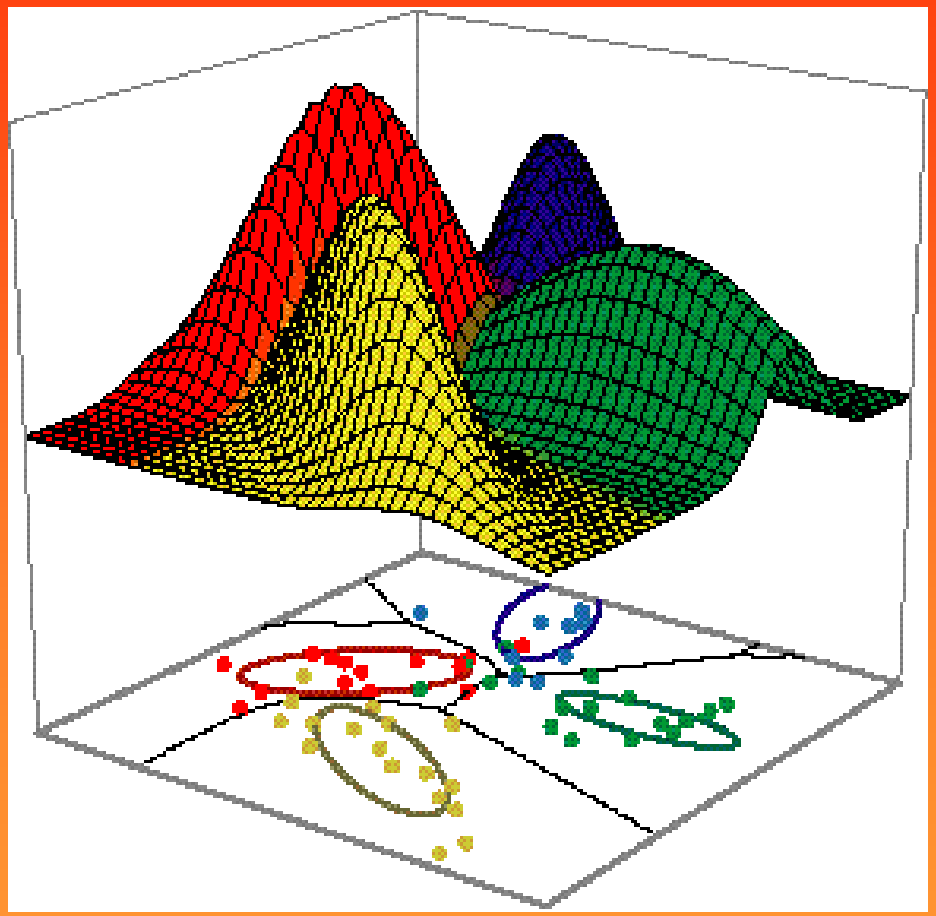


Pattern Classification

All materials in these slides were taken from Pattern Classification (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000
with the permission of the authors and the publisher

Chapter 2 (Part 1): Bayesowska teoria decyzji (Sections 2.1-2.2)

- Wstęp
- Bayesowska teoria decyzji – cechy ciągłe

Wstęp

- Przykład zadania rozpoznawania okoń / łosoś
  - Informacja aprioryczna
    - Wartość cechy obiektu jest realizacją zmiennej losowej
    - Prawdopodobieństwo złowienia ryