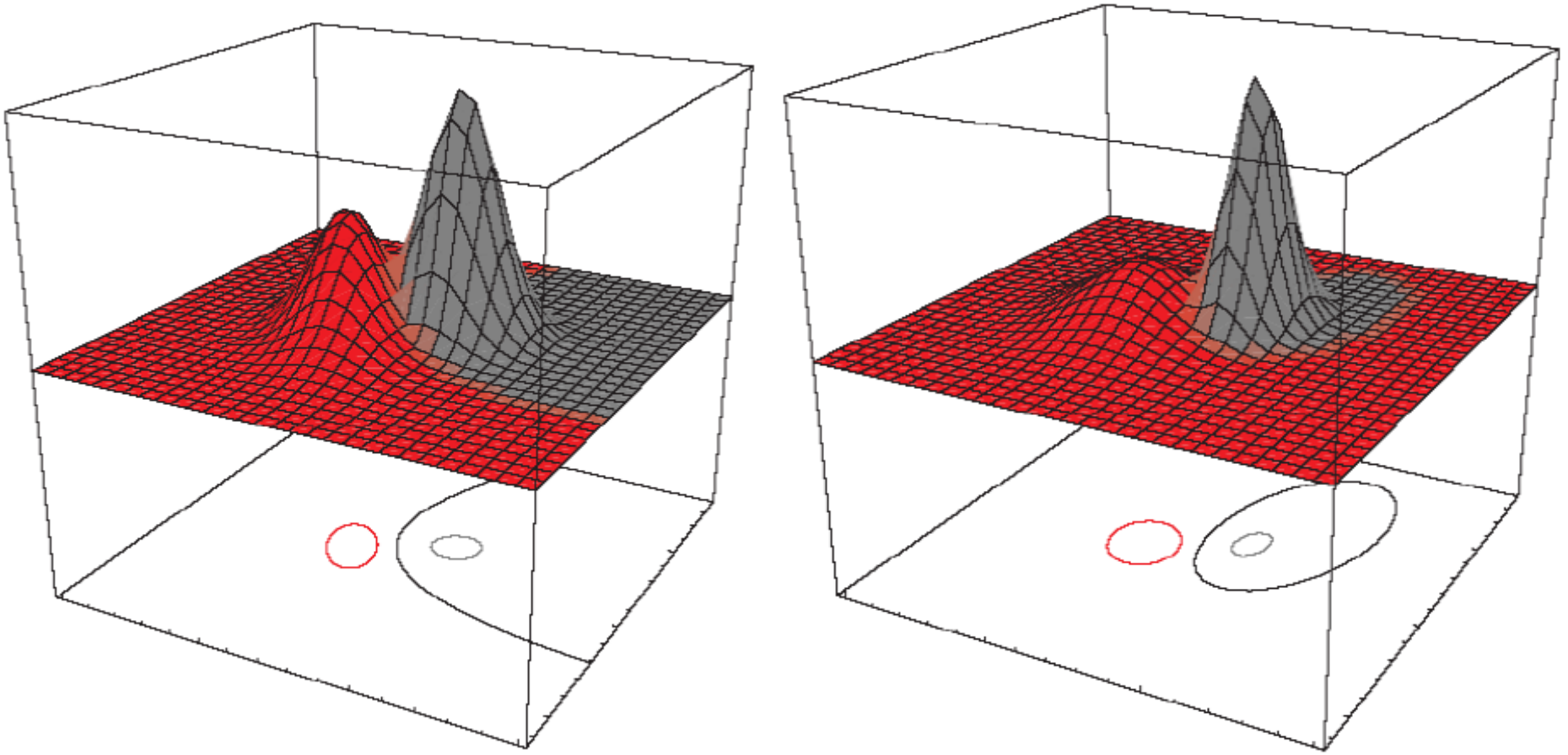
[Quelle: Duda et al., 2001]



[Quelle: Duda et al., 2001]

