3. Bayes'sche Klassifikation

3.1. Allgemeine Überlegungen zur Klassifikation

Definitionen:

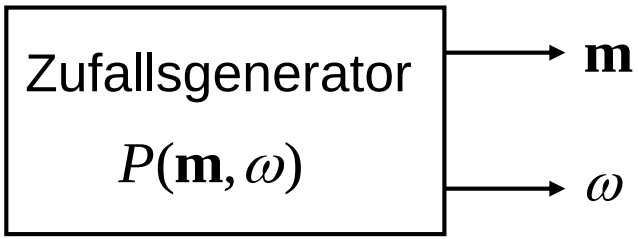
D: Lernstichprobe

T: Teststichprobe

S: Gesamtstichprobe $S = D \cup T$ mit $D \cap T = \varnothing$

Probabilistische Beschreibung von Klassen und Merkmalen

Mathematisches Modell: Umwelt als Zufallsgenerator, der Paare ($\mathbf{m}$, ω) erzeugt. Die Gesamtheit (das Ensemble) dieser Paare mit ihrer Variabilität und ihren Häufigkeiten wird durch eine Wahrscheinlichkeitsverteilung (WV) beschrieben.



Der Zufallsgenerator liefere dabei unabhängige Ausgaben in dem Sinne, dass die Verbundwahrscheinlichkeit von N Ausgaben ($\mathbf{m}^n$, ω^n) gleich dem Produkt $\Pi_{n=1}^{N} P(\mathbf{m}^n, \omega^n)$ ist.

3.1. Allgemeine Überlegungen zur Klassifikation

Probabilistische Beschreibung von Klassen und Merkmalen

$$P(\mathbf{m},\omega) = p(\mathbf{m}\,|\,\omega)P(\omega) = P(\omega\,|\,\mathbf{m})\,p(\mathbf{m})$$

Verbund-WV

$$P(\omega) = \int_{\mathbb{M}} P(\mathbf{m},\omega)\,\mathrm{d}\mathbf{m}$$

A Priori WV

$$p(\mathbf{m}\,|\,\omega)$$

Klassenspezifische Merkmals-WV, „Likelihood“

$$p(\mathbf{m}) = \sum_{\omega=\omega_1}^{\omega_c} P(\mathbf{m},\omega) = \sum_{\omega=\omega_1}^{\omega_c} p(\mathbf{m}\,|\,\omega)P(\omega)$$

Gesamt-WV der Merkmale

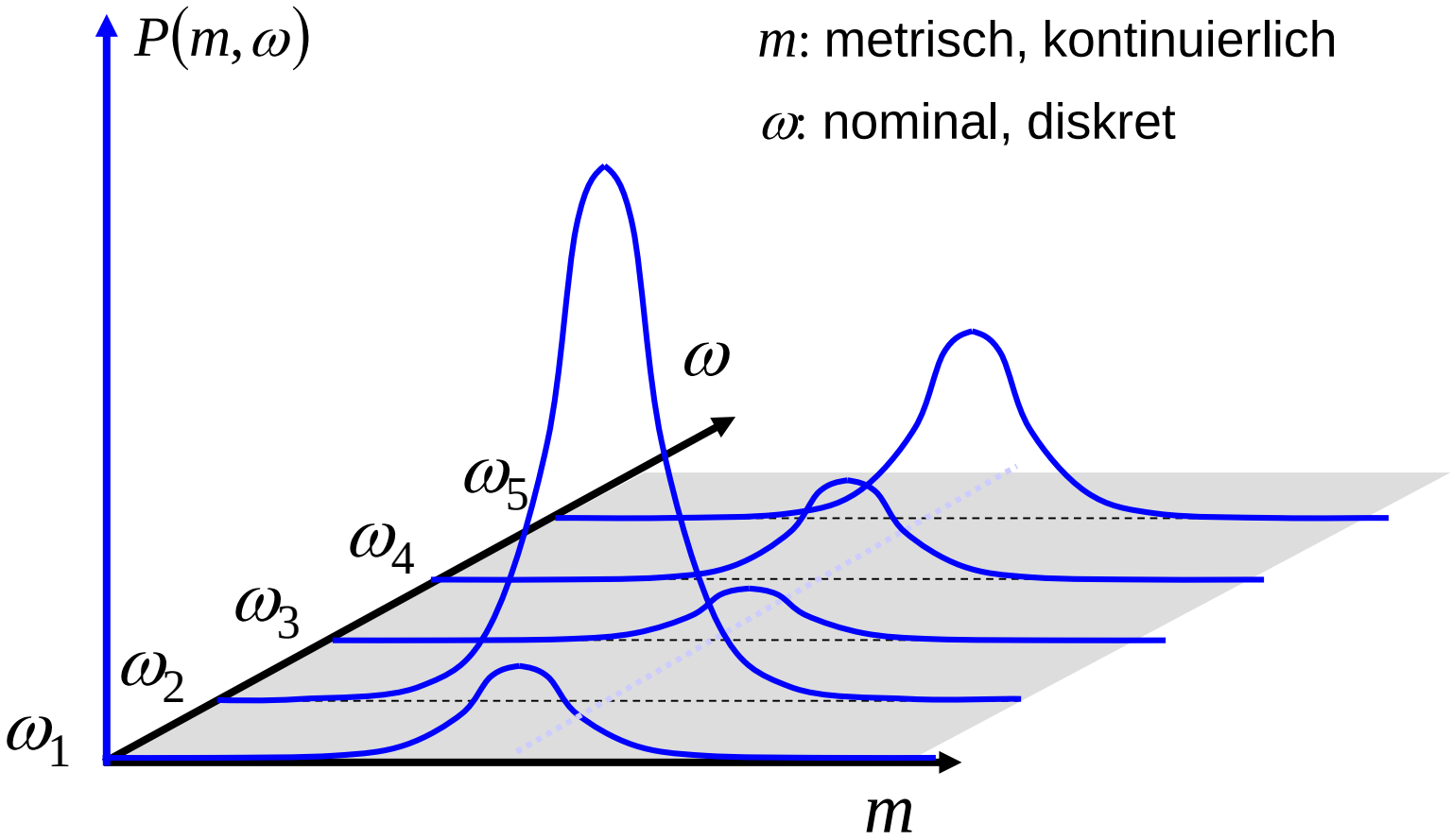
Bayes'sche Formel:

$$P(\omega\,|\,\mathbf{m}) = \frac{p(\mathbf{m}\,|\,\omega)P(\omega)}{p(\mathbf{m})}$$

A Posteriori WV

3.1. Allgemeine Überlegungen zur Klassifikation

Beispiel: Wahrscheinlichkeitsverteilung gemischter Größen



Entscheidungsraum (Klassifikationsraum): $\mathbb{K}$

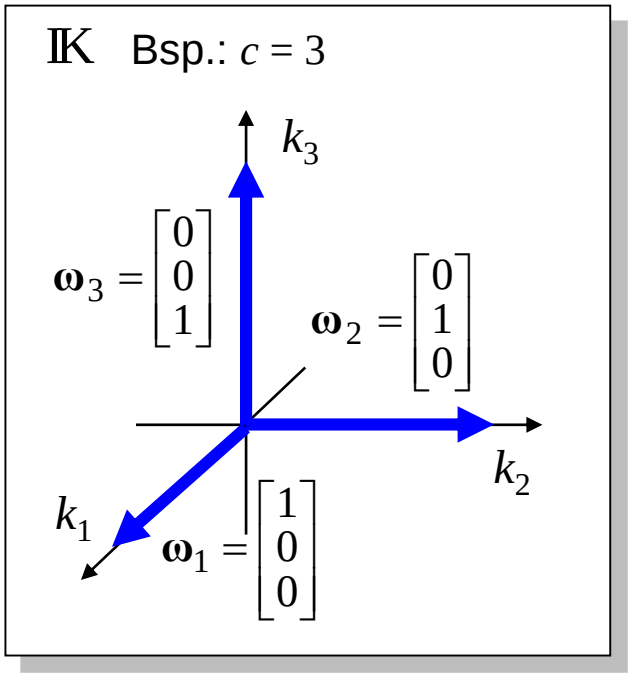
$$\Omega/\sim \; = \{\omega_1, ..., \omega_c\}$$

$$\Omega/\sim \; \rightarrow \; \{\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_c\}$$

$$\omega_i \mapsto \boldsymbol{\omega}_i, \quad i = 1, ..., c$$

$$\omega_i \quad \Leftrightarrow \quad \boldsymbol{\omega}_i := [0, ..., 1, ..., 0]^{\mathrm{T}} \in \mathbb{R}^c$$

i-te Komponente



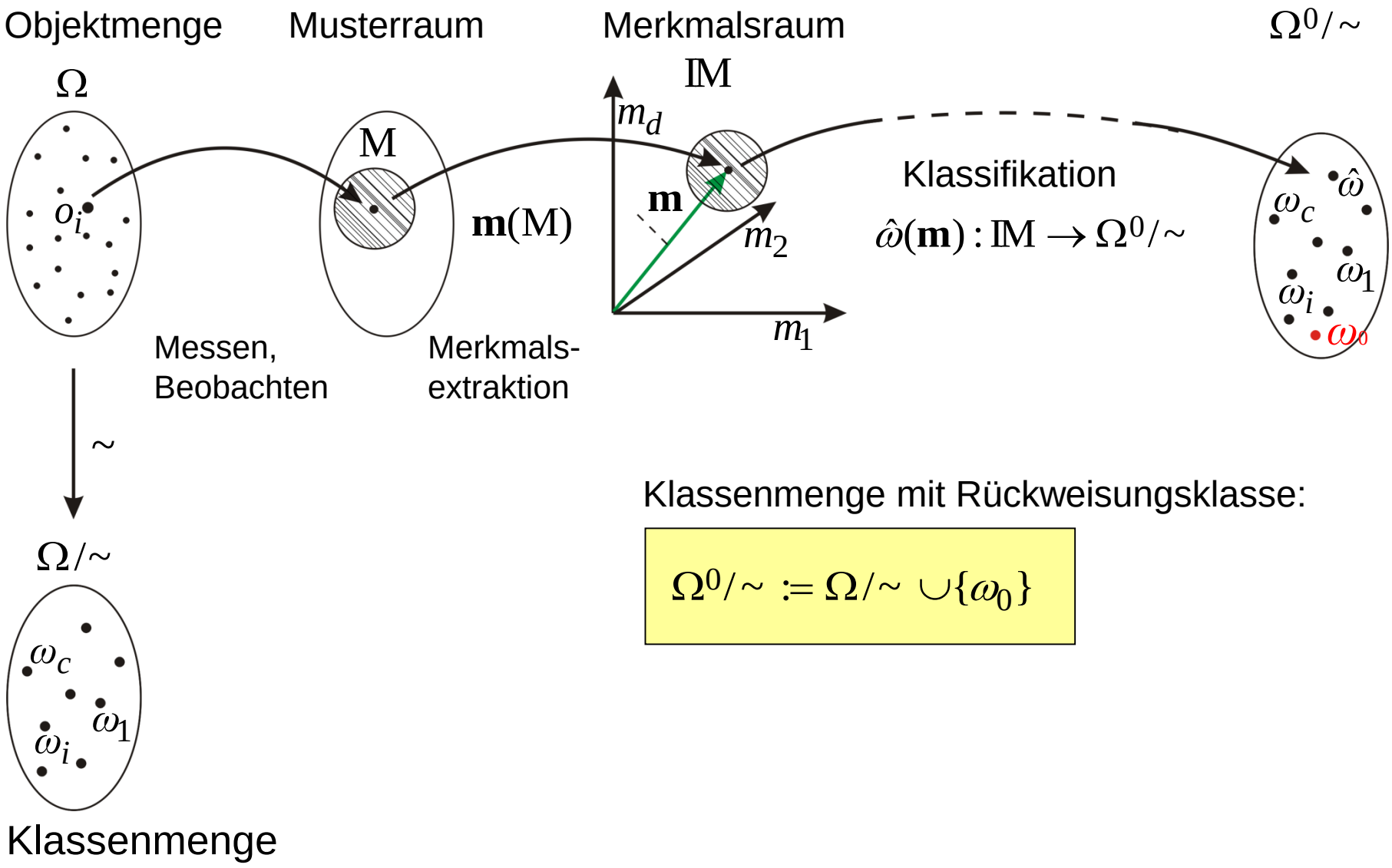
$\boldsymbol{\omega}_i$: Zielvektor (target vector)

$$\| \boldsymbol{\omega}_i \| = 1 \qquad \boldsymbol{\omega}_i^{\mathrm{T}} \boldsymbol{\omega}_j = \delta_i^j$$

Ordinalität der Nummerierung der ω_i suggeriert Nachbarschaften zwischen den Klassen, die mit Falle einer nominalen Klassenstruktur keine Bedeutung haben.

Die Vektorrepräsentation eliminiert diesen Anschein.

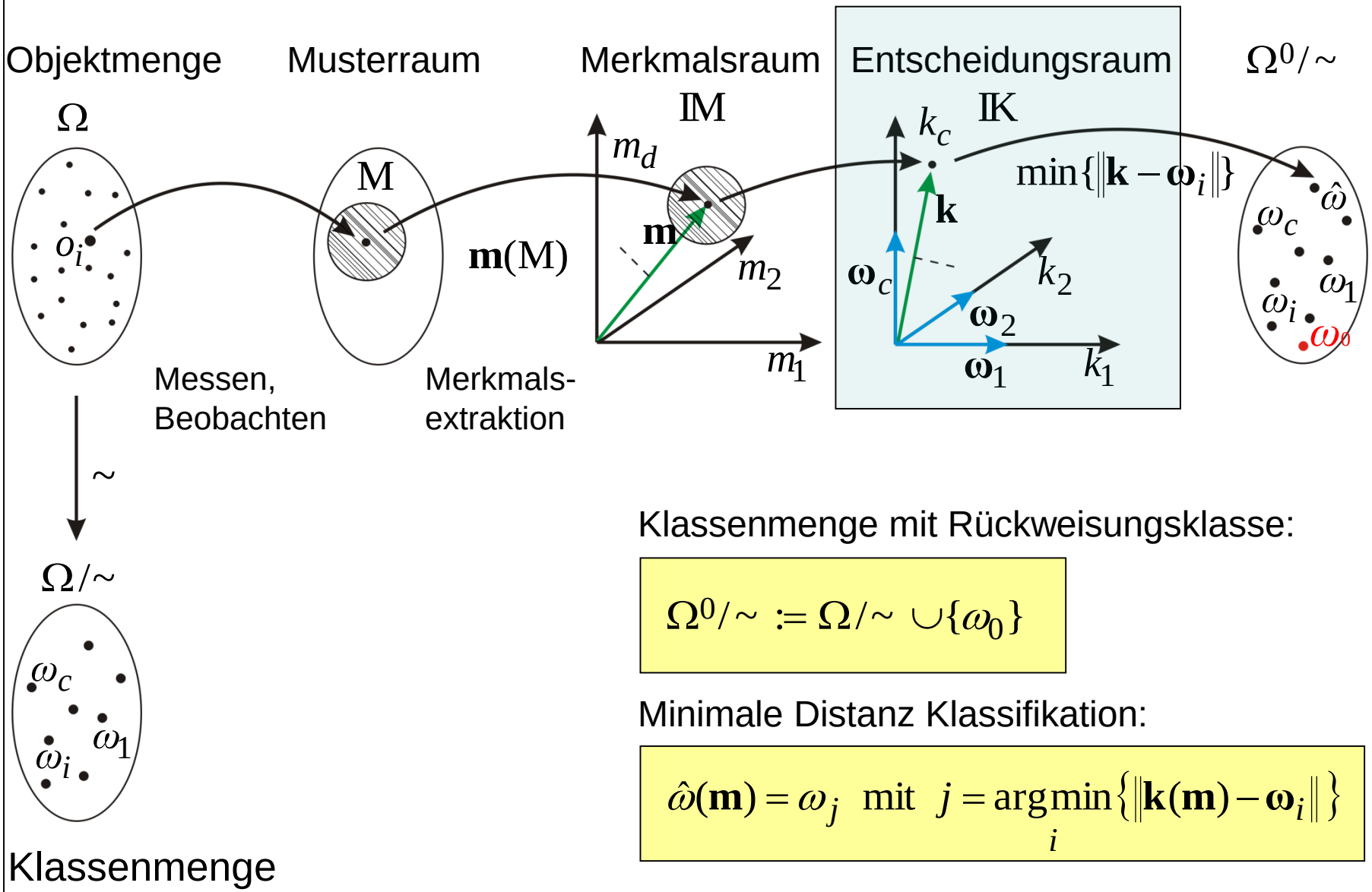
3.1. Allgemeine Überlegungen zur Klassifikation



Klassenmenge mit Rückweisungsklasse:

$$\Omega^0/\sim \; := \Omega/\sim \cup\{\omega_0\}$$

3.1. Allgemeine Überlegungen zur Klassifikation



Klassenmenge mit Rückweisungsklasse:

$$\Omega^0/\sim \;:=\; \Omega/\sim \cup\{\omega_0\}$$

Minimale Distanz Klassifikation:

$$\hat{\omega}(\mathbf{m}) = \omega_j \quad \text{mit} \quad j = \arg\min_i \left\{ \|\mathbf{k}(\mathbf{m}) - \boldsymbol{\omega}_i\| \right\}$$

3.1. Allgemeine Überlegungen zur Klassifikation

Entscheidungsvektor (Klassifikationsvektor):

$$\mathbf{k}(\mathbf{m}) : \mathbb{M} \rightarrow \mathbb{K} = \text{span}(\boldsymbol{\omega}_1,..., \boldsymbol{\omega}_c) = \mathbb{R}^c$$

Komponenten: Entscheidungsfunktionen

$$k_i(\mathbf{m}), \; i = 1,..., c$$

Alle Klassifikatoren (und mithin alle Entscheidungsgrenzen im Merkmalsraum) lassen sich mit diesem Formalismus beschreiben.

Parametrische Entscheidungsfunktionen: $\mathbf{k}(\mathbf{m};\boldsymbol{\theta})$

$\boldsymbol{\theta} = [\theta_1,..., \theta_k]^T \in \Theta$: Parametervektor

Θ : Parameterraum

$\boldsymbol{\theta}$ anhand der Daten D bestimmen bedeutet: Lernen aus Beispielen.

Ansatz: Lernen anhand der Stichprobe $\mathrm{D} = \{\mathbf{m}_1,..., \mathbf{m}_N\}$ mit bekannter Klassenzugehörigkeit $\Leftrightarrow$ Bestimmung der Parameter $\boldsymbol{\theta}$ derart, dass $\mathbf{k}(\mathbf{m})$ die wahren $\boldsymbol{\omega}(\mathbf{m}_i) \in \{\boldsymbol{\omega}_1,...,\boldsymbol{\omega}_c\}$ für alle $\mathbf{m}_i \in \mathrm{D}$ *in gewissem Sinne* möglichst gut approximiert.

3.1. Allgemeine Überlegungen zur Klassifikation

Zusammenhang:

Entscheidungsgebiete R_i im Merkmalsraum:

$$\mathrm{R}_i \subset \mathbb{M} \Leftrightarrow \left\{ \mathbf{k} \middle| \| \mathbf{k} - \boldsymbol{\omega}_i \| < \| \mathbf{k} - \boldsymbol{\omega}_j \|, \forall j \neq i \right\} \subset \mathbb{K}$$

$$\mathrm{R}_i = \left\{ \mathbf{m} \middle| \| \mathbf{k}(\mathbf{m};\boldsymbol{\theta}) - \boldsymbol{\omega}_i \| < \| \mathbf{k}(\mathbf{m};\boldsymbol{\theta}) - \boldsymbol{\omega}_j \|, \forall j \neq i \right\}$$

$$\partial \mathrm{R}_i = \left\{ \mathbf{m} \middle| \| \mathbf{k}(\mathbf{m};\boldsymbol{\theta}) - \boldsymbol{\omega}_i \| \leq \| \mathbf{k}(\mathbf{m};\boldsymbol{\theta}) - \boldsymbol{\omega}_j \|, \forall j \neq i \right\} \setminus \mathrm{R}_i$$

$\mathbf{k}$ ist parametrisiert durch $\boldsymbol{\theta} \Rightarrow \partial \mathrm{R}_i$ sind parametrisiert durch $\boldsymbol{\theta}$.

Struktur und Parameter von $\mathbf{k}(.,\boldsymbol{\theta})$ legen die Grenzen $\partial \mathrm{R}_i$ der Entscheidungsgebiete in $\mathbb{M}$ fest.

Entscheidungsvektor mit kleinster mittlerer quadratischer Abweichung

$$f := \mathrm{E}\{\| \mathbf{k}(\mathbf{m}) - \boldsymbol{\omega} \|^2\} = \sum_{i=1}^{c} \int_{\mathrm{MI}} \| \mathbf{k}(\mathbf{m}) - \boldsymbol{\omega}_i \|^2 \, P(\mathbf{m}, \boldsymbol{\omega}_i) \, \mathrm{d}\mathbf{m} \xrightarrow{!} \underset{\mathbf{k}(\mathbf{m})}{\text{minimal}}$$

Problem der Variationsrechnung.

Sei $\mathbf{k}$ die optimale Lösung. Dann gilt: $f(\mathbf{k} + \delta\mathbf{k}) > f(\mathbf{k}), \ \forall \delta\mathbf{k} \neq \mathbf{0}$

$\mathbf{k}$ und $\delta\mathbf{k}$ sind Funktionen von $\mathbf{m}$.

$$f(\mathbf{k} + \delta\mathbf{k}) = \mathrm{E}\left\{ \| \mathbf{k} + \delta\mathbf{k} - \boldsymbol{\omega} \|^2 \right\}$$

$$= \mathrm{E}\left\{ \| \mathbf{k} - \boldsymbol{\omega} \|^2 \right\} + 2\,\mathrm{E}\left\{ \delta\mathbf{k}^{\mathrm{T}} (\mathbf{k} - \boldsymbol{\omega}) \right\} + \mathrm{E}\left\{ \| \delta\mathbf{k} \|^2 \right\}$$

$$f(\mathbf{k}) = \mathrm{E}\left\{ \| \mathbf{k} - \boldsymbol{\omega} \|^2 \right\}$$

Entscheidungsvektor mit kleinster mittlerer quadratischer Abweichung

$$f(\mathbf{k}+\delta\mathbf{k}) > f(\mathbf{k}) \quad \Leftrightarrow \quad \mathrm{E}\left\{ \| \delta\mathbf{k} \|^2 \right\} + 2\mathrm{E}\left\{ \delta\mathbf{k}^{\mathrm{T}}(\mathbf{k}-\boldsymbol{\omega}) \right\} > 0, \quad \forall \delta\mathbf{k} \neq \mathbf{0}$$

Die Ungleichung ist erfüllt, wenn gilt: $\mathrm{E}\left\{ \delta\mathbf{k}^{\mathrm{T}}(\mathbf{k}-\boldsymbol{\omega}) \right\} = 0$

$$\mathrm{E}\left\{ \delta\mathbf{k}^{\mathrm{T}}(\mathbf{k}-\boldsymbol{\omega}) \right\} = \int_{\mathbb{M}} \sum_{i=1}^{c} \delta\mathbf{k}^{\mathrm{T}}(\mathbf{k}-\boldsymbol{\omega}_i) P(\mathbf{m},\boldsymbol{\omega}_i)\,\mathrm{d}\mathbf{m} =$$

$$= \int_{\mathbb{M}} \delta\mathbf{k}^{\mathrm{T}} \left[\sum_{i=1}^{c} (\mathbf{k}-\boldsymbol{\omega}_i) P(\boldsymbol{\omega}_i \mid \mathbf{m}) \right] p(\mathbf{m})\,\mathrm{d}\mathbf{m} = 0$$

Dieser Ausdruck wird für alle möglichen $\delta\mathbf{k}$ zu Null, falls der geklammerte Ausdruck Null ist.

Entscheidungsvektor mit kleinster mittlerer quadratischer Abweichung

$$\sum_{i=1}^{c}(\mathbf{k}-\boldsymbol{\omega}_i)P(\boldsymbol{\omega}_i\,|\,\mathbf{m}) = \mathbf{k}\sum_{i=1}^{c}P(\boldsymbol{\omega}_i\,|\,\mathbf{m}) - \sum_{i=1}^{c}\boldsymbol{\omega}_i P(\boldsymbol{\omega}_i\,|\,\mathbf{m}) = 0$$

Der optimale Entscheidungsvektor lautet somit:

$$\mathbf{k}(\mathbf{m}) = \sum_{i=1}^{c}\boldsymbol{\omega}_i P(\boldsymbol{\omega}_i\,|\,\mathbf{m}) = \mathrm{E}\{\boldsymbol{\omega}\,|\,\mathbf{m}\}$$

Setzt man nun die Definitionsgleichungen der Vektoren $\boldsymbol{\omega}_i$ ein und beachtet $P(\boldsymbol{\omega}_i\,|\,\mathbf{m}) = P(\omega_i\,|\,\mathbf{m})$, so folgt:

$$\mathbf{k}(\mathbf{m}) = \begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix}P(\omega_1\,|\,\mathbf{m}) + \begin{bmatrix}0\\1\\\vdots\\0\end{bmatrix}P(\omega_2\,|\,\mathbf{m}) + \ldots + \begin{bmatrix}0\\0\\\vdots\\1\end{bmatrix}P(\omega_c\,|\,\mathbf{m}) = \begin{bmatrix}P(\omega_1\,|\,\mathbf{m})\\P(\omega_2\,|\,\mathbf{m})\\\vdots\\P(\omega_c\,|\,\mathbf{m})\end{bmatrix} =: \mathbf{p}$$

Entscheidungsvektor mit kleinster mittlerer quadratischer Abweichung

Der optimale Entscheidungsvektor ist der Vektor der A Posteriori Wahrscheinlichkeiten der Klassen ω bei gegebenem Merkmalsvektor $\mathbf{m}$.

Entscheidung für Klasse ω_i mit minimalem Abstand zwischen $\boldsymbol{\omega}_i$ und $\mathbf{k}(\mathbf{m})$.

$$\left\|\mathbf{k}(\mathbf{m})-\boldsymbol{\omega}_i\right\|^2 = \left\|\mathbf{p}-\boldsymbol{\omega}_i\right\|^2 \xrightarrow{!} \min_i$$

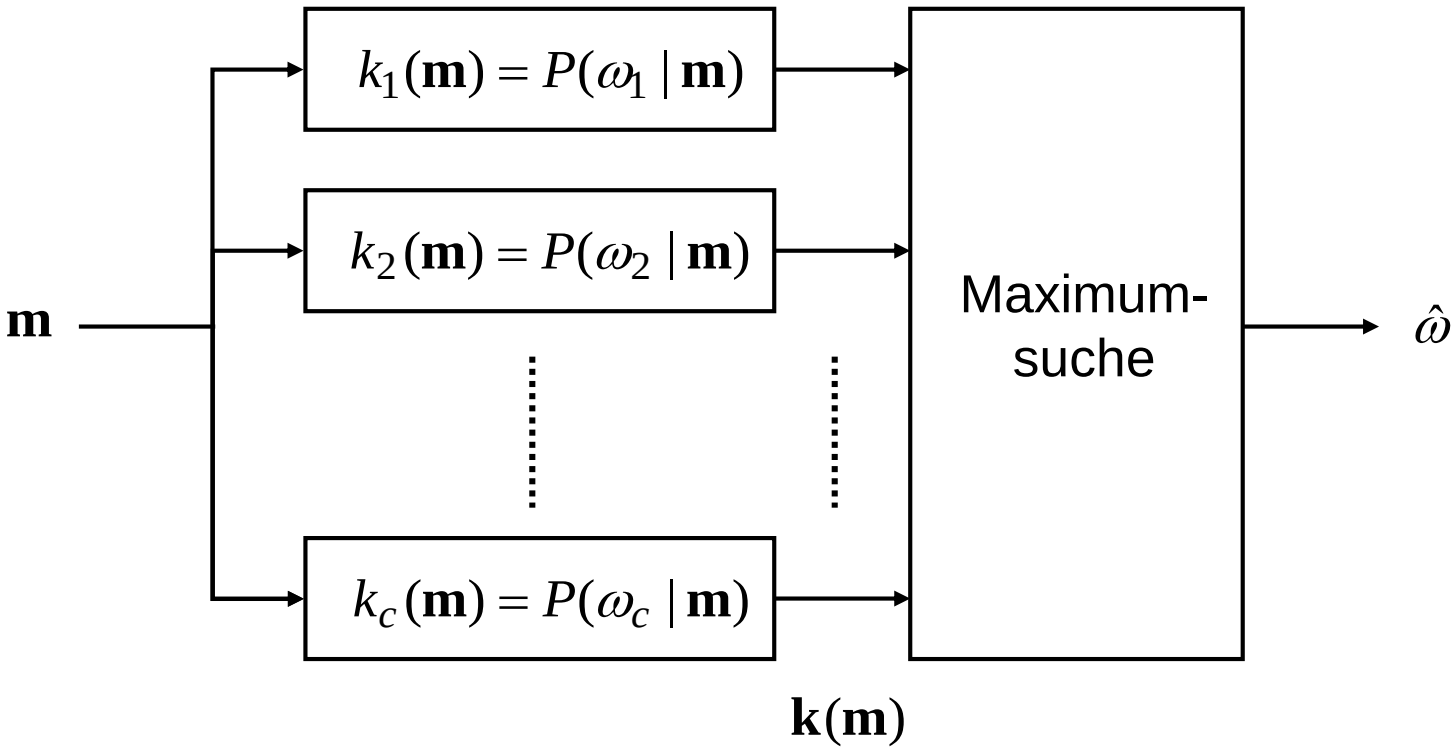
$$\left\|\mathbf{p}-\boldsymbol{\omega}_i\right\|^2 = \left\| \begin{bmatrix} P_1 \\ \vdots \\ P_i - 1 \\ \vdots \\ P_c \end{bmatrix} \right\|^2 = \sum_{j\neq i} P_j^2 + (P_i-1)^2 = \sum_{j=1}^{c} P_j^2 - 2P_i + 1 = \left\|\mathbf{p}\right\|^2 - 2P_i + 1$$

Ausdruck wird minimal für das größte $P_i := P(\omega_i \,|\, \mathbf{m})$.

Optimalentscheidung: Klasse mit maximaler A Posteriori Wahrscheinlichkeit

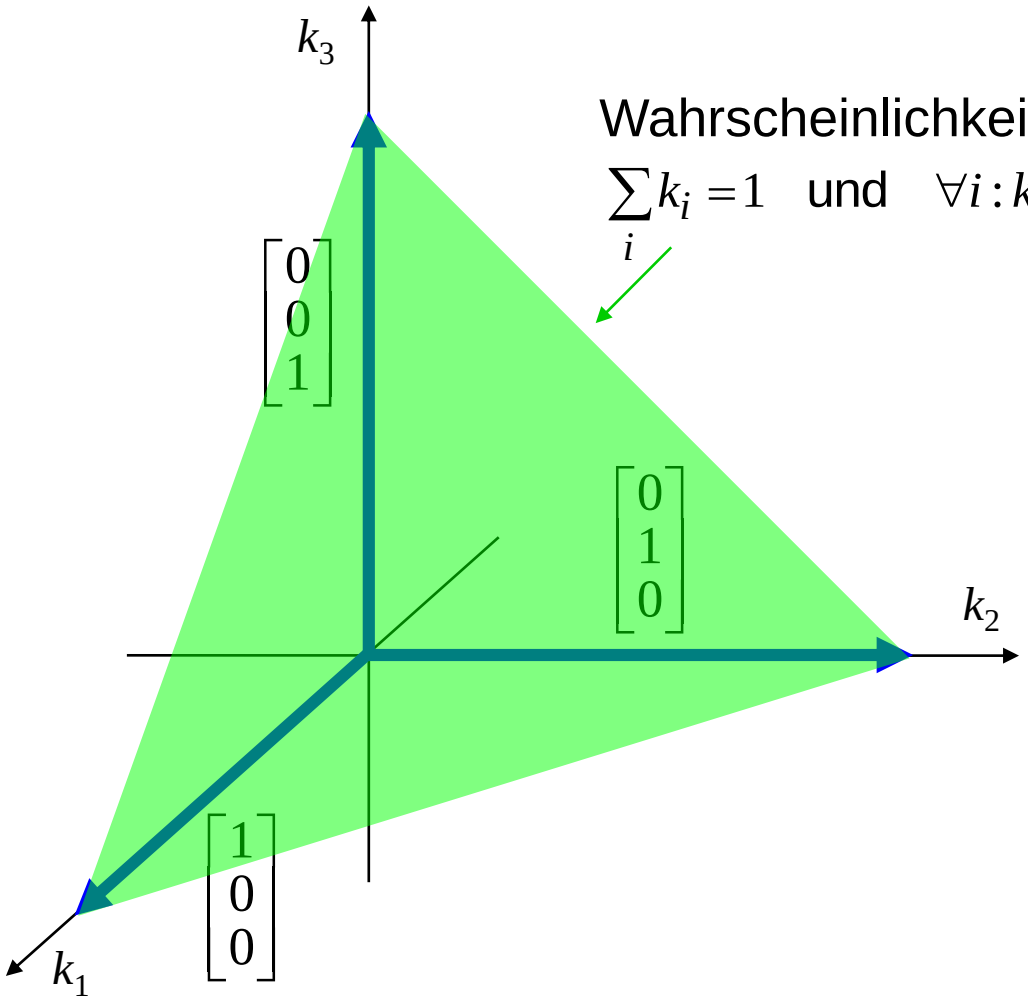
3.1. Allgemeine Überlegungen zur Klassifikation

Maximum A Posteriori Entscheidung (MAP)



3.1. Allgemeine Überlegungen zur Klassifikation

Wahrscheinlichkeiten im Entscheidungsraum

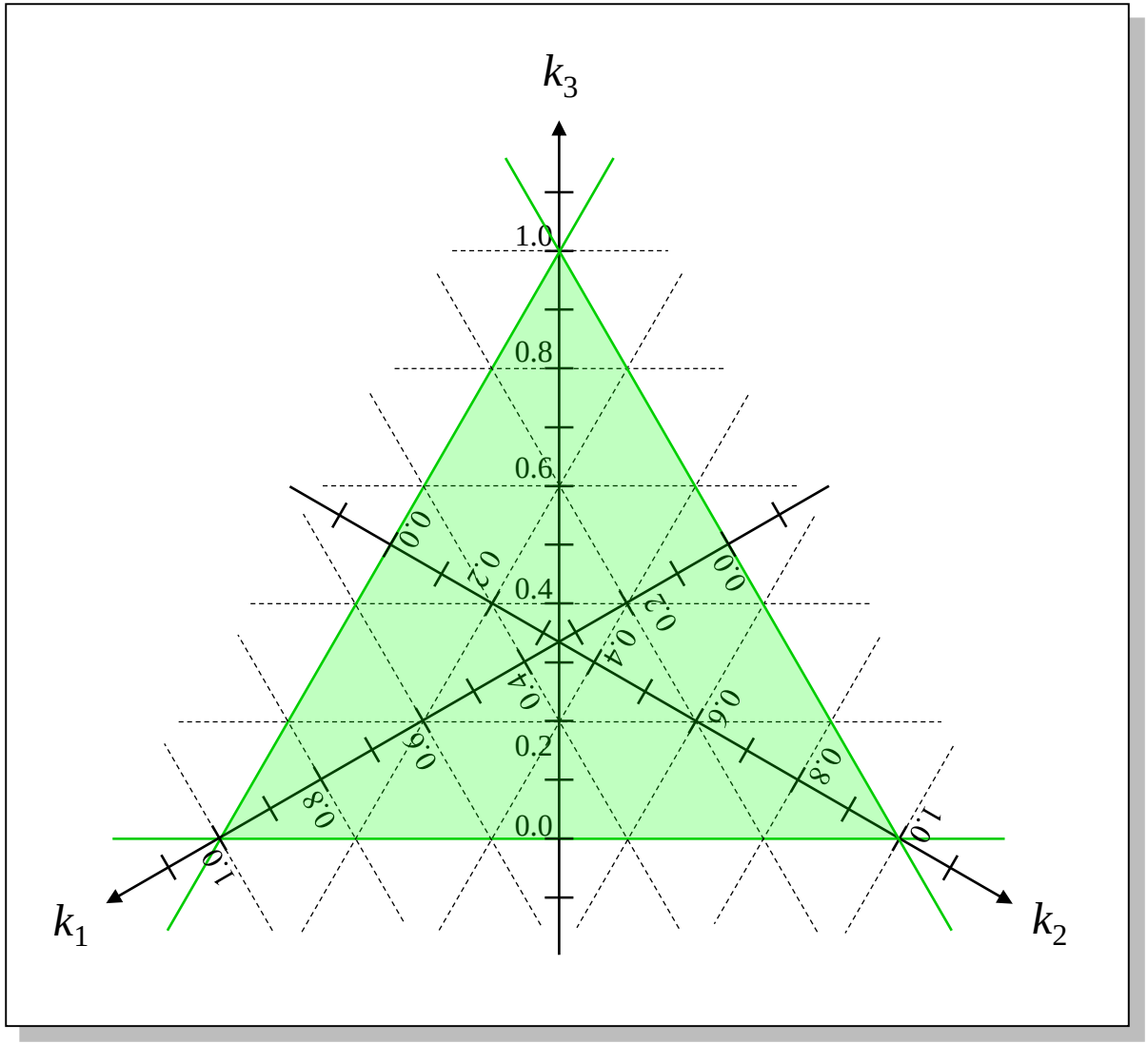


Wahrscheinlichkeitssimplex:

$$\sum_i k_i = 1 \quad \text{und} \quad \forall i : k_i \geq 0, k_i \leq 1$$

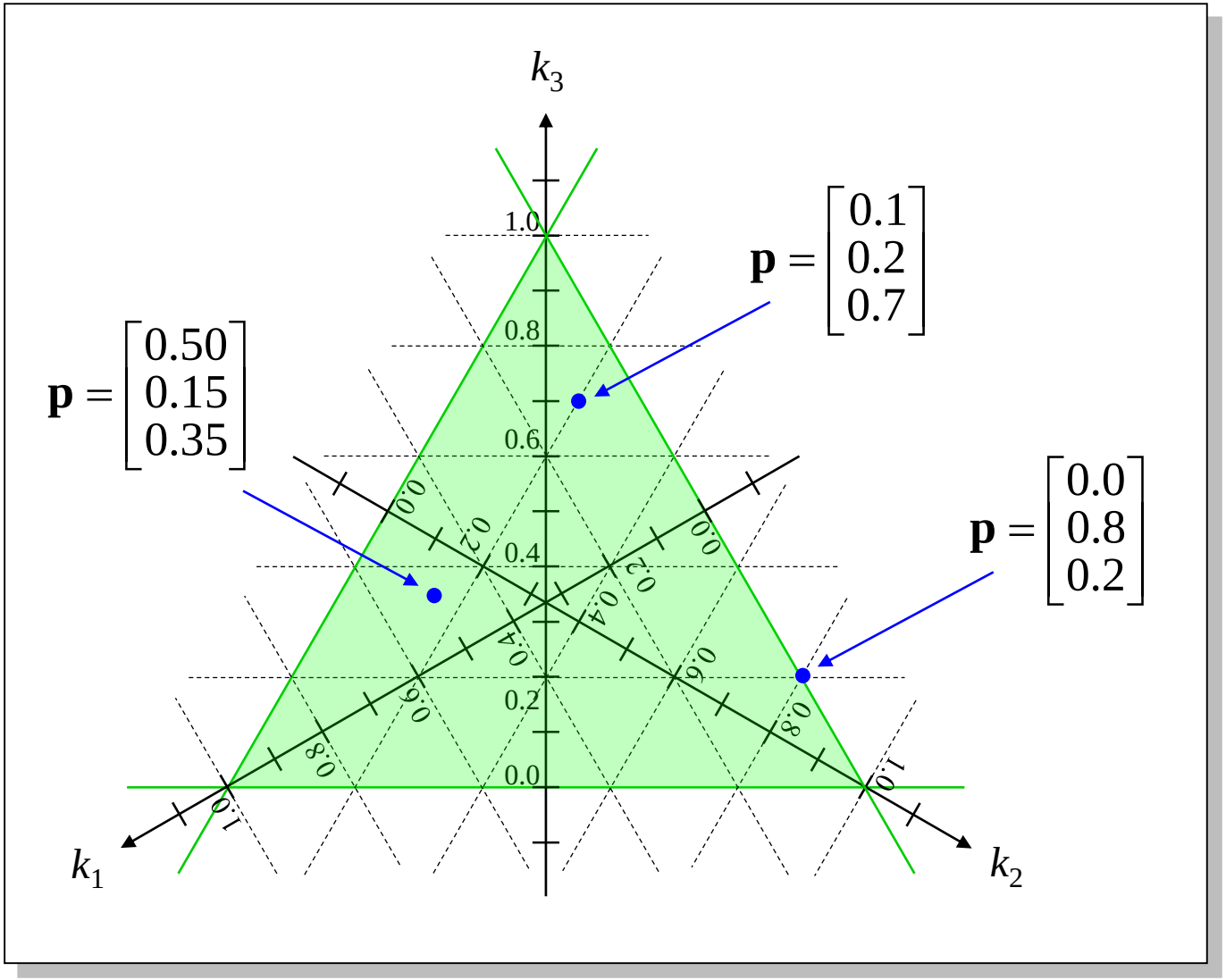
3.1. Allgemeine Überlegungen zur Klassifikation

Wahrscheinlichkeiten im Entscheidungsraum



3.1. Allgemeine Überlegungen zur Klassifikation

Wahrscheinlichkeiten im Entscheidungsraum



Allgemeiner Fall:

Kostenfunktion: $l(\hat{\omega},\omega) : \Omega^0/\sim \times \Omega/\sim \rightarrow \mathbb{R}$

Beschreibt die Kosten für die Entscheidung für $\hat{\omega}$,
wenn tatsächlich die Klasse ω zugrunde liegt.

Kostenmatrix:

$$\mathbf{L} := \begin{bmatrix} l(\omega_0,\omega_1) & l(\omega_0,\omega_2) & \cdots & l(\omega_0,\omega_c) \\ \hline l(\omega_1,\omega_1) & l(\omega_1,\omega_2) & \cdots & l(\omega_1,\omega_c) \\ \vdots & \vdots & \ddots & \vdots \\ l(\omega_c,\omega_1) & \cdots & \cdots & l(\omega_c,\omega_c) \end{bmatrix}$$

Ziel: Risiko $R = \mathrm{E}\{l(\hat{\omega},\omega)\} \overset{!}{=} \underset{\hat{\omega}(\mathbf{m})}{\text{minimal}}$ = minimale mittlere Kosten

Allgemeiner Fall:

$$R = \mathrm{E}\{l(\hat{\omega},\omega)\} = \int_{\mathbb{M}} \sum_{\omega=\omega_1}^{\omega_c} l(\hat{\omega}(\mathbf{m}),\omega) P(\mathbf{m},\omega)\,\mathrm{d}\mathbf{m} =$$

$$= \int_{\mathbb{M}} \underbrace{\left(\sum_{\omega=\omega_1}^{\omega_c} l(\hat{\omega}(\mathbf{m}),\omega) P(\omega\,|\,\mathbf{m}) \right)}_{} p(\mathbf{m})\,\mathrm{d}\mathbf{m} \quad \overset{!}{=} \quad \min_{\hat{\omega}(\mathbf{m})}$$

$$R(\hat{\omega}\,|\,\mathbf{m}) = \sum_{\omega=\omega_1}^{\omega_c} l(\hat{\omega}(\mathbf{m}),\omega) P(\omega\,|\,\mathbf{m}) \overset{!}{=} \min_{\hat{\omega}(\mathbf{m})}$$

A Posteriori Risiko (bedingtes Risiko)

$$\mathbf{r} = [r_0, r_1, \ldots, r_c]^{\mathrm{T}} := \left[R(\omega_0\,|\,\mathbf{m}), R(\omega_1\,|\,\mathbf{m}), \ldots, R(\omega_c\,|\,\mathbf{m})\right]^{\mathrm{T}} = \mathbf{L}\mathbf{p}$$

A Posteriori Risiken:

$$r_i = R(\omega_i\,|\,\mathbf{m}) = \sum_{j=1}^{c} l(\omega_i,\omega_j) P(\omega_j\,|\,\mathbf{m})$$

Optimale Entscheidung:

$$r_i = \min\{r_0, r_1, \ldots, r_c\} \quad \Leftrightarrow \quad \hat{\omega} = \omega_i$$

Diskussion:

Ingredienzen Bayes'sche Entscheidungstheorie:

- Klassenbedingte WV der Merkmale und A Priori WV der Klassen
  → A Posteriori WV der Klassen
- Kostenfunktion

Bayes'sche Entscheidungstheorie nutzt im Kontext des probabilistischen Ansatzes alles, was man über die Merkmale und die Klassen für die Gesamtheit aller Objekte der Domäne überhaupt wissen kann.

→ Leistung des Bayes'schen Optimalklassifikators stellt die obere Grenze für die Leistung anderer Klassifikatoren dar.

→ Bayes'sche Optimalklassifikation dient als Referenz für andere Klassifikatoren.

Problem: A Posteriori WV der Klassen $P(\omega|\mathbf{m})$ bzw. deren Bestimmungsstücke: die klassenspezifischen WVen der Merkmale $p(\mathbf{m}|\omega)$ und die A Priori WV $P(\omega)$ der Klassen sind i.d.R. nicht bekannt.

Praktische Vorgehensweise: Z.B. parametrische Modelle insbesondere für $p(\mathbf{m};\boldsymbol{\theta}|\omega)$ und Schätzung der Parameter $\boldsymbol{\theta}$ anhand der Lernstichprobe D.

Beispiel für 2 Klassen:

Kosten: $l_{ij} = l(\hat{\omega} = \omega_i , \omega_j) \quad i, j = 1, 2$

Risiko: $R(\omega_1 \mid \mathbf{m}) = l_{11} P(\omega_1 \mid \mathbf{m}) + l_{12} P(\omega_2 \mid \mathbf{m})$

$R(\omega_2 \mid \mathbf{m}) = l_{21} P(\omega_1 \mid \mathbf{m}) + l_{22} P(\omega_2 \mid \mathbf{m})$

Entscheidung:

$\hat{\omega} = \omega_1$ wenn $R(\omega_1 \mid \mathbf{m}) < R(\omega_2 \mid \mathbf{m})$

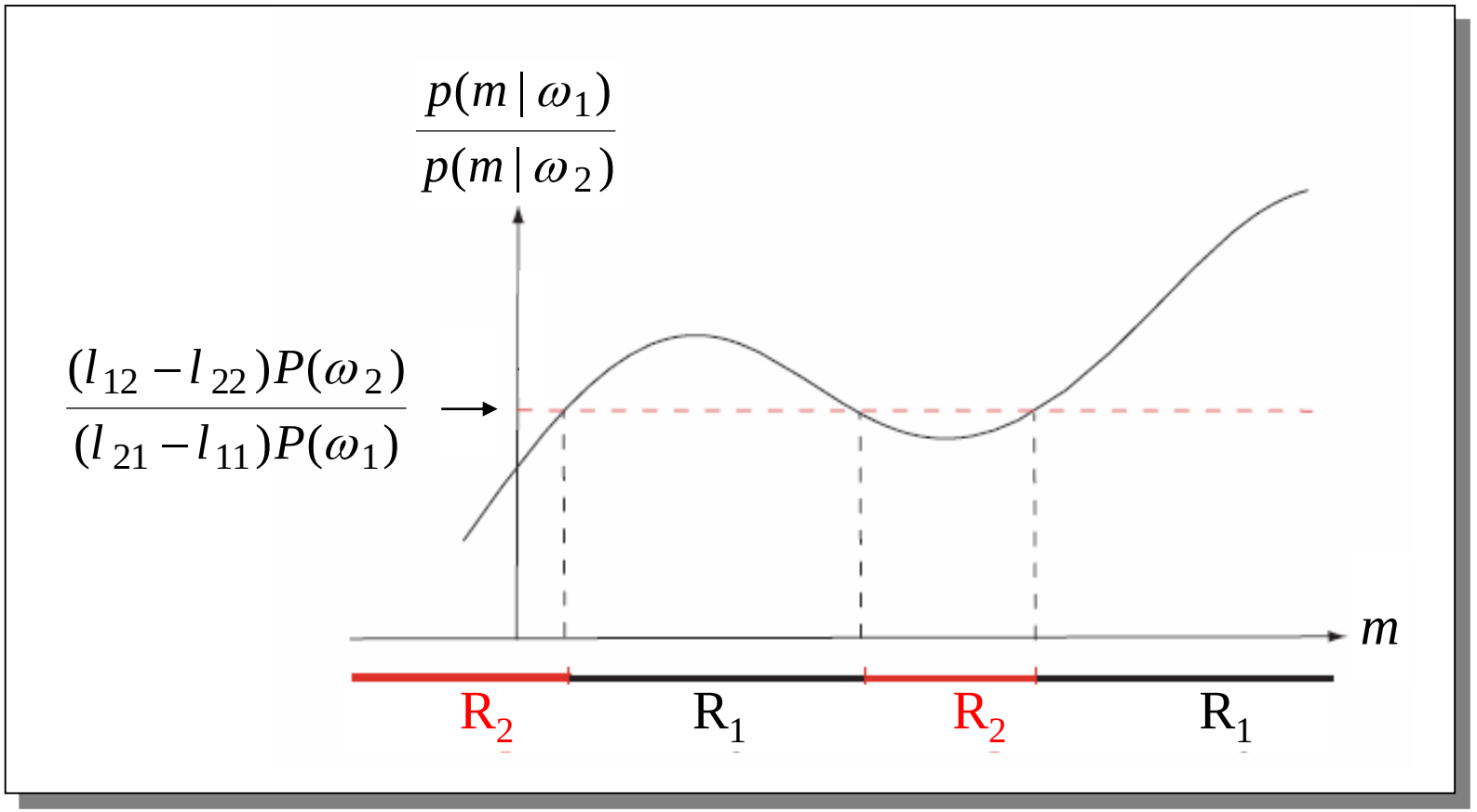
$\Leftrightarrow \quad (l_{21} - l_{11}) P(\omega_1 \mid \mathbf{m}) > (l_{12} - l_{22}) P(\omega_2 \mid \mathbf{m})$

$\Leftrightarrow \quad (l_{21} - l_{11}) p(\mathbf{m} \mid \omega_1) P(\omega_1) > (l_{12} - l_{22}) p(\mathbf{m} \mid \omega_2) P(\omega_2)$

sonst $\hat{\omega} = \omega_2$

$$\hat{\omega} = \omega_1 \quad \text{wenn} \quad \underbrace{\frac{p(\mathbf{m} \mid \omega_1)}{p(\mathbf{m} \mid \omega_2)}}_{\text{Likelihood-Verhältnis}} > \frac{(l_{12} - l_{22}) P(\omega_2)}{(l_{21} - l_{11}) P(\omega_1)} \quad \text{sonst} \quad \hat{\omega} = \omega_2$$

Beispiel für 2 Klassen, 1 Merkmal:



Quelle: R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification

Minimum Fehler Klassifikation

Kosten: $l(\omega_i, \omega_j) := 1 - \delta_i^j = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1,\ldots, c$

$$\Rightarrow \quad R(\omega_i \mid \mathbf{m}) = \sum_{j=1}^{c} l(\omega_i, \omega_j) P(\omega_j \mid \mathbf{m}) = \sum_{j \neq i} P(\omega_j \mid \mathbf{m}) = 1 - P(\omega_i \mid \mathbf{m})$$

MAP-Entscheidung: „Maximum A Posteriori Entscheidung“

$$\hat{\omega} = \omega_i \quad \text{wenn} \quad P(\omega_i \mid \mathbf{m}) > P(\omega_j \mid \mathbf{m}) \quad \forall j \neq i$$

Das Risiko ist dann gleich der Fehlerwahrscheinlichkeit.

