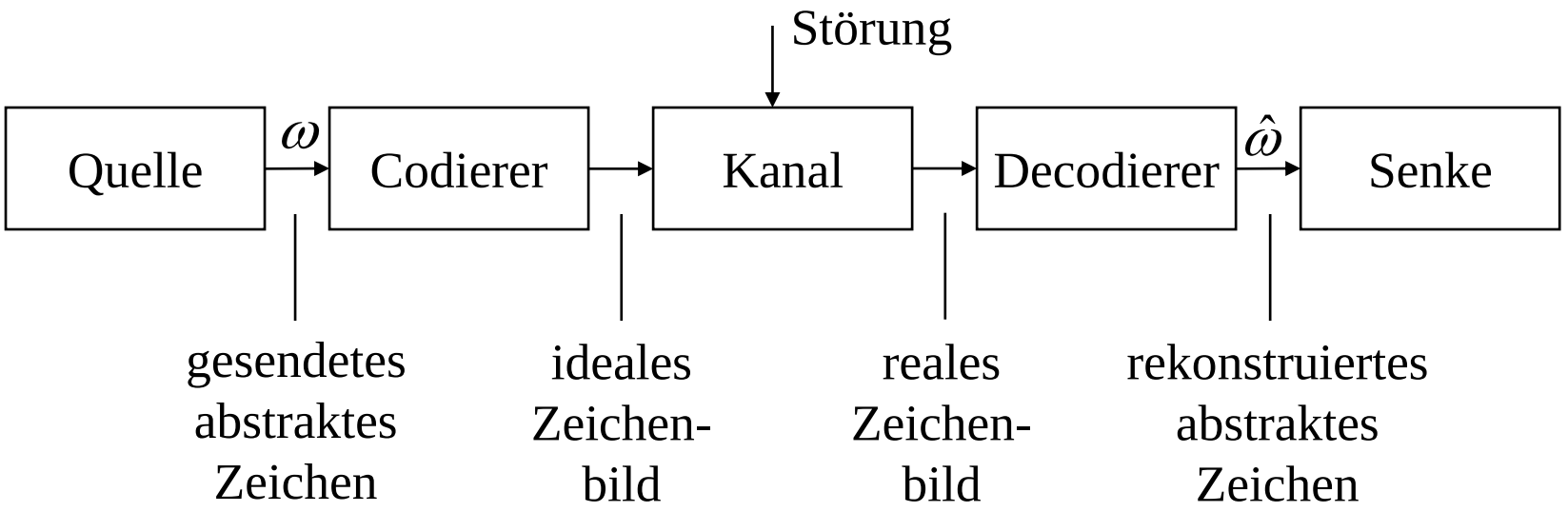
Kapitel 7

Bayes- oder Optimalklassifikator

Der Entwurf eines optimalen Klassifikators

- Letztes Glied in der Mustererkennungskette
- Der Klassifikator hat die Aufgabe einer optimalen Zuordnung eines Merkmalsvektors zu einer Bedeutungsklasse
- Grundlage für den Entwurf: Statistische Entscheidungstheorie
- Beschreibung des Erkennungssystems in Analogie zu einem Nachrichtenübertragungssystem:



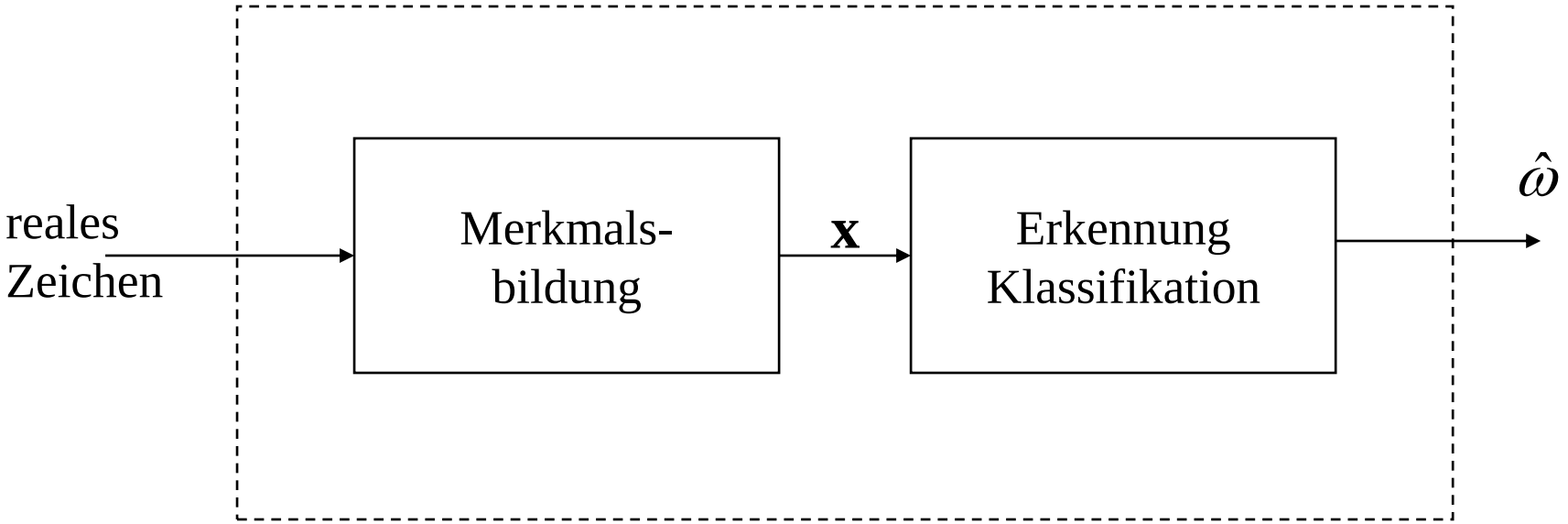
Über die einzelnen Komponenten

Codierer: aus einem abstrakten Quellzeichen *K* entsteht ein Schriftzeichen: z.Bsp. OCR-B

Störung: beinhaltet alle Veränderungen wie z.B. eigentlicher Druck- oder Schreibvorgang, mögliche Verschmutzungen, Fehler des Scanners usw.

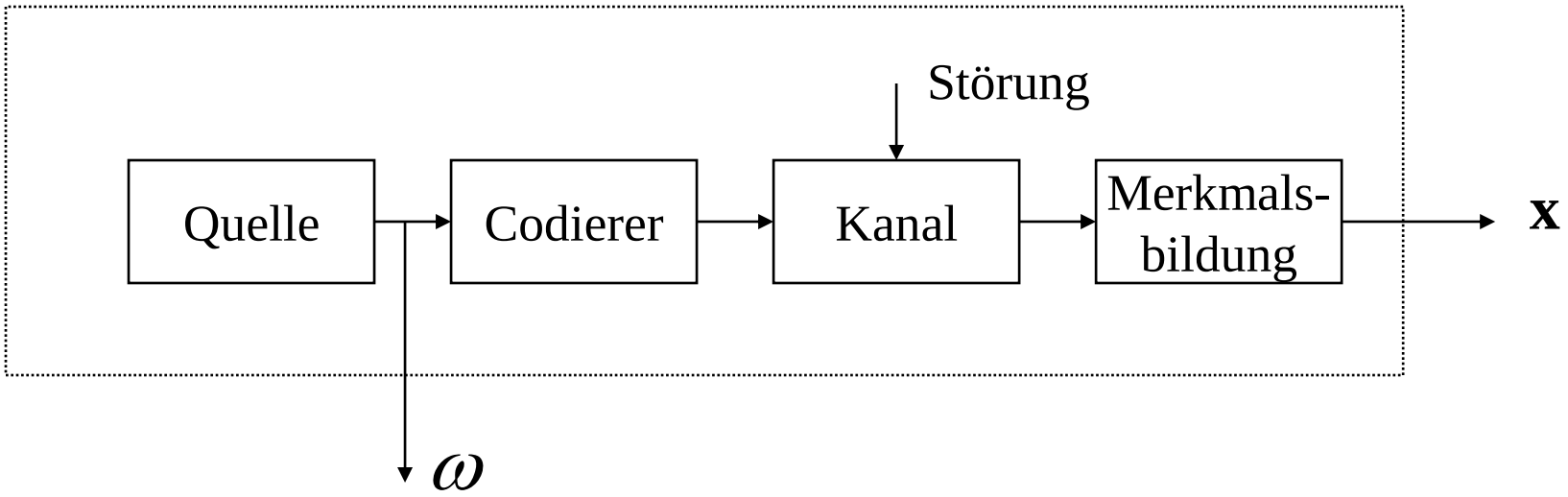
Decodierer: Lesemaschine, Entscheider, rekonstruiert das gesendete Zeichen, übergibt Entscheidung an Senke

Aufbau des Decodierers



Stochastisches Modell

Der zeichenerzeugende Prozeß erzeugt jeweils miteinander verbundene Wertepaare (ω,**x**)



Seine statistischen Eigenschaften werden vollständig durch die **Verbundverteilung** beschrieben:

$$p(\omega, \mathbf{x}) = p(\mathbf{x}, \omega) \qquad \omega \in \{\omega_i\} \quad i = 1, 2, \ldots K$$

ω diskret, x kontinuierlich

Optimalitätskriterium

Gesucht ist ein Klassifikator, welcher „bestmöglich“ nach einem vorgegebenen statistischen Gütekriterium klassifiziert (Optimalklassifikator)

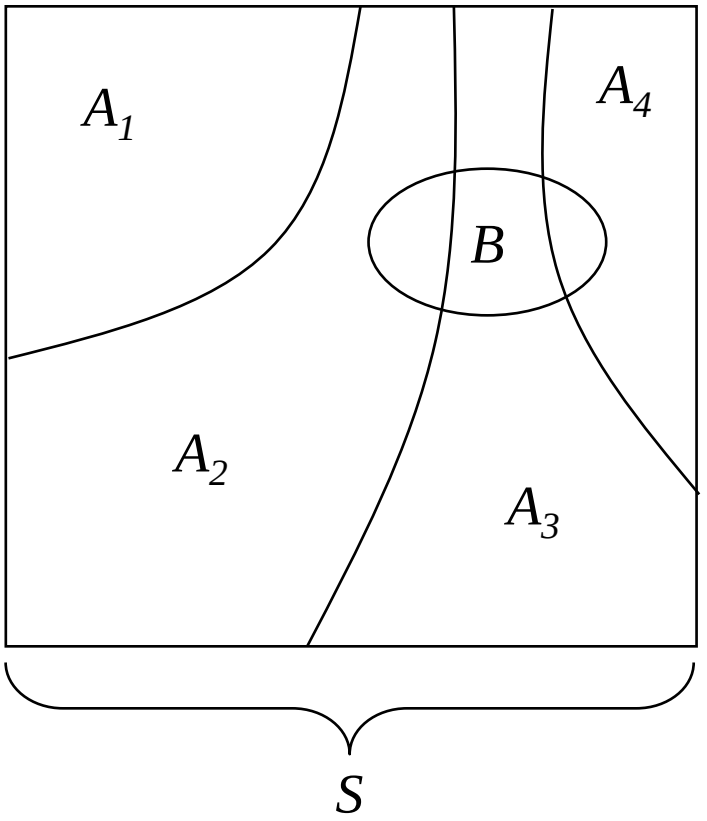
Wählt man als Optimalitätskriterium die ***Minimierung von Fehlentscheidungen*** bei vielen Versuchen, so ergibt sich ein Klassifikator, welcher die A-posteriori-Wahrscheinlichkeit maximiert (Maximum-A-Posteriori-Klassifikator):

$$\max_K \{P(\omega_k \mid \mathbf{x})\} \quad \text{MAP- oder Bayes-Klassifikator}$$

Theorem von Bayes

A_i seien sich gegenseitig ausschliessende (disjunkte) Ereignisse

$S = \bigcup_{i=1}^{n} A_i$ Ereignisraum (sample space)



B sei ein beliebiges Ereignis. Dann gilt:

$$P(A_i \mid B) = \frac{P(A_i, B)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)}$$

$$\Rightarrow P(A_i \mid B)P(B) = P(A_i, B) = P(B \mid A_i)P(A_i)$$

Beispiel

Die Anwendung des Theorems soll nun durch ein Beispiel veranschaulicht werden. Im Beispiel tritt das Ereignis B „Die Reifen eines Autos quietschen" mit der Wahrscheinlichkeit P(B)=0,05 auf; die Hypothese A „Die Bremsen des Autos sind schlecht eingestellt" mit der Wahrscheinlichkeit P(A)=0,02.

Nehmen wir des weiteren an, dass schlecht eingestellte Bremsen oft, aber nicht immer, ein Quietschen der Räder verursachen. Die bedingte Wahrscheinlichkeit dafür ist P(B|A)=0,7. Wenn man nun ein Quietschen der Reifen beobachtet, so kann man mit Hilfe des Bayes'schen Theorem die Wahrscheinlichkeit berechnen, dass die Bremsen schlecht eingestellt sind:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} = \frac{0,7 \cdot 0,02}{0,05} = 0,28$$

Man führt die W. *P*(*Ursache* | *Ereignis*) auf die einfacher zu berechnende W. *P*(*Ereignis* | *Ursache*) zurück.

Durch die Beobachtung des Ereignisses B hat sich die Wahrscheinlichkeit der Hypothese A von 0,02 auf 0,28 erhöht. Die Berechnung von P(A|B) ausgehend von P(A) kann als Neubewertung der Hypothese A beim Eintreten des Ereignisses B aufgefaßt werden.
Darin liegt die Stärke dieses Theorems. Mit ihm läßt sich die Fortpflanzung der Unsicherheit berechnen. Der Nachteil ist jedoch, dass zu jedem Ereignis und zu jeder Hypothese die Wahrscheinlichkeit und die entsprechenden bedingten Wahrscheinlichkeiten gespeichert werden müssen. Dies erfordert eine große Datenmenge, die häufig schwer zu beschaffen ist und auch oft nicht mit mathematischer Exaktheit beschafft werden kann. [Gottlob1990]

Grundlage der Optimalentscheidung ist die **a-posteriori-** oder ***Rückschlusswahrscheinlichkeit*** $P(\omega_k|\mathbf{x})$. Dies ist die ***bedingte Wahrscheinlichkeit***, dass bei bei einem beobachteten Wert **x** das Zeichen ω_k vorlag.

Diese kann mit dem *Bayes-Theorem* wie folgt umgewandelt werden:

$$P(\omega_k \mid \mathbf{x}) = \frac{p(\mathbf{x}, \omega_k)}{p(\mathbf{x})} = \frac{p(\mathbf{x} \mid \omega_k) P(\omega_k)}{p(\mathbf{x})}$$

mit der Randverteilung:

$$p(\mathbf{x}) = \sum_K p(\mathbf{x}, \omega_k) = \sum_K p(\mathbf{x} \mid \omega_k) P(\omega_k)$$

Die a-posteriori oder Rückschlusswahrscheinlichkeit zur Entscheidung der Klassenzugehörigkeit wird zurückgeführt auf die klassenspezifischen Verteilungsdichten, welche gemessen werden können!

Bayes- oder Maximum-A-Posteriori (MAP) Klassifikator

Die Maximierung der a-posteriori oder Rückschlusswahrscheinlichkeit zur Entscheidung der Klassenzugehörigkeit wird zurückgeführt auf die klassenspezifischen Verteilungsdichten, welche gemessen werden können:

$$\max_{\omega_i} P(\omega_i \mid \mathbf{x})$$

$$\Rightarrow P(\omega_i \mid \mathbf{x}) \overset{?}{\gtrless} P(\omega_j \mid \mathbf{x})$$

$p(\mathbf{x})$ auf beiden Seiten gleich!

$$\Rightarrow \boxed{\frac{p(\mathbf{x} \mid \omega_i)P(\omega_i)}{\cancel{p(\mathbf{x})}} \overset{?}{\gtrless} \frac{p(\mathbf{x} \mid \omega_j)P(\omega_j)}{\cancel{p(\mathbf{x})}}}$$

Bayes- oder MAP-Klassifikator

Bei gleichwahrscheinlicher a-priori-Verteilung $P(\omega_i) = P(\omega_j)$ erhält man daraus:

$$\boxed{p(\mathbf{x} \mid \omega_i) \overset{?}{\gtrless} p(\mathbf{x} \mid \omega_j)}$$

MLE-Klassifikator
(Maximum-Likelihood Estimation)

Optimalklassifikatoren

$$\max_k \{P(\omega_k \mid \mathbf{x})\}$$

$$\sim \max_k \{p(\mathbf{x} \mid \omega_k) P(\omega_k)\}$$

Bayes- oder MAP-Klassifikator

Das MAP-Kriterium wird zurückgeführt auf klassenspezifische Verteilungsdichten, welche gemessen werden können.

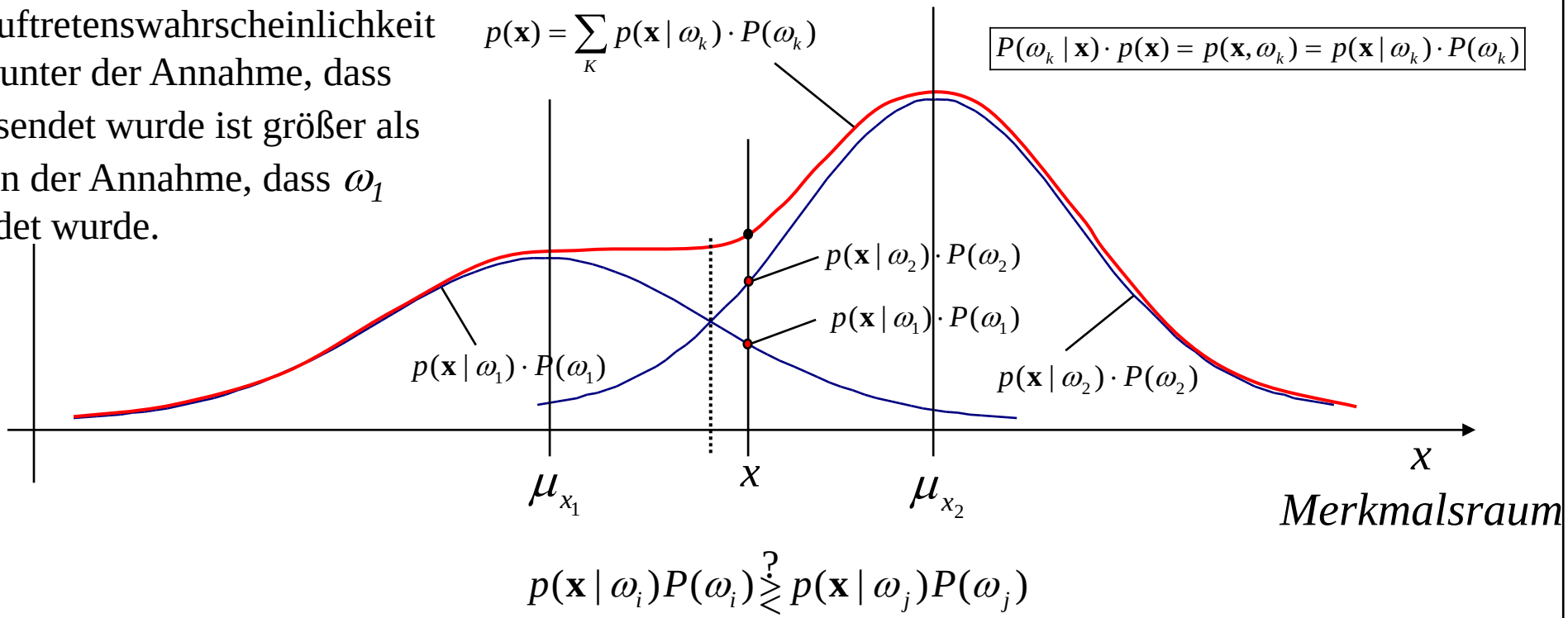
$$\max_K \{p(\mathbf{x} \mid \omega_k)\}$$

Maximum-Likelihood-Klassifikator

(p(x|ωk) Bed. Wahrscheinlichkeit von x bzgl. ω_k)

Zwei-Klassen-Problem mit Gaußverteilungsdichten und einem skalaren Merkmal *x*

Die Auftretenswahrscheinlichkeit von *x* unter der Annahme, dass ω_2 gesendet wurde ist größer als die von der Annahme, dass ω_1 gesendet wurde.



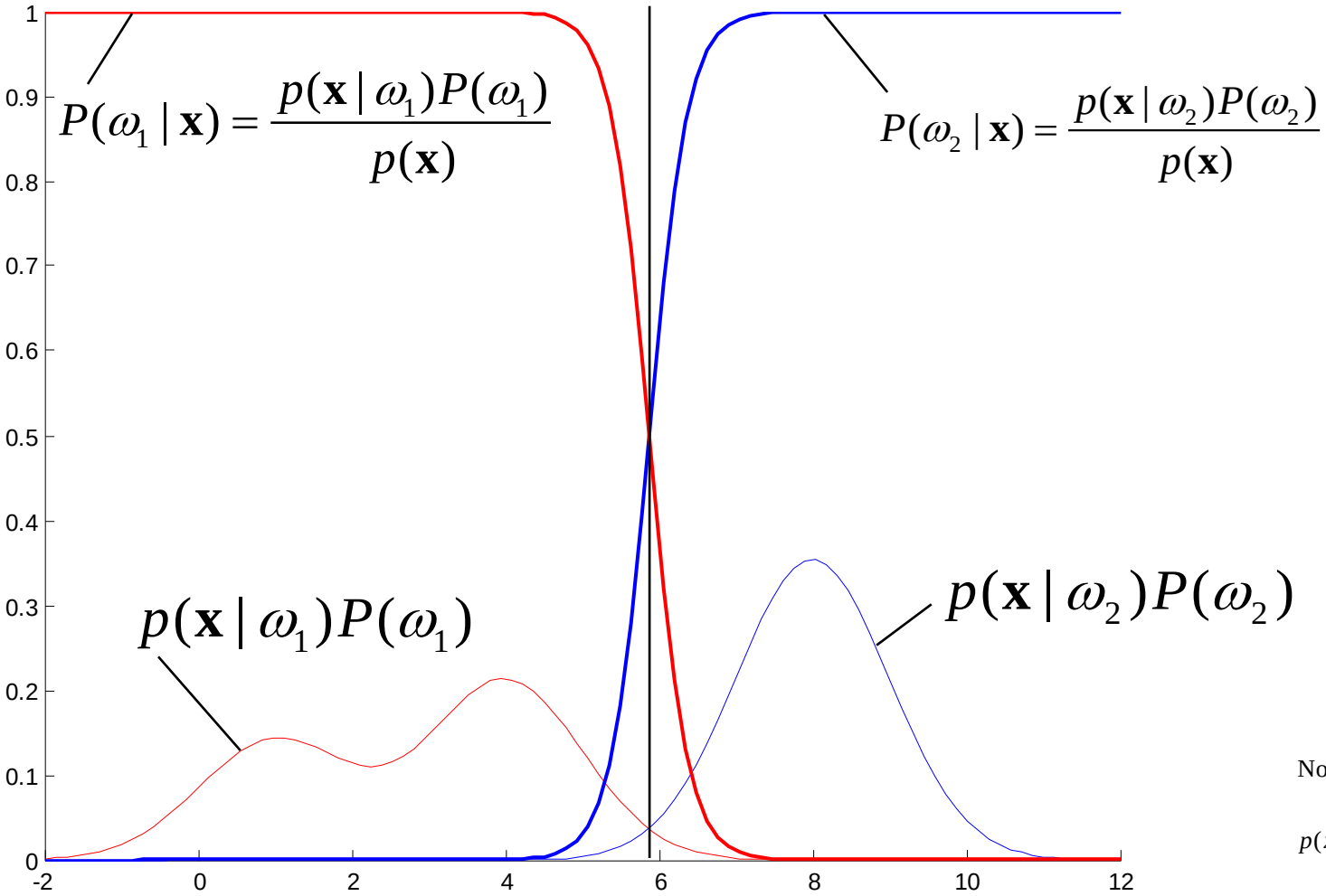
$$p(\mathbf{x}\,|\,\omega_i)P(\omega_i) \gtrless^{?} p(\mathbf{x}\,|\,\omega_j)P(\omega_j)$$

$p(\mathbf{x}|\omega_k)$ *Klassenspezifische Verteilungsdichte* für den Merkmalsvektor **x**, die der Klasse *k* zuzuordnen sind.

A-priori-W. für die Häufigkeit der Quellsymbole: $P(\omega_k)$
(Quellstatistik, Auftretens-W. für die Ereignisse ω_k, z.Bsp. Buchstaben in einer Sprache)

Die W., dass x gemessen wird ergibt sich aus der Überlagerung der Auswirkungen, dass ω_k gesendet wurde: $p(\mathbf{x}) = p(\mathbf{x}\,|\,\omega_1)P(\omega_1) + p(\mathbf{x}\,|\,\omega_2)P(\omega_2) + \cdots = \sum_K p(\mathbf{x}\,|\,\omega_k)P(\omega_k)$

MAP-Entscheidung



Normalverteilung:

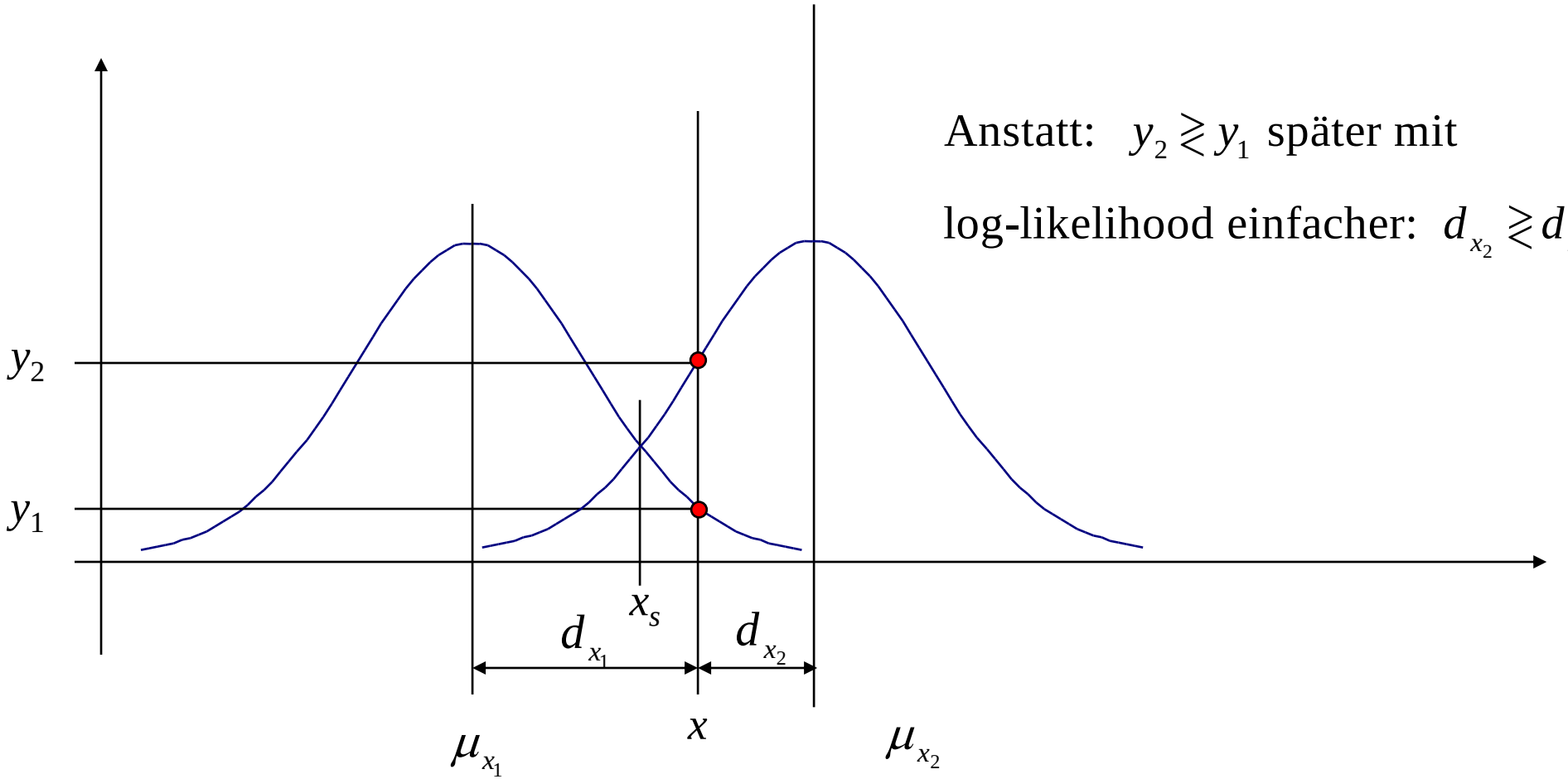
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2}$$

$$p(\mathbf{x}) = \sum_K p(\mathbf{x}, \omega_k) = \sum_K p(\mathbf{x} \mid \omega_k)P(\omega_k)$$

$$p(\mathbf{x} \mid \omega_1)P(\omega_1) = 0,2\, p(x-1) + 0,3\, p(x-4)$$

$$p(\mathbf{x} \mid \omega_2)P(\omega_2) = 0,5\, p(x-8)$$

Entscheidung mit Log-Likelihood

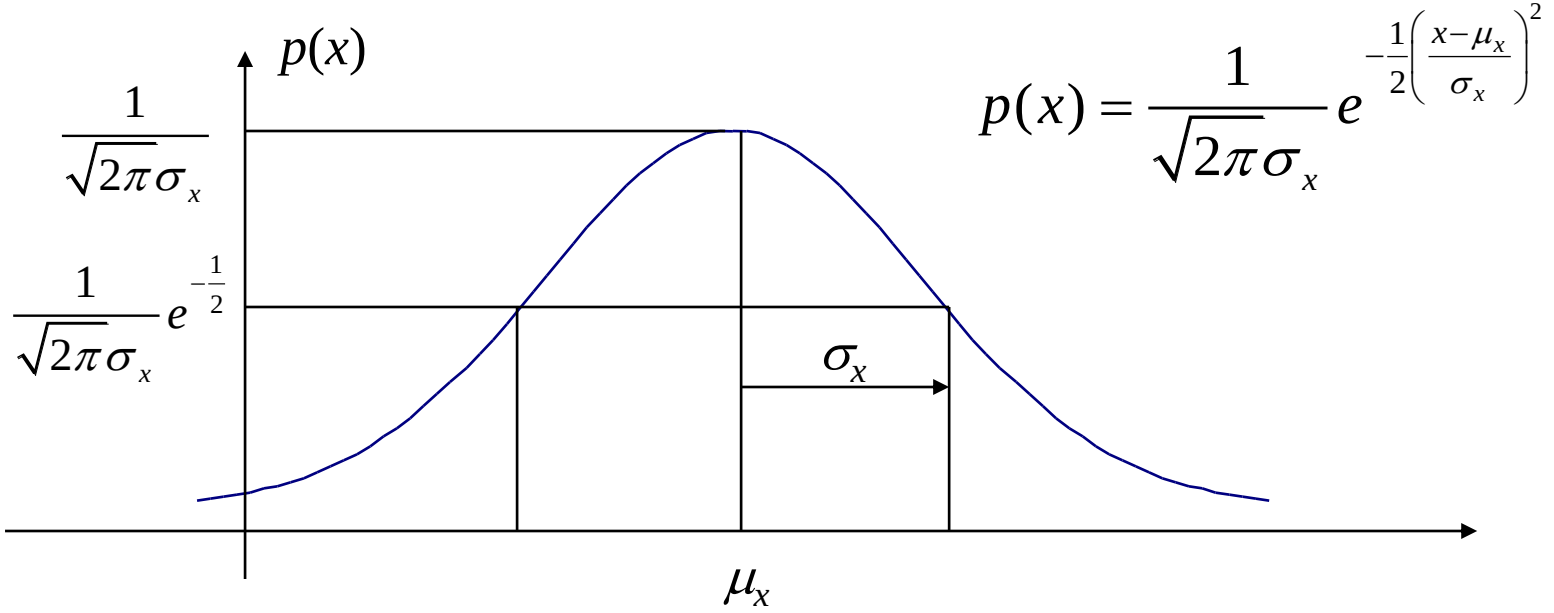


Anstatt: $y_2 \gtrless y_1$ später mit

log-likelihood einfacher: $d_{x_2} \gtrless d_{x_1}$

Normalverteilte klassenspezifische Merkmale $p(\mathbf{x}|\omega_k)$

Eindimensionaler Fall (nur ein Merkmal):



Erwartungswert von x: $\mu_x = E\{x\} = \int_{x=-\infty}^{x=+\infty} x \cdot p(x)dx$

Varianz: $\mathrm{var}(x) = \sigma_x^2 = E\{(x-\mu_x)^2\} = \int_{x=-\infty}^{x=+\infty} (x-\mu_x)^2 \cdot p(x)dx$

Standardabweichung: $\sigma_x = \sqrt{\mathrm{var}(x)}$

N-dimensionale Normalverteilung

Erwartungswert: $\boldsymbol{\mu}_x = E(\mathbf{x})$ (Vektor)

Statt Varianz σ^2 nun Autokovarianzmatrix:

$$\mathbf{K} = \mathbf{C}_{\mathbf{xx}} = E\{(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})^T\} = \mathbf{R}_{\mathbf{xx}} - \boldsymbol{\mu}_\mathbf{x}\boldsymbol{\mu}_\mathbf{x}^T$$

N-dimensionale Normalverteilung: $p(\mathbf{x}) = \dfrac{1}{\sqrt{(2\pi)^N \det(\mathbf{K})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})^T \mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})}$

$$\mathbf{K} = \begin{bmatrix} K_{1,1} & K_{1,2} & \cdots & K_{1,N} \\ K_{2,1} & K_{2,2} & \cdots & K_{2,N} \\ \vdots & \vdots & \vdots & \vdots \\ K_{N,1} & K_{N,2} & \cdots & K_{N,N} \end{bmatrix}$$

$$K_{m,n} = E\{(x_m - \mu_{x_m})(x_n - \mu_{x_n})\}$$

$$K_{n,n} = E\{(x_n - \mu_{x_n})^2\}$$

$\mathbf{K}$: a) symmetrisch

b) positiv semidefinit

N-dimensionale Normalverteilung

Aus der Positiv-Semidefinitheit folgt: $\mathbf{a}^T\mathbf{K}\mathbf{a} \geq 0$ für beliebige $\mathbf{a} \neq 0$

Falls eine oder mehr Komponenten Linearkombinationen von anderen sind, ist **K** semidefinit, andernfalls positiv definit (soll hier i.allg. angenommen werden).

Falls **K** pos. definit, dann auch $\mathbf{K}^{-1}$ => det(**K**)>0 und det($\mathbf{K}^{-1}$)>0 .

Ortskurven konstanter Wahrscheinlichkeitsdichten $p(\mathbf{x})$:

$$Q = (\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})^T \mathbf{K}^{-1} (\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}) = const.$$

Diese quadratische Form ergibt Kegelschnitte und für pos. def. $\mathbf{K}^{-1}$ erhält man *N*-dimensionale Ellipsoide.

N=2: Ellipsen

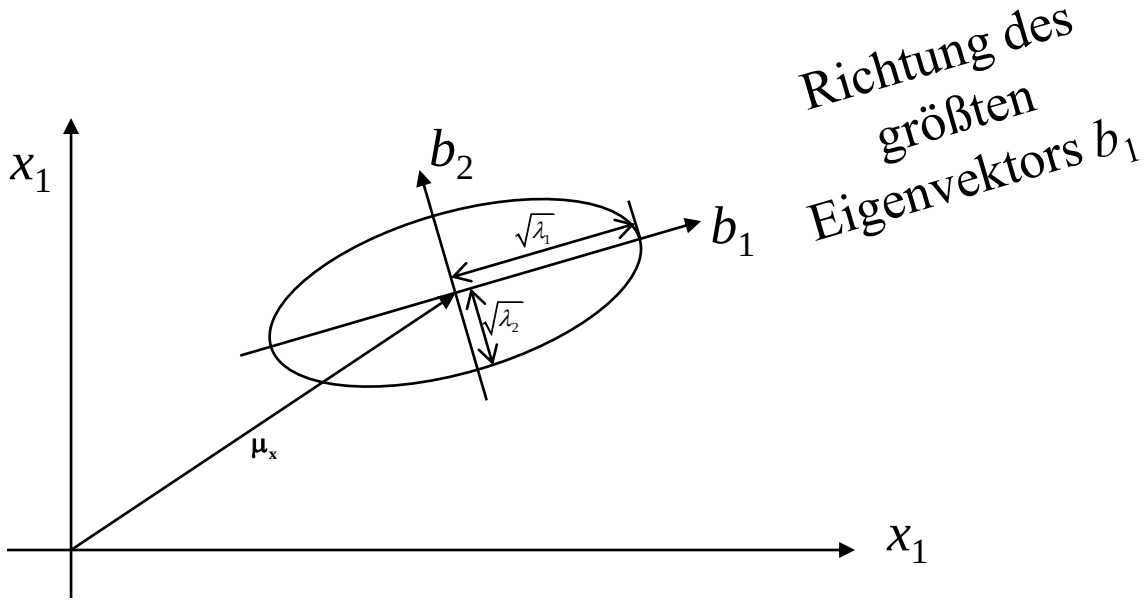
Aus der Eigenwertgleichung:

$$\mathbf{Kb} = \lambda\mathbf{b} \Rightarrow [\mathbf{K} - \lambda\mathbf{I}]\mathbf{b} = 0$$