Beispiele

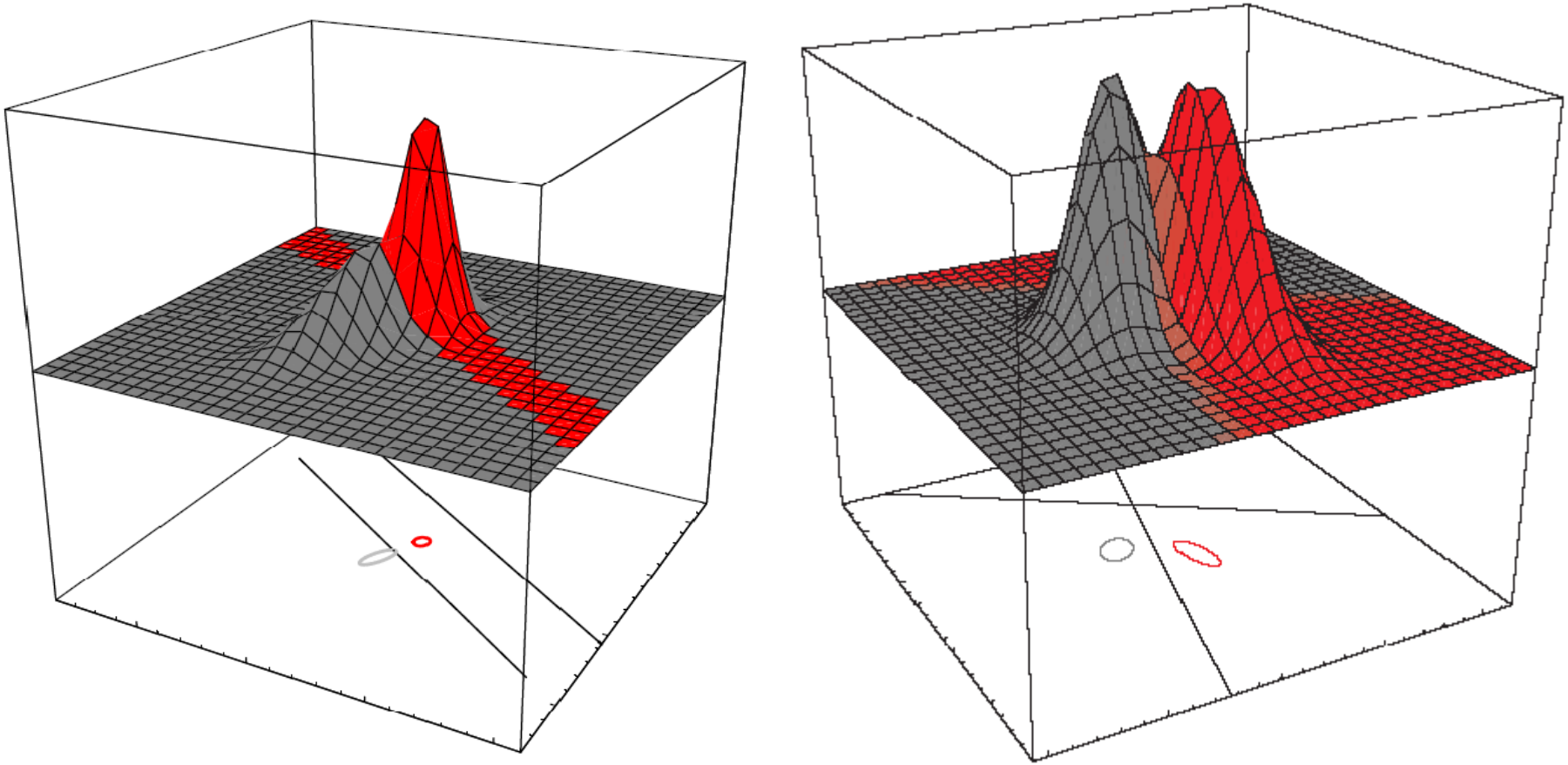
… für 2D-Normalverteilungen mit beliebigen Kovarianzmatrizen



