Überblick

- **Grundlagen**
  - Einführung in die automatische Mustererkennung
  - Grundlagen der Wahrscheinlichkeitsrechnung
- **Klassifikation bei bekannter Wahrscheinlichkeitsverteilung**
  - Entscheidungstheorie
  - Bayes-Klassifikator
  - Entscheidungsfunktionen bei gaußverteilten Daten
- **Überwachtes Lernen bei unbekannter Verteilung der Daten**
  - Nichtparametrische Klassifikation
  - Probleme bei hochdimensionalen Daten
  - Lineare Klassifikation, Perzeptron
  - Nicht linear trennbare Systeme
  - Nichtlineare Klassifikatoren
  - Vergleich von Klassifikatoren
- **Unüberwachtes Lernen**
  - Hauptkomponentenanalyse
  - K-Means-Clustering
  - Agglomeratives Clustern

Überblick

- **Wiederholung: Diskriminantenfunktionen**
- **Kontinuierliche Zufallsvariablen**
- **Gaußverteilung**
- **Diskriminantenfunktionen bei gaußverteilten Daten**

Diskriminantenfunktionen

Diskriminantenfunktionen sind die häufigste Art, einen Klassifikator zu beschreiben.

Vorgehensweise:

- Zu jeder Klasse ω_i wird eine Diskriminantenfunktion $g_i(\boldsymbol{x})$ definiert.
- Der Klassifikator klassifiziert den Inputvektor $\boldsymbol{x}$ als zur Klasse ω_i zugehörig, wenn

$$g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}) \quad \text{für alle } j \neq i$$

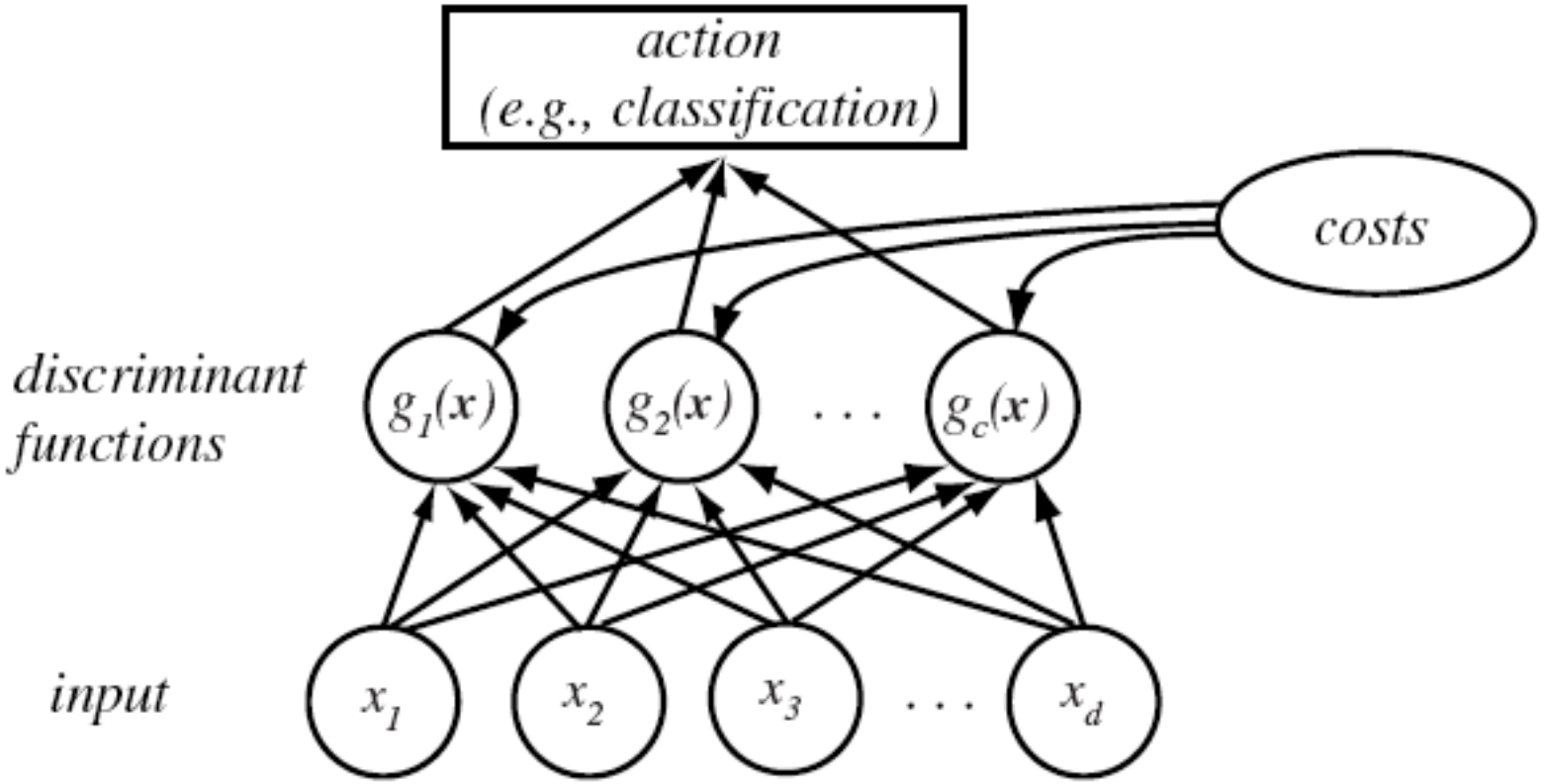
Beispiel Bayes-Klassifikator: „Wähle immer die Aktion α_i, die das bedingte Risiko für die gegebene Beobachtung **x** minimiert.“, d.h. entscheide für Klasse ω_i, wenn gilt

$$R(\alpha_i \mid \mathbf{x}) < R(\alpha_j \mid \mathbf{x}) \; bzw. \; p(\omega_i \mid \mathbf{x}) > p(\omega_j \mid \mathbf{x}) \quad für\, alle \quad j \neq i$$

d.h. die Diskriminantenfunktion ist hier

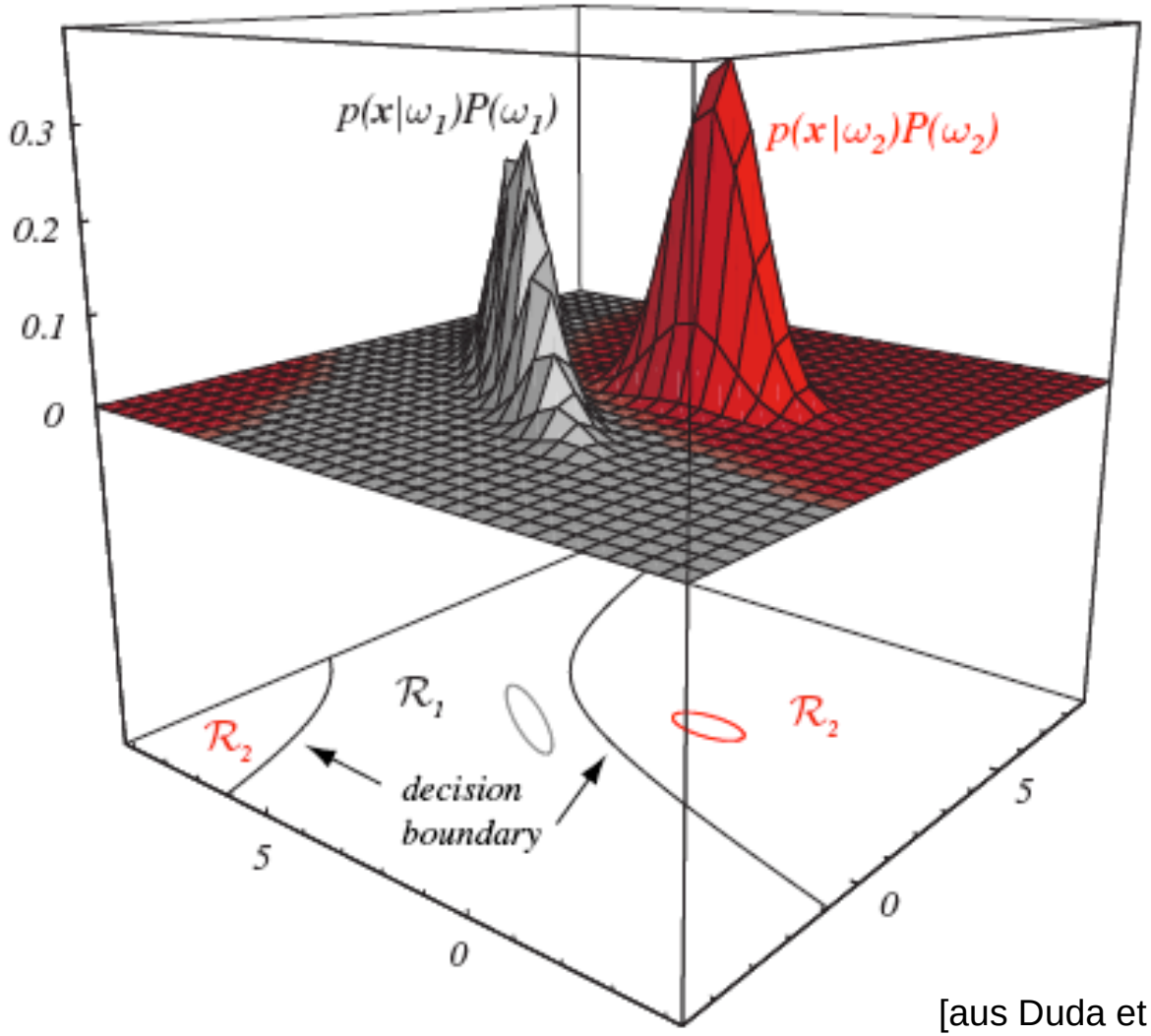
$$g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x}) \quad bzw. \quad g_i(\mathbf{x}) = p(\omega_i \mid \mathbf{x})$$

Allgemeiner Aufbau eines statistischen Klassifikators



[aus Duda et al., 2001]

Entscheidungsgrenzen und Entscheidungsregionen



Die Entscheidungsregel unterteilt den Inputraum in Entscheidungsregionen.

Die Entscheidungsregionen sind durch Entscheidungsgrenzen voneinander getrennt.

[aus Duda et al., 2001]

Überblick

- **Wiederholung: Diskriminantenfunktionen**
- **Kontinuierliche Zufallsvariablen**
- **Gaußverteilung**
- **Diskriminantenfunktionen bei gaußverteilten Daten**

Kontinuierliche Zufallsvariablen (1)

• Bisher: Zufallsvariablen, bei denen ein diskreter Wert a_i aus einem Alphabet mit einer Wahrscheinlichkeit p_i angenommen wird.

• Jetzt: Erweiterung auf Zufallsvariablen mit kontinuierlichen Werten

• Die Wahrscheinlichkeit für einen bestimmten Wert a_i ist bei kontinuierlichen Zufallsvariablen 0!

• Es macht mehr Sinn, davon zu sprechen, daß die ZV einen bestimmten Wert im Intervall $[a,b]$ annimmt.

• Bisher: $$p(x \in [a,b]) = \sum_{x_i \in [a,b]} p(x)$$

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

(kontinuierliche Wahrscheinlichkeitssdichte)

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928

Kontinuierliche Zufallsvariablen (2)

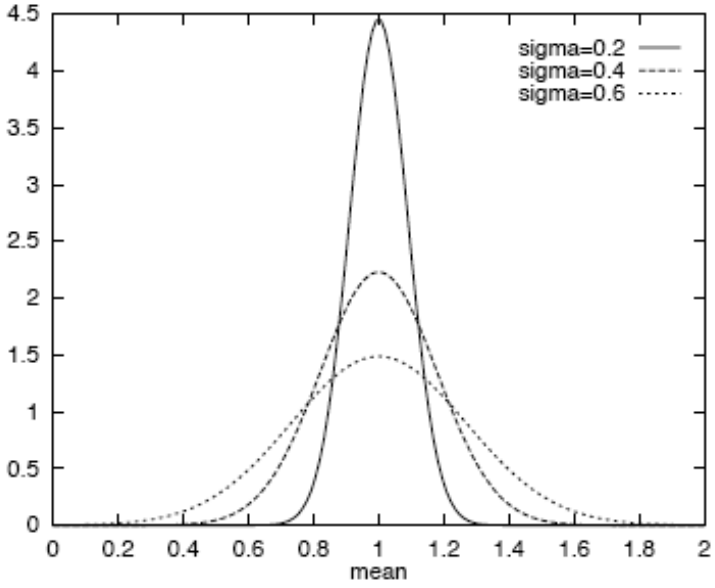
Die bisher benutzten Formeln für diskrete ZV können direkt übernommen werden, indem die Summierung durch ein Integral ersetzt wird:

- Positivität und Normierung: $p(x) \geq 0 \quad und \quad \int_{-\infty}^{\infty} p(x)\, dx = 1$
- Erwartungswert: $E[f(x)] = \int_{-\infty}^{\infty} f(x)\, p(x)\, dx$
- Mittelwert: $\mu = E[x] = \int_{-\infty}^{\infty} x p(x)\, dx$
- Varianz: $\sigma^2 = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2\, p(x)\, dx$
- Bayes-Formel: $p(x \mid y) = \dfrac{p(y \mid x)\, p(x)}{\int_{-\infty}^{\infty} p(y \mid x)\, p(x)\, dx}$

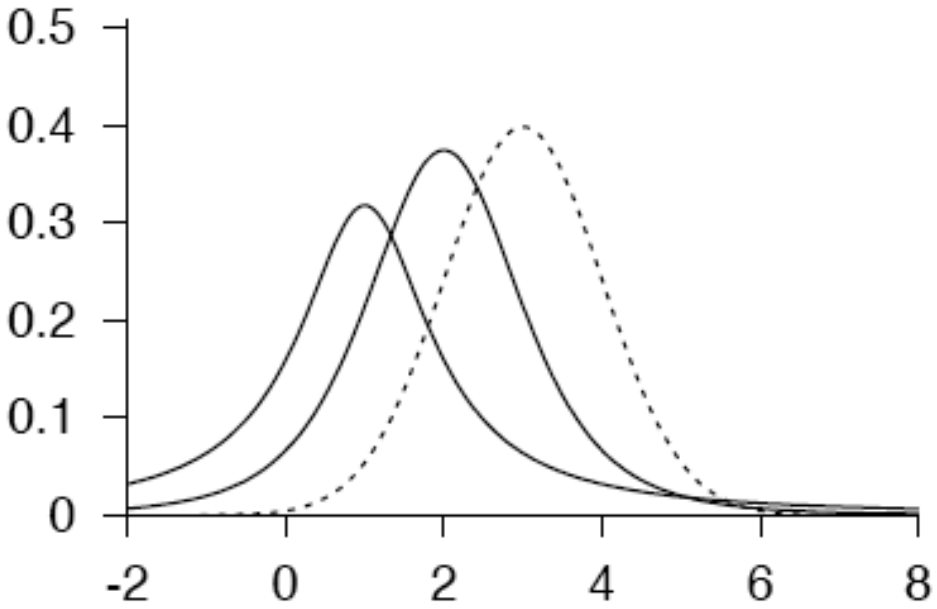
Häufig benutzte Verteilungen (1)

Da im Normalfall die genaue Verteilung der Daten nicht bekannt sind, benutzt man bestimmte Verteilungsfunktionen als Modell der Daten, um damit Vorwissen oder Vorannahmen zu beschreiben.

Beispiel: Verteilungen für ZV, deren Werte unbeschränkt sind:



Gaußverteilung

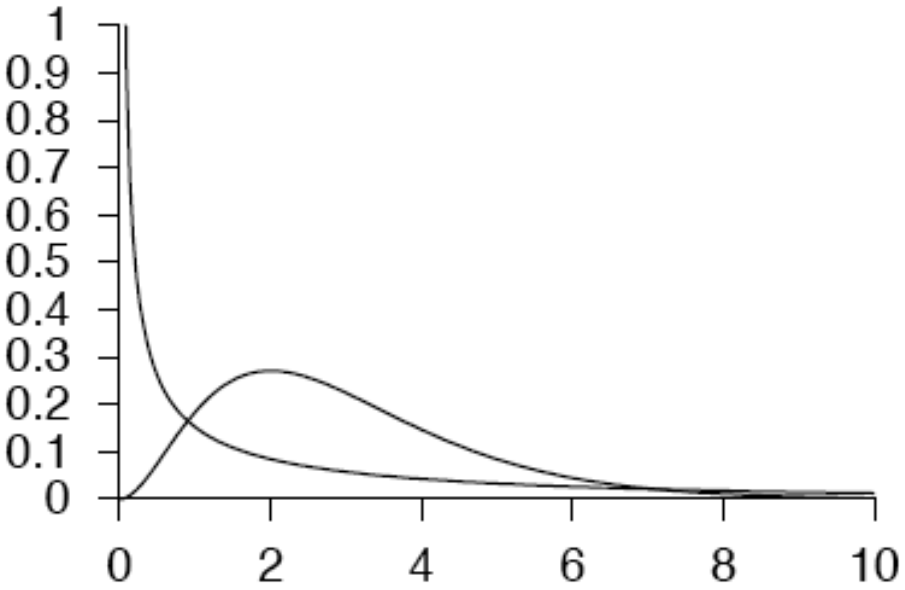


Student-t-Verteilung
(fällt langsamer ab)

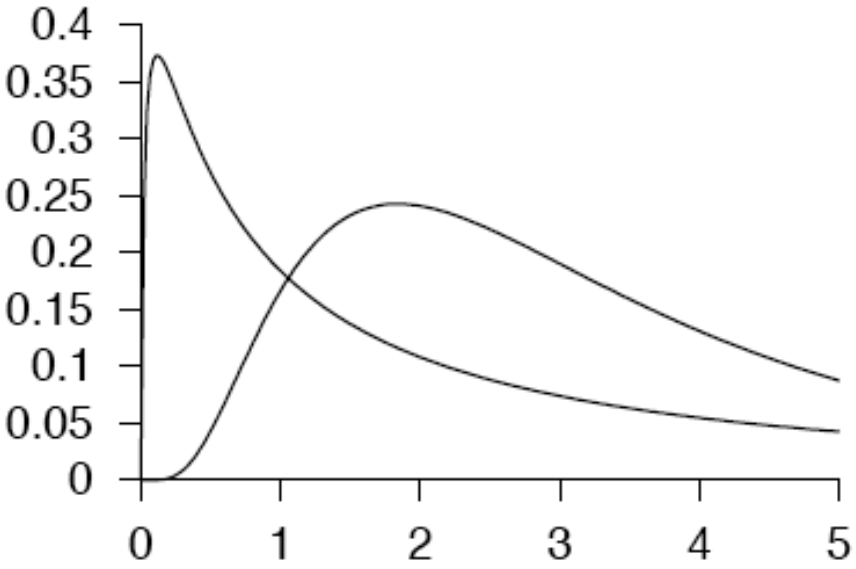
[aus McKay, 2003]

Häufig benutzte Verteilungen (2)

Verteilungen für ZV mit positiven Werten:



Gammaverteilung (x >= 0)



Log-Normalverteilung (x > 0)

Auswahl nach:
- Art der Daten (positiv, diskret, ...)
- unimodal, bimodal
- einfache Berechenbarkeit

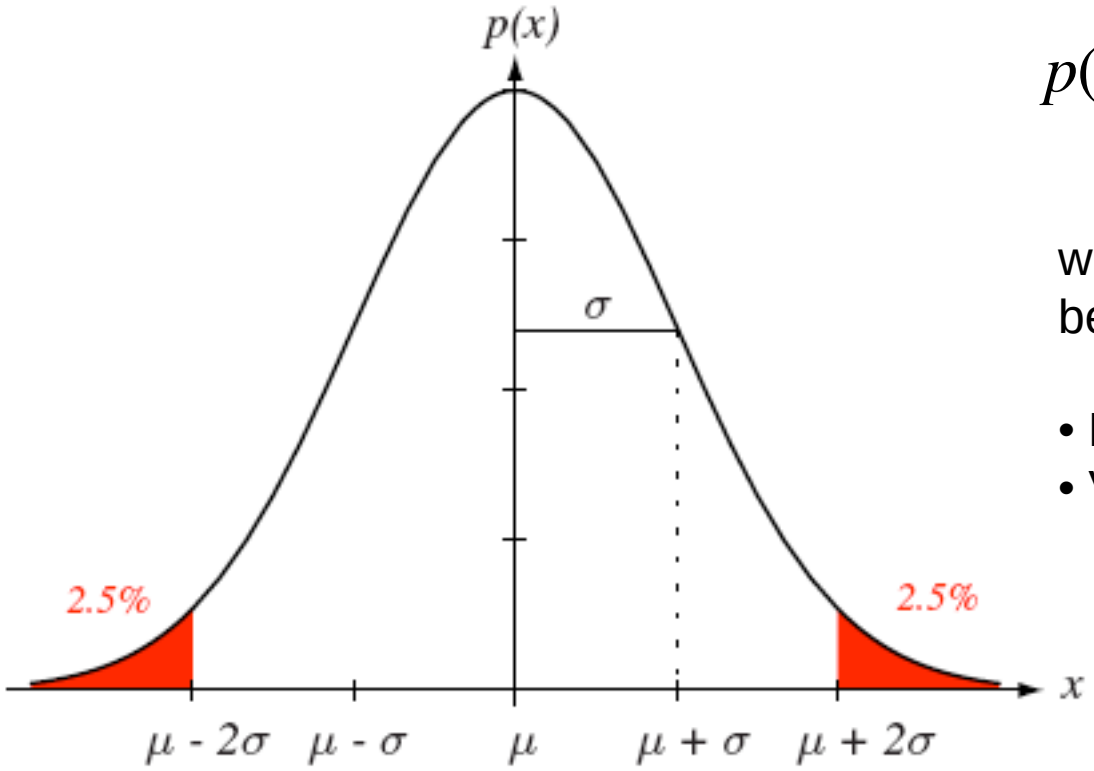
[aus McKay, 2003]

Überblick

- **Wiederholung: Diskriminantenfunktionen**
- **Kontinuierliche Zufallsvariablen**
- **Gaußverteilung**
- **Diskriminantenfunktionen bei gaußverteilten Daten**

Univariate Gaußverteilung (1)

Zentraler Grenzwertsatz der Wahrscheinlichkeitsrechnung: die Summe einer großen Zahl von (beinahe) beliebig verteilten Zufallsvariablen nähert sich der Gaußverteilung.



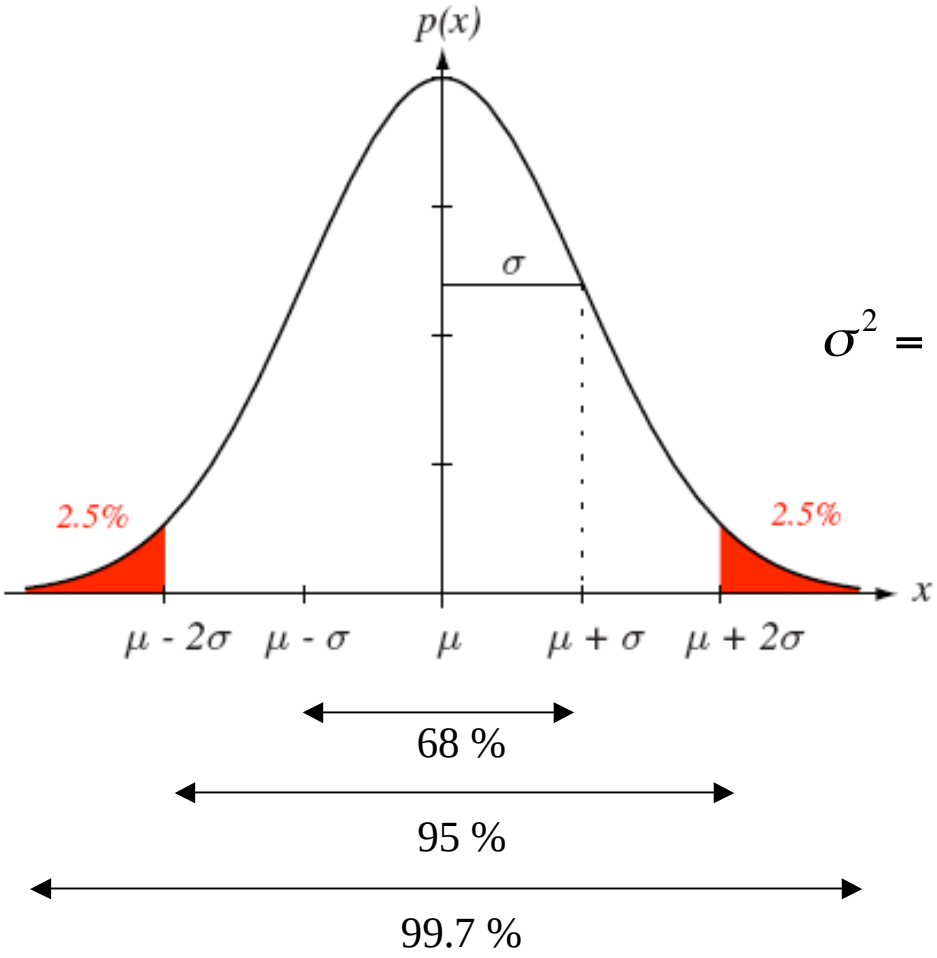
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

wird durch 2 Parameter beschrieben:

- Mittelwert μ
- Varianz σ^2

Schreibweise: $N(\mu, \sigma^2)$

Gaußverteilung (2)



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

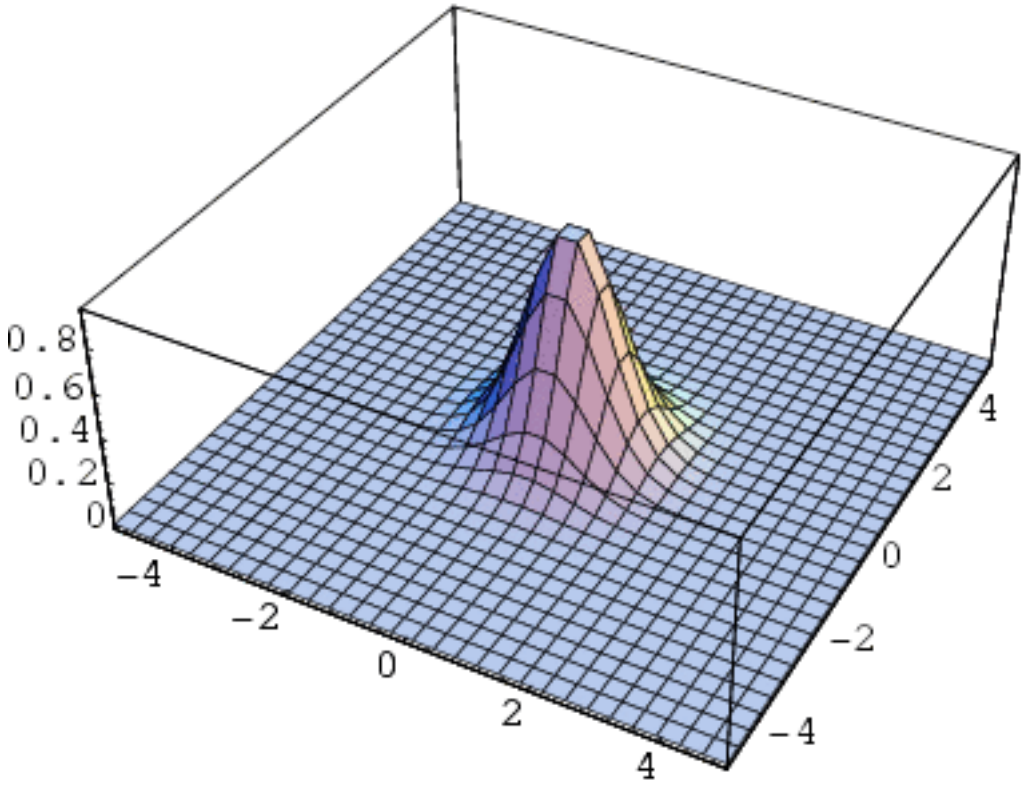
$$\mu = E[x] = \int_{-\infty}^{\infty} xp(x)\,dx$$

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

- Aufgrund des zentralen Grenzwertsatzes ist die Gaußverteilung oft ein gutes Modell.
- Viele Integrale lassen sich geschlossen lösen => häufig in probabilistischen Modellen.

Multivariate Gaußverteilung

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N \det \Sigma}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu_x)\right]$$



mit

μ: Mittelwertvektor
mit $\mu_i = E[x_i]$

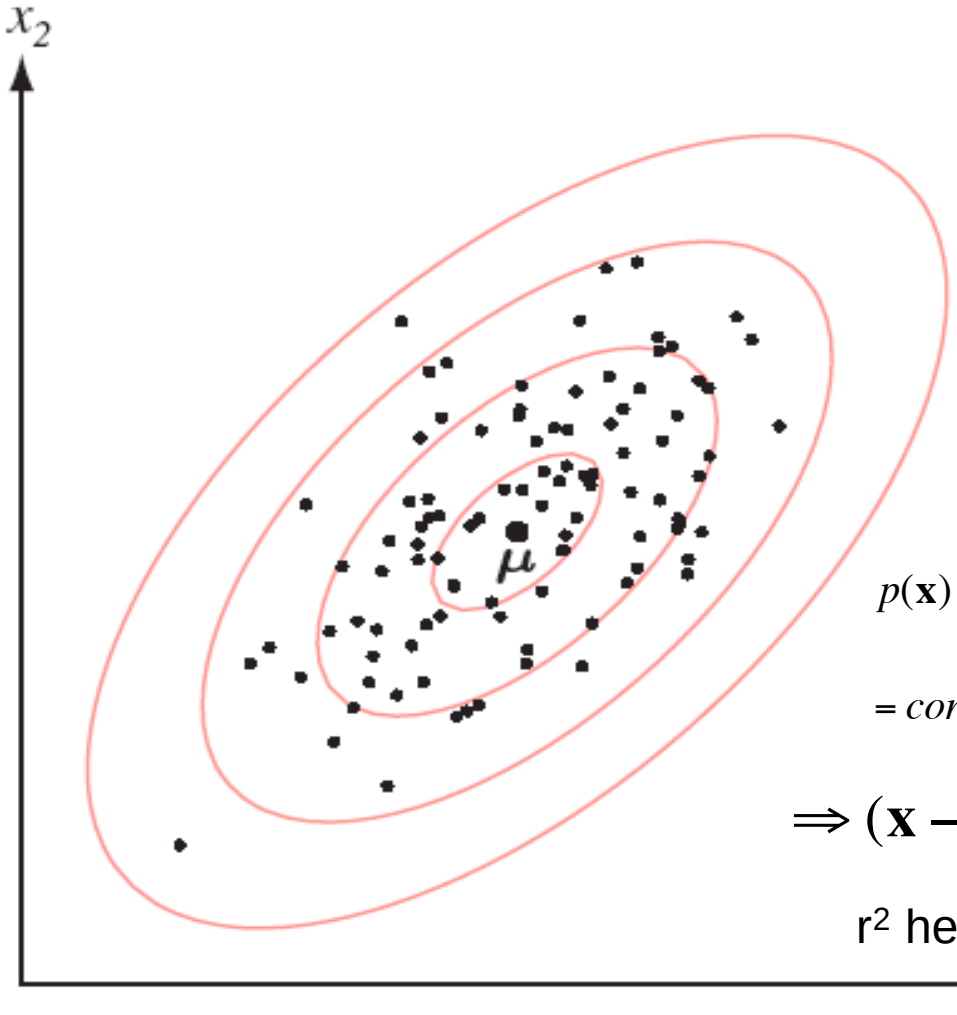
Σ: Kovarianzmatrix
mit
$\sigma_{ij} = E[(x_i - \mu_i)\,(x_j - \mu_j)]$

Die Diagonalelemente von Σ sind die Varianzen der x_i .