Ergeben sich die

Eigenwerte: λ_1, λ_2

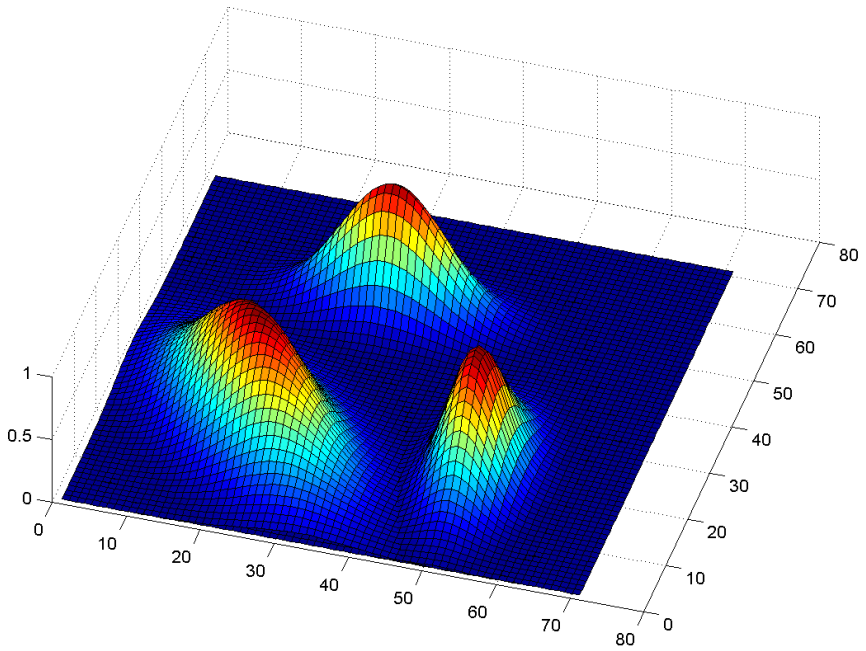
und die Eigenvektoren: b_1, b_2

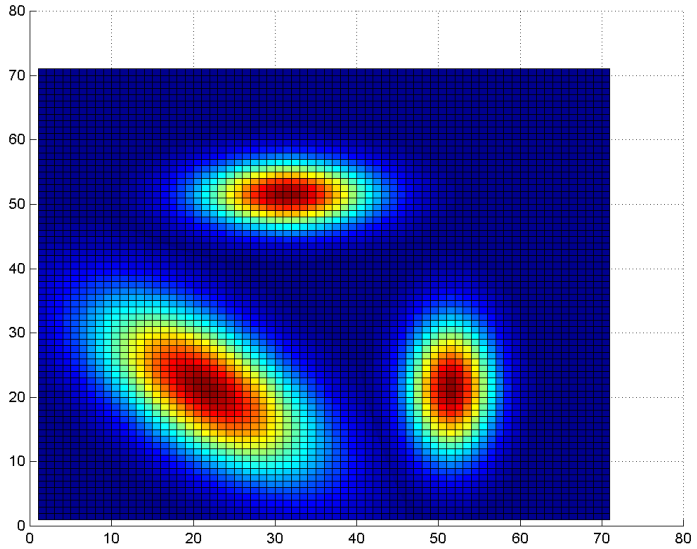


Der Bayes-Klassifikator

$$\max_{\omega_i} P(\omega_i \mid \mathbf{x})$$

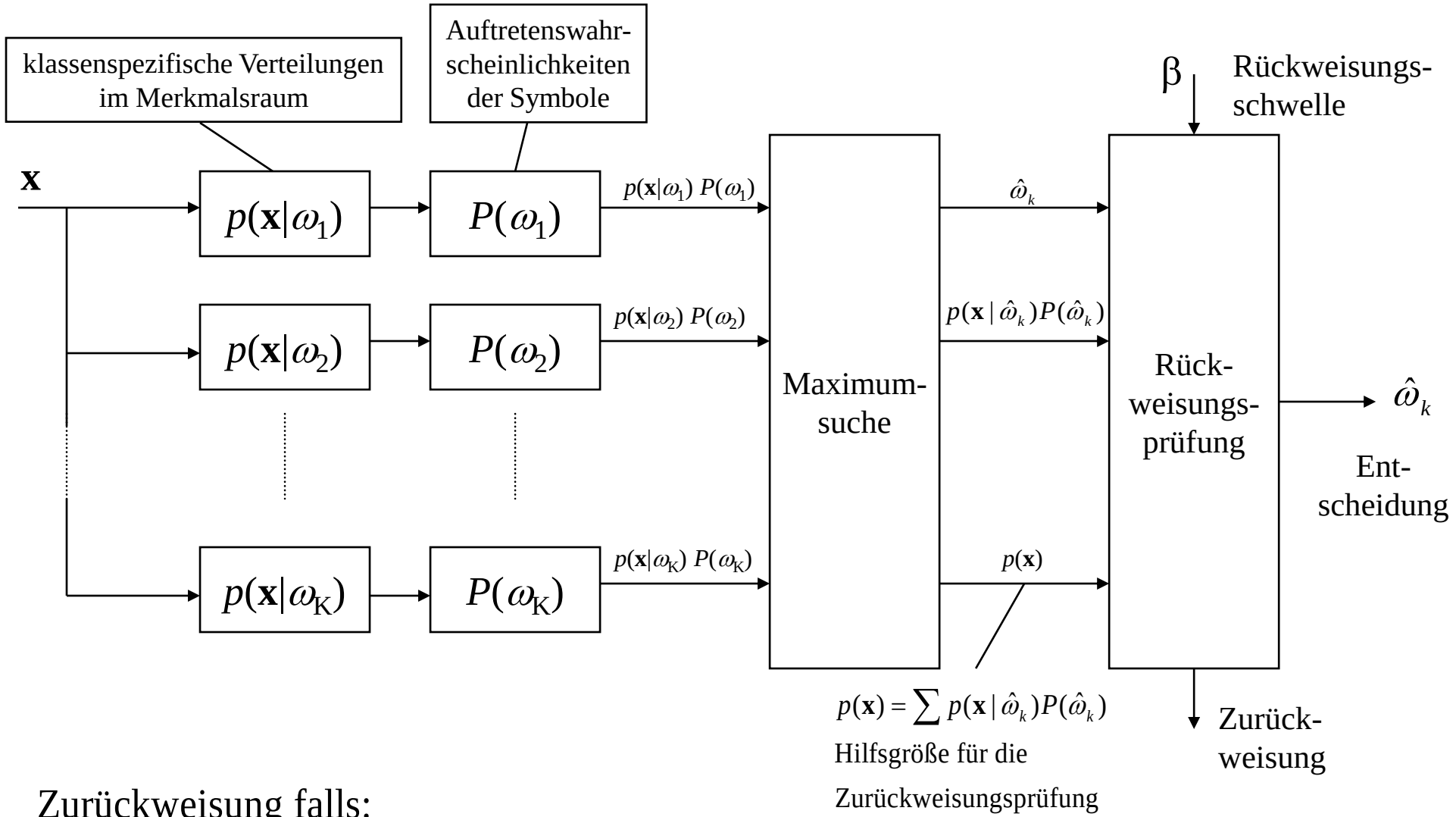
Annahme: Klassenspezifische Gauß-Verteilungen





Zweidimensionaler Merkmalsraum

Optimales Erkennungssystem



Zurückweisung falls:

$p(\mathbf{x}\,|\,\hat{\omega}_k)P(\hat{\omega}_k) < \beta\, p(\mathbf{x})$

d.h. $P(\hat{\omega}_k\,|\,\mathbf{x}) < \beta$

Falls maximale klassenspez. W. im Vergleich zur absoluten Auftretens-W. (rel. Mass) einen zu geringen Wert hat
➔Zurückweisung (sonst wäre Entscheidung sehr unsicher)
(typisch in der Nähe der Klassengrenzen !)

Zur Positiv-Definitheit der Kovarianzmatrix **K**

Man erwartet, dass die Beobachtungen des Zufallsprozesses unabhängig sind.

Beh.: $Q = \mathbf{z}^T \mathbf{K}\mathbf{z} > 0 \quad \text{für} \quad \forall \mathbf{z} \neq \mathbf{0}$

Bew:

$$Q = \mathbf{z}^T E\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\}\mathbf{z}$$

$$= E\{\mathbf{z}^T(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{z}\} = E\{\underbrace{[(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{z}]}_{=w\ \text{(Skalar)}}^2\}$$

$$= E\{w^2\} > 0 \quad \text{für} \quad w \neq 0$$

Der singuläre Fall Q=0 bedeutet, dass der Zufallsprozess $(\mathbf{x}\text{-}\boldsymbol{\mu}) \perp \mathbf{z}$ steht, d.h. er belegt nur einen linearen Unterraum des N-dimensionalen Beobachtungsraumes $\mathbb{R}^N$. Dies ist dann der Fall, wenn die Zufallsvariablen nicht den ganzen Raum aufspannen, d.h. wenn ein Vektor linear abhängig ist von anderen (z.B. wenn die Beobachtungen im dreidimensionalen Raum immer nur in einer Ebene liegen).

Für einzelne Vektoren darf die Orthogonalität $(\mathbf{x}\text{-}\boldsymbol{\mu}) \perp \mathbf{z}$ gegeben sein, jedoch nicht für das ganze Ensemble, so dass E{...}=0.

Konsequenzen der Positiv-Definitheit von **K**

- **K** ist regulär und es existiert $\mathbf{K}^{-1}$
- $\det(\mathbf{K})>0$
- $\mathbf{K}^{-1}$ ist ebenfalls positiv definit
- $\det(\mathbf{K}^{-1})>0$
- Die Eigenwerte von **K** sind positiv

Fall 1: Klassenweise beliebige normalverteilte Merkmale

Mit dieser Annahme kann das MAP-Kriterium weiter spezifiziert werden:

$$p(\mathbf{x}, \omega_k) = p(\mathbf{x} \mid \omega_k) \cdot P(\omega_k)$$

Der zeichenerzeugende Prozess zerfällt in K voneinander unabhängige Teilprozesse $\{p(\mathbf{x}|\omega_k)\}$ mit den Kenngrößen:

$\boldsymbol{\mu}_{x_k} = E\{\mathbf{x} \mid \omega_k\}$ klassenspezifischer Erwartungswert

$\mathbf{K}_k = E\{(\mathbf{x} - \boldsymbol{\mu}_{x_k})(\mathbf{x} - \boldsymbol{\mu}_{x_k})^T \mid \omega_k\}$ klassenspezifische Kovarianzmatrix

Berechnet man die k-te Entscheidungsfunktion des MAP-Kriteriums, so ergibt sich:

$$D_k(\mathbf{x}) = p(\mathbf{x} \mid \boldsymbol{\omega}_k) \cdot P(\boldsymbol{\omega}_k) = \frac{P(\omega_k)}{\sqrt{(2\boldsymbol{\pi})^N \det(\mathbf{K}_k)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{x_k})^T \mathbf{K}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_{x_k})}$$

Mit Hilfe einer monotonen Abbildung ln(...), welche die Größenverhältnisse nicht verändert, ergibt sich:

$$D'_k(\mathbf{x}) = \ln P(\omega_k) - \frac{1}{2}\ln(\det(\mathbf{K}_k)) - \frac{1}{2}[(\mathbf{x}-\boldsymbol{\mu}_{x_k})^T \mathbf{K}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_{x_k})]$$

mit abschließendem Maximumvergleich.

Die Grenzen zwischen den Klassengebieten ergeben sich zu:

$$D_i'(\mathbf{x}) \overset{!}{=} D_j'(\mathbf{x})$$

daraus ergibt sich die Grenzfläche $g_{ij}(\mathbf{x}) = 0$, mit:

$$g_{ij}(\mathbf{x}) = \ln \frac{\det \mathbf{K}_i}{\det \mathbf{K}_j} - 2 \ln \frac{P(\omega_i)}{P(\omega_j)}$$
$$+ (\mathbf{x} - \boldsymbol{\mu}_{x_i})^T \mathbf{K}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_{x_i}) - (\mathbf{x} - \boldsymbol{\mu}_{x_j})^T \mathbf{K}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_{x_j})$$

Aus der Differenz zweier quadratischen Formen ergibt sich eine gemeinsame quadratische Form der Gestalt:

$$\boxed{g_{ij}(\mathbf{x}) = g_0 + (\mathbf{x} - \mathbf{x}_0)^T \mathbf{M}^{-1} (\mathbf{x} - \mathbf{x}_0)}$$

mit:

$$g_0 = \ln \frac{\det \mathbf{K}_i}{\det \mathbf{K}_j} - 2 \ln \frac{P(\omega_i)}{P(\omega_j)} + \boldsymbol{\mu}_{x_i}^T \mathbf{K}_i^{-1} \boldsymbol{\mu}_{x_i}$$

$$- \boldsymbol{\mu}_{x_j}^T \mathbf{K}_j^{-1} \boldsymbol{\mu}_{x_j} + \mathbf{x}_0^T \mathbf{M}^{-1} \mathbf{x}_0$$

$$\mathbf{x}_0 = \mathbf{M}[\mathbf{K}_i^{-1} \boldsymbol{\mu}_{x_i} - \mathbf{K}_j^{-1} \boldsymbol{\mu}_{x_j}]$$