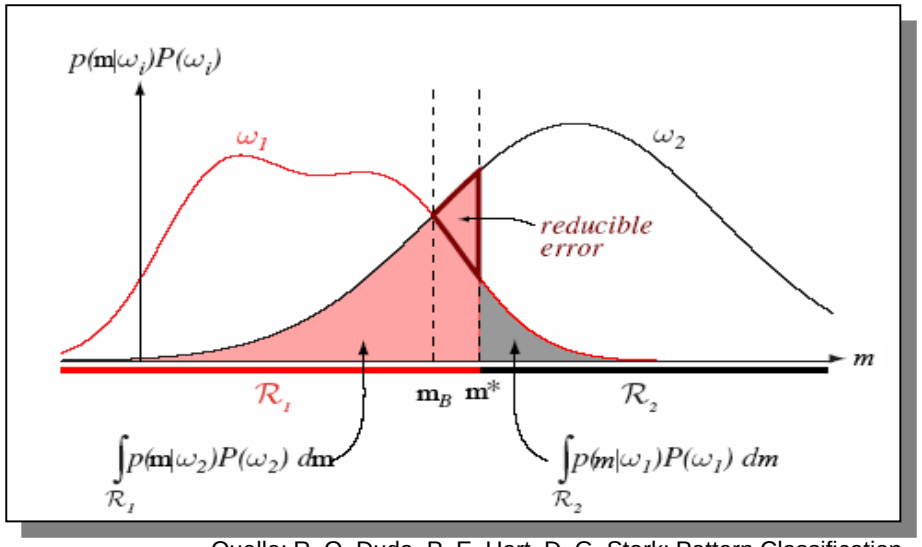
Fehlerwahrscheinlichkeit

$$P(\text{Richtig}) = \sum_{i=1}^{c} P(\mathbf{m} \in \mathrm{R}_i, \omega_i) = \sum_{i=1}^{c} P(\mathbf{m} \in \mathrm{R}_i \mid \omega_i) P(\omega_i) = \sum_{i=1}^{c} \int_{\mathrm{R}_i} p(\mathbf{m} \mid \omega_i) P(\omega_i) \mathrm{d}\mathbf{m}$$

$$P(\text{Fehler}) = 1 - P(\text{Richtig})$$

Für $c = 2$ Klassen:

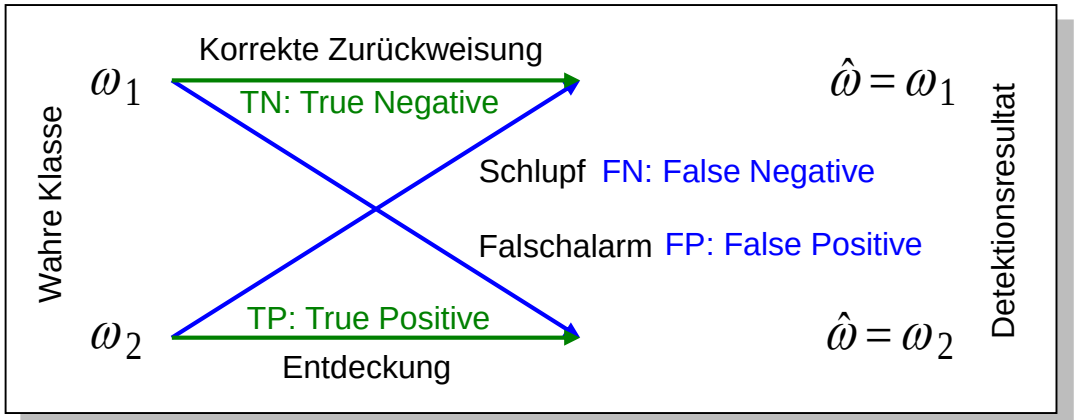
$$\begin{aligned} P(\text{Fehler}) &= P(\mathbf{m} \in \mathrm{R}_2, \omega_1) + P(\mathbf{m} \in \mathrm{R}_1, \omega_2) \\ &= P(\mathbf{m} \in \mathrm{R}_2 \mid \omega_1) P(\omega_1) + P(\mathbf{m} \in \mathrm{R}_1 \mid \omega_2) P(\omega_2) \\ &= \int_{\mathrm{R}_2} p(\mathbf{m} \mid \omega_1) P(\omega_1) \mathrm{d}\mathbf{m} + \int_{\mathrm{R}_1} p(\mathbf{m} \mid \omega_2) P(\omega_2) \mathrm{d}\mathbf{m} \end{aligned}$$



Quelle: R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification

ROC (Receiver operating characteristic)

Aufgabenstellung Detektion: $c=2$; speziell *Klasse* ω_2 *soll „entdeckt“ werden.*

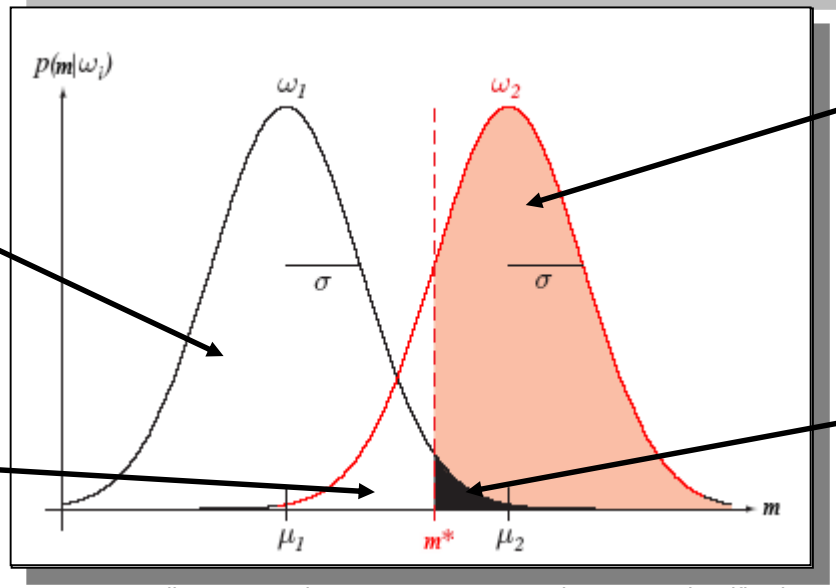


Korrekte Zurückweisung

$P(m<m^*|\omega_1)$

Schlupf

$P(m<m^*|\omega_2)$



Entdeckung

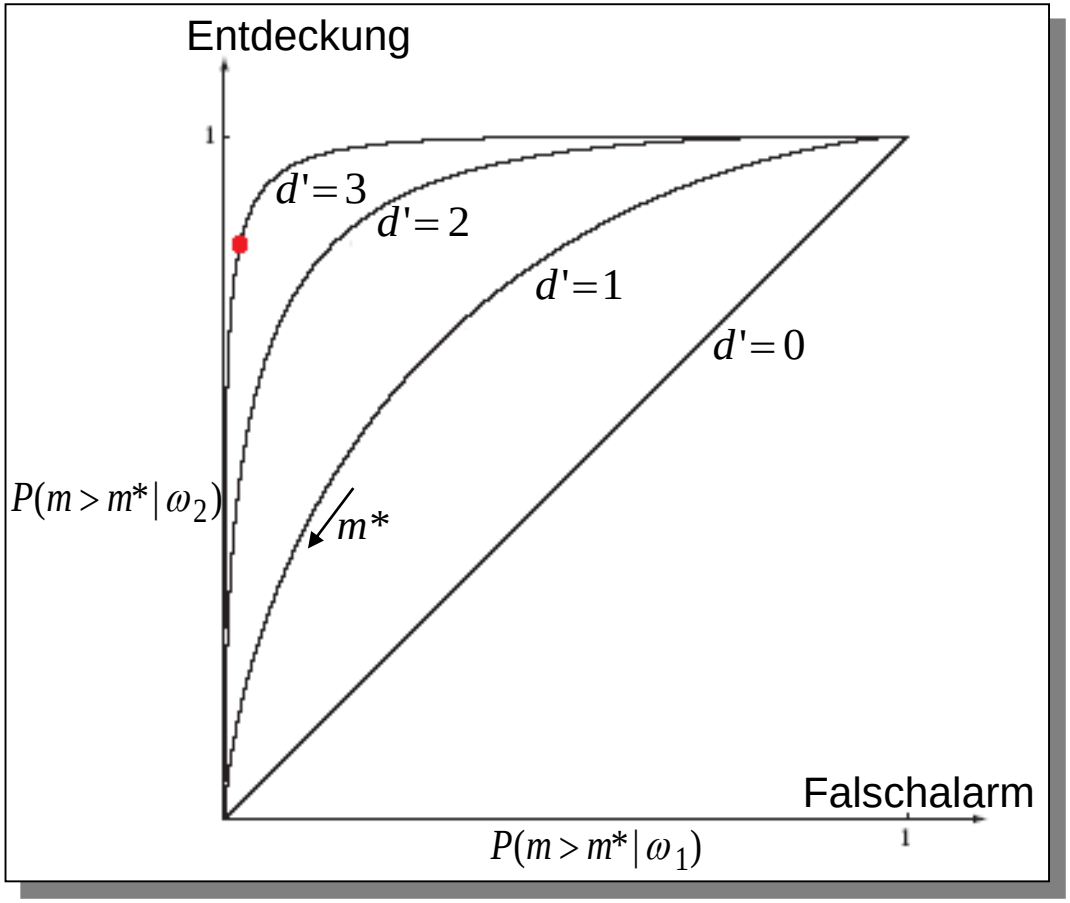
$P(m>m^*|\omega_2)$

Falschalarm

$P(m>m^*|\omega_1)$

Quelle: R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification

ROC (Receiver operating characteristic)



Quelle: R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification

Die Kurven werden mit fallendem m^* durchlaufen.

$$d' := \frac{|\mu_2 - \mu_1|}{\sigma}$$

Referenzbeispiel: Optimale Entscheidungsgrenze nach Bayes

Zahl der Klassen: $c = 2$

Klassenbedingte WDFen der Merkmale als Gauß'sche Mixtur:

$$p(\mathbf{m} \mid \omega_j) = \sum_{i=1}^{7} \frac{1}{7} \, N\left(\mathbf{m}; \boldsymbol{\mu}_{ji}, \begin{pmatrix} \sigma_j^2 & 0 \\ 0 & \sigma_j^2 \end{pmatrix}\right) \quad j = 1,2$$

$$\sigma_1^2 = \sigma_2^2 := 1$$

A priori Wahrscheinlichkeiten der Klassen: $P(\omega_j) := \frac{1}{2} \quad j = 1,2$

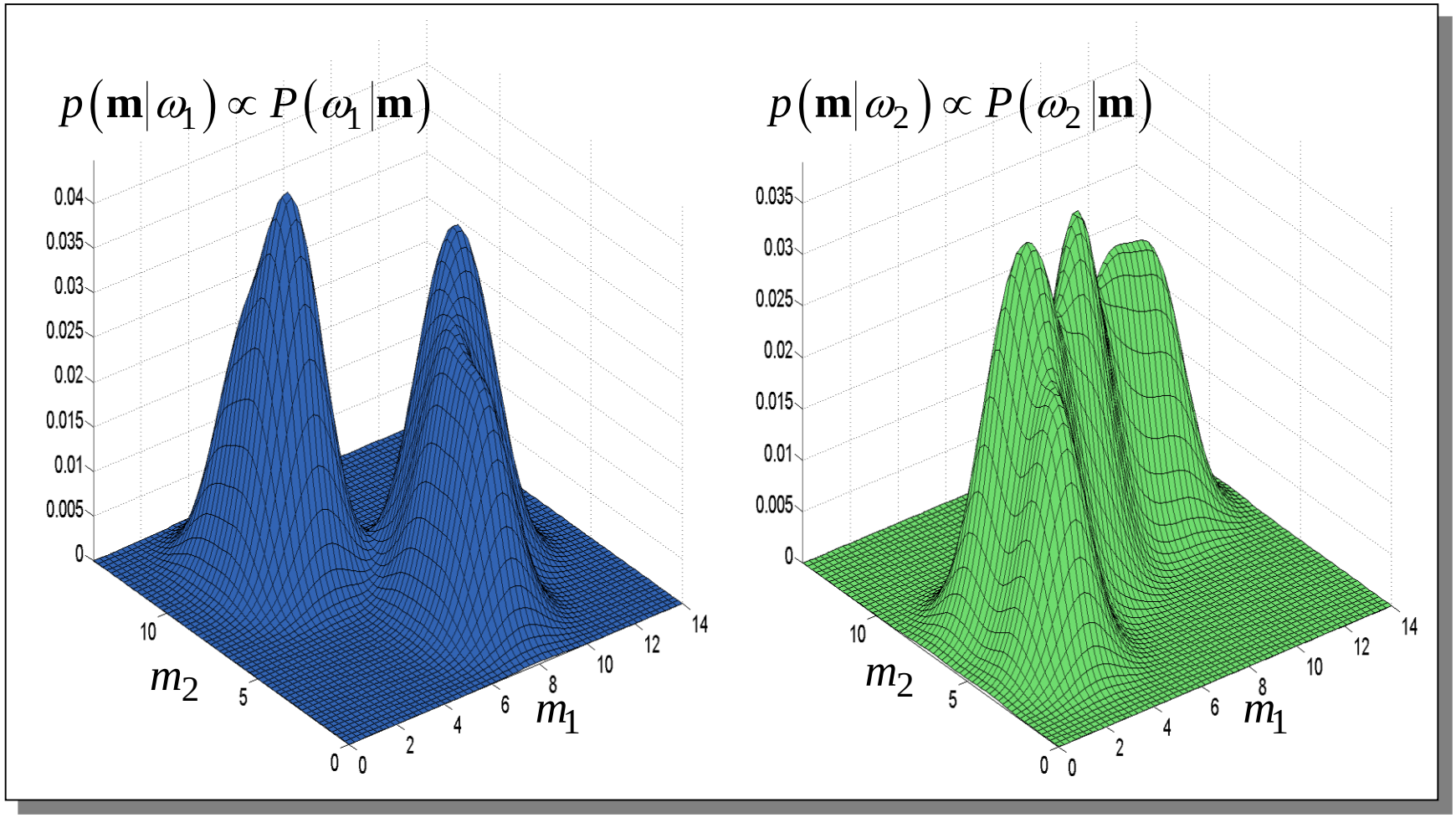
Theoretischer Klassifikationsfehler = Fehlerwahrscheinlichkeit $\approx 10{,}85\%$

Mit diesem Modell wurden Stichproben erzeugt:

Lernstichprobe: $\mathrm{D}: \; N = 200, \; N_1 = 100, \; N_2 = 100,$

Teststichprobe: $\mathrm{T}: \; N = 200, \; N_1 = 100, \; N_2 = 100,$

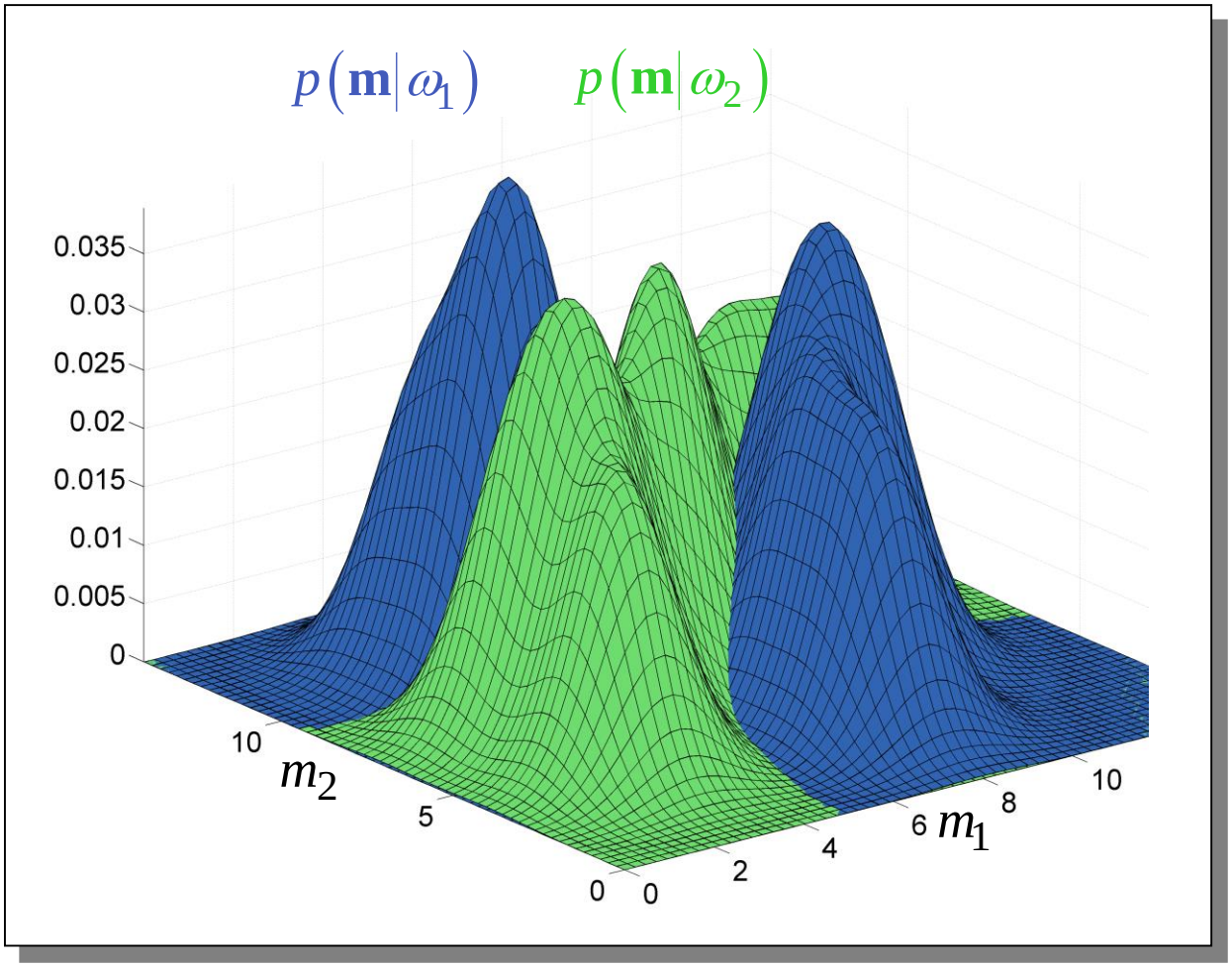
Referenzbeispiel: Optimale Entscheidungsgrenzen nach Bayes



Bedingte WDF der Merkmale im Falle der Klasse 1

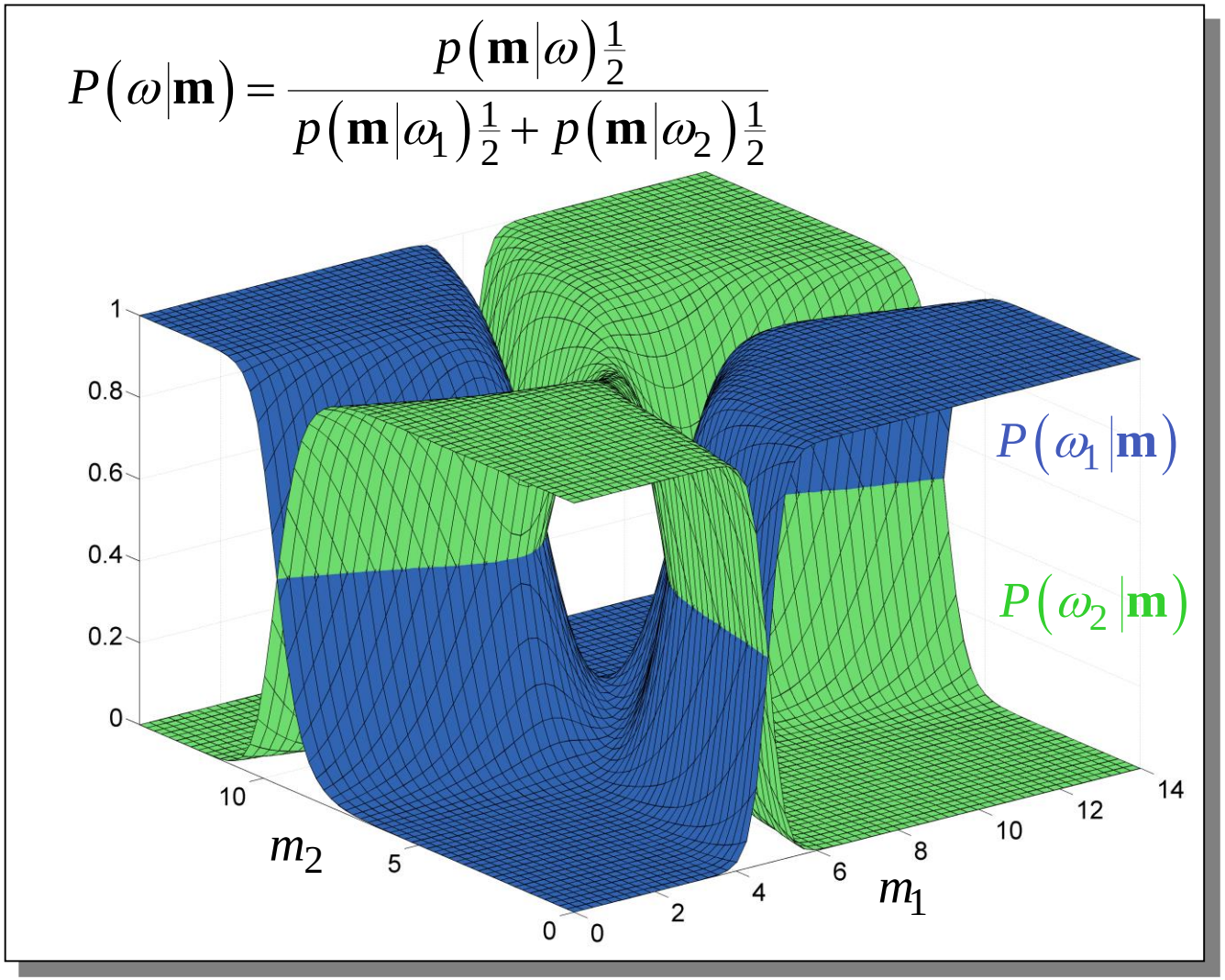
Bedingte WDF der Merkmale im Falle der Klasse 2

Referenzbeispiel: Optimale Entscheidungsgrenzen nach Bayes



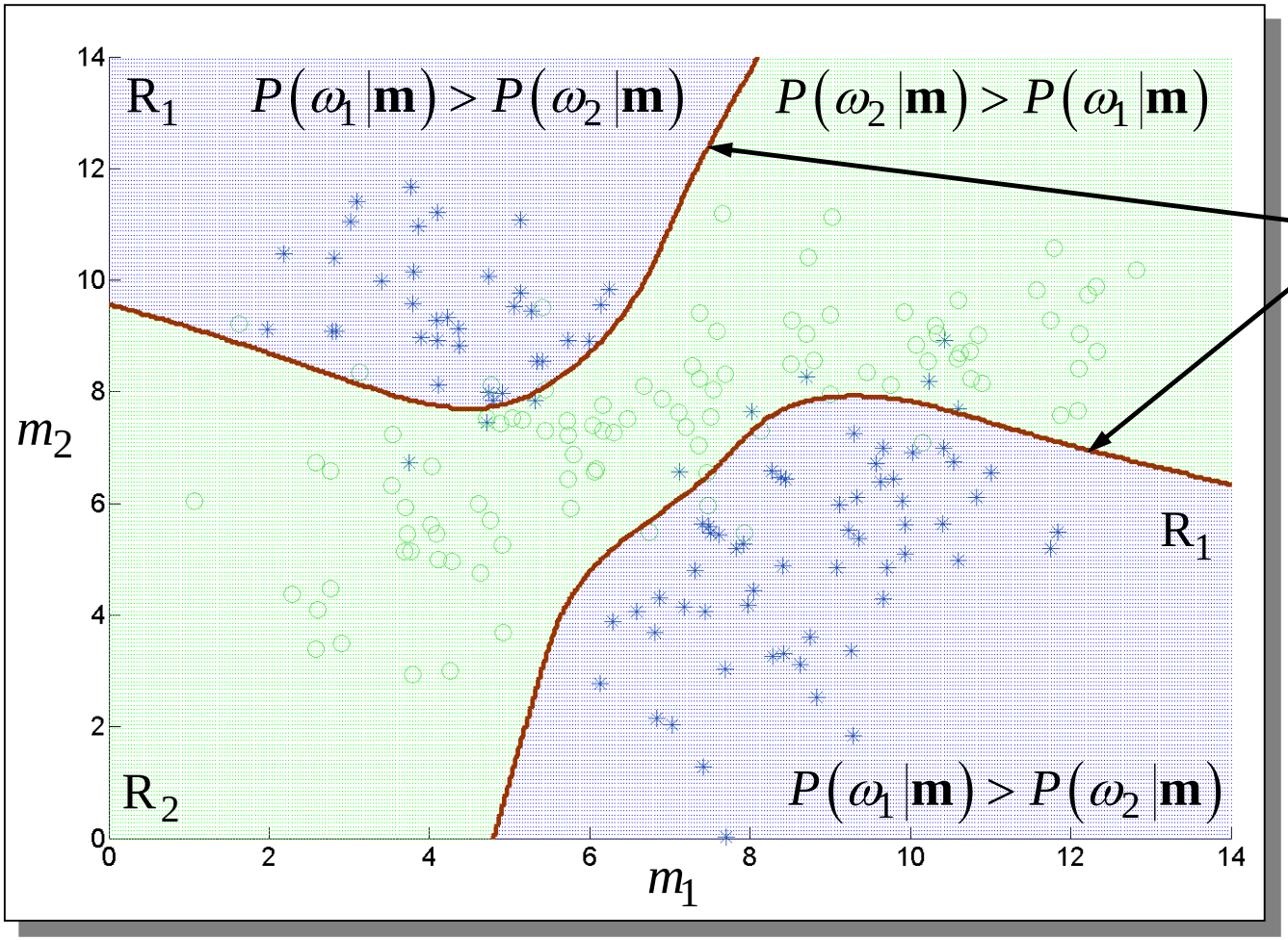
Bedingte WDFen der Merkmale für beide Klassen

Referenzbeispiel: Optimale Entscheidungsgrenzen nach Bayes



A posteriori Wahrscheinlichkeiten für beide Klassen

Referenzbeispiel: Optimale Entscheidungsgrenzen nach Bayes



Stichprobe: $N_1 = 100$, $N_2 = 100$, Empirischer Klassifikationsfehler: 9%

Minimax Entscheidungen

Ziel: Klassifikation, die *robust* ist bezüglich einer Variation der A Priori WV $P(\omega)$ der Klassen.

Beispiel für 2 Klassen:

Risiko:

$$R = \int R(\hat{\omega}(\mathbf{m}) \mid \mathbf{m}) p(\mathbf{m}) \mathrm{d}\,\mathbf{m} = \int_{\mathrm{R}_1} [l_{11} P(\omega_1) p(\mathbf{m} \mid \omega_1) + l_{12} P(\omega_2) p(\mathbf{m} \mid \omega_2)] \mathrm{d}\,\mathbf{m} + \int_{\mathrm{R}_2} [l_{21} P(\omega_1) p(\mathbf{m} \mid \omega_1) + l_{22} P(\omega_2) p(\mathbf{m} \mid \omega_2)] \mathrm{d}\,\mathbf{m}$$

Mit $P(\omega_1) = 1 - P(\omega_2)$ und $\int_{\mathrm{R}_i} p(\mathbf{m} \mid \omega_i) \mathrm{d}\,\mathbf{m} = 1 - \int_{\mathrm{R}_j} p(\mathbf{m} \mid \omega_i) \mathrm{d}\,\mathbf{m}, \forall i \neq j$ folgt: $i, j = 1,2$

Minimax Entscheidungen, Beispiel für 2 Klassen

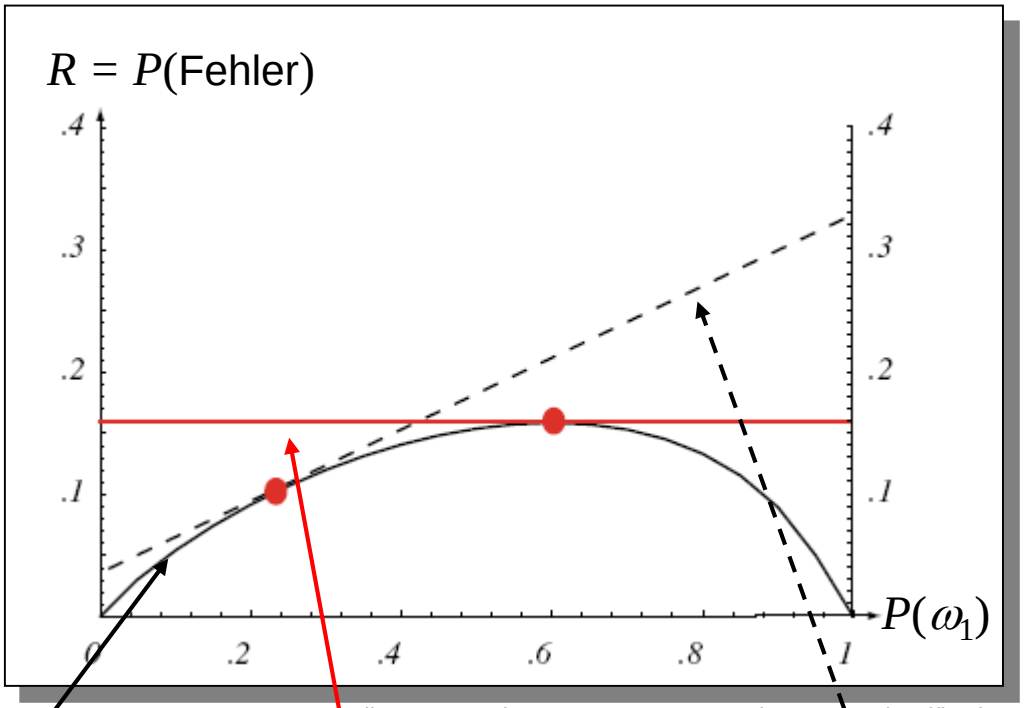
$$R = \overbrace{l_{22} + (l_{12} - l_{22}) \int_{R_1} p(\mathbf{m} \mid \omega_2) \mathrm{d}\mathbf{m} +}^{R_{\text{Minimax}}}$$

$$+ P(\omega_1) \underbrace{\left[\left(l_{11} - l_{22}\right) + \left(l_{21} - l_{11}\right) \int_{R_2} p(\mathbf{m} \mid \omega_1) \mathrm{d}\mathbf{m} - \left(l_{12} - l_{22}\right) \int_{R_1} p(\mathbf{m} \mid \omega_2) \mathrm{d}\mathbf{m} \right]}_{\overset{!}{=} 0}$$

Beim Design des Minimax-Klassifikators werden die Entscheidungsgebiete R_i so festgelegt, dass die Abhängigkeit von der A Priori WV $P(\omega)$ der Klassen verschwindet.

$$R_{\text{Minimax}} = l_{22} + (l_{12} - l_{22}) \int_{R_1} p(\mathbf{m} \mid \omega_2) \mathrm{d}\mathbf{m} = l_{11} + \left(l_{21} - l_{11}\right) \int_{R_2} p(\mathbf{m} \mid \omega_1) \mathrm{d}\mathbf{m}$$

Minimax Entscheidungen, Beispiel für 2 Klassen, $l(\omega_i, \omega_j) = 1 - \delta_i^j$



Quelle: R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification

Bayes'sche Klassifikation

Bayes'scher Klassifikator $P_{\text{Entwurf}}(\omega_1) = 0{,}25$

Minimax Klassifikator $\triangleq$ *worst case* Bayes-Klassifikator

Minimax Entscheidungen

Diskussion:

Der Minimax-Klassifikator entspricht dem Bayes'schen Klassifikator für die *ungünstigste* A Priori WV der Klassen, d.h. für die A Priori WV mit dem maximalen Bayes'schen Risiko.

Der Minimax-Klassifikator minimiert das maximale Risiko.

Minimax-Entscheidungen sind vor allem in der Spieltheorie von Bedeutung, wo ein intelligenter Gegner versucht, maximal zu schaden. Bei Mustererkennungsaufgaben hingegen ist i.d.R. die Umwelt der „Gegner“.

Normalverteilung eindimensional

Eine kontinuierliche Zufallsvariable m heißt (univariat) normalverteilt mit Erwartungswert μ und Varianz σ^2, wenn für ihre Wahrscheinlichkeitsdichtefunktion (WDF) gilt:

$$m \sim \mathrm{N}(m;\mu,\sigma^2) \qquad p(m) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{m-\mu}{\sigma}\right)^2\right]$$

$$\mu := \mathrm{E}\{m\} = \int_{-\infty}^{\infty} m\,p(m)\,\mathrm{d}\,m \qquad \sigma^2 := \mathrm{E}\left\{(m-\mu)^2\right\} = \int_{-\infty}^{\infty} (m-\mu)^2\,p(m)\,\mathrm{d}\,m$$

Zur Bedeutung der Normalverteilung (Gauß'sche Verteilung):
Zentraler Grenzwertsatz: Unter sehr allgemeinen Voraussetzungen für eine Folge stochastisch unabhängiger Zufallsvariablen ist die Summe dieser Folge asymptotisch normalverteilt;
Details findet man z.B. im Satz von Lindeberg-Feller, „Wahrscheinlichkeitsrechnung und mathematische Statistik", Fisz, VEB-Verlag, Berlin 1989.

Normalverteilung mehrdimensionnal

Eine kontinuierliche Zufallsvariable $\mathbf{m} \in \mathbb{R}^d$ heißt (multivariat) normalverteilt mit Erwartungswert $\boldsymbol{\mu}$ und Kovarianzmatrix $\boldsymbol{\Sigma}$, wenn für ihre WDF gilt:

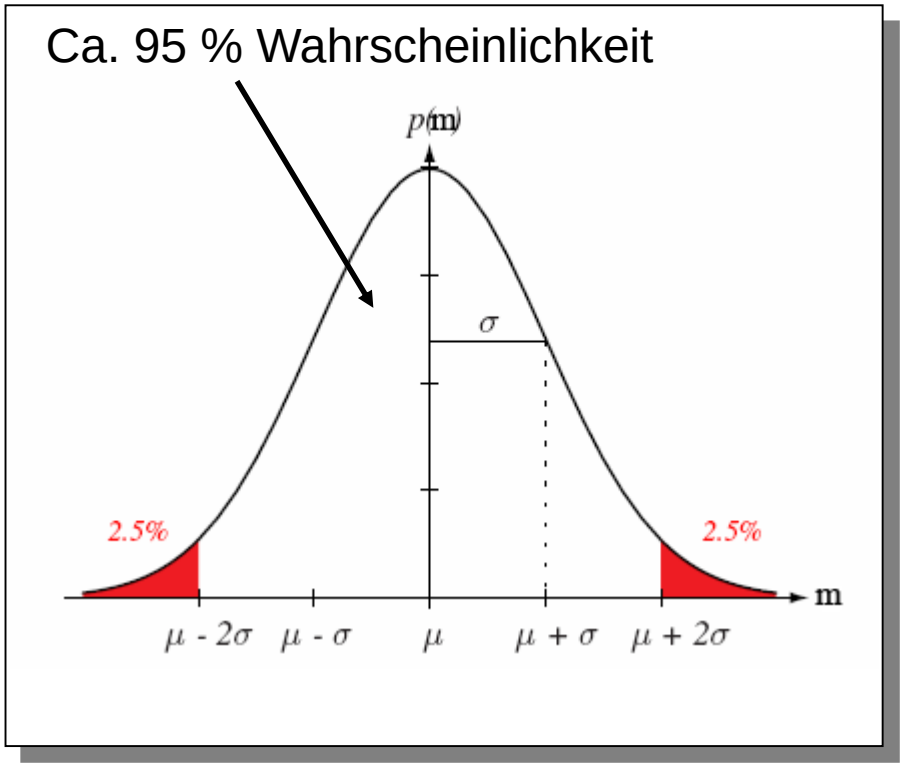
$$\mathbf{m} \sim \mathrm{N}(\mathbf{m}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$p(\mathbf{m}) = \frac{1}{(2\pi)^{d/2} \, |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{m}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{m}-\boldsymbol{\mu}) \right]$$

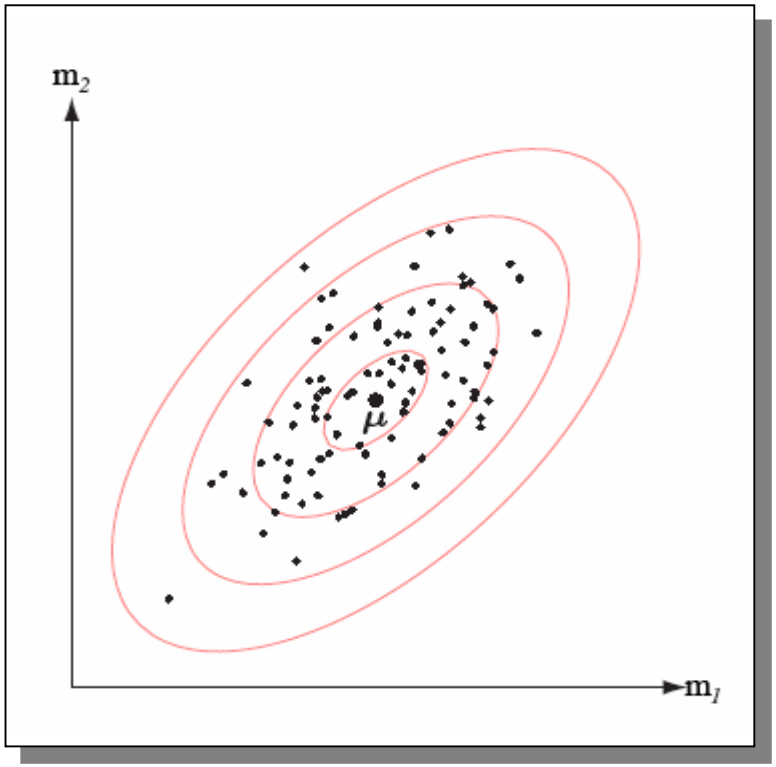
$$\boldsymbol{\mu} := \mathrm{E}\{\mathbf{m}\} = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \mathbf{m} p(\mathbf{m}) \, \mathrm{d}\mathbf{m}$$

$$\boldsymbol{\Sigma} := \mathrm{E}\left\{ (\mathbf{m}-\boldsymbol{\mu})(\mathbf{m}-\boldsymbol{\mu})^{\mathrm{T}} \right\} = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} (\mathbf{m}-\boldsymbol{\mu})(\mathbf{m}-\boldsymbol{\mu})^{\mathrm{T}} \, p(\mathbf{m}) \, \mathrm{d}\mathbf{m}$$

Normalverteilung - Beispiele



Univariate Normalverteilung



Quelle: R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification

Multivariate Normalverteilung

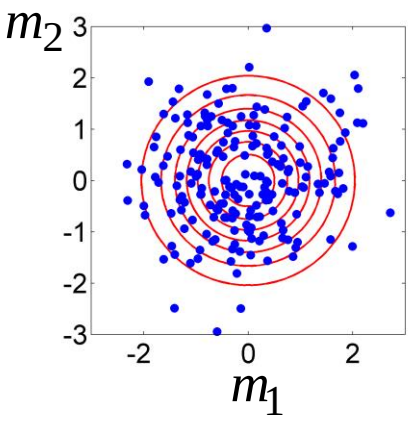
Ellipsen: Punkte konstanter Wahrscheinlichkeitssdichte.

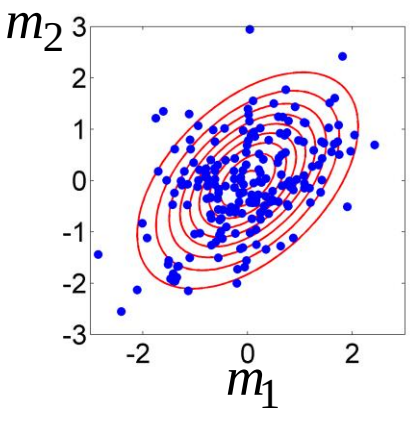
3.3. Normalverteilte Merkmale

Bsp.: $\mathbf{m} \sim \mathrm{N}(\mathbf{m}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathrm{N}\left(\mathbf{m}; \boldsymbol{\mu}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right) = \mathrm{N}\left(\mathbf{m}; \mathbf{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$

$\rho \in [-1, 1]$: Korrelationskoeffizient

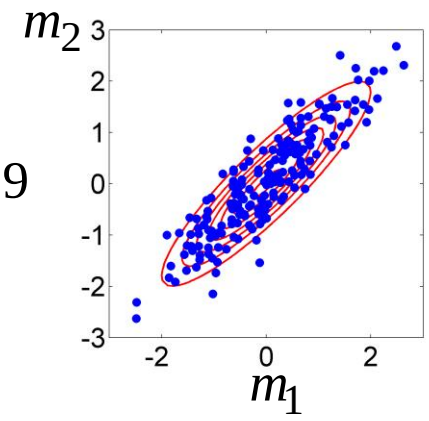
$\rho = 0$

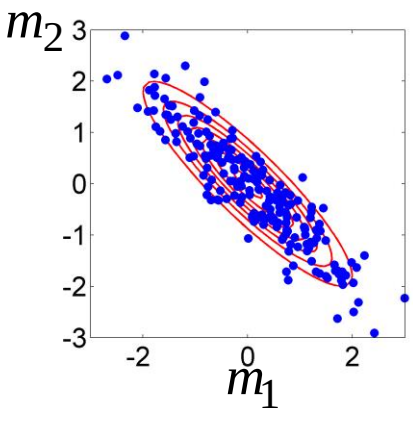




$\rho = 0{,}5$

$\rho = 0{,}9$





$\rho = -0{,}9$

MAP-Klassifikation bei normalverteilten Merkmalen:

$$\hat{\omega} = \omega_i \text{ wenn } P(\omega_i|\mathbf{m}) > P(\omega_j|\mathbf{m}) \quad \forall j \neq i \qquad P(\omega|\mathbf{m}) = \frac{p(\mathbf{m}|\omega)P(\omega)}{p(\mathbf{m})}$$

Normalverteilungsannahme für die klassenbedingte WV der Merkmale:

$$p(\mathbf{m}|\omega_i) = \mathrm{N}(\mathbf{m};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) \quad i = 1,\ldots,c$$

Normalverteilungsannahme macht nur bei mindestens *intervall-skalierten* Merkmalen Sinn.

MAP-Klassifikation bei normalverteilten Merkmalen:

$$\hat{\omega} = \omega_i \text{ wenn } P(\omega_i|\mathbf{m}) > P(\omega_j|\mathbf{m}) \quad \forall j \neq i$$

Wegen der strengen Monotonie der Logarithmusfunktion gilt:

$$P(\omega_i|\mathbf{m}) > P(\omega_j|\mathbf{m}) \quad \Leftrightarrow \quad \ln P(\omega_i|\mathbf{m}) > \ln P(\omega_j|\mathbf{m})$$

Bei normalverteilten Merkmalen rechnet es sich leichter mit den Logarithmen:

$$k_i(\mathbf{m}) := \ln p(\mathbf{m}|\omega_i) + \ln P(\omega_i)$$

$$\downarrow \quad p(\mathbf{m}|\omega_i) = \mathrm{N}(\mathbf{m};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$$

$$k_i(\mathbf{m}) = -\frac{1}{2}(\mathbf{m}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\mathbf{m}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

3.3. Normalverteilte Merkmale

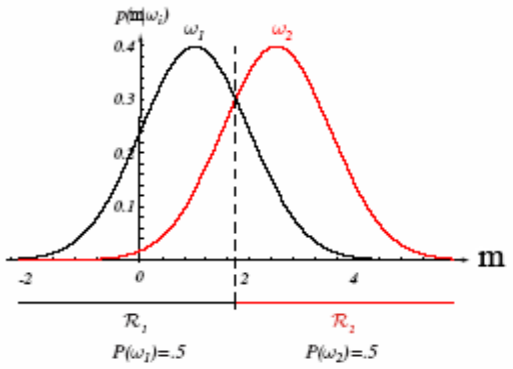
Beispiel: $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}$

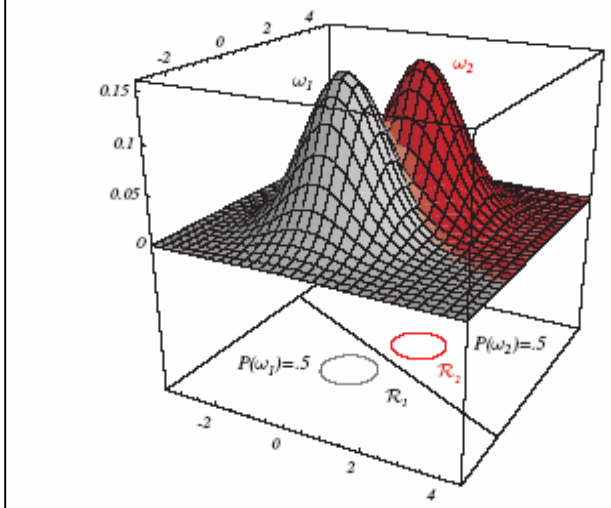
$$k_i(\mathbf{m}) = -\frac{1}{2}(\mathbf{m}-\boldsymbol{\mu}_i)^{\mathrm{T}} \underbrace{\boldsymbol{\Sigma}_i^{-1}}_{(1/\sigma^2)\mathbf{I}} (\mathbf{m}-\boldsymbol{\mu}_i) - \overbrace{\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln \underbrace{|\boldsymbol{\Sigma}_i|}_{\sigma^{2d}} + \ln P(\omega_i)}^{\text{unabhängig von } i}$$

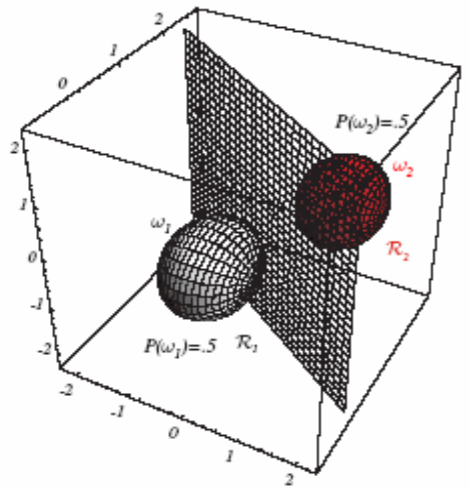
Entscheidungsrelevant: $k_i(\mathbf{m}) := -\frac{1}{2\sigma^2} \| \mathbf{m}-\boldsymbol{\mu}_i \|^2 + \ln P(\omega_i)$

Bsp.: $c = 2$,

$P(\omega_1) = P(\omega_2)$







Quelle: R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification

Verteilungen sind *sphärisch* in d-Dimensionen, Grenzen sind d–1-dimensionale Ebenen (Hyperebenen) senkrecht zu den Verbindungsgeraden der Erwartungswerte.

