12. Mai 2017

Bayes‘sche Theorie II

Bayes‘sche Entscheidungstheorie

Das Verhältnis

$$\frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)}$$

wird likelihood ratio genannt und hängt nur von $\mathbf{x}$ ab.

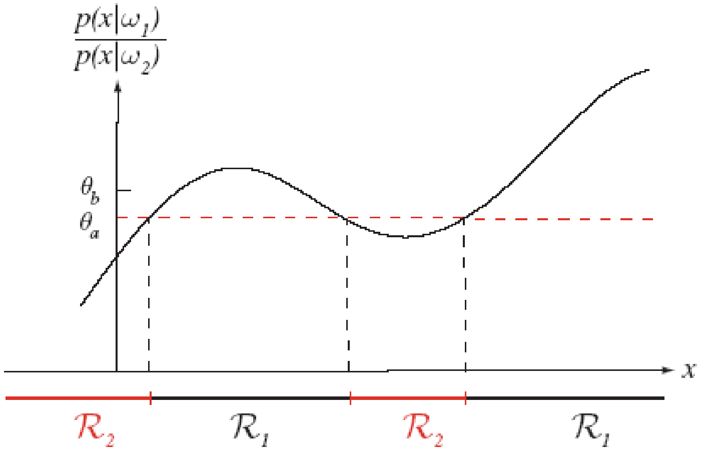


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the larger threshold θ_b, and hence $\mathcal{R}_1$ becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Minimum Error Classification

Verwenden wir als Verlustfunktion die symmetrische Verlustfunktion (oder zero-one loss function),

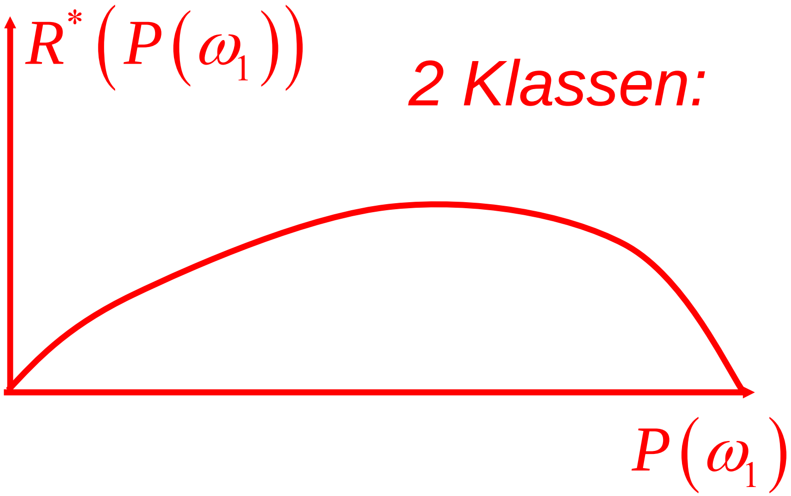
$$\lambda\left(\alpha_i \mid \omega_j\right) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

so fällt jede Fehlklassifizierung gleich ins Gewicht, und wir minimieren die Gesamtfehlerwahrscheinlichkeit.

Minimax Estimation

Gesamtrisiko hängt ab von:

- $p(\mathbf{x} \mid \omega_i)$
- $\lambda(\alpha_i \mid \omega_j)$
- $P(\omega_i)$



Sind a-priori-Wahrscheinlichkeiten nicht bekannt, so kann man das maximale Gesamtrisiko minimieren, das bei einer beliebigen a-priori-Wahrscheinlichkeit auftreten kann.

Minimax Estimation

Beispiel: 2 Klassen

Gesamtrisiko kann geschrieben werden als:

$$\boxed{\begin{aligned} R\left(P\left(\omega_1\right)\right) = & \int_{\mathcal{R}_1} \lambda_{11} P\left(\omega_1\right) p\left(\mathbf{x} \mid \omega_1\right) + \lambda_{12} P\left(\omega_2\right) p\left(\mathbf{x} \mid \omega_2\right) d\mathbf{x} \\ & + \int_{\mathcal{R}_2} \lambda_{21} P\left(\omega_1\right) p\left(\mathbf{x} \mid \omega_1\right) + \lambda_{22} P\left(\omega_2\right) p\left(\mathbf{x} \mid \omega_2\right) d\mathbf{x} \end{aligned}}$$

Umformung mit:

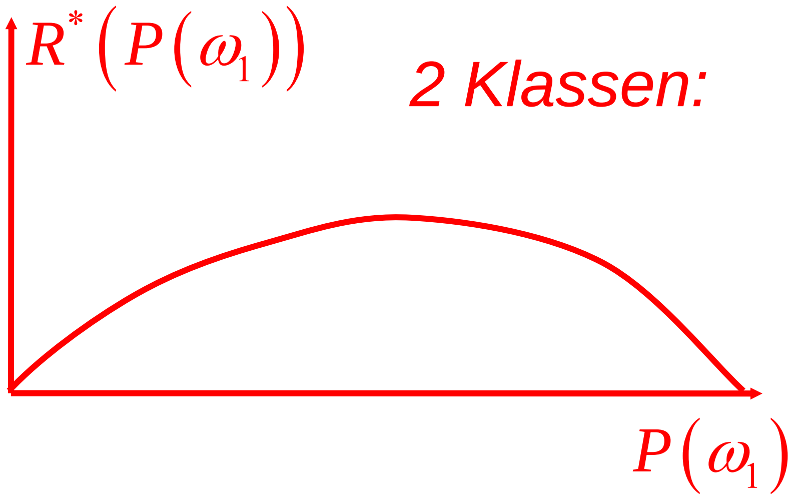
$$P\left(\omega_1\right) + P\left(\omega_2\right) = 1 \quad \text{und} \quad \int_{\mathcal{R}_1} p\left(\mathbf{x} \mid \omega_i\right) d\mathbf{x} + \int_{\mathcal{R}_2} p\left(\mathbf{x} \mid \omega_i\right) d\mathbf{x} = 1$$