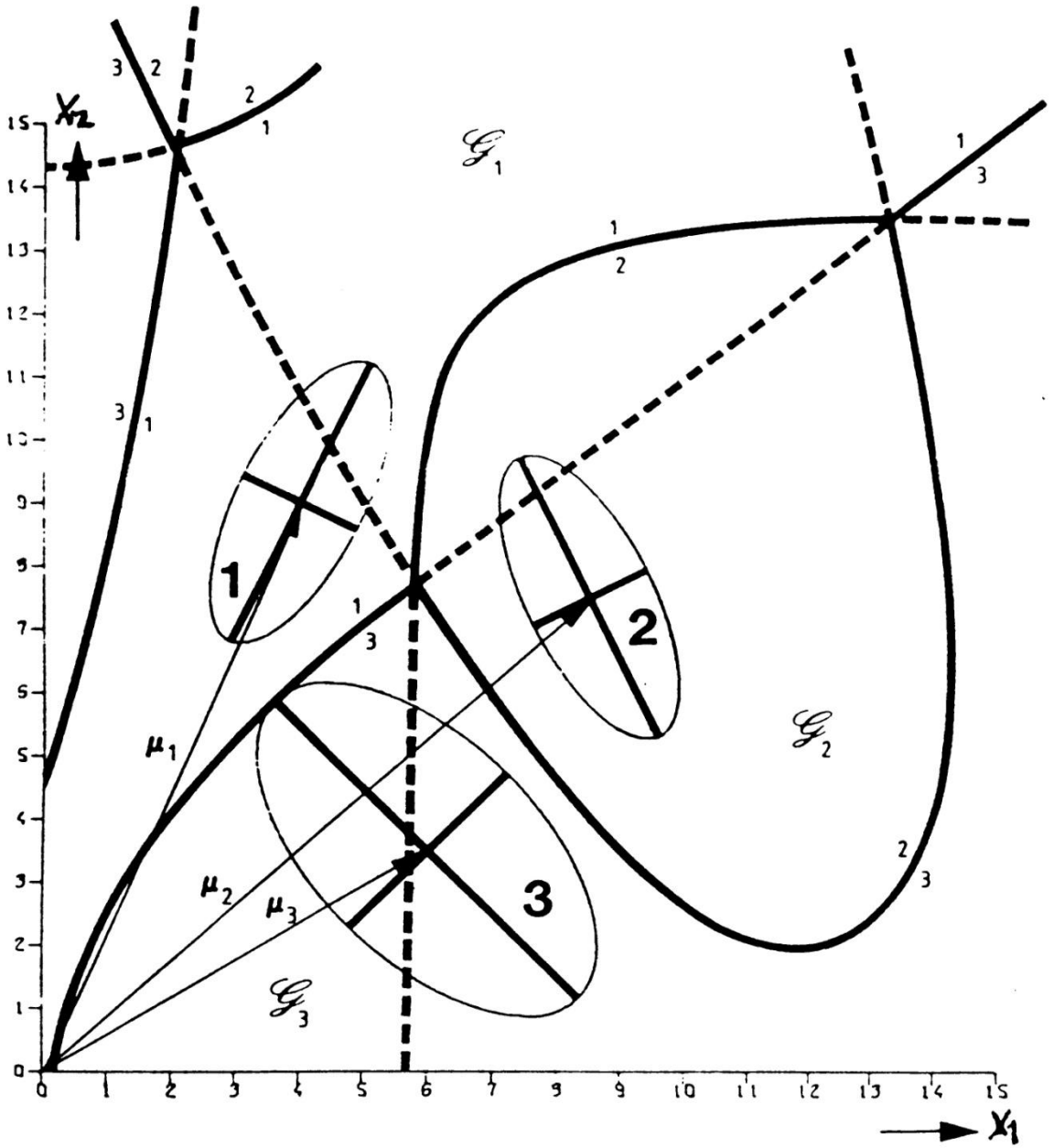
$$\mathbf{M} = [\mathbf{K}_i^{-1} - \mathbf{K}_j^{-1}]^{-1} = \mathbf{K}_i [\mathbf{K}_j - \mathbf{K}_i]^{-1} \mathbf{K}_j$$

$$= \mathbf{K}_j [\mathbf{K}_j - \mathbf{K}_i]^{-1} \mathbf{K}_i$$

Die die quadratische Form charakterisierende Matrix $\mathbf{M}^{-1}$ ist nun nicht mehr zwingend pos. definit => die Grenzflächen zwischen den Gebieten sind *allgemeine* Kegelschnitte (bei N=2: Ellipsen, Parabeln, Hyperbeln, Geraden)

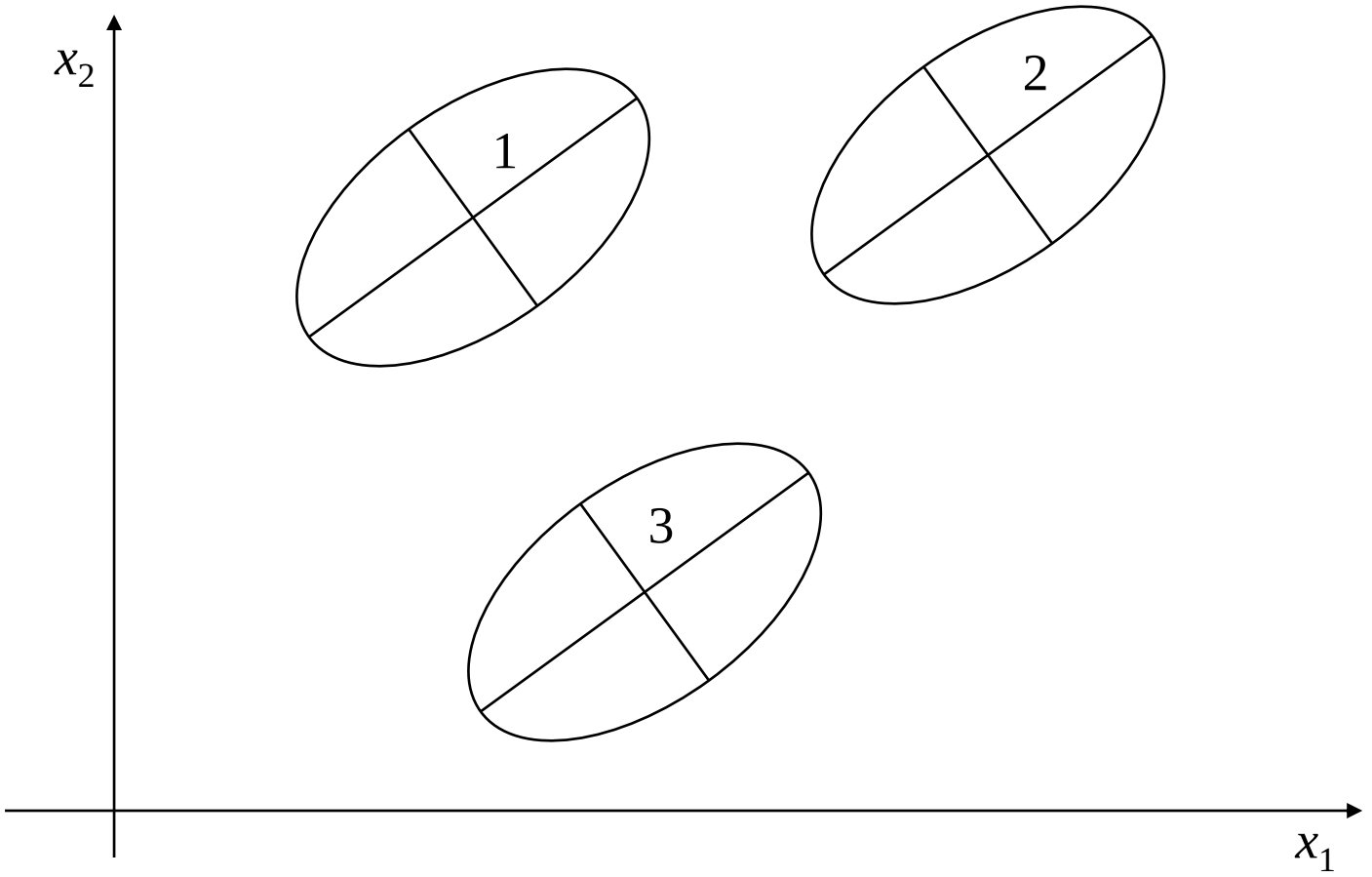
Die Unterscheidungsfunktionen $D'_k(\mathbf{x})$ sind in Bezug auf den Merkmalsvektor quadratische Funktionen oder Polynome zweiten Grades (***quadratischer oder Polynomklassifikator***)

Klassenweise normalverteilte Merkmale



(aus J. Schürmann: „Polynomklassifikatoren für die Zeichenerkennung“, Oldenbourg Verlag)

Fall 2: Klassenweise normalverteilte Merkmale mit *identischen* Kovarianzmatrizen **K**



Fall 2: Klassenweise normalverteilte Merkmale mit identischen Kovarianzmatrizen **K**

Entscheidungsfunktion:

$$D'_k(\mathbf{x}) = \ln P(\omega_k) - \frac{1}{2}\ln(\det \mathbf{K}) - \frac{1}{2}[(\mathbf{x} - \boldsymbol{\mu}_{x_k})^T \mathbf{K}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{x_k})]$$

$$D''_k = -2D'_k - \ln(\det \mathbf{K})$$

$$\Rightarrow D''_k = -2\ln P(\omega_k) + (\mathbf{x} - \boldsymbol{\mu}_{x_k})^T \mathbf{K}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{x_k})$$

Für gleiche Auftretenswahrscheinlichkeiten $P(\omega_k) = 1/K$ folgt:

$$D'''_k = D''_k - 2\ln K$$

$$\Rightarrow \boxed{D'''_k = d_M^2 = (\mathbf{x} - \boldsymbol{\mu}_{x_k})^T \mathbf{K}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{x_k})}$$

Mahalanobis-Abstands-Klassifikator

Dies ist eine allgemeine gewichtete quadratische Metrik.

Formulierung als *linearer* Klassifikator

Die Entscheidungsfunktion enthält noch einen quadratischen Term. Dieser ist jedoch für jede Klasse identisch und kann somit eliminiert werden. Damit kommt man zu einer linearen Formulierung des Klassifikators.

Alternativ ergibt sich:

$$D_k'' = D_k' + \frac{1}{2}[\ln(\det \mathbf{K}) + \mathbf{x}^T \mathbf{K}^{-1} \mathbf{x}]$$

$$\Rightarrow D_k'' = \ln P(\omega_k) - \frac{1}{2} \boldsymbol{\mu}_{x_k}^T \mathbf{K}^{-1} \boldsymbol{\mu}_{x_k} + \boldsymbol{\mu}_{x_k}^T \mathbf{K}^{-1} \mathbf{x}$$

Dieser Ausdruck ist linear in $\mathbf{x}$!

$$\Rightarrow \boxed{D_k''(\mathbf{x}) = a_{0k} + \mathbf{a}_k^T \mathbf{x} = a_{0k} + \langle \mathbf{a}_k, \mathbf{x} \rangle}$$

Hyperebene als Trennfläche !

mit:

$$a_{0k} = \ln P(\omega_k) - \frac{1}{2} \boldsymbol{\mu}_{x_k}^T \mathbf{K}^{-1} \boldsymbol{\mu}_{x_k}$$

$$\mathbf{a}_k = \mathbf{K}^{-1} \boldsymbol{\mu}_{x_k}$$

Fall 3: Klassenweise normalverteilte Merkmale mit der Einheitsmatrix als Kovarianzmatrix $\mathbf{K}=\sigma^2\mathbf{I}$

(sphärisch invariante Verhältnisse, Hyperkugeln)

Entscheidungsfunktion:

$$D_k'(\mathbf{x}) = \ln P(\omega_k) - N \ln \sigma - \frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu}_{x_k})^T(\mathbf{x}-\boldsymbol{\mu}_{x_k})$$

Für konstante A-priori-W. folgt:

$$\Rightarrow \boxed{D_k'' = \left\|\mathbf{x}-\boldsymbol{\mu}_{x_k}\right\|^2}$$

Euklidsche Metrik
Minimum-Abstands-Klassifikator

Auch hier formuliert als linearer Klassifikator:

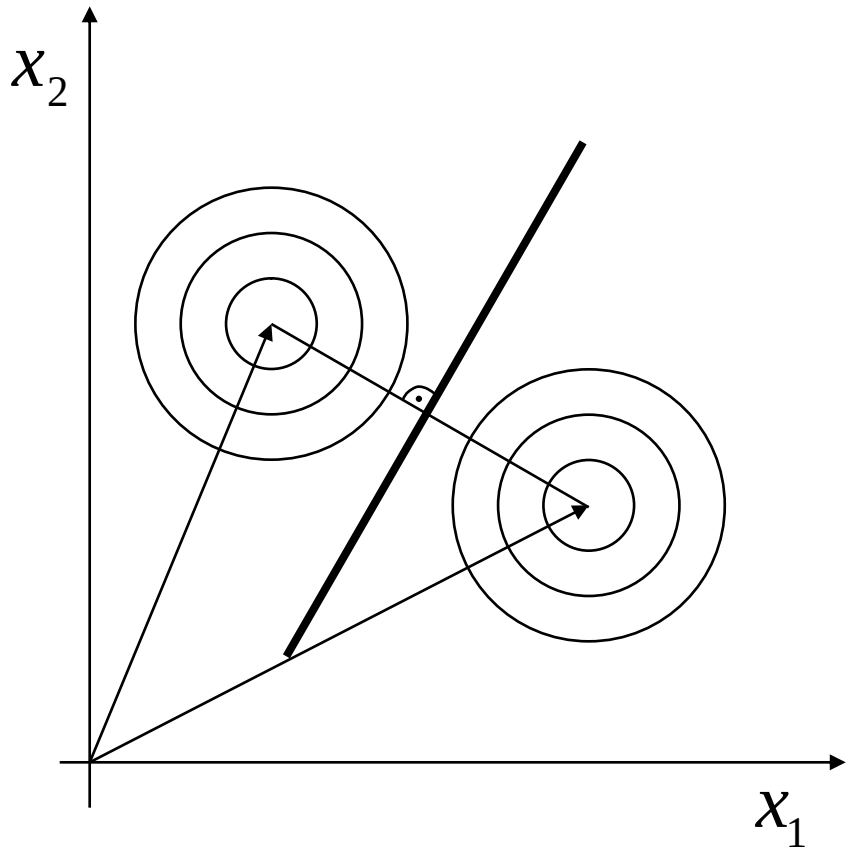
$$D_k''' = \frac{1}{2}(\|\mathbf{x}\|^2 - D_k'')$$

$$\Rightarrow \boxed{D_k'''(\mathbf{x}) = a_{0k} + \mathbf{a}_k^T\mathbf{x}}$$

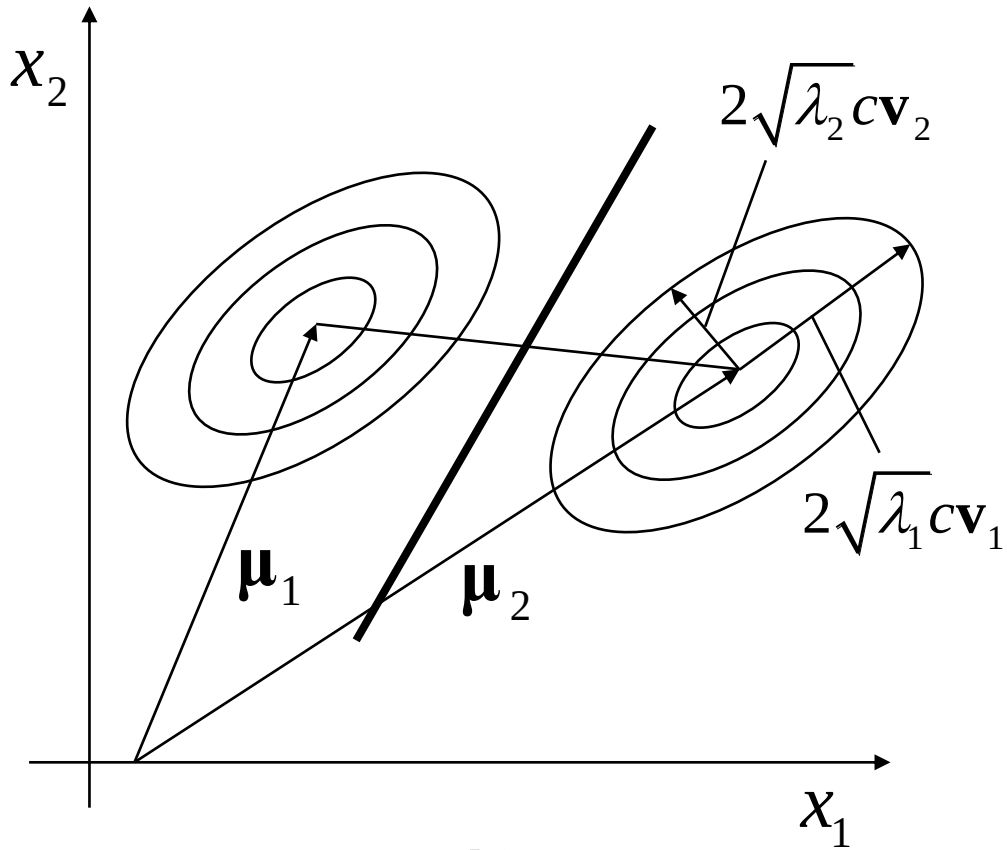
mit:

$$D_k''' = \frac{1}{2}(\|\mathbf{x}\|^2 - (\|\mathbf{x}\|^2 - \left\|\boldsymbol{\mu}_{x_k}\right\|^2 - 2 < \mathbf{x}, \boldsymbol{\mu}_{x_k} >)) = \underbrace{-\frac{1}{2}\left\|\boldsymbol{\mu}_{x_k}\right\|^2}_{a_{0k}} + < \underbrace{\boldsymbol{\mu}_{x_k}}_{\mathbf{a}_k}, \mathbf{x} >$$

Kurven konstanter a) Euklidscher- und b) Mahalanobis-Distanz d_M zum Erwartungswert der jeweiligen Klasse



a)



b)

Bayes-Entscheidungsgrenzen für normalverteilte Musterklassen

- Allgemeiner Fall: matlab-Bayes-Fall1.bat
- Mahalanobis: matlab-Bayes-Fall2.bat
- Euklid: matlab-Bayes-Fall3.bat

Transformation der Mahalanobis-Metrik auf sphärisch invariante Verhältnisse

Die Kovarianzmatrix kann mit einer KLT diagonalisiert werden:

$$\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N) = \mathbf{A}^T \mathbf{K} \mathbf{A}$$

bzw:

$$\mathbf{K} = \mathbf{A} \mathbf{\Lambda} \mathbf{A}^T$$

Die unitäre Matrix erfüllt: $\mathbf{A}^T = \mathbf{A}^{-1}$

Die Eigenwerte sowie die Eigenvektoren von K definieren die Diagonalmatrix und die Eigenvektoren die Tansformationsmatrix:

$$\mathbf{A} = \left[\mathbf{e}'_1, \mathbf{e}'_2, \ldots, \mathbf{e}'_N \right]$$

Kurven konstanter Mahalanobisdistanz ergeben sich zu:

$$(\mathbf{x} - \boldsymbol{\mu}_i)^T \mathbf{A} \mathbf{\Lambda}^{-1} \mathbf{A}^T (\mathbf{x} - \boldsymbol{\mu}_i) = c^2$$

Transformation auf sphärisch invariante Verhältnisse

Einführung einer Koordinatentransformation: $\mathbf{x}' = \mathbf{A}^T \mathbf{x}$

Damit werden die ursprünglichen Koordinaten auf die Eigenvektoren projeziert und man kommt dann zu den folgenden Kurven konstanter Mahalanobis-Distanz:

$$\frac{(x_1' - \mu_{i1}')^2}{\lambda_1} + \cdots + \frac{(x_N' - \mu_{iN}')^2}{\lambda_N} = c^2$$

Dies ist ein Hyperellipsoid in dem neuen Koordinatensystem.

