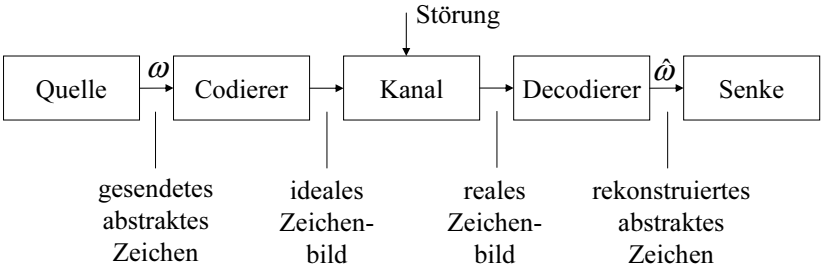
Kapitel 7

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:



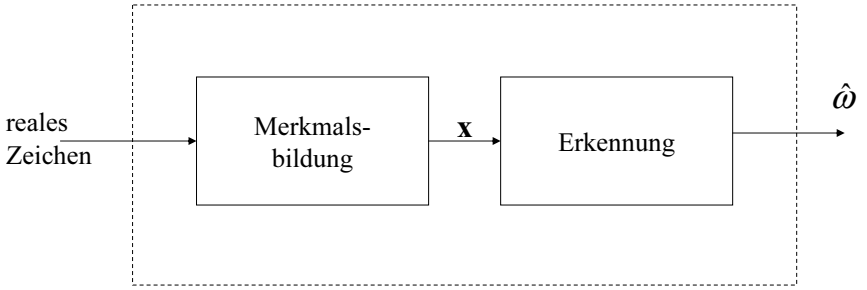
Über die einzelnen Komponenten

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

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

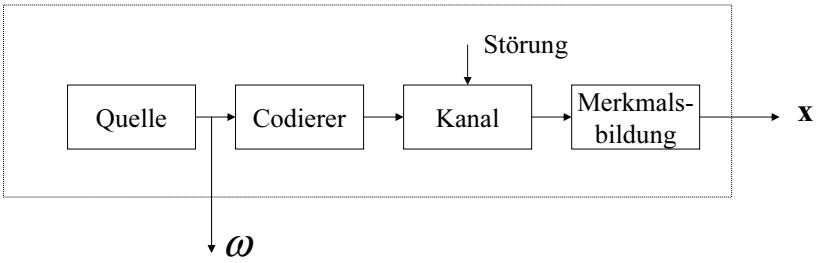
Decodierer: Lesemaschine, rekonstruiert das gesendete Zeichen, übergibt Entscheidung an Senke

Aufbau des Decodierers



Stochastisches Modell

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



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

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

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):

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

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:

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

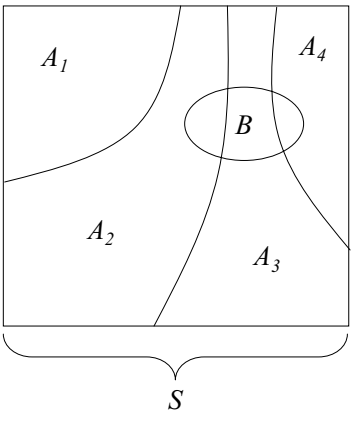
mit der Randverteilung:

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

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:

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

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

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

Maximierung der a-posteriori oder Rückschlusswahrscheinlichkeit zur Entscheidung der Klassenzugehörigkeit:

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

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

$p(\mathbf{x})$ 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)

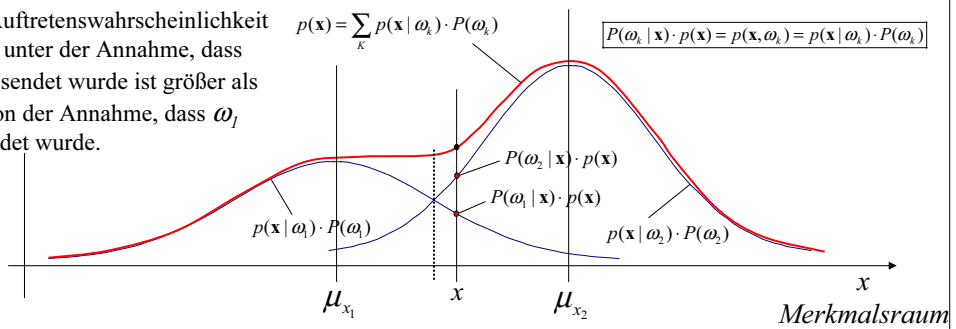
Optimalklassifikatoren

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

$\max_K \{p(\mathbf{x} \mid \omega_k\}$ Maximum-Likelihood-Klassifikator
($p(\mathbf{x}|\omega_k)$ Likelihood-Fkt.)