Minimax Estimation

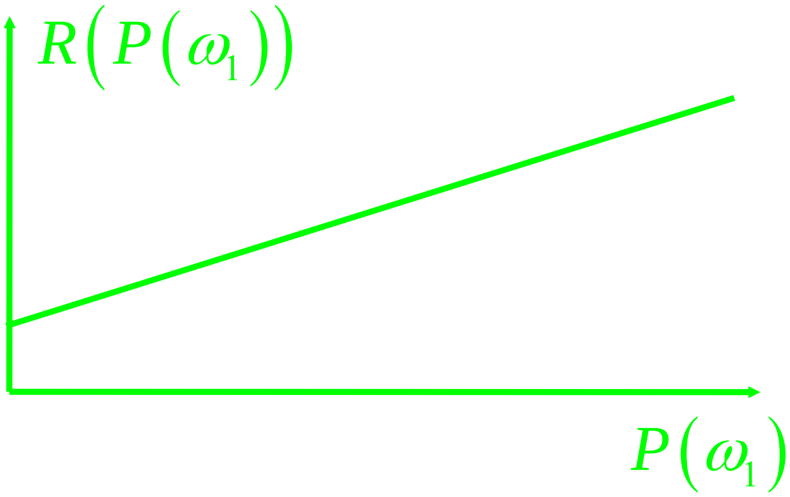
Gesamtrisiko kann geschrieben werden als:

$$R\left(P\left(\omega_1\right)\right)=\boxed{\lambda_{22}+\left(\lambda_{12}-\lambda_{22}\right)\int\limits_{\mathcal{R}_1} p\left(\mathbf{x}\mid\omega_2\right)d\mathbf{x}}$$

$$+P\left(\omega_1\right)\boxed{\left[\left(\lambda_{11}-\lambda_{22}\right)+\left(\lambda_{21}-\lambda_{11}\right)\int\limits_{\mathcal{R}_2} p\left(\mathbf{x}\mid\omega_1\right)d\mathbf{x}-\left(\lambda_{12}-\lambda_{22}\right)\int\limits_{\mathcal{R}_1} p\left(\mathbf{x}\mid\omega_2\right)d\mathbf{x}\right]}$$



2 Klassen:



Minimax Estimation

12. Mai 2017

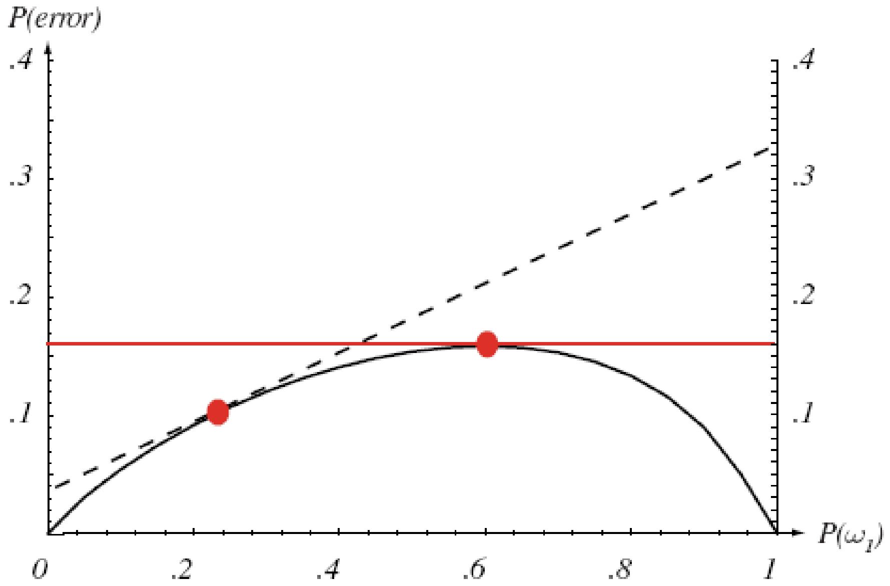


FIGURE 2.4. The curve at the bottom shows the minimum (Bayes) error as a function of prior probability $P(\omega_1)$ in a two-category classification problem of fixed distributions. For each value of the priors (e.g., $P(\omega_1) = 0.25$) there is a corresponding optimal decision boundary and associated Bayes error rate. For any (fixed) such boundary, if the priors are then changed, the probability of error will change as a linear function of $P(\omega_1)$ (shown by the dashed line). The maximum such error will occur at an extreme value of the prior, here at $P(\omega_1) = 1$. To minimize the maximum of such error, we should design our decision boundary for the maximum Bayes error (here $P(\omega_1) = 0.6$), and thus the error will not change as a function of prior, as shown by the solid red horizontal line. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Minimax Estimation

Gesamtrisiko kann geschrieben werden als:

Minimax Risiko

$$R\left(P\left(\omega_1\right)\right)=\boxed{\lambda_{22}+\left(\lambda_{12}-\lambda_{22}\right)\int\limits_{\mathcal{R}_1} p\left(\mathbf{x}\,|\,\omega_2\right)d\mathbf{x}}$$

$$+P\left(\omega_1\right)\boxed{\left[\left(\lambda_{11}-\lambda_{22}\right)+\left(\lambda_{21}-\lambda_{11}\right)\int\limits_{\mathcal{R}_2} p\left(\mathbf{x}\,|\,\omega_1\right)d\mathbf{x}-\left(\lambda_{12}-\lambda_{22}\right)\int\limits_{\mathcal{R}_1} p\left(\mathbf{x}\,|\,\omega_2\right)d\mathbf{x}\right]}$$