[Quelle: Duda et al., 2001]

Beispiele

… für 2D-Normalverteilungen mit beliebigen Kovarianzmatrizen

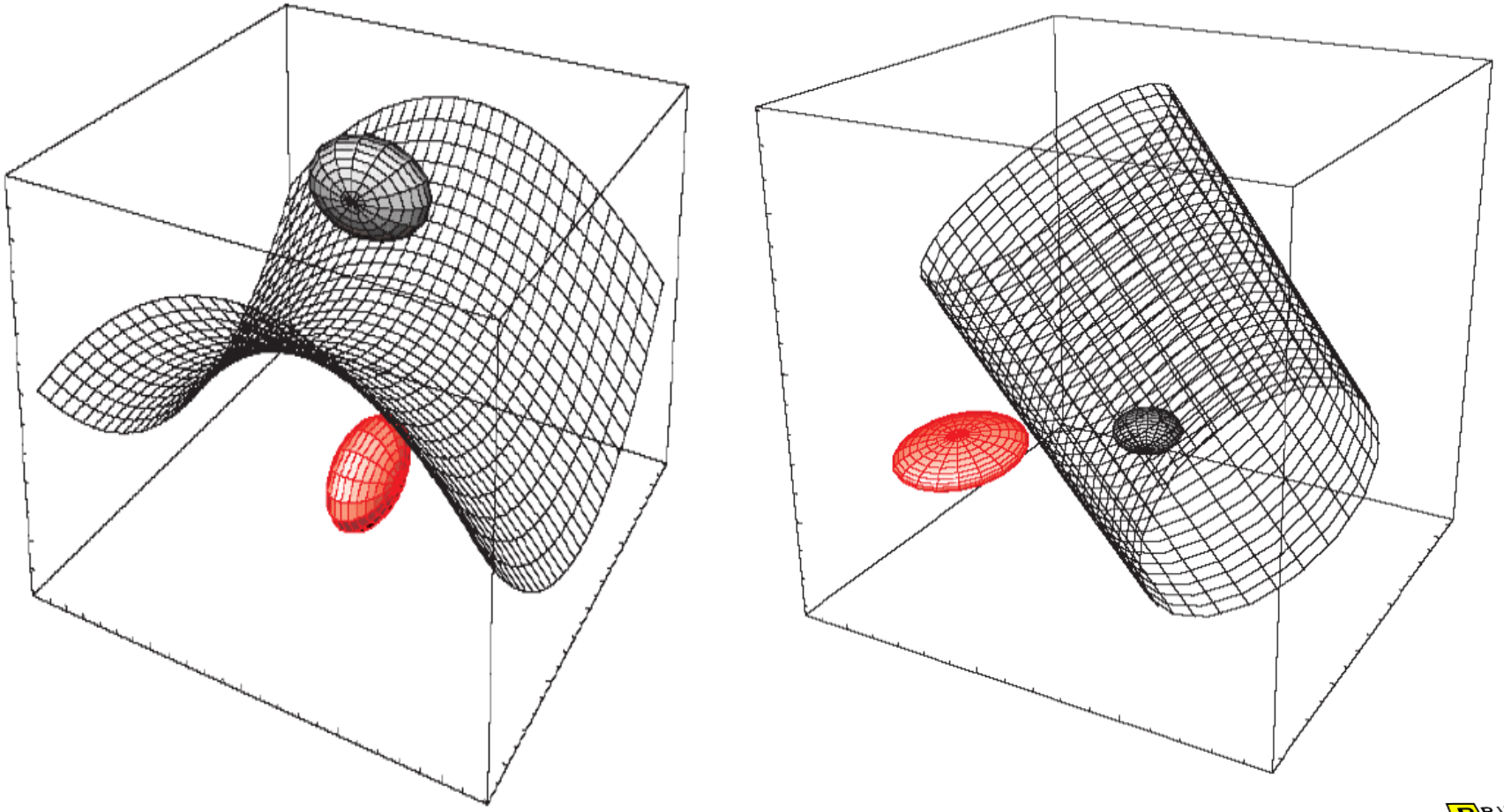


© Nicole M. Artner

[Quelle: Duda et al., 2001]