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.



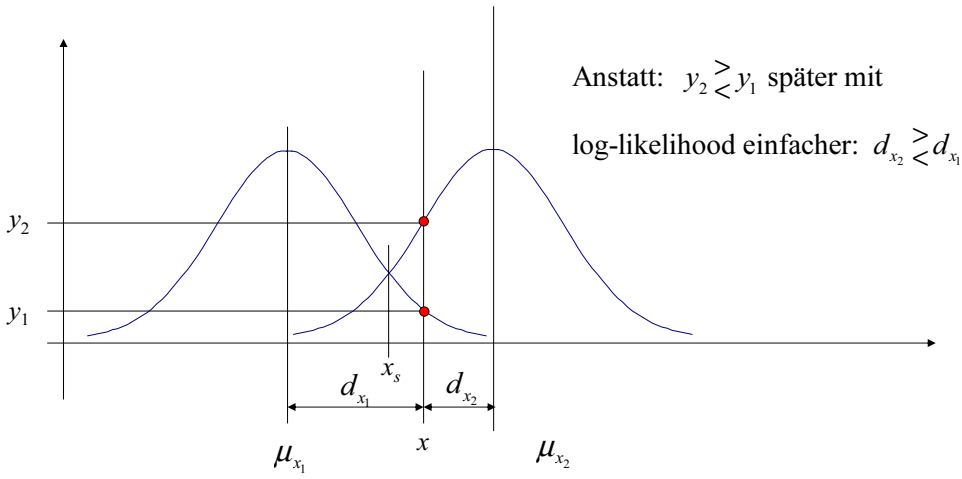
$$p(\mathbf{x} \mid \omega_i) P(\omega_i) \gtrless^{?} p(\mathbf{x} \mid \omega_j) P(\omega_j)$$

$p(\mathbf{x}|\omega_k)$ *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: $p(\mathbf{x}) = p(\mathbf{x} \mid \omega_1) P(\omega_1) + p(\mathbf{x} \mid \omega_2) P(\omega_2) + \cdots = \sum_K p(\mathbf{x} \mid \omega_k) P(\omega_k)$

Entscheidung mit Log-Likelihood

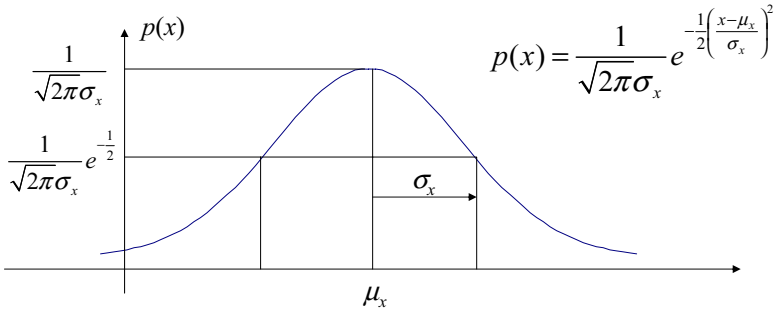


Anstatt: $y_2 \gtrless y_1$ später mit

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

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

Eindimensionaler Fall:



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

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:

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

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

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

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

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

$\mathbf{K}$: 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:

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

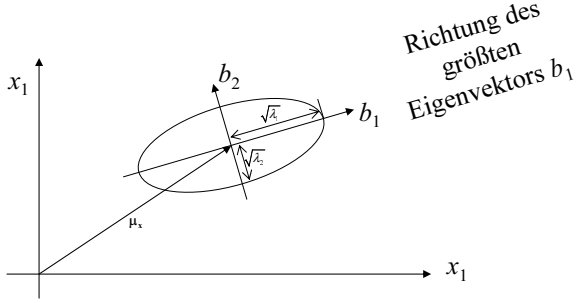
Diese quadratische Form ergibt Kegelschnitte und für pos. def. $\mathbf{K}^{-1}$erhält man *N*-dimensionale Ellipsoide.

N=2: Ellipsen

Aus der Eigenwertgleichung:

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

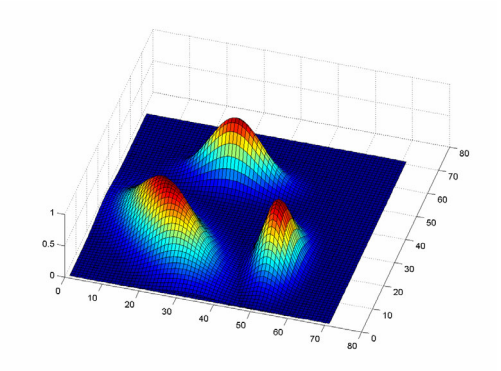
Ergeben sich die
Eigenwerte: λ_1, λ_2
und die Eigenvektoren: b_1, b_2

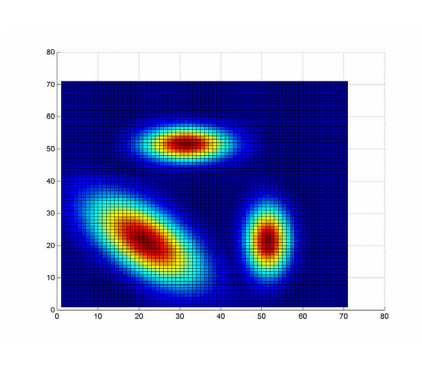


Der Bayes-Klassifikator

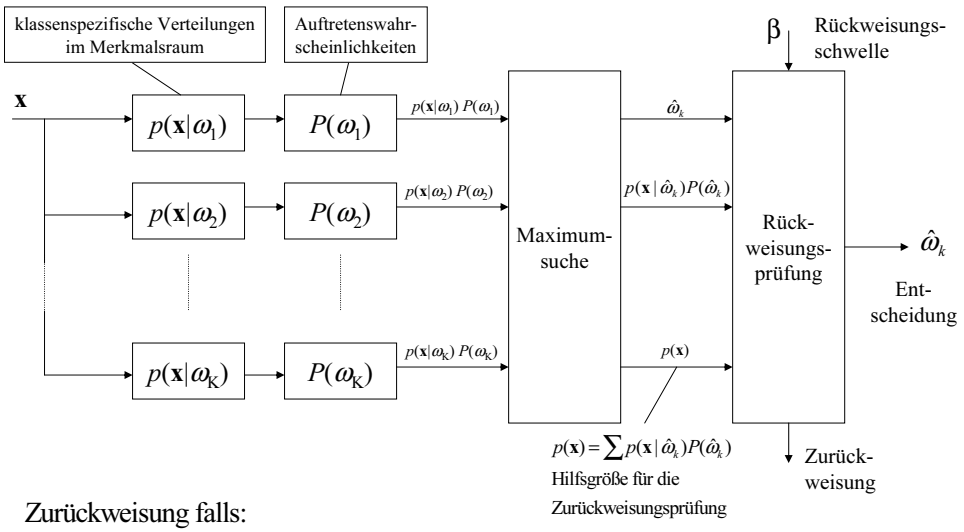
$$\max_{\omega_i} P(\omega_i \mid x)$$

Annahme: Klassenspezifische Gauß-Verteilungen





Optimales Erkennungssystem



Zurückweisung falls:

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

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

Falls Wahrscheinlichkeit einen zu geringen Wert hat
➔ Zurückweisung (sonst wäre Entscheidung sehr unsicher)

Zur Positiv-Definitheit der Kovarianzmatrix **K**

Man erwartet, dass die Beobachtungen des Zufallsprozesses unabhängig sind.

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

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

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

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

Der singuläre Fall Q=0 bedeutet, dass der Zufallsprozess $(\mathbf{x}-\boldsymbol{\mu}) \perp \mathbf{z}$ 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 $(\mathbf{x}-\boldsymbol{\mu}) \perp \mathbf{z}$ 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(**K**)>0
- $\mathbf{K}^{-1}$ ist ebenfalls positiv definit
- det($\mathbf{K}^{-1}$)>0
- Die Eigenwerte von **K** sind positiv

Klassenweise 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 $\{p(\mathbf{x}|\omega_k)\}$:

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

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

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:

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

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, Linien)

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