=0 für Minimax-Lösung

Diskriminanzfunktionen

Entscheidung für ω_i , wenn $g_i(\mathbf{x}) > g_j(\mathbf{x})$

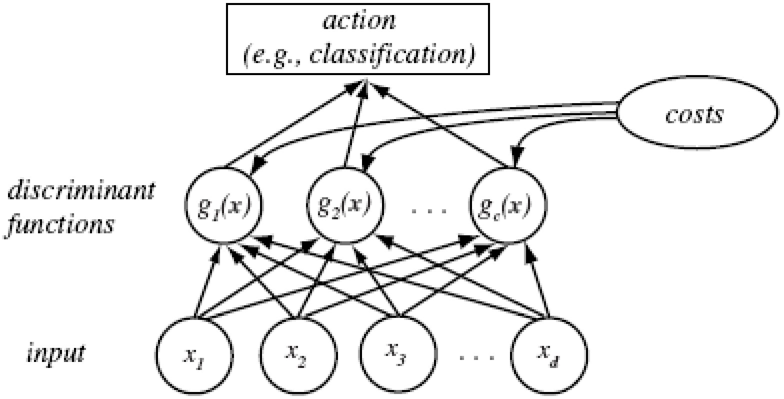


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bayes Klassifikation z.B. mit: $\boxed{g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x})}$

Diskriminanzfunktionen

12. Mai 2017

Für kleinste Fehlerwahrscheinlichkeit (zero-one risk):

$$\boxed{g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i) P(\omega_i)}{\sum_j p(\mathbf{x} \mid \omega_j) P(\omega_j)}}$$

oder auch: $\boxed{g_i(\mathbf{x}) = p(\mathbf{x} \mid \omega_i) P(\omega_i)}$

oder sogar: $\boxed{g_i(\mathbf{x}) = \ln p(\mathbf{x} \mid \omega_i) + \ln P(\omega_i)}$

Diskriminanzfunktionen

Diskriminanzfunktion im Fall von zwei Kategorien (dichotomizer):

- es genügt eine Diskriminanzfunktion:

$$\boxed{g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})}$$

- Entscheidung für ω_1 , wenn $g(\mathbf{x}) > 0$

- Aus den Funktionen auf S.10 wird dann:

$$\boxed{g(\mathbf{x}) = P(\omega_1 \mid \mathbf{x}) - P(\omega_2 \mid \mathbf{x})}$$ bzw. $$\boxed{g(\mathbf{x}) = \ln \frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}}$$

Diskriminanzfunktionen

Beispiel für zwei Kategorien:

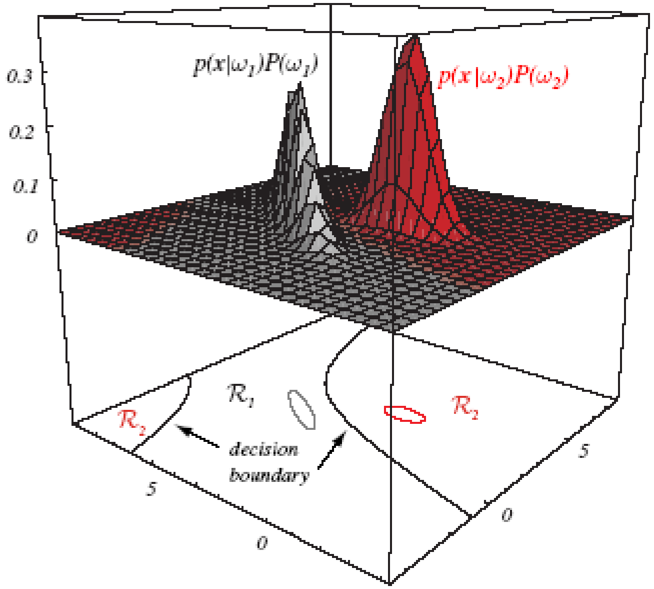


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Diskriminanzfunktionen bei Normalverteilung

12. Mai 2017

Diskriminanzfunktionen bei d-dimensionalen Normalverteilungen

$$\boxed{p(\mathbf{x}\,|\,\omega_i)=\frac{1}{(2\pi)^{d/2}\,|\mathbf{C}_i|^{1/2}}\,e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\mathbf{C}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}=N(\boldsymbol{\mu}_i,\mathbf{C}_i)}$$

- wir verwenden die logarithmierte Version:

$$\boxed{g_i(\mathbf{x})=\ln p(\mathbf{x}\,|\,\omega_i)+\ln P(\omega_i)}$$

- und erhalten:

$$\boxed{g_i(\mathbf{x})=-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\,\mathbf{C}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)-\cancel{\frac{d}{2}\ln(2\pi)}-\frac{1}{2}\ln|\mathbf{C}_i|+\ln P(\omega_i)}$$

Diskriminanzfunktionen bei Normalverteilung

Spezialfall 1: $\boxed{\mathbf{C}_i = \sigma^2 \mathbf{I}}$

- statistische Unabhängigkeit der Merkmale,
- gleiche Varianzen

$$\boxed{g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{C}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) \xcancel{-\frac{d}{2}\ln(2\pi)} \xcancel{-\frac{1}{2}\ln|\mathbf{C}_i|} + \ln P(\omega_i)}$$

$$\boxed{g_i(\mathbf{x}) = -\frac{\|\mathbf{x}-\boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(\omega_i)}$$

Diskriminanzfunktionen bei Normalverteilung

12. Mai 2017

Berechnung von $\left\|\mathbf{x}-\boldsymbol{\mu}_i\right\|^2$ ist nicht notwendig.
Zwar ist $g_i(\mathbf{x})$ quadratisch in $\mathbf{x}$:

$$\boxed{\begin{aligned} g_i(\mathbf{x}) &= -\frac{\left\|\mathbf{x}-\boldsymbol{\mu}_i\right\|^2}{2\sigma^2} + \ln P(\omega_i) \\ &= -\frac{1}{2\sigma^2}\left[\mathbf{x}^T\mathbf{x} - 2\boldsymbol{\mu}_i^T\mathbf{x} + \boldsymbol{\mu}_i^T\boldsymbol{\mu}_i\right] + \ln P(\omega_i) \end{aligned}}$$

Der quadratische Term $\mathbf{x}^T\mathbf{x}$ ist allerdings unabhängig von i, und kann daher weggelassen werden.

Diskriminanzfunktionen bei Normalverteilung

Ergebnis ist eine lineare Diskriminanzfunktion:

$$\boxed{g_i(\mathbf{x}) = \mathbf{w}_i{}^T \mathbf{x} + w_{i0}}$$

wobei

$$\boxed{\begin{aligned} \mathbf{w}_i &= \frac{1}{\sigma^2} \boldsymbol{\mu}_i \\ w_{i0} &= -\frac{1}{2\sigma^2} \boldsymbol{\mu}_i{}^T \boldsymbol{\mu}_i + \ln P(\omega_i) \end{aligned}}$$

w_{i0} wird Schwelle genannt (threshold / bias)

Diskriminanzfunktionen bei Normalverteilung

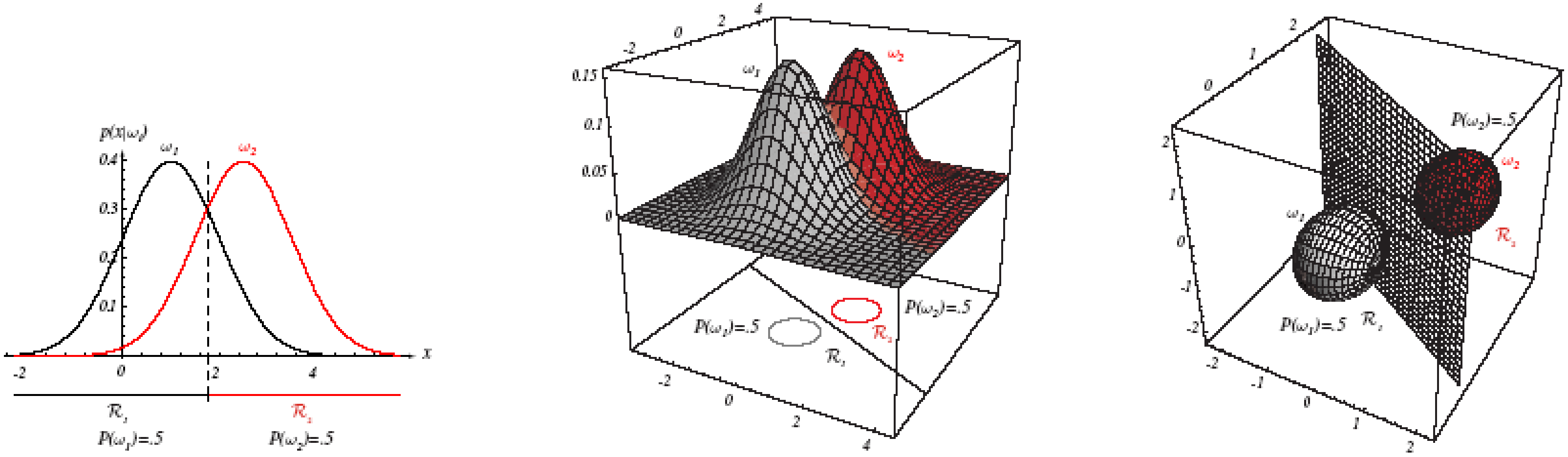


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-, and three-dimensional examples, we indicate $p(\mathbf{x}|\omega_i)$ and the boundaries for the case $P(\omega_1) = P(\omega_2)$. In the three-dimensional case, the grid plane separates $\mathcal{R}_1$ from $\mathcal{R}_2$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Diskriminanzfunktionen bei Normalverteilung

Für die Entscheidungsfunktion bei (je) zwei Kategorien erhalten wir die Differenzfunktion

$$\boxed{\begin{aligned} g(\mathbf{x}) &= g_i(\mathbf{x}) - g_j(\mathbf{x}) = \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) \\ \mathbf{w} &= \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \\ \mathbf{x}_0 &= \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \end{aligned}}$$