Beispiele

… für 3D-Normalverteilungen mit beliebigen Kovarianzmatrizen

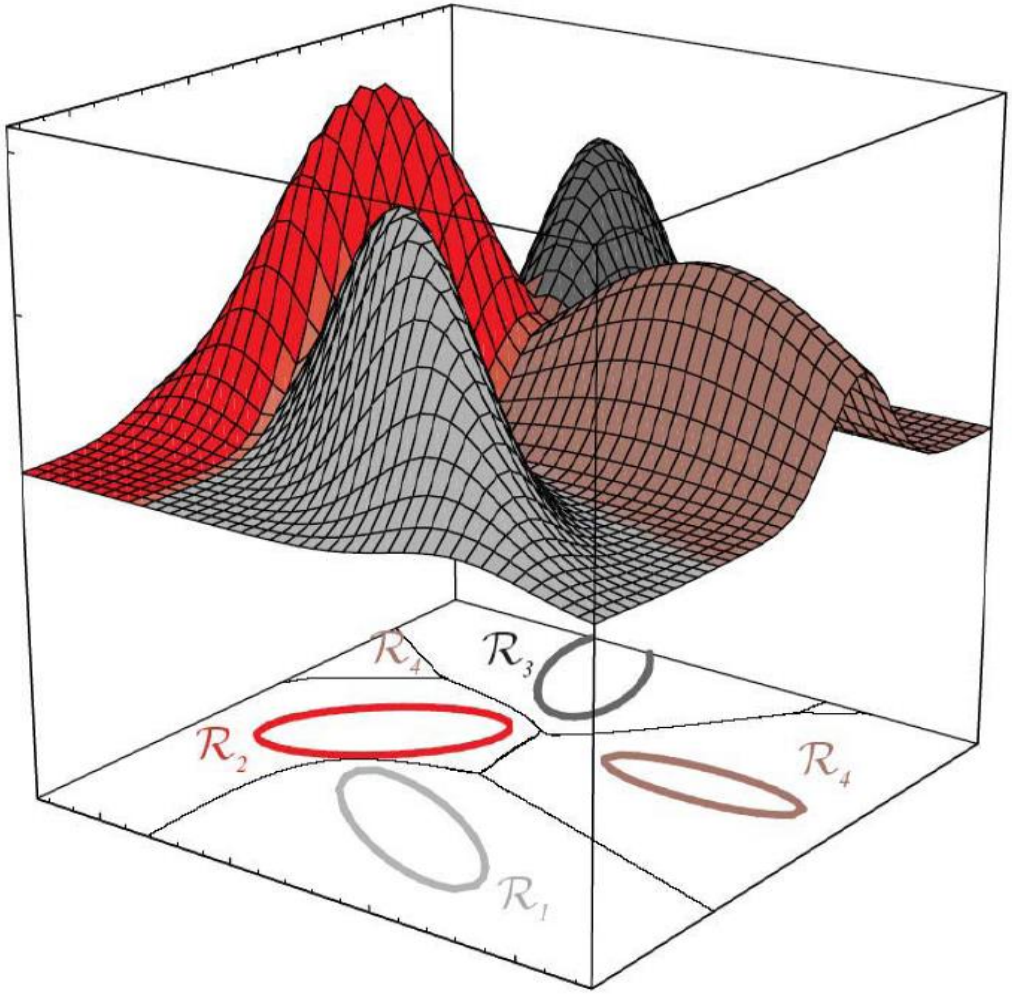


© Nicole M. Artner

[Quelle: Duda et al., 2001]