3.3. Normalverteilte Merkmale

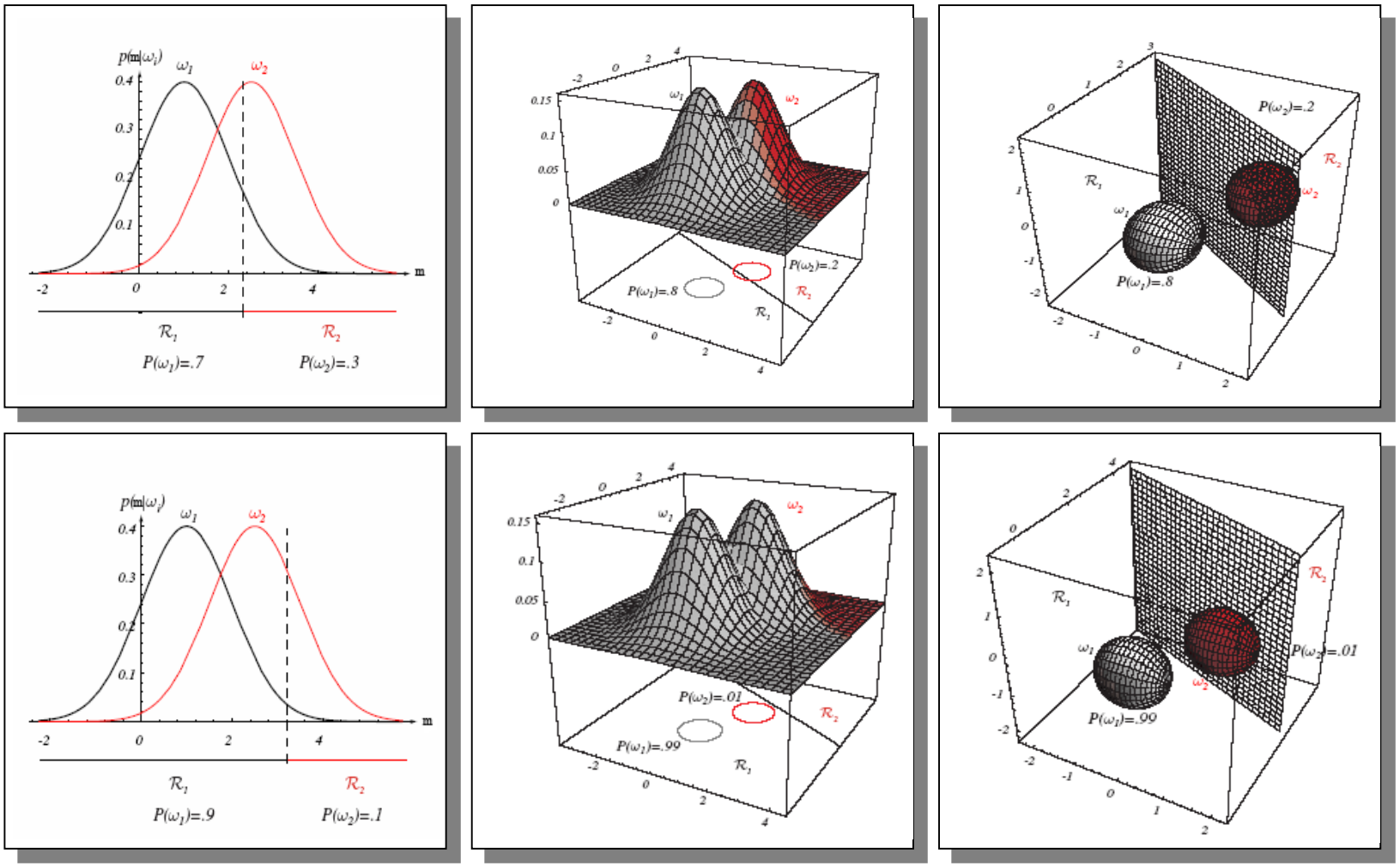
Entscheidungsgebietsgrenze zwischen R_i und R_j :

$$k_i(\mathbf{m}) - k_j(\mathbf{m}) = \underbrace{\frac{1}{\sigma^2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \mathbf{m} - \frac{1}{2\sigma^2}\left(\|\boldsymbol{\mu}_i\|^2 - \|\boldsymbol{\mu}_j\|^2\right) + \ln\frac{P(\omega_i)}{P(\omega_j)} = 0}_{\text{Hyperebene im Merkmalsraum}}$$

3.3. Normalverteilte Merkmale

Beispiel: $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}$

Für $P(\omega_1) \neq P(\omega_2)$: Grenze nicht mehr mittig zwischen den Erwartungswerten.



Quelle: R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification

3.3. Normalverteilte Merkmale

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

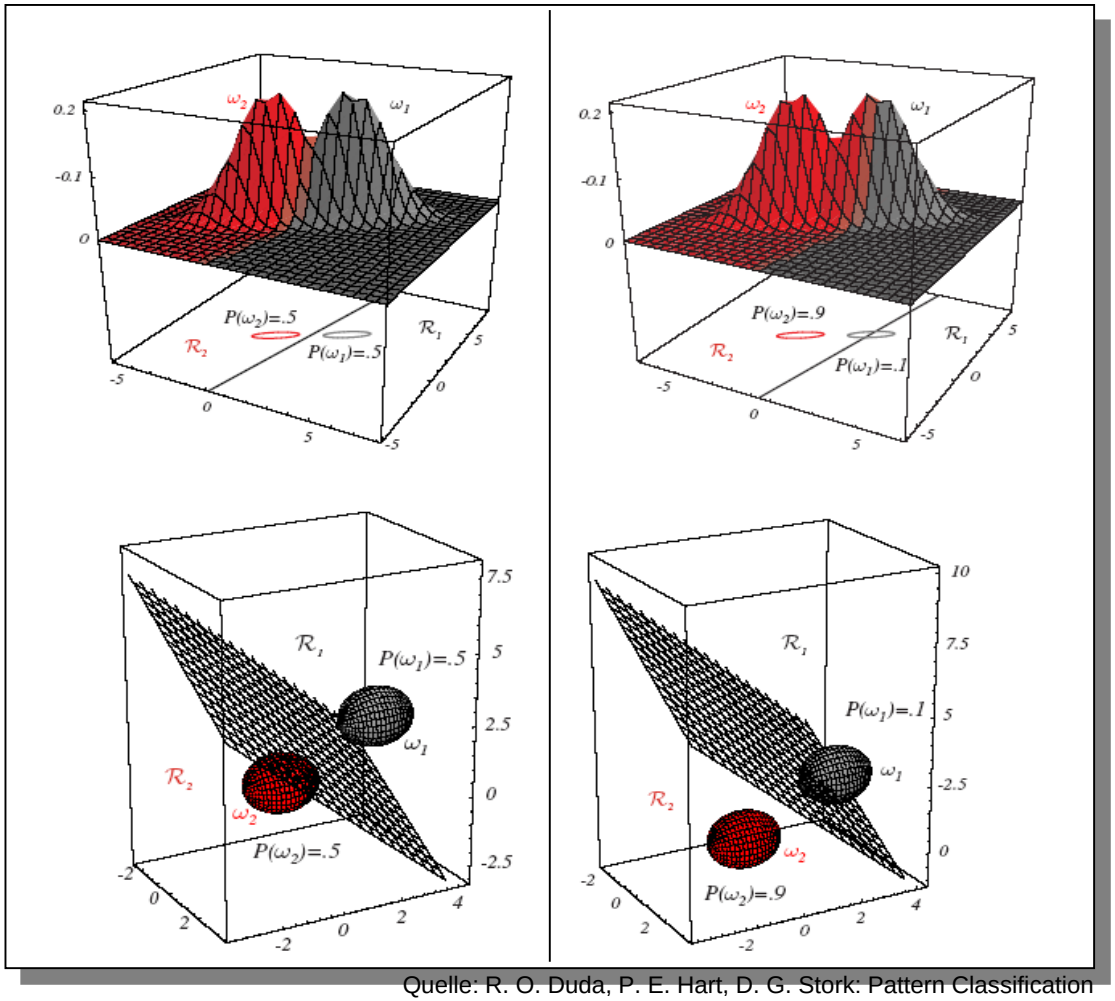
$$k_i(\mathbf{m}) = -\frac{1}{2}(\mathbf{m}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{m}-\boldsymbol{\mu}_i) - \underbrace{\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}|}_{\text{unabhängig von } i} + \ln P(\omega_i)$$

Entscheidungsrelevant:

$$k_i(\mathbf{m}) := -\frac{1}{2}(\mathbf{m}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{m}-\boldsymbol{\mu}_i) + \ln P(\omega_i)$$

3.3. Normalverteilte Merkmale

Beispiel: $\Sigma_i = \Sigma$ $\quad P(\omega_1) = P(\omega_2)$ $\quad P(\omega_1) \neq P(\omega_2)$



Quelle: R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification

WV sind „*elliptisch*“. Grenzen sind d–1 dim. Ebenen (Hyperebenen); i.Allg. nicht senkrecht zu den Verbindungsgeraden der Erwartungswerte.

3.3. Normalverteilte Merkmale

Allgemeiner Fall:

$$k_i(\mathbf{m}) = -\frac{1}{2}(\mathbf{m}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\mathbf{m}-\boldsymbol{\mu}_i) - \underbrace{\frac{d}{2}\ln 2\pi}_{\text{unabhängig von } i} - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

Entscheidungsrelevant:

$$k_i(\mathbf{m}) := \mathbf{m}^{\mathrm{T}}\underbrace{\left(-\tfrac{1}{2}\right)\boldsymbol{\Sigma}_i^{-1}}_{\mathbf{W}_i}\mathbf{m} + \tfrac{1}{2}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{m} + \tfrac{1}{2}\mathbf{m}^{\mathrm{T}}\underbrace{\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i}_{\mathbf{w}_i} \underbrace{- \tfrac{1}{2}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \tfrac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)}_{w_{i0}}$$

$$k_i(\mathbf{m}) = \mathbf{m}^{\mathrm{T}}\mathbf{W}_i\mathbf{m} + \mathbf{w}_i^{\mathrm{T}}\mathbf{m} + w_{i0}$$

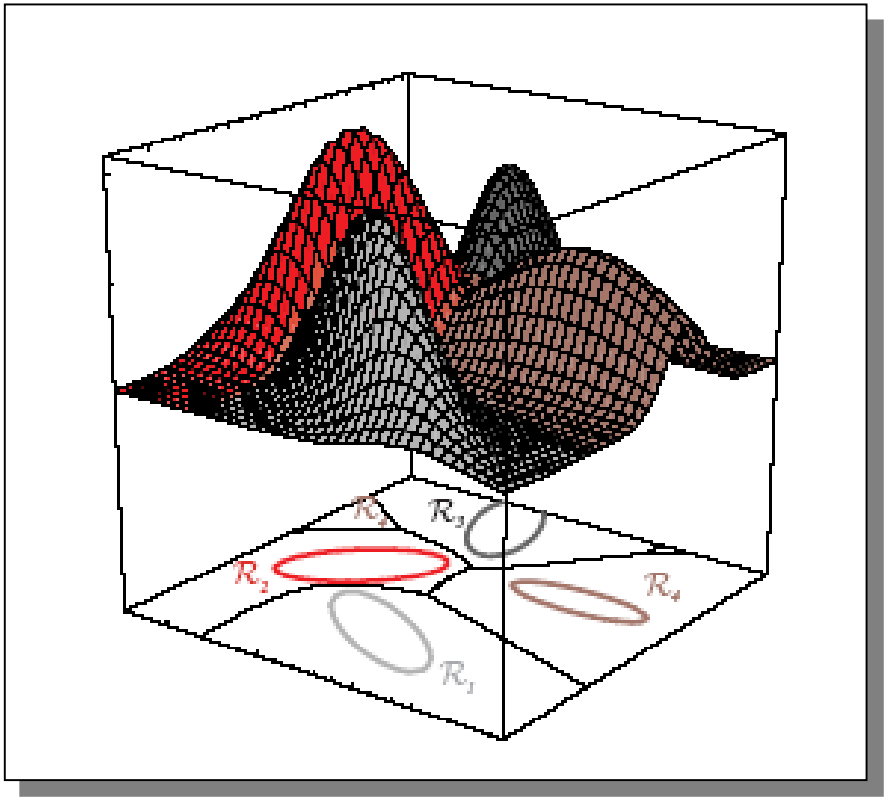
Entscheidungsgebietsgrenze zwischen R_i und R_j :

$$k_i(\mathbf{m}) - k_j(\mathbf{m}) = \mathbf{m}^{\mathrm{T}}(\mathbf{W}_i - \mathbf{W}_j)\mathbf{m} + (\mathbf{w}_i - \mathbf{w}_j)^{\mathrm{T}}\mathbf{m} + w_{i0} - w_{j0} = 0$$

Die Entscheidungsgebietsgrenzen im Merkmalsraum sind Hyperquadriken.

3.3. Normalverteilte Merkmale

Bsp: $c = 4, d = 2$

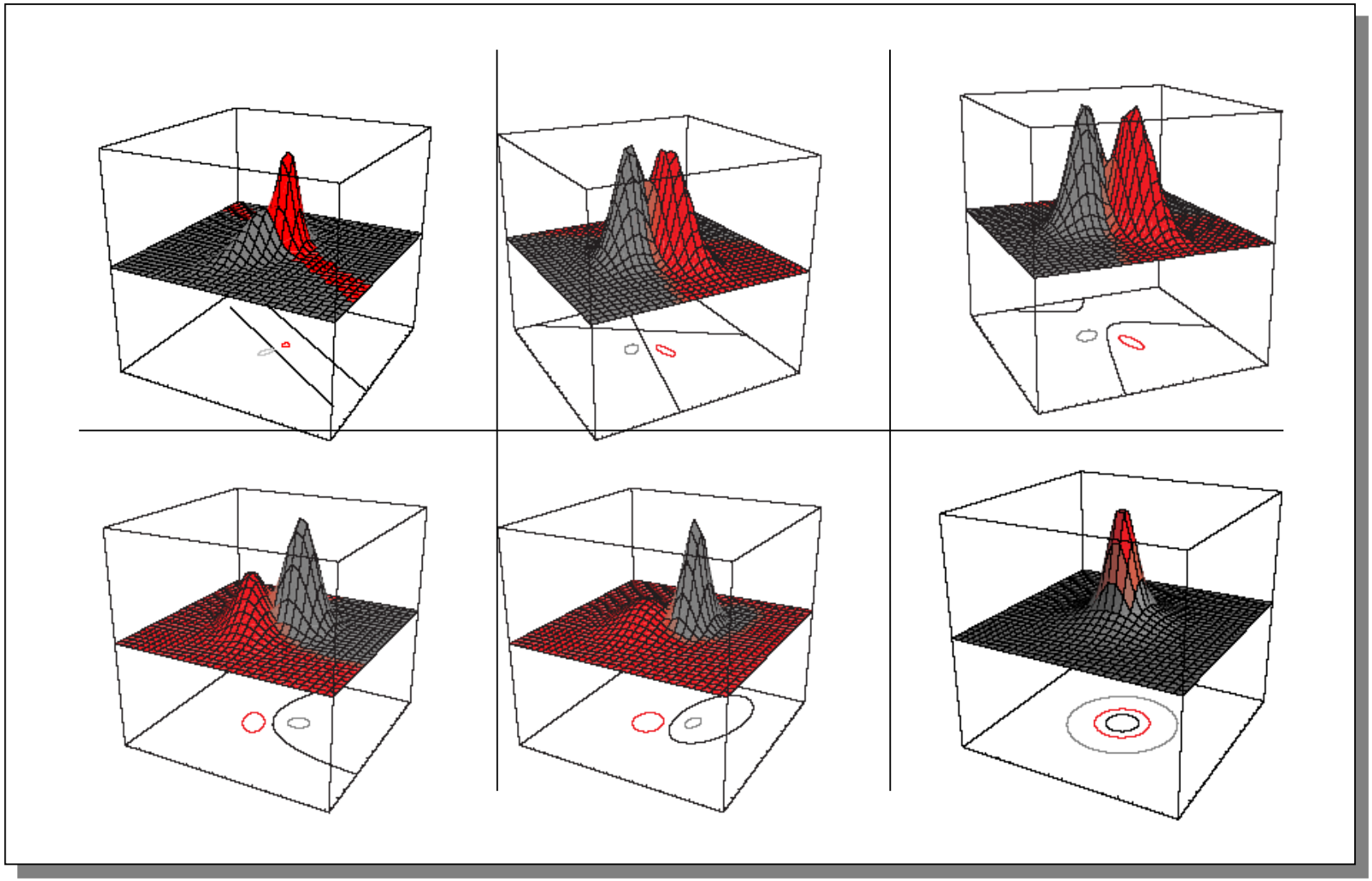


Quelle: R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification

$$\partial \mathrm{R}_i = \bigcup_{j \neq i} \left\{ \mathbf{m} \middle| k_i(\mathbf{m}) = k_j(\mathbf{m}) \wedge k_j(\mathbf{m}) \geq k_l(\mathbf{m}) \forall l \neq i, j \right\}$$

3.3. Normalverteilte Merkmale

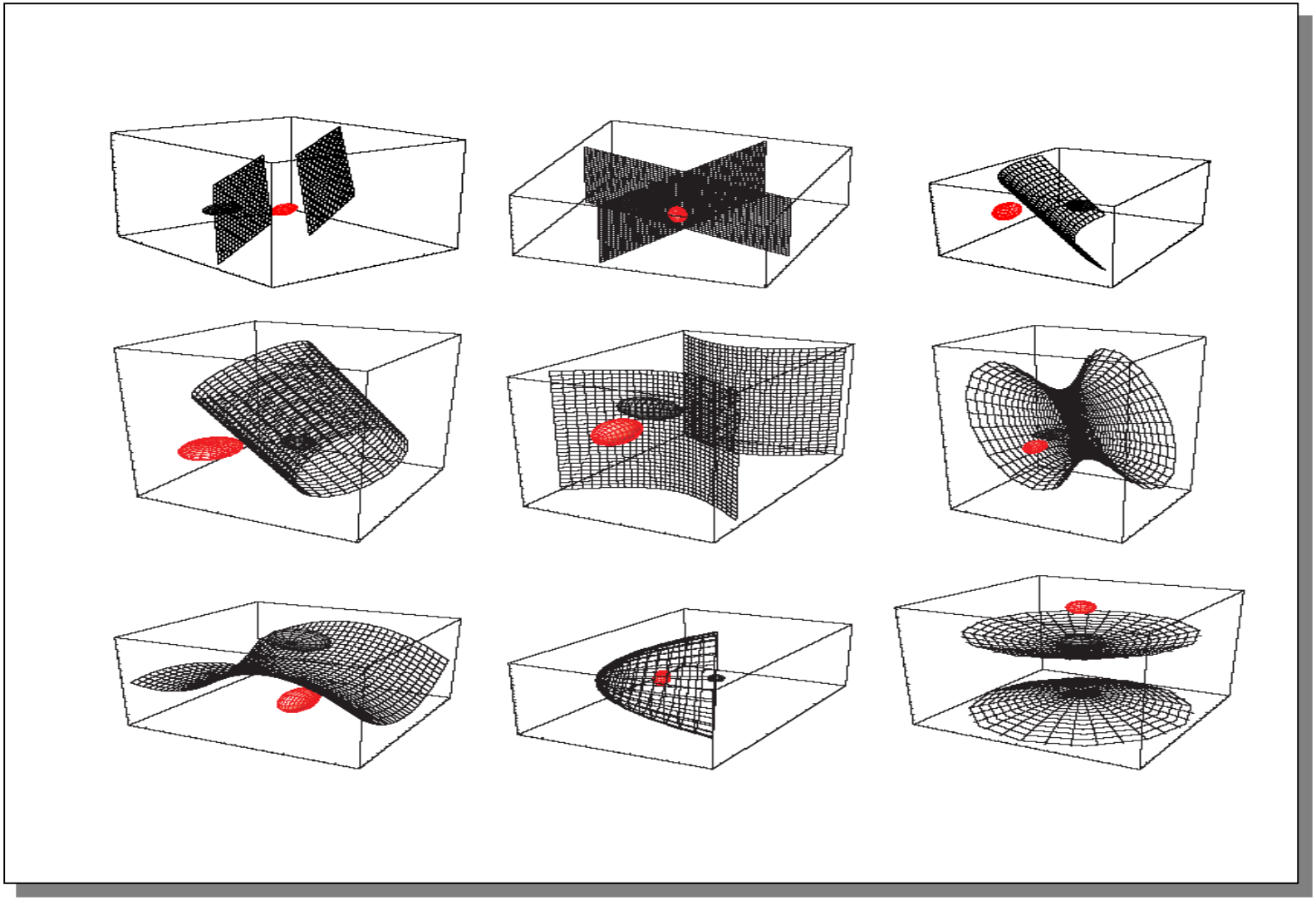
Beispiele: $c = 2$, $d = 2$



Quelle: R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification

3.3. Normalverteilte Merkmale

Beispiele: $c = 2$, $d = 3$

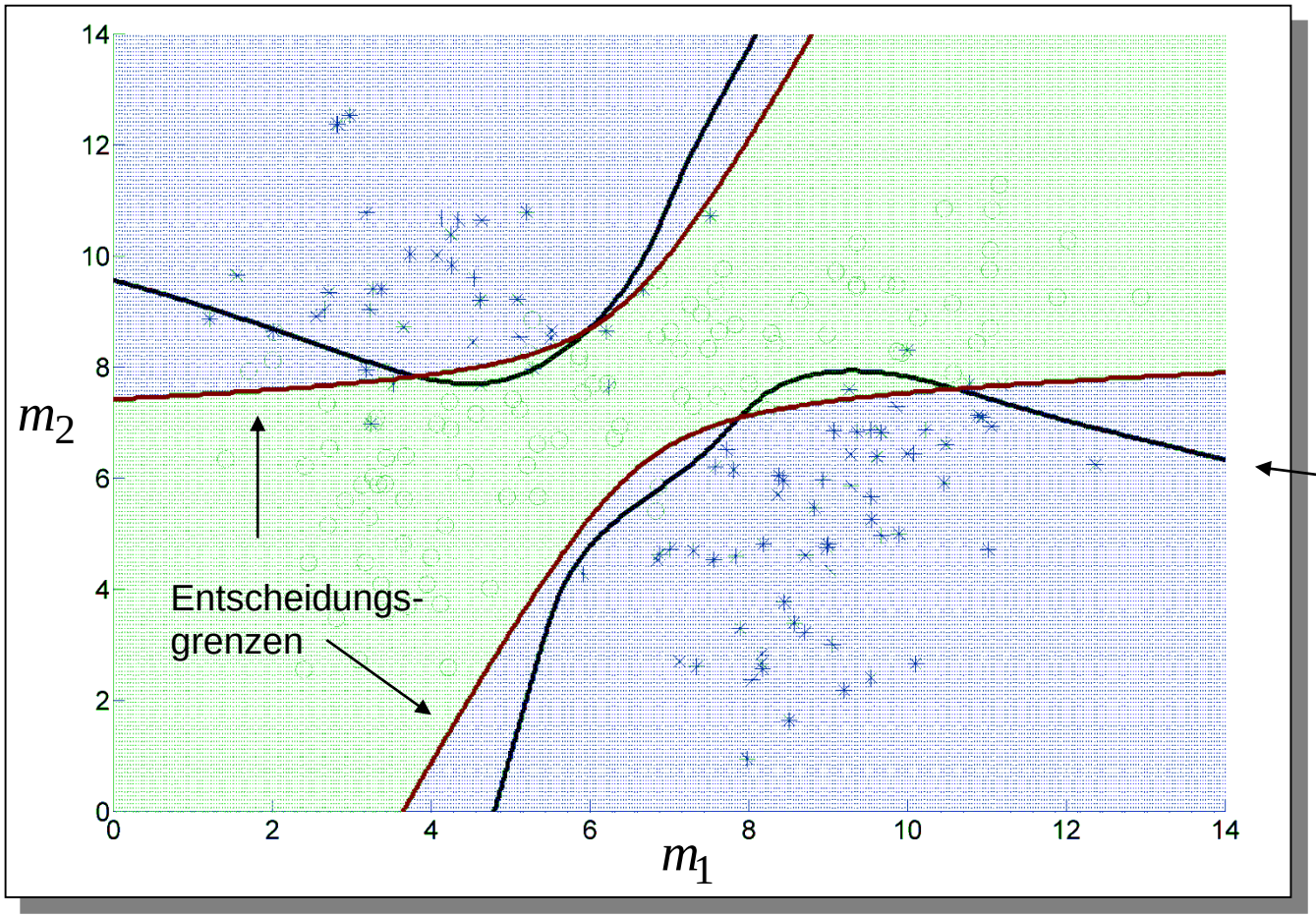


Quelle: R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification

3.3. Normalverteilte Merkmale

Beispiel: Klassifikator für das Referenzbeispiel auf der Basis einer Gaußdichte für jede Klasse:

$$p(\mathbf{m} \mid \omega_j) = \mathrm{N}(\mathbf{m}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad j = 1,2$$

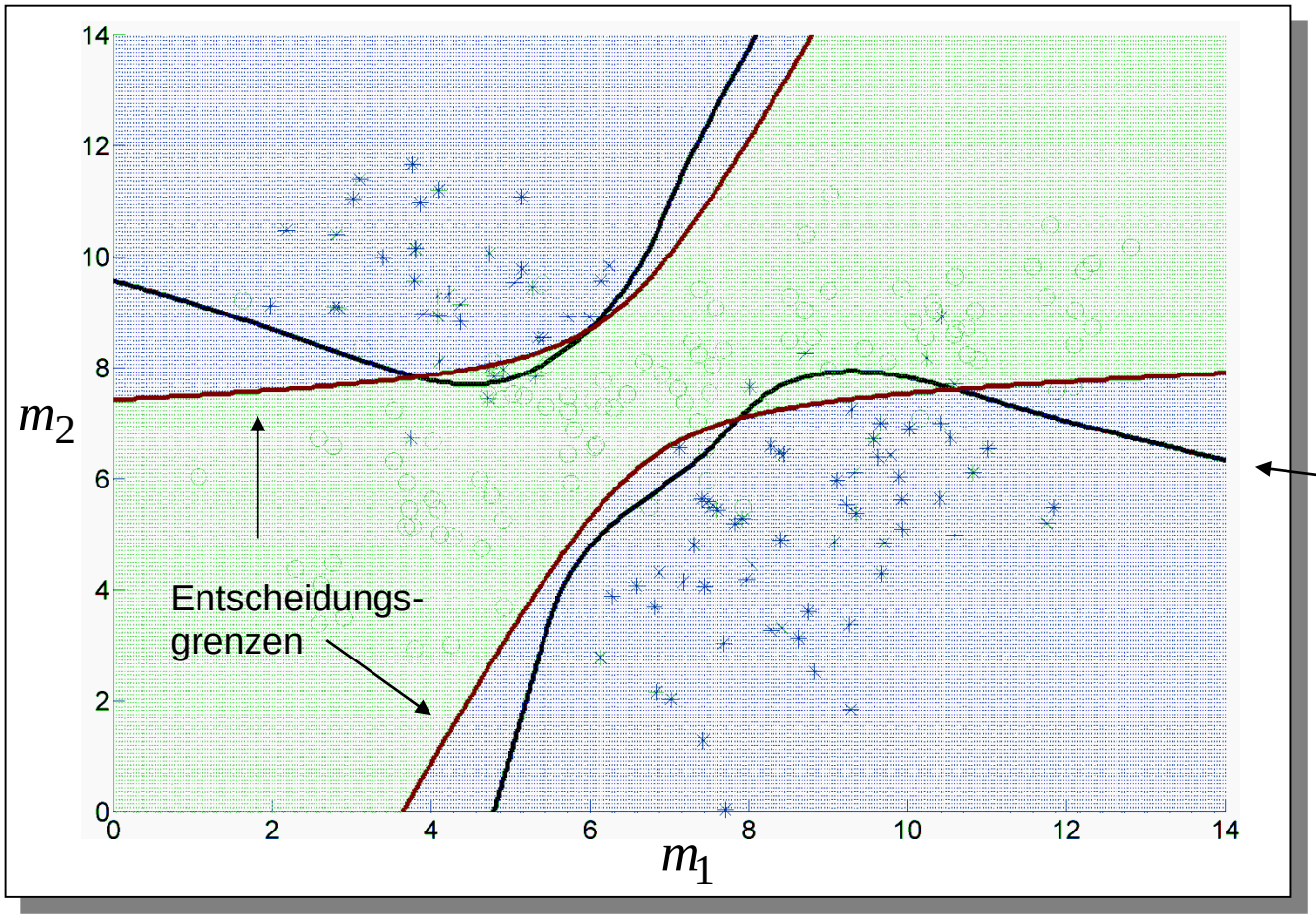


Parameter aus der Lernstichprobe geschätzt. → Trainingsfehler = 9,5%

3.3. Normalverteilte Merkmale

Beispiel: Klassifikator für das Referenzbeispiel auf der Basis einer Gaußdichte für jede Klasse:

$$p(\mathbf{m} \mid \omega_j) = \mathrm{N}(\mathbf{m}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad j = 1,2$$



Parameter aus Lernstichprobe geschätzt.

Teststichprobe → Testfehler = 11% Asymptotischer Testfehler ≈ 12,37%

3.4. Merkmale mit beliebiger Verteilung

Beliebige Verteilungen können mittels Gauß‘schen Mixturen approximiert werden. Definition einer Gauß‘schen Mixtur:

$$\mathbf{m} \sim p(\mathbf{m}) = \sum_{k=1}^{K} \pi_k \, \mathrm{N}\left(\mathbf{m}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \qquad \pi_k \geq 0, \; \sum_{k=1}^{K} \pi_k = 1$$

Gauß‘sche Mixturen sind universelle Approximatoren. Mit entsprechend vielen Komponenten lassen sich beliebige WDFen beliebig genau nähern.

Jede klassenbedingte WDF der Merkmale $p(\mathbf{m}|\omega_j)$, $j = 1,...,c$ kann mittels einer klassenindividuellen Gauß‘schen Mixtur approximiert werden. Komponentenanzahlen K_j für jede Klasse individuell definierbar.

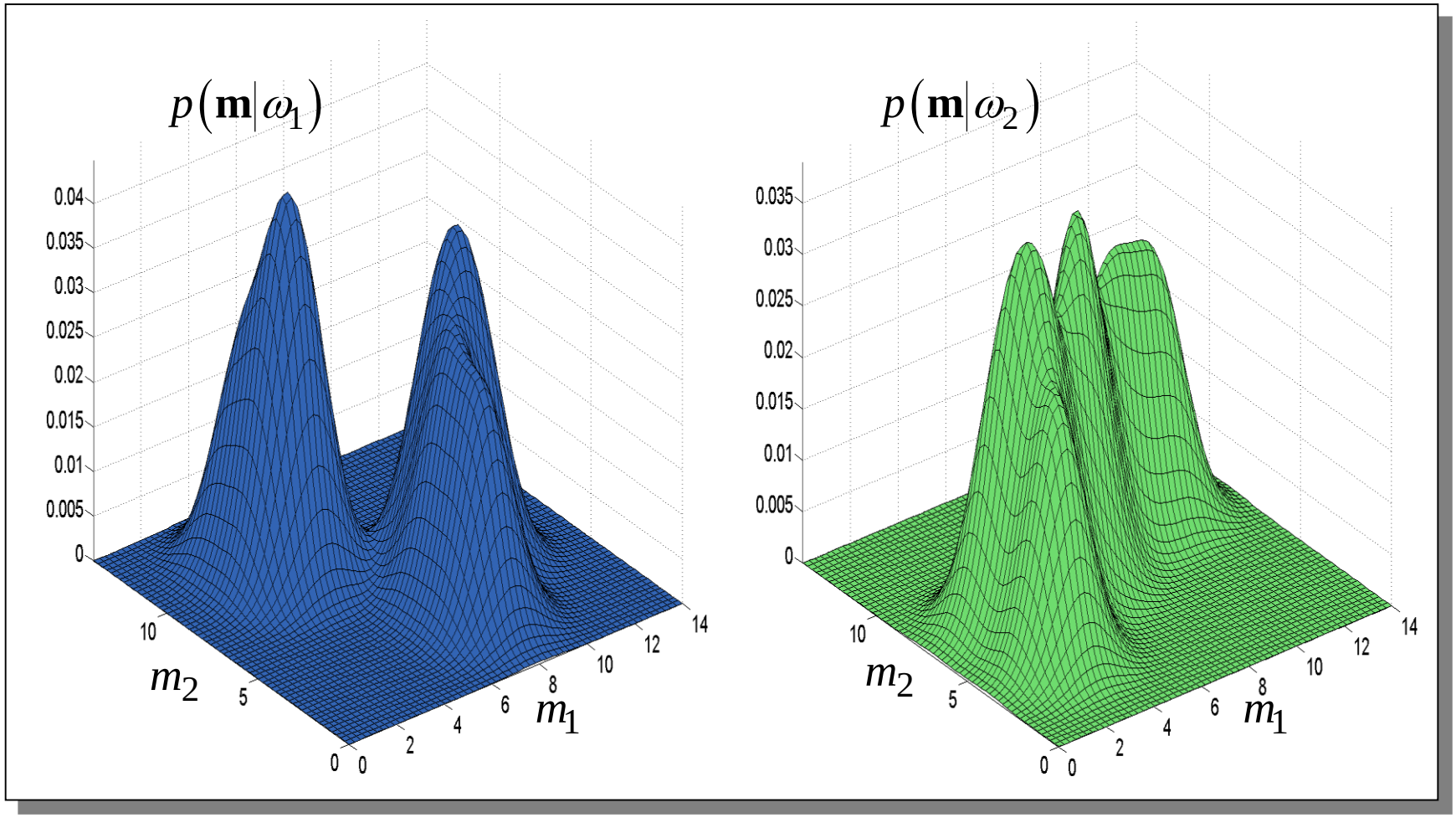
$$\mathbf{m}|\omega_j \sim p(\mathbf{m}|\omega_j) = \sum_{k=1}^{K_j} \pi_{jk} \, \mathrm{N}\left(\mathbf{m}; \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}\right) \qquad \pi_{jk} \geq 0, \; \sum_{k=1}^{K_j} \pi_{jk} = 1$$

Betrachtung für $K_j = K$: Pro Klasse sind auf Basis der Lerndaten $^1/_2 K(d^2+3d+2)-1$ Parameter zu schätzen.

$$\underbrace{\pi_{j1},...,\pi_{jK}}_{K-1}, \underbrace{\boldsymbol{\mu}_{j1},...,\boldsymbol{\mu}_{jK}}_{Kd}, \underbrace{\boldsymbol{\Sigma}_{j1},...,\boldsymbol{\Sigma}_{jK}}_{\frac{1}{2}Kd(d+1)} \qquad j = 1,...,c$$

3.4. Merkmale mit beliebiger Verteilung

Bsp.: Referenzbeispiel benutzt Gauß‘sche Mixturen mit jeweils $K = 7$.



$$p(\mathbf{m}|\omega_j) = \sum_{k=1}^{7} \frac{1}{7} \, N\left(\mathbf{m}; \boldsymbol{\mu}_{jk}, \mathbf{I}\right) \quad j = 1,2$$

4.3. Bayes'sche Klassifikation – ergänzende Bemerkungen

Fehler bei der Bayes'schen Klassifikation

- Bayes'scher Fehler: ergibt sich durch die Überlappung der der klassenbedingten Verteilungsdichten der Merkmale.

- Modellfehler: ergibt sich durch ein nicht passendes Modell.
  Auswahl eines passenden Modells mit Hilfe eines Anpassungstests für WVen, z.B. Chi-Quadrat-Test; Details in statistischer Grundlagenliteratur.

- Schätzfehler: ergibt sich aufgrund endlicher Datensätze.