$g(\mathbf{x}) = 0$ ist eine Ebenengleichung

Diskriminanzfunktionen bei Normalverteilung

Verschiebung der Trennebene bei verschiedenen a-priori-Wahrscheinlichkeiten

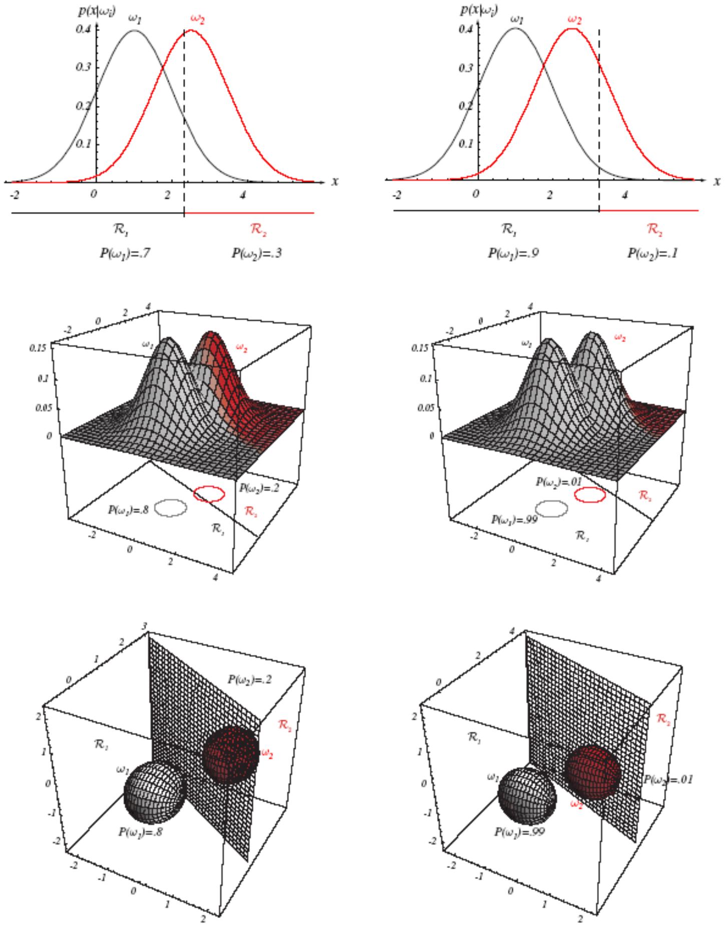


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficient disparate priors the boundary will not lie between the means of these one-, two- ar three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley Sons, Inc.

Diskriminanzfunktionen bei Normalverteilung

Spezialfall 2: $\boxed{\mathbf{C}_i = \mathbf{C}}$

- gleiche Kovarianzmatrizen

$$\boxed{g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \cancel{\frac{d}{2}\ln(2\pi)} - \cancel{\frac{1}{2}\ln|\mathbf{C}|} + \ln P(\omega_i)}$$

Auch hier ist der quadratische Term $\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}$

unabhängig von i und kann weggelassen werden

Diskriminanzfunktionen bei Normalverteilung

Ergebnis ist wieder eine lineare Diskriminanzfunktion:

$$\boxed{g_i(\mathbf{x}) = \mathbf{w}_i{}^T \mathbf{x} + w_{i0}}$$

wobei

$$\boxed{\begin{aligned} \mathbf{w}_i &= \mathbf{C}^{-1}\boldsymbol{\mu}_i \\ w_{i0} &= -\frac{1}{2}\boldsymbol{\mu}_i{}^T \mathbf{C}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i) \end{aligned}}$$

Diskriminanzfunktionen bei Normalverteilung

Wieder erhalten wir Entscheidungsebenen:

$$g(\mathbf{x}) = g_i(\mathbf{x}) - g_j(\mathbf{x}) = \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$$

$$\mathbf{w} = \mathbf{C}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{1}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \mathbf{C}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)} \ln\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

Diskriminanzfunktionen bei Normalverteilung

Schräge Trennebenen bei beliebigen (aber untereinander gleichen) mehrdimensionalen Gaußverteilungen

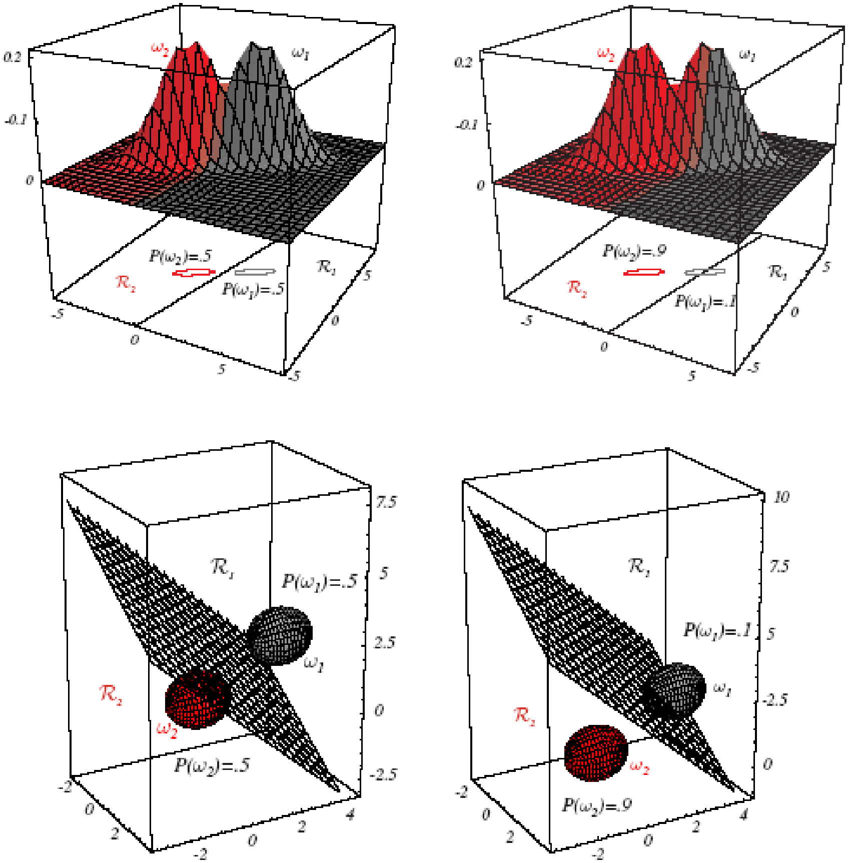


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions a ellipsoidal surfaces in three dimensions) and decision regions for equal but asymn ric Gaussian distributions. The decision hyperplanes need not be perpendicular to line connecting the means. From: Richard O. Duda, Peter E. Hart, and David G. St *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Diskriminanzfunktionen bei Normalverteilung

Allgemeiner Fall (3): $\boxed{\mathbf{C}_i = \text{beliebig}}$

- Verschiedene Gaußverteilungen

$$\boxed{g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{C}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \xcancel{\frac{d}{2}\ln(2\pi)} - \frac{1}{2}\ln|\mathbf{C}_i| + \ln P(\omega_i)}$$

Diskriminanzfunktion bleibt quadratische Funktion:

$\boxed{g_i(\mathbf{x}) = \mathbf{x}^T \mathbf{W}^{-1}\mathbf{x} + \mathbf{w}_i^T \mathbf{x} + w_{i0}}$, wobei $\boxed{\mathbf{W}^{-1} = \mathbf{C}^{-1}}$

Diskriminanzfunktionen bei Normalverteilung

12. Mai 2017

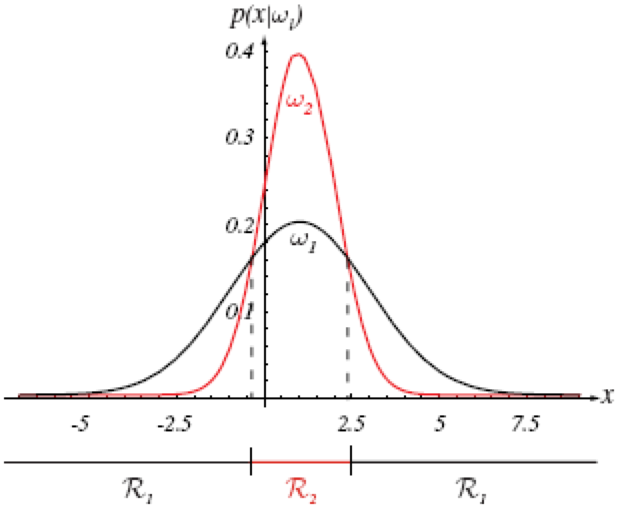


FIGURE 2.13. Non-simply connected decision regions can arise in one dimensions for Gaussians having unequal variance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Diskriminanzfunktionen bei Normalverteilung

12. Mai 2017

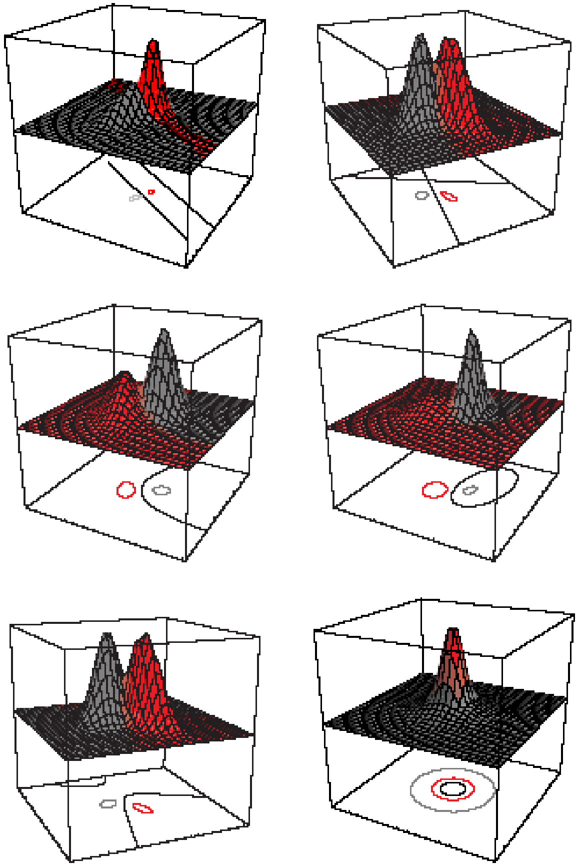


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability density. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

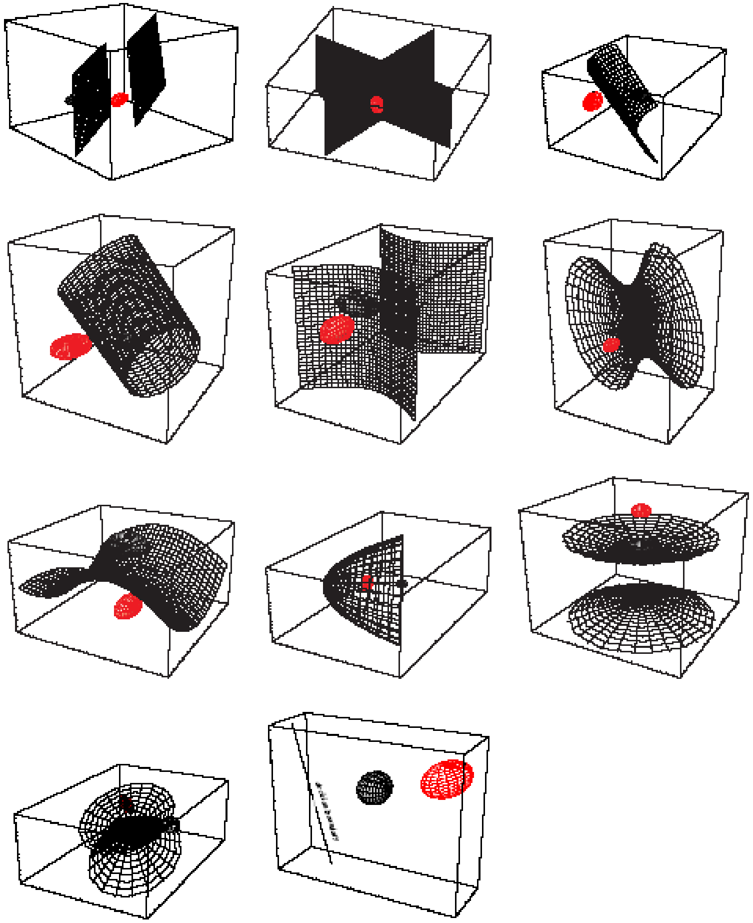


FIGURE 2.15. Arbitrary three-dimensional Gaussian distributions yield Bayes decision boundaries that are two-dimensional hyperquadrics. There are even degenerate cases in which the decision boundary is a line. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Fehler bei Klassifizierung