Die nichtdiagonalen Elemente von Σ sind die Kovarianzen von x_i und x_j.

Die d-dimensionale Gaußverteilung ist durch d + d(d + 1)/2 Größen bestimmt.

Beispiel: von einer bivariaten Gaußverteilung erzeugt Daten



[aus Duda et al., 2001]

Linien gleicher Wahrscheinlichkeitsdichte sind Ellipsen.

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N \det \Sigma}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu_x)\right]$$

$$= const.$$

$$\Rightarrow (\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu_x) = const. = r^2$$

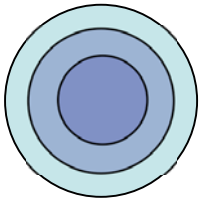
r^2 heißt Mahalanobis-Distanz

d.h. die Ellipsen haben konstante Mahalanobis-Distanz zum Mittelwert μ.

2D-Beispiele für Kovarianzmatrizen

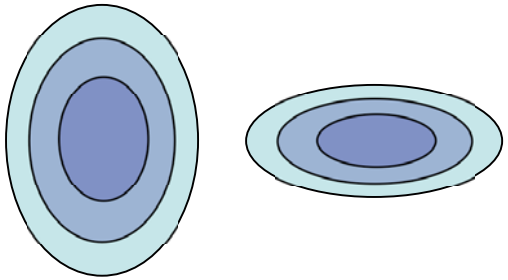
$$\begin{pmatrix} \sigma & 0 \\ 0 & \sigma \end{pmatrix}$$

Diagonalelemente gleich, nichtdiagonale Elemente 0:
- Linien konstanter Dichte sind Kreise
- Verteilung isotrop
- beide Variablen sind statistisch unabhängig (es gilt $p(x_1,x_2) = p(x_1)\, p(x_2)$).



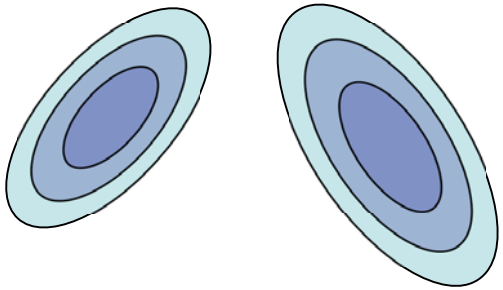
$$\begin{pmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{pmatrix}$$

Diagonalelemente ungleich, nichtdiagonale Elemente 0:
- Linien konstanter Dichte sind Ellipsen, die an den Koordinatenachsen ausgerichtet sind
- beide Variablen sind immer noch statistisch unabhängig.

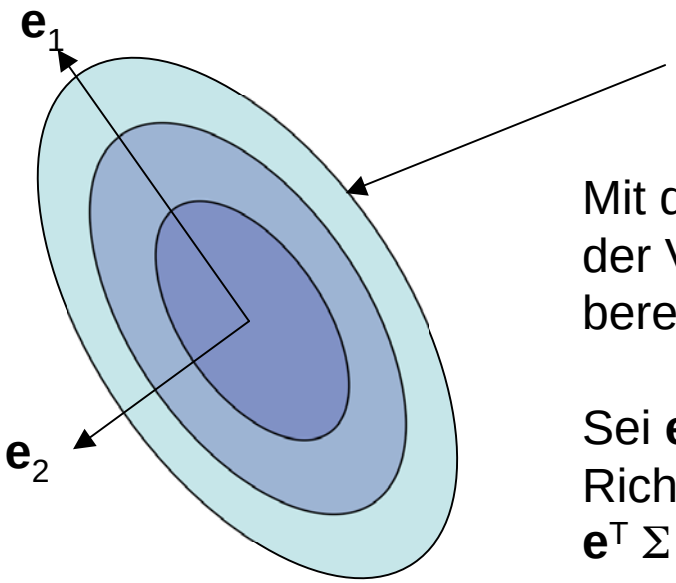


$$\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

Alle Einträge von Σ sind ungleich 0, $\sigma_{12} = \sigma_{21}$:
- Linien konstanter Dichte sind beliebig ausgerichtete Ellipsen
- die Variablen sind statistisch abhängig.



Eigenwerte der Kovarianzmatrix



$$(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu_x) = const.$$

Mit der Kovarianzmatrix läßt sich die Varianz der Verteilung in einer beliebigen Richtung berechnen:

Sei **e** ein Einheitsvektor in der gewünschten Richtung, dann ist die Varianz in dieser Richtung $\mathbf{e}^T \Sigma \mathbf{e}$.

Die Ausrichtung der Ellipsen wird durch die Eigenvektoren beschrieben. Für einen Eigenvektor gilt:

$$\Sigma \mathbf{e} = \lambda \mathbf{e}$$

bzw. $(\Sigma - \lambda I)\,\mathbf{e} = 0$

mit dem Eigenwert λ.

Die Eigenwerte λ_1 und λ_2 geben die Varianz der Gaußverteilung entlang der Richtungen der Eigenvektoren $\mathbf{e}_1$ und $\mathbf{e}_2$ an.

Überblick

- **Wiederholung: Diskriminantenfunktionen**
- **Kontinuierliche Zufallsvariablen**
- **Gaußverteilung**
- **Diskriminantenfunktionen bei gaußverteilten Daten**

Diskriminantenfunktionen für gaußverteilte Daten

Der Bayesklassifikator hat im Falle der Minimierung der Fehlerwahrscheinlichkeit die Diskriminantenfunktion:

$$g_j(\mathbf{x}) = p(\omega_j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_j)\, p(\omega_j)}{p(\mathbf{x})}$$

bzw. $$g_j(\mathbf{x}) = p(\mathbf{x} \mid \omega_j)\, p(\omega_j)$$

Besonders einfach für gaußverteilte Daten ist die log-transformierte Diskriminantenfunktion:

$$g_j(\mathbf{x}) = \ln p(\mathbf{x} \mid \omega_j) + \ln p(\omega_j)$$

Für $$p(\mathbf{x} \mid \omega_j) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma_j}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_j)^T \Sigma_j^{-1} (\mathbf{x}-\mu_j)\right]$$

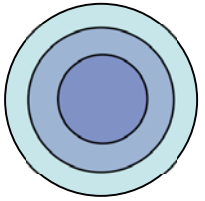
ergibt sich

$$g_j(\mathbf{x}) = -\tfrac{1}{2}(\mathbf{x}-\mu_j)^T \Sigma_j^{-1} (\mathbf{x}-\mu_j) - \tfrac{d}{2}\ln 2\pi - \tfrac{1}{2}\ln\det\Sigma_j + \ln p(\omega_j)$$

Spezialfall: $\Sigma_j = \sigma^2 I$

$$g_j(\mathbf{x}) = -\tfrac{1}{2}(\mathbf{x}-\mu_j)^T \Sigma_j^{-1}(\mathbf{x}-\mu_j) - \tfrac{d}{2}\ln 2\pi - \tfrac{1}{2}\ln\det\Sigma_j + \ln p(\omega_j)$$

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \quad \Rightarrow \boxed{g_j(\mathbf{x}) = -\frac{1}{2\sigma^2}\left\|\mathbf{x}-\mu_j\right\|^2 + \ln p(\omega_j)}$$



Die Diskriminantenfunktion ist also einfach der quadratische Abstand zum Mittelwert der Klasse (gewichtet durch σ^2) und einem Offset abhängig von der A-priori-Wahrscheinlichkeit der Klasse.

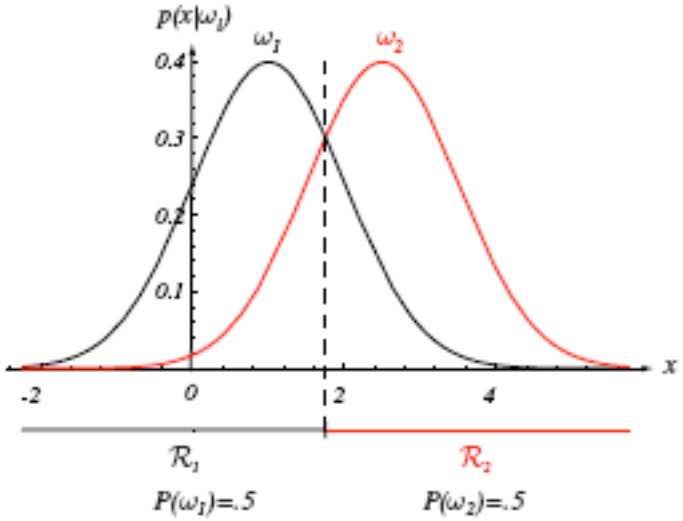
Noch einfachere Form: mit $\left\|\mathbf{x}-\mu_j\right\|^2 = (x-\mu)^T(x-\mu) = x^T x - 2\mu^T x + \mu^T\mu$

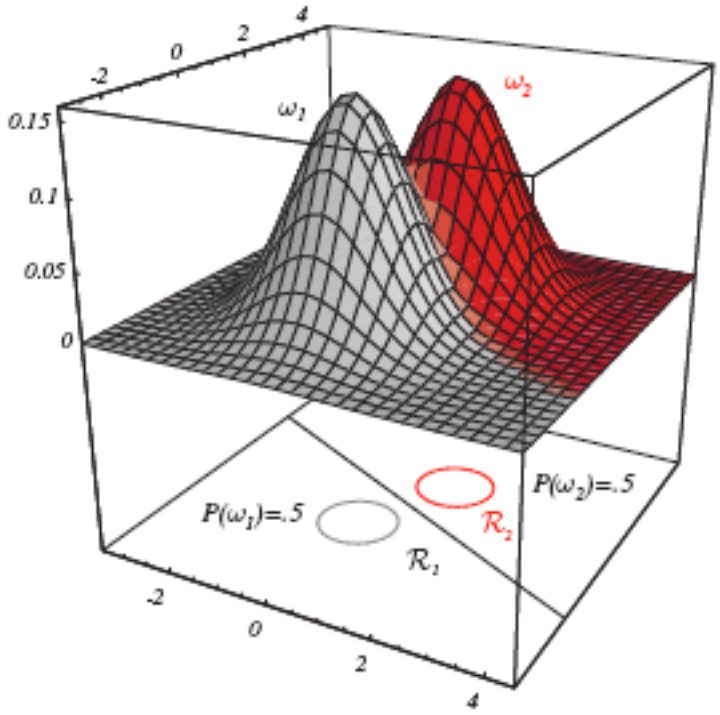
ergibt sich $g_j(\mathbf{x}) = w_j^T x + w_{j0}$

mit $w_j = \frac{1}{\sigma^2}\mu_j$ und $w_{j0} = -\frac{1}{\sigma^2}\mu_j^T\mu_j + \ln p(\omega_j)$

d.h. eine einfache lineare Diskriminantenfunktion. Der daraus konstruierte Klassifikator ist eine lineare Maschine.