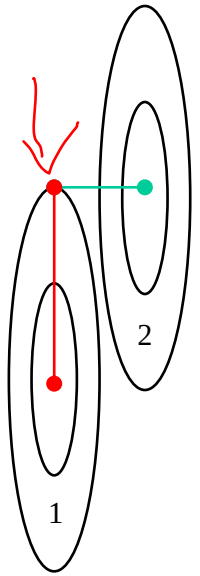
Mit $x_k'' = x_k' / \lambda_k$ und $\mu_{ik}'' = \mu_{ik}' / \lambda_k$ erhält man sphärisch invariante (Euklidsche) Verhältnisse:

$(x_1'' - \mu_{i1}'')^2 + \cdots + (x_N'' - \mu_{iN}'')^2 = c^2$ (Kugeln)

Beispiel: Zweiklassenproblem der Dimension 2

Die Kovarianzmatrix und die Erwartungswerte seien gegeben mit:

$$\mathbf{K} = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix} \quad \boldsymbol{\mu}_1 = \begin{bmatrix} 0 & 0 \end{bmatrix}^T \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 3 & 3 \end{bmatrix}^T$$



Klassifiziere die Beobachtung $\mathbf{x} = \begin{bmatrix} 1.0 & 2.2 \end{bmatrix}^T$ nach Bayes.

Die soll durch Berechnung der Mahalanobisdistanz zu den beiden Erwartungswerten geschehen:

$$d_M^2(\boldsymbol{\mu}_1, \mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_1)^T \mathbf{K}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) = \begin{bmatrix} 1.0 & 2.2 \end{bmatrix} \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix} \begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix} = 2.952$$

enstprechend:

$$d_M^2(\boldsymbol{\mu}_2, \mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_2)^T \mathbf{K}^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) = \begin{bmatrix} -2.0 & -0.8 \end{bmatrix} \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix} \begin{bmatrix} -2.0 \\ -0.8 \end{bmatrix} = 3.672$$

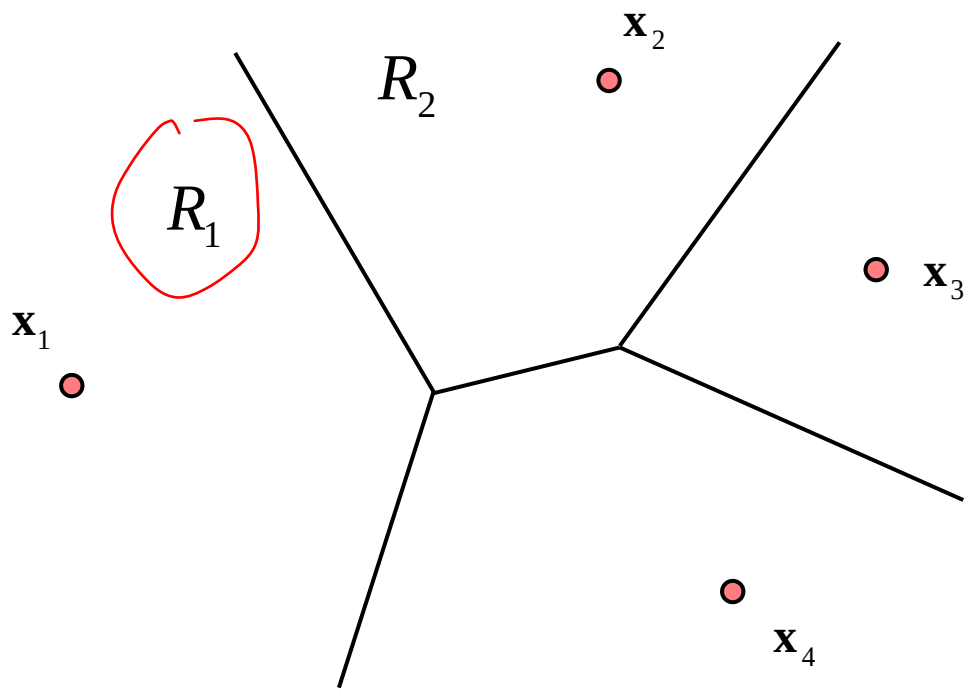
D.h. die Beobachtung wird der Klasse 1 zugeordnet.

Man beachte, dass die Beobachtung bzgl. der Euklidschen Distanz näher an Klasse 2 liegt!!

$$\|\mathbf{x} - \boldsymbol{\mu}_1\|^2 = 5.84 \qquad \|\mathbf{x} - \boldsymbol{\mu}_2\|^2 = 4.64$$

Siehe dazu auch:
http://www.cs.mcgill.ca/~mcleish/644/main.html
Applet für Bayessche Entscheidungen

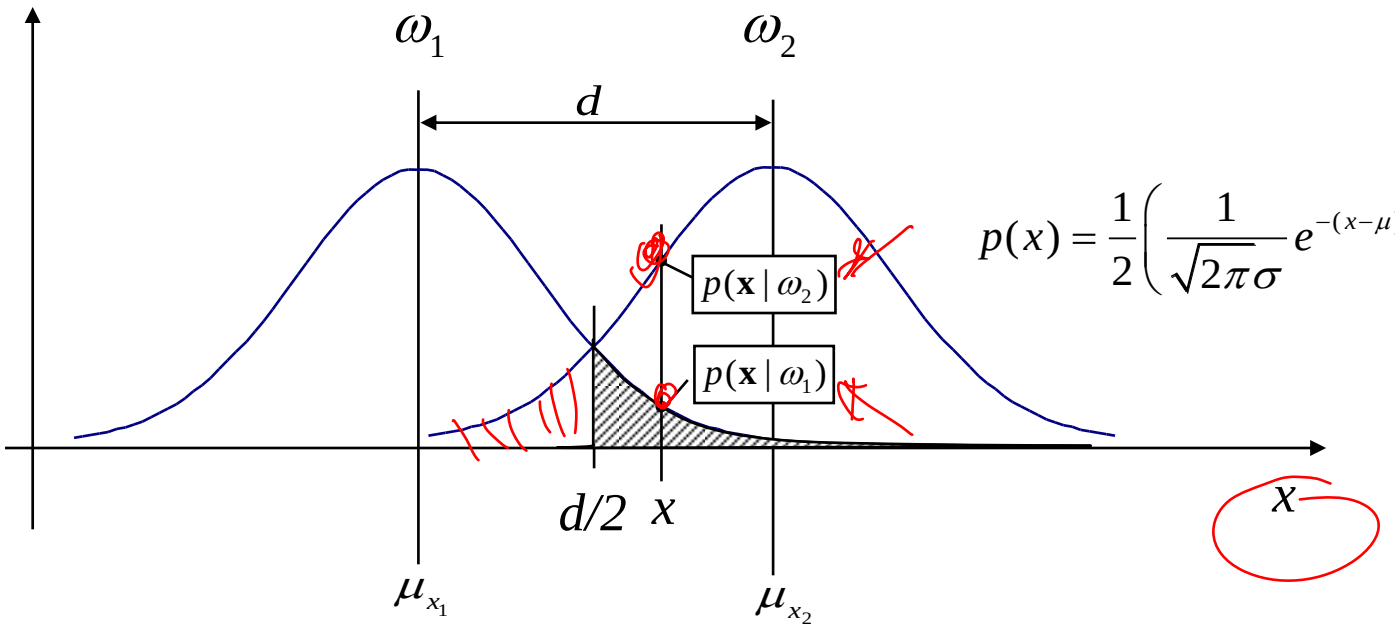
Das Voronoi-Diagramm im zweidimensionalen Raum als Entscheidungsgrenzen für die Euklidsche Distanz



Zerlegung der Ebene in Regionen R_i für eine Menge von Punkten $\mathbf{x}_i$. Jede Region enthält genau die Punkte, welche näher sind zu den jeweiligen Punkten (Erwartungswerten) als irgend ein anderer Punkt:

$$R_i = \{\mathbf{x} : d(\mathbf{x}, \mathbf{x}_i) < d(\mathbf{x}, \mathbf{x}_j) \quad \text{für} \quad i \neq j\}$$

Wahrscheinlichkeit einer Fehlklassifikation bei einem Zwei-Klassen-Problem mit Normalverteilungen und einem skalaren Merkmal



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

Die W. dass ω_1 gesendet wurde und bei gemessenen Werten x zugunsten von ω_2 entschieden wird, entspricht der W. dass x oberhalb von $d/2$ liegt (schraffierte Fläche). Ebenso für $\omega_2 \rightarrow \omega_1$. D.h. die W. einen Fehler bei der Klassifikation zu machen ergibt sich aus der doppelten Fläche:

$$P(E) = 2 \cdot \frac{1}{2} \int_{x=d/2}^{\infty} g(u)du$$

Mit der kumulativen Verteilungsfunktion $F(x_0) = P(x \le x_0)$ gilt:

$$F(x) = \int_{-\infty}^{x} g(u)du = \frac{1}{2}(1 + \text{erf}\left(\frac{x - \mu_x}{\sqrt{2}\sigma}\right))$$

und mit der Gauß'schen Fehlerfunktion

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt \quad \text{(analytisch nicht integrierbar)}$$

bzw. komplementären Fehlerfunktion

$$\text{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-t^2} dt = 1 - \text{erf}(x)$$

erhält man die Fehler-W. zu:

$$P(E) = \int_{x=d/2}^{\infty} g(u)du = 1 - F(x = d/2)$$

$$= \frac{1}{2}(1 - \text{erf}\left(\frac{x - \mu_x}{\sqrt{2}\sigma}\right))\Big|_{\substack{x=d/2 \\ \mu_x=0}} = \frac{1}{2}\text{erfc}\left(\frac{x - \mu_x}{\sqrt{2}\sigma}\right)\Big|_{\substack{x=d/2 \\ \mu_x=0}}$$

$$\boxed{P(E) = \frac{1}{2}\text{erfc}\left(\frac{d/2}{\sqrt{2}\sigma}\right)}$$

Die W. für eine Fehlklassifikation sinkt mit wachsendem Klassenabstand und steigt mit wachsender Streuung der Merkmale

Die Gesamtfehlerwahrscheinlichkeit im N-dimensionalen Merkmalsraum berechnet sich aus der minimalen Distanz $d_{\min}$ zu (Forney):

$$P(E) = const \cdot \frac{1}{2} \operatorname{erfc}\left(\frac{d_{\min} / 2}{\sqrt{2\sigma}} \right)$$

(ohne Beweis)