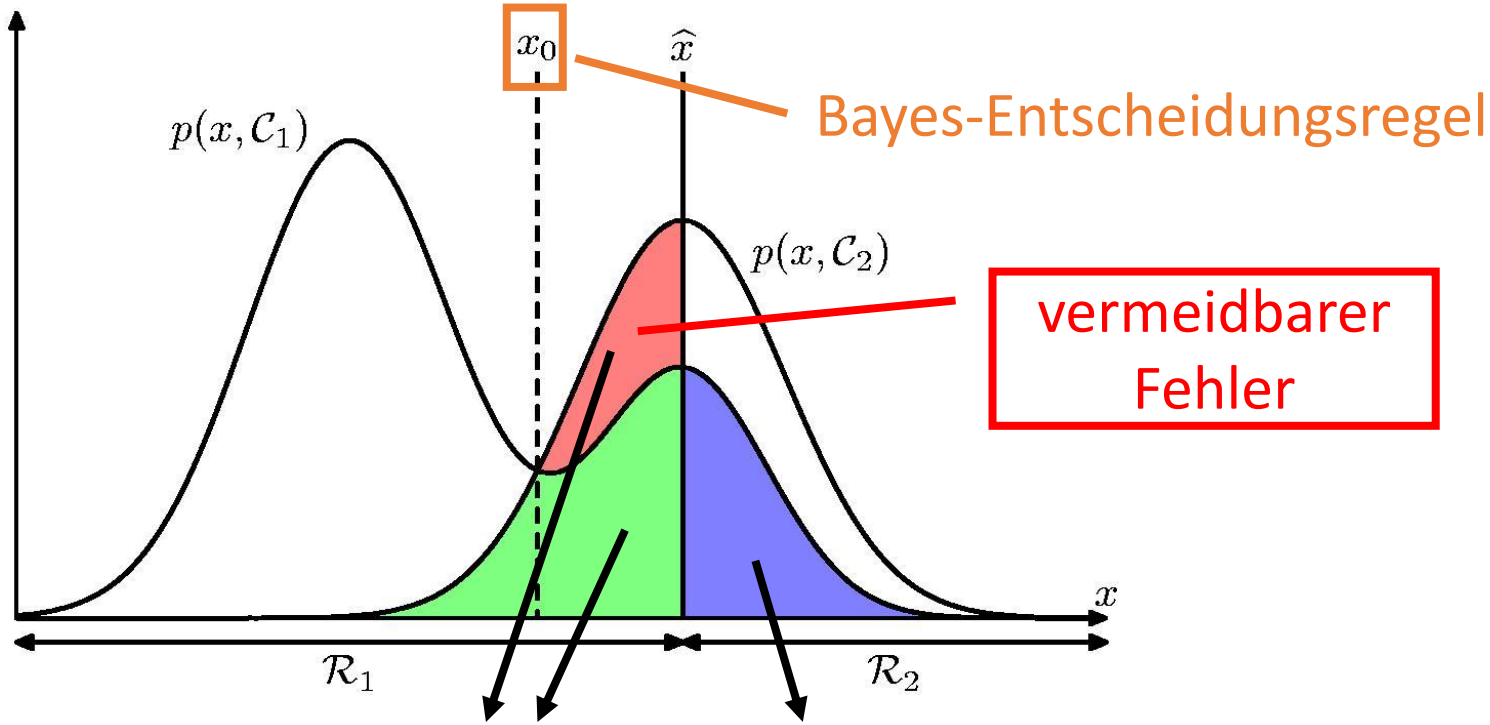
Beispiele

- vier 2D-Normal-verteilungen
- komplexe Entscheidungsregionen und -grenzen



© Nicole M. Artner

Fehlerrate



$$P(error) = P(x \in \mathcal{R}_2, w_1) + P(x \in \mathcal{R}_1, w_2)$$

$$= P(x \in \mathcal{R}_2 | w_1) P(w_1) + P(x \in \mathcal{R}_1 | w_2) P(w_2)$$

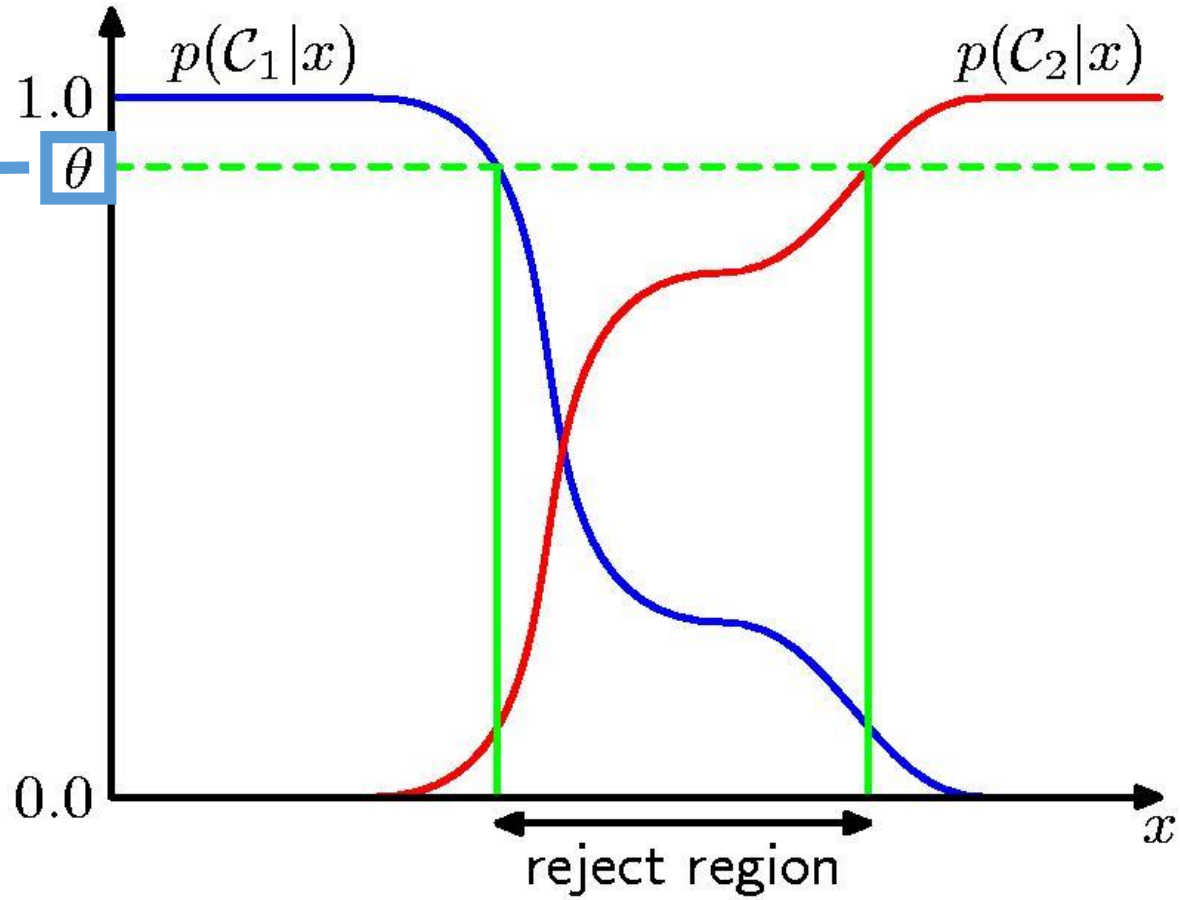
$$= \int_{\mathcal{R}_2} p(x|w_1)\, P(w_1)\, dx + \int_{\mathcal{R}_1} p(x|w_2)\, P(w_2)\, dx$$

© Nicole M. Artner

Reject Option

… Ablehnungsoption (frei ins Deutsche übersetzt)

Schwellwert für Ablehnung



Reject region:
alle x deren größter Posterior kleiner θ ist, werden abgelehnt

© Nicole M. Artner

THE END
für
Bayes-Theorem

© Nicole M. Artner