Spezialfall: $\Sigma_j = \sigma^2 I$ in 1 und 2 Dimensionen



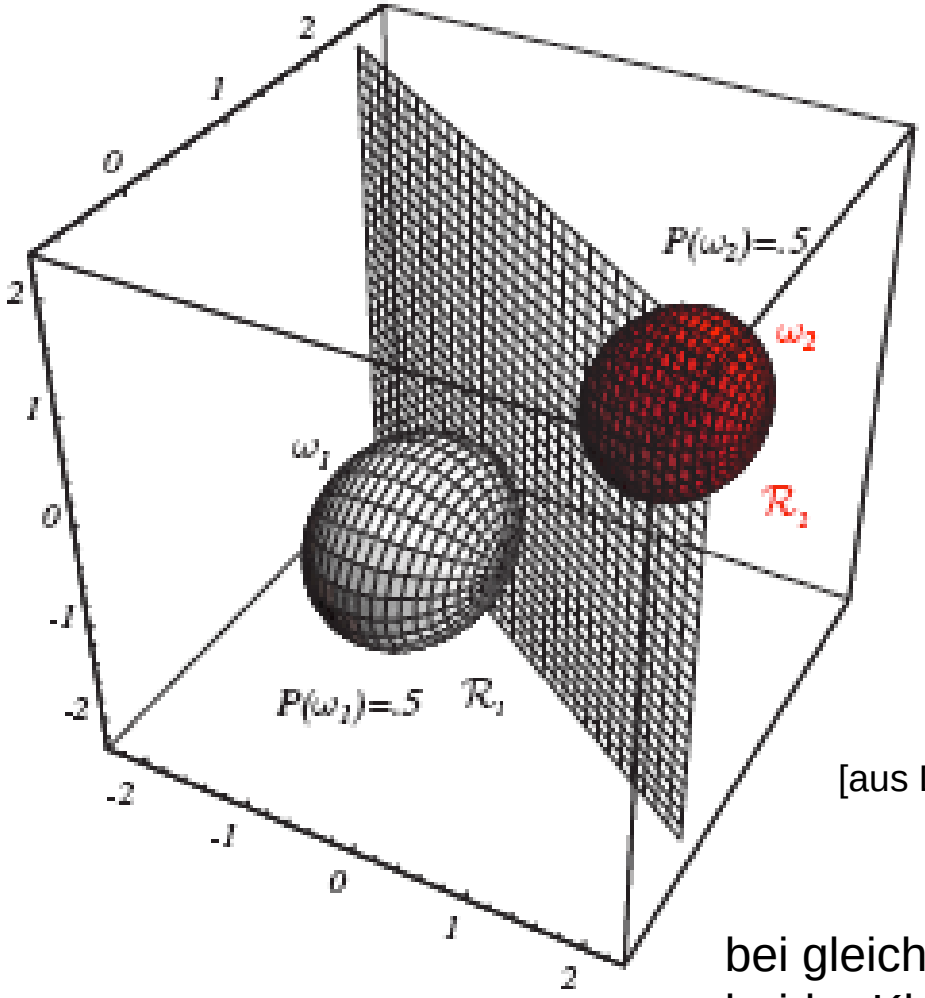


bei gleichen A-priori-Wahrscheinlichkeiten beider Klassen

Minimaldistanz-Klassifikator, Template Matching mit Mittelwert als Klassenprototyp

[aus Duda et al., 2001]

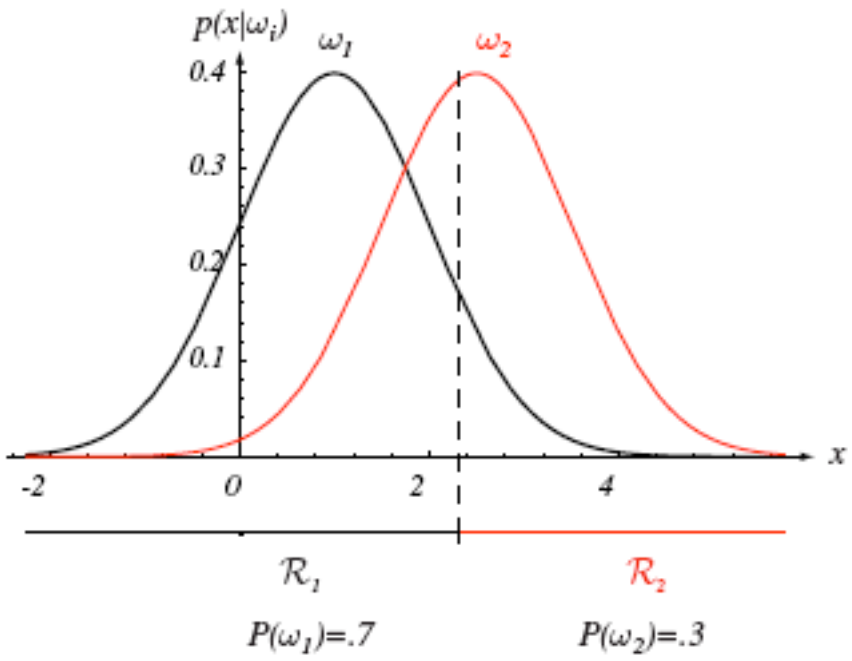
Spezialfall: $\Sigma_j = \sigma^2 I$ in 3 Dimensionen

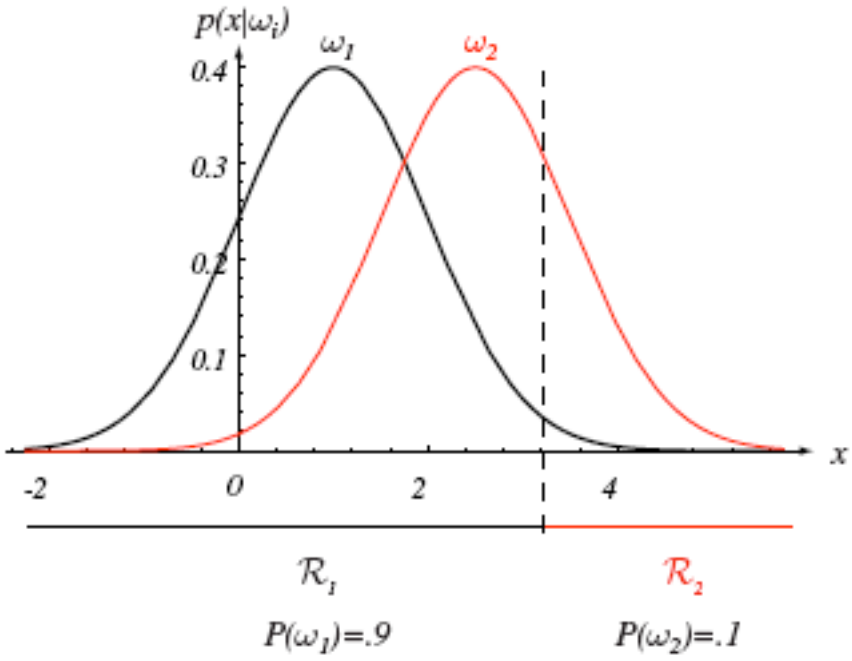


[aus Duda et al., 2001]

bei gleichen A-priori-Wahrscheinlichkeiten beider Klassen

Spezialfall: $\Sigma_j = \sigma^2 I$ in einer Dimension

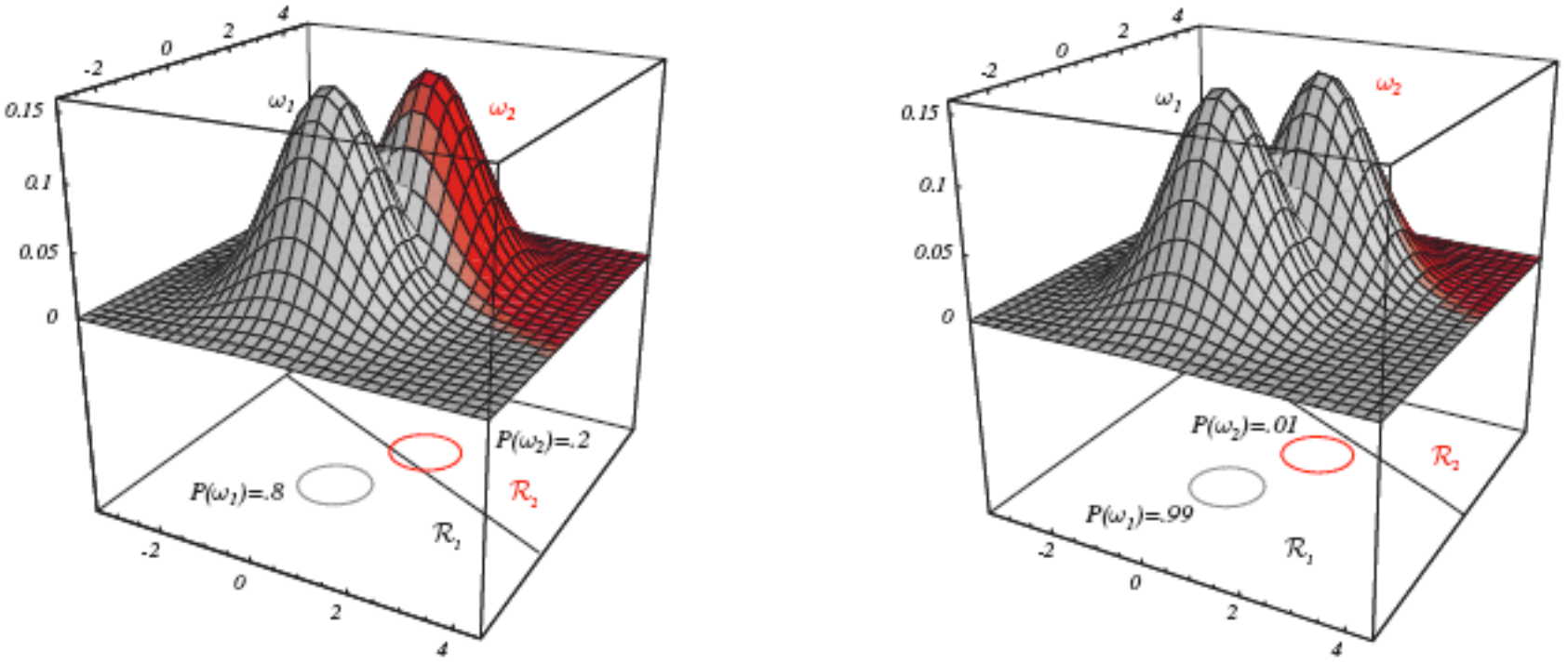




[aus Duda et al., 2001]

bei ungleichen A-priori-Wahrscheinlichkeiten beider Klassen

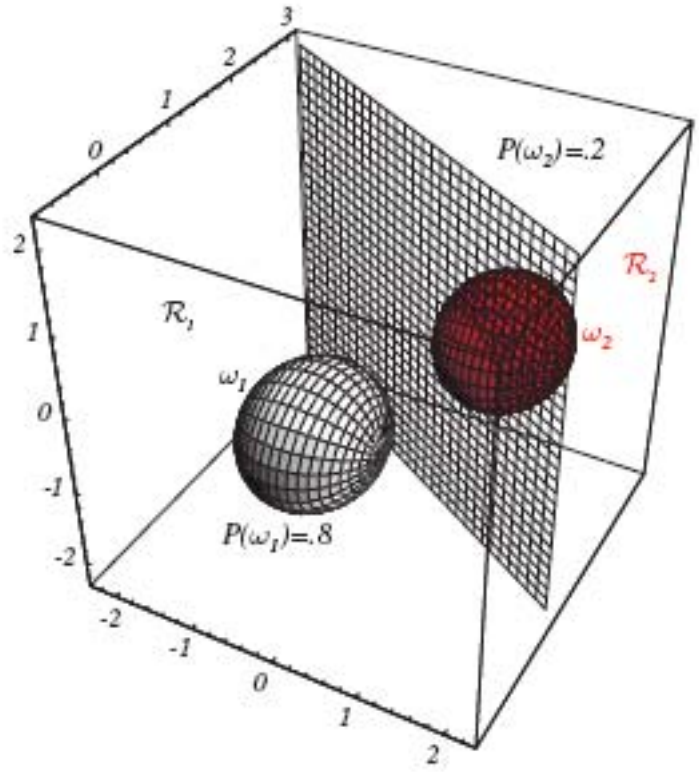
Spezialfall: $\Sigma_j = \sigma^2 I$ in zwei Dimensionen

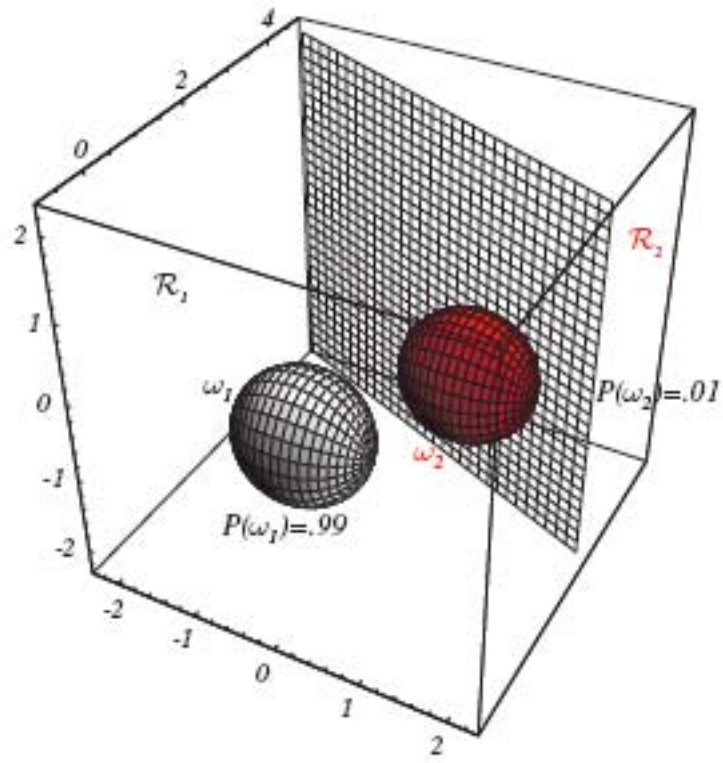


[aus Duda et al., 2001]

bei ungleichen A-priori-Wahrscheinlichkeiten beider Klassen

Spezialfall: $\Sigma_j = \sigma^2 I$ in drei Dimensionen





[aus Duda et al., 2001]

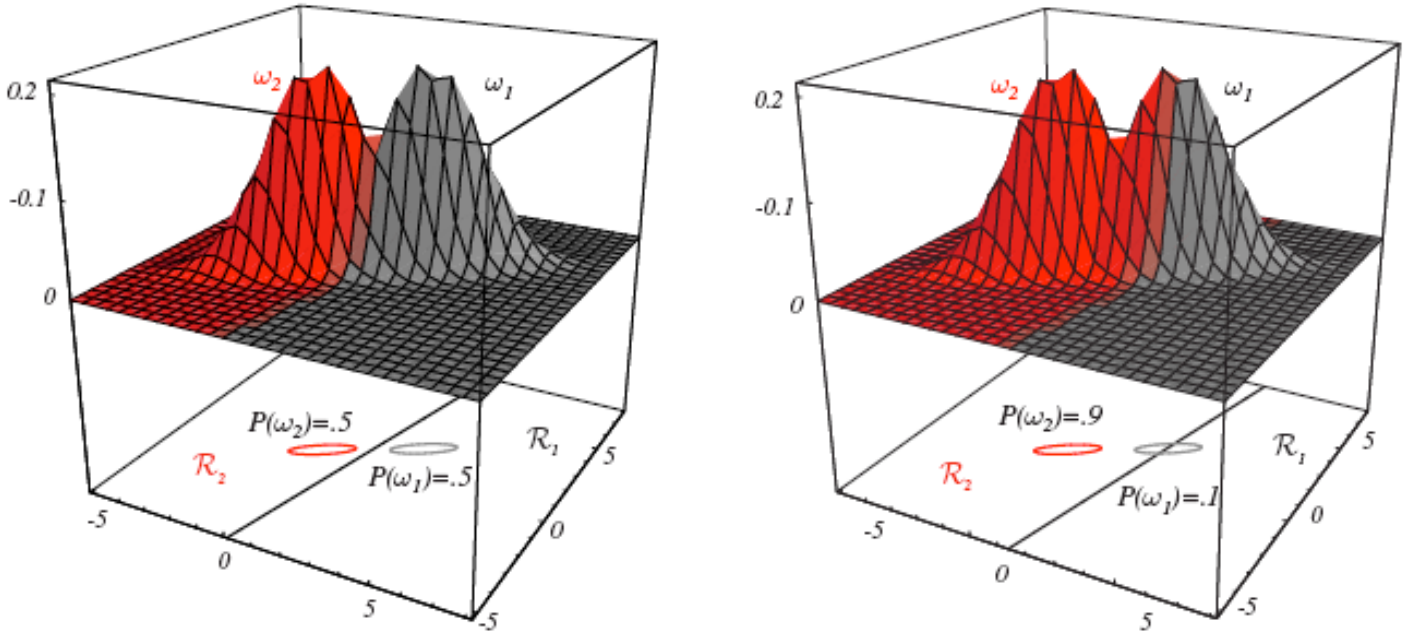
bei ungleichen A-priori-Wahrscheinlichkeiten beider Klassen

Spezialfall: beliebige, aber gleiche Kovarianzen

$$g_j(\mathbf{x}) = -\tfrac{1}{2}(\mathbf{x}-\mu_j)^T \Sigma_j^{-1}(\mathbf{x}-\mu_j) - \tfrac{d}{2}\ln 2\pi - \tfrac{1}{2}\ln\det\Sigma_j + \ln p(\omega_j)$$

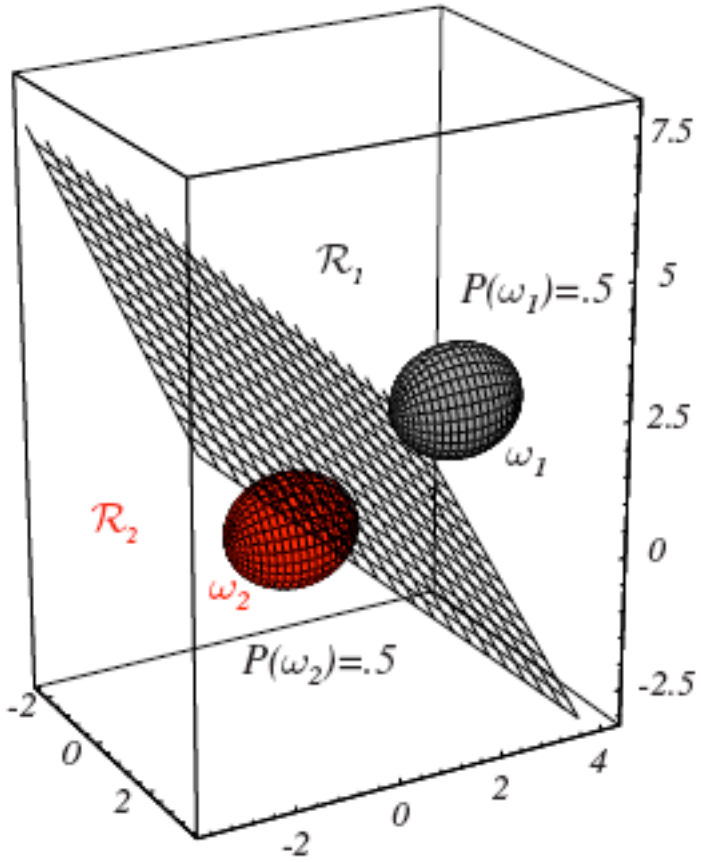
$$\Rightarrow g_j(\mathbf{x}) = -\frac{1}{2\sigma^2}(\mathbf{x}-\mu_j)^T \Sigma^{-1}(\mathbf{x}-\mu_j) + \ln p(\omega_j)$$

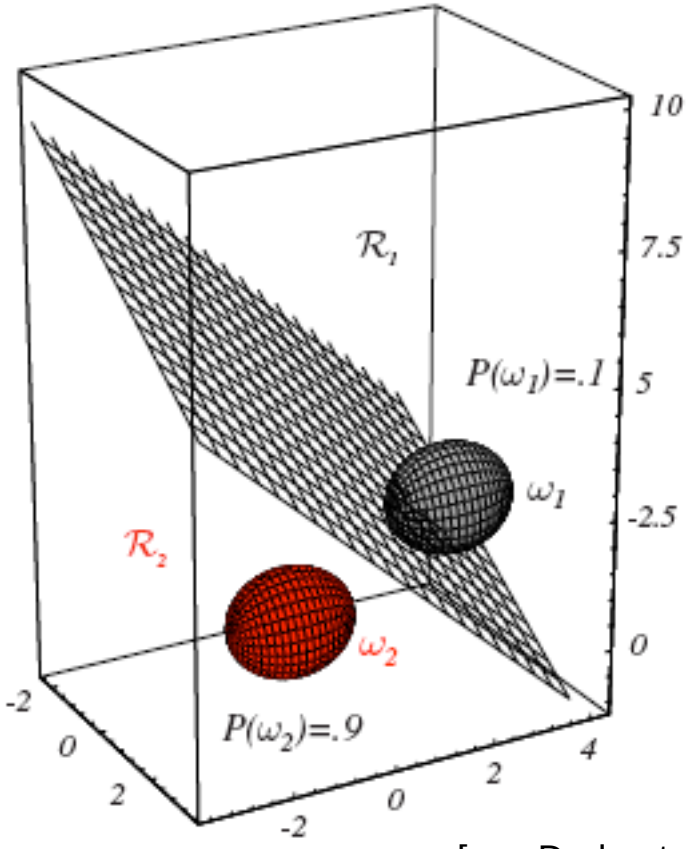
Läßt sich ebenfalls in eine lineare Form bringen: $g_j(\mathbf{x}) = w_j^T x + w_{j0}$



[aus Duda et al., 2001]

Spezialfall: beliebige, aber gleiche Kovarianzen (3D)





[aus Duda et al., 2001]

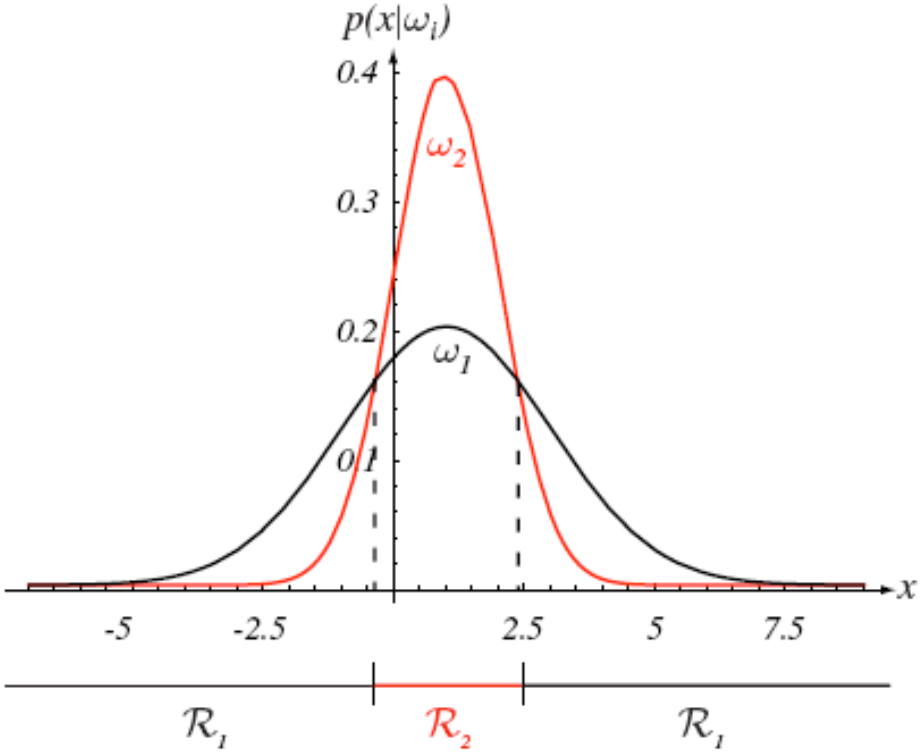
Trennflächen müssen nicht mehr senkrecht auf der Verbindungslinie der Klassenmittelwerte liegen

Beliebige, unterschiedliche Kovarianzen in allen Klassen

$$g_j(\mathbf{x}) = -\tfrac{1}{2}(\mathbf{x}-\mu_j)^T \Sigma_j^{-1}(\mathbf{x}-\mu_j) - \tfrac{d}{2}\ln 2\pi - \tfrac{1}{2}\ln\det\Sigma_j + \ln p(\omega_j)$$

$$\Rightarrow g_j(\mathbf{x}) = -\tfrac{1}{2}(\mathbf{x}-\mu_j)^T \Sigma_j^{-1}(\mathbf{x}-\mu_j) - \tfrac{1}{2}\ln\det\Sigma_j + \ln p(\omega_j)$$

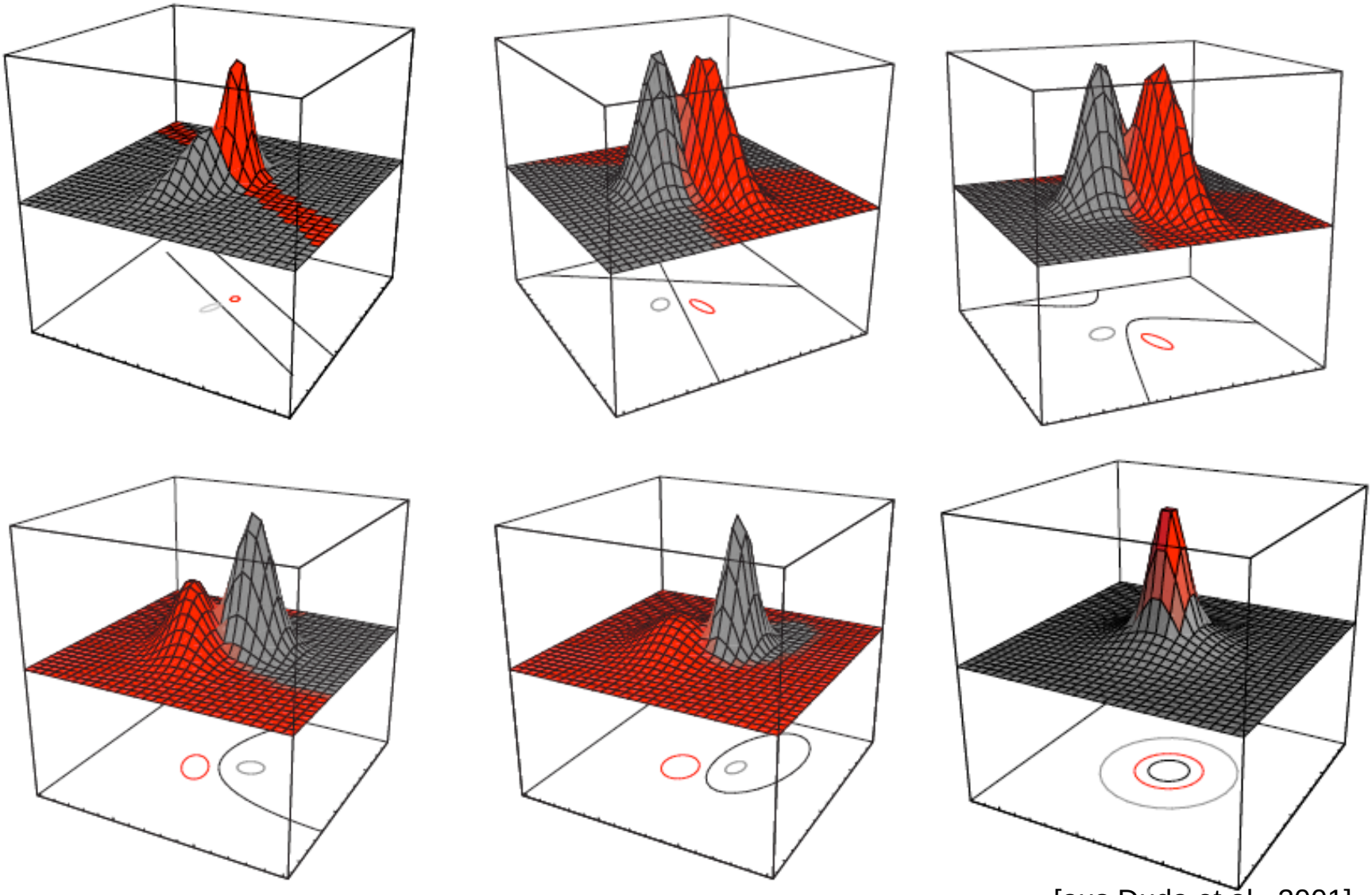
In 1 Dimension:



[aus Duda et al., 2001]

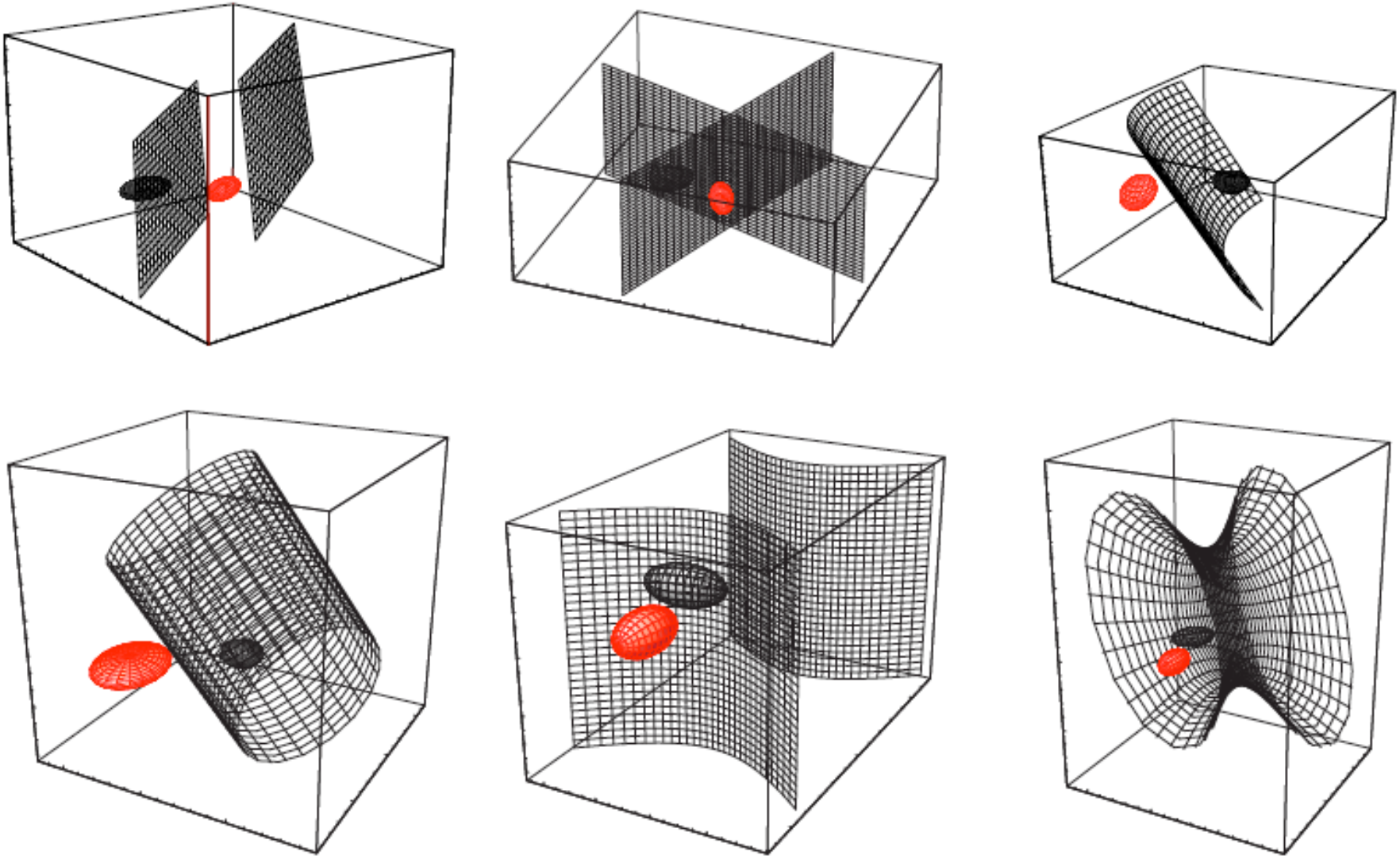
Entscheidungsregionen können aus unverbundenen Teilgebieten bestehen.

Bei 2 Kategorien: Quadriken als Trennflächen (2D)



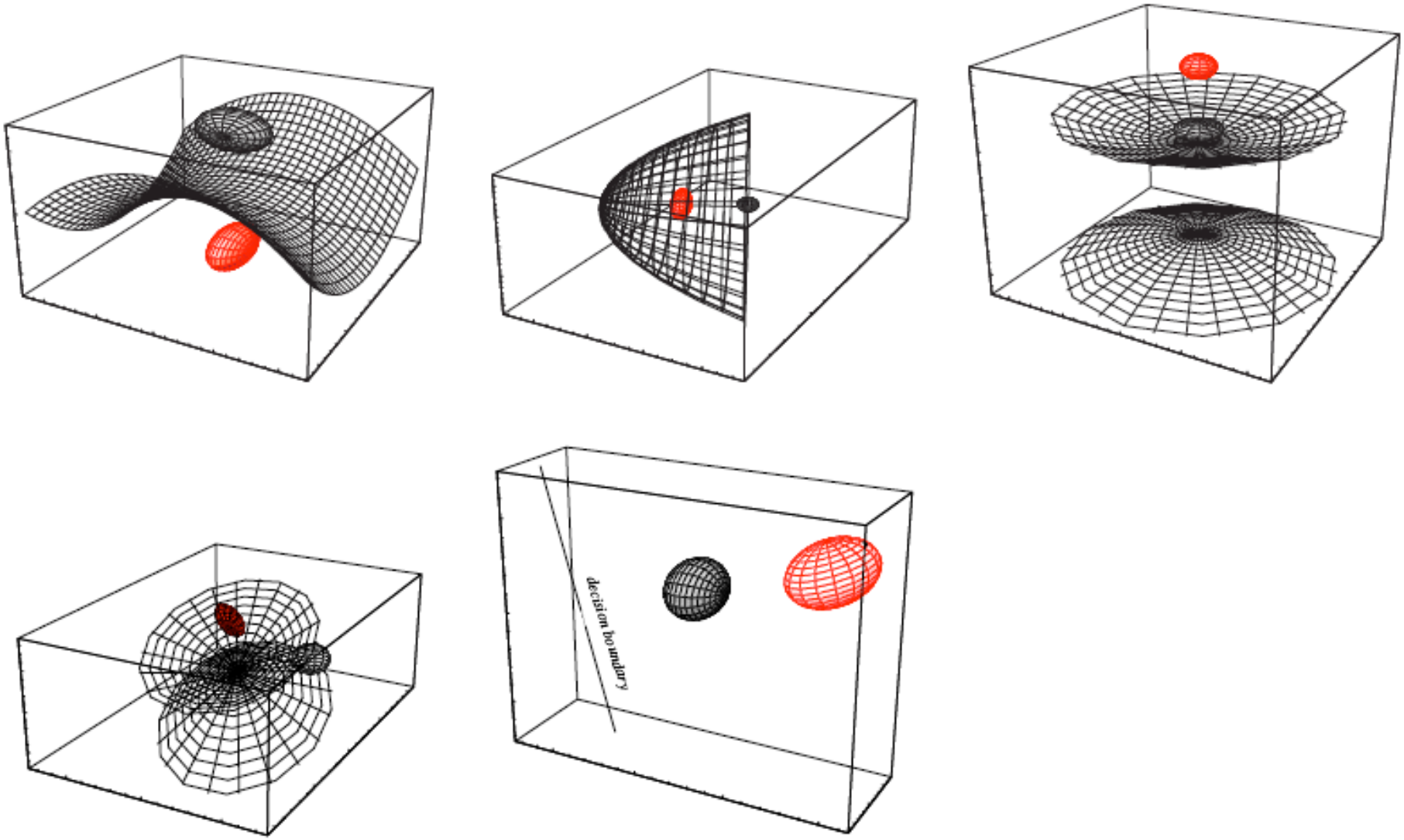
[aus Duda et al., 2001]

Bei 2 Kategorien: Quadriken als Trennflächen (3D)



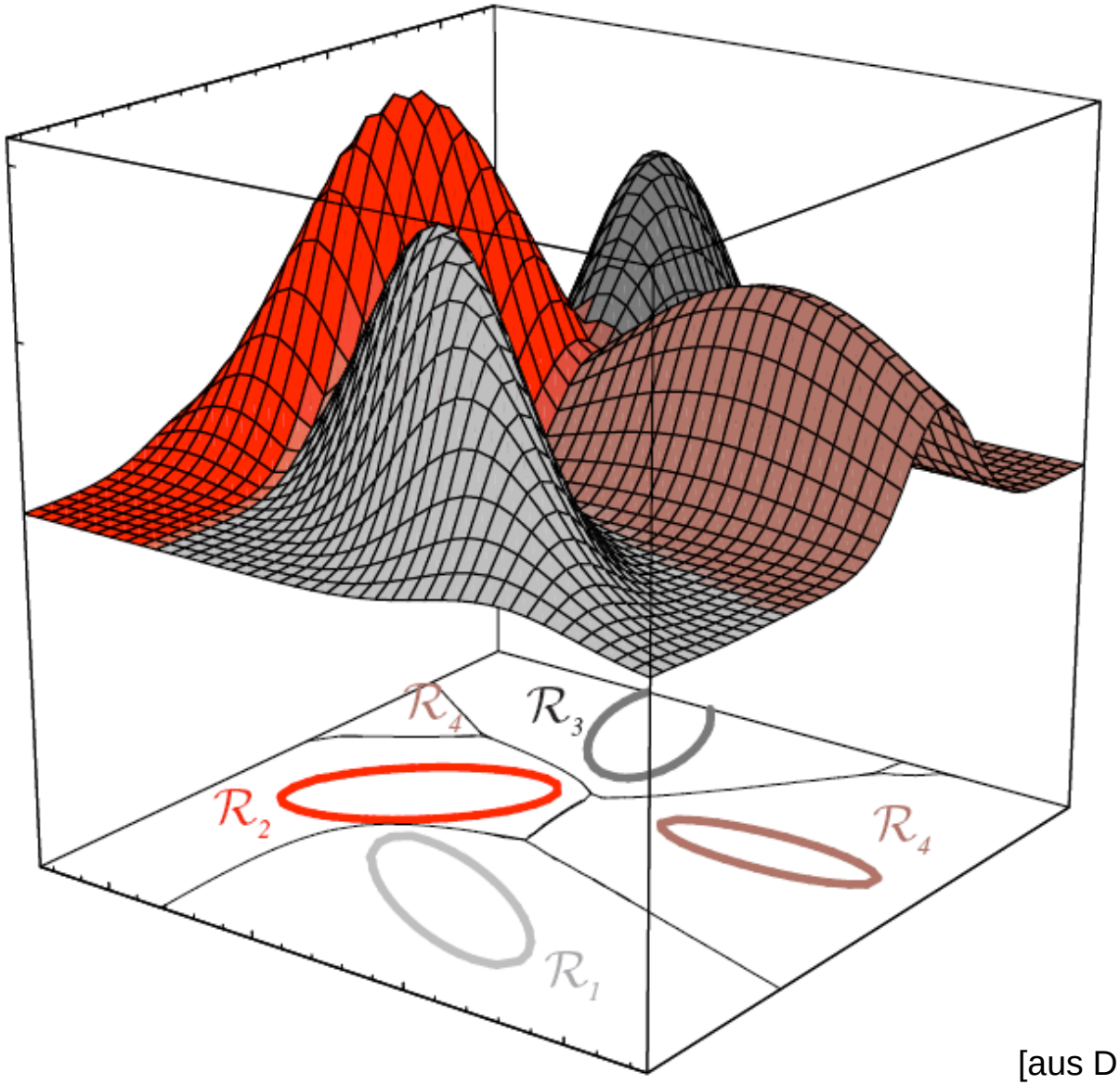
[aus Duda et al., 2001]

Bei 2 Kategorien: Quadriken als Trennflächen (3D) - 2



[aus Duda et al., 2001]

Bei mehr als 2 Kategorien...



[aus Duda et al., 2001]