12. Mai 2017

Vermeidbare und unvermeidliche Fehler

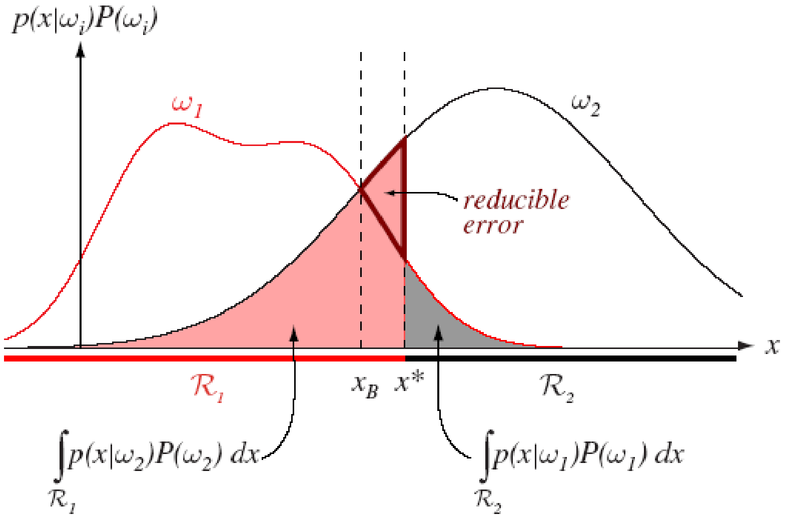


FIGURE 2.17. Components of the probability of error for equal priors and (nonoptimal) decision point x^*. The pink area corresponds to the probability of errors for deciding ω_1 when the state of nature is in fact ω_2; the gray area represents the converse, as given in Eq. 70. If the decision boundary is instead at the point of equal posterior probabilities, x_B, then this reducible error is eliminated and the total shaded area is the minimum possible; this is the Bayes decision and gives the Bayes error rate. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Signalentdeckungstheorie

Annahme:

- Zwei Normalverteilungen, gleiche Varianz.
- rot: Signal
- schwarz: kein Signal

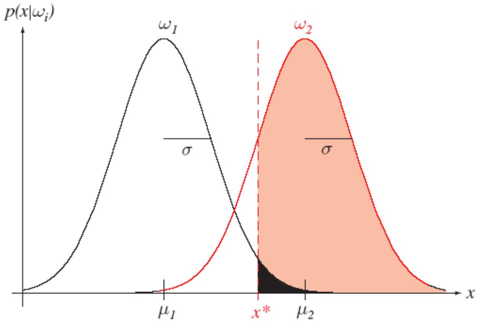


FIGURE 2.19. During any instant when no external pulse is present, the probability density for an internal signal is normal, that is, $p(x|\omega_1) \sim N(\mu_1, \sigma^2)$; when the external signal is present, the density is $p(x|\omega_2) \sim N(\mu_2, \sigma^2)$. Any decision threshold x^* will determine the probability of a hit (the pink area under the ω_2 curve, above x^*) and of a false alarm (the black area under the ω_1 curve, above x^*). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Maß für Unterscheidbarkeit: Diskriminationsfähigkeit (discriminability)

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

Signalentdeckungstheorie

Gegeben: Zustand + Entscheidung

Unbekannt: $\mu_1, \mu_2, \sigma, x^*$

Wir messen

1. Trefferrate:

$$P\left(x > x^* \mid x \in \omega_2\right)$$

2. Rate falscher Alarme:

$$P\left(x > x^* \mid x \in \omega_1\right)$$

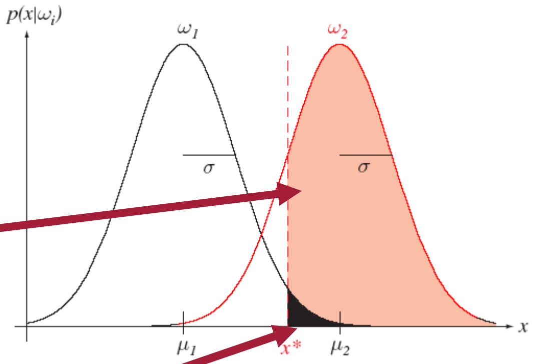


FIGURE 2.19. During any instant when no external pulse is present, the probability density for an internal signal is normal, that is, $p(x|\omega_1) \sim N(\mu_1, \sigma^2)$; when the external signal is present, the density is $p(x|\omega_2) \sim N(\mu_2, \sigma^2)$. Any decision threshold x^* will determine the probability of a hit (the pink area under the ω_2 curve, above x^*) and of a false alarm (the black area under the ω_1 curve, above x^*). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Signalentdeckungstheorie

Änderung der Schwelle ergibt verschiedene Punkte auf der Receiver Operating Characteristic (ROC), die charakteristisch für Systemeigenschaften ist.

Wenn Verteilung Normalverteilung ist läßt sich d' berechnen.

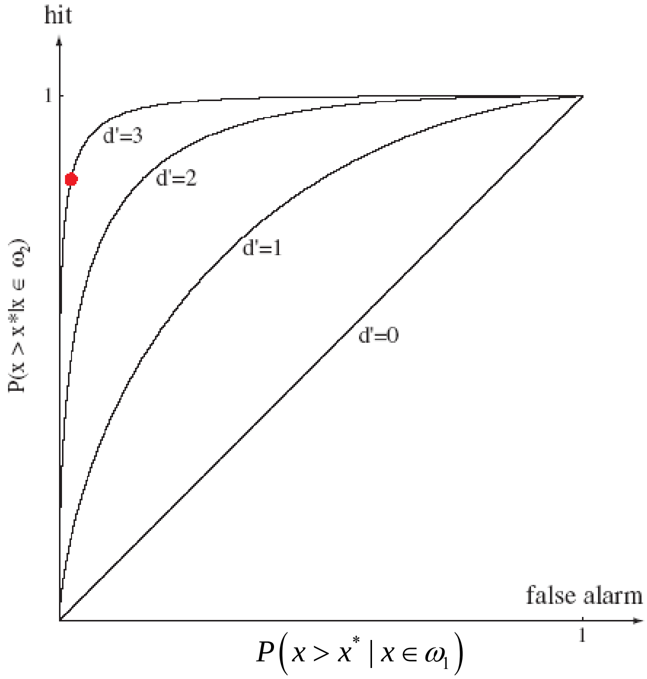


FIGURE 2.20. In a receiver operating characteristic (ROC) curve, the abscissa is the probability of false alarm, $P(x > x^*|x \in \omega_1)$, and the ordinate is the probability of hit, $P(x > x^*|x \in \omega_2)$. From the measured hit and false alarm rates (here corresponding to x^* in Fig. 2.19 and shown as the red dot), we can deduce that $d' = 3$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Signalentdeckungstheorie

Allgemeine Receiver Operating Characteristic:

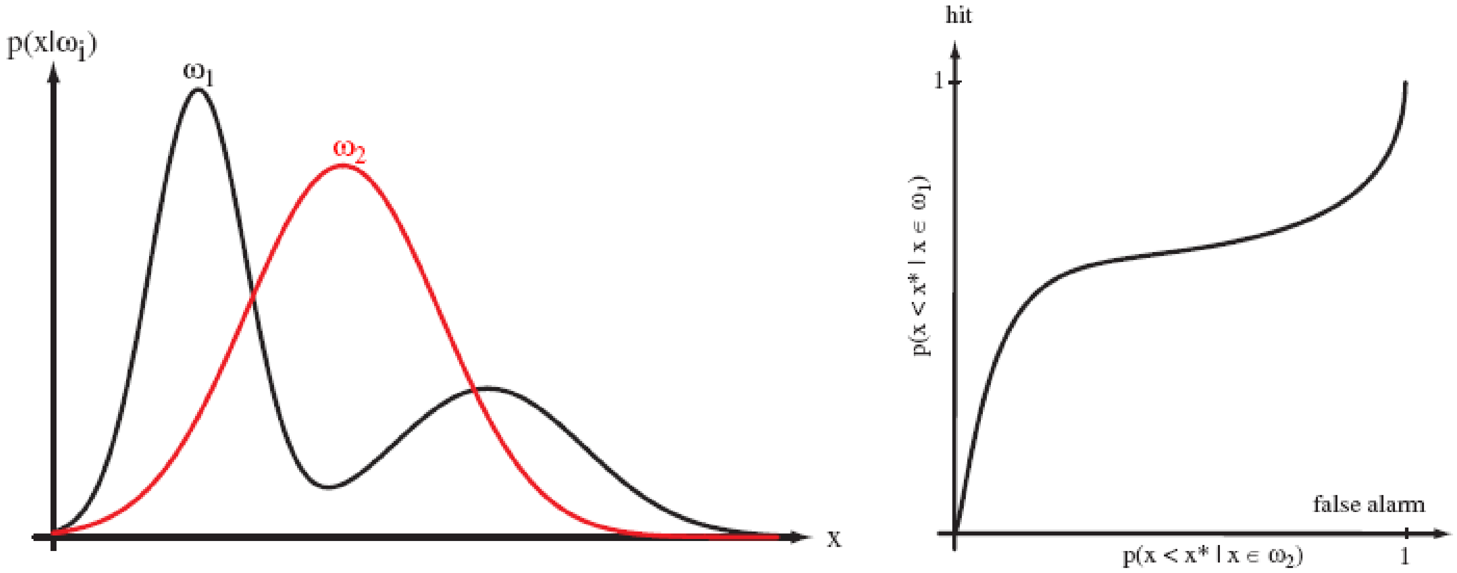


FIGURE 2.21. In a general operating characteristic curve, the abscissa is the probability of false alarm, $P(x \in \mathcal{R}_2|x \in \omega_1)$, and the ordinate is the probability of hit, $P(x \in \mathcal{R}_2|x \in \omega_2)$. As illustrated here, operating characteristic curves are generally not symmetric, as shown at the right. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Marginalisierung

Was tun bei fehlenden Features?

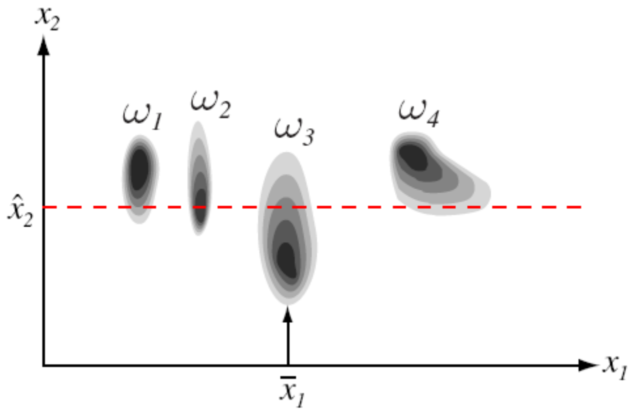


FIGURE 2.22. Four categories have equal priors and the class-conditional distributions shown. If a test point is presented in which one feature is missing (here, x_1) and the other is measured to have value $\hat{x}_2$ (red dashed line), we want our classifier to classify the pattern as category ω_2, because $p(\hat{x}_2|\omega_2)$ is the largest of the four likelihoods. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Lösung: Nicht Mittelwert annehmen, sondern

Marginalisierung

$$\boxed{p\left(\mathbf{x}_g\right) = \int p\left(\mathbf{x}_g, \mathbf{x}_b\right) d\mathbf{x}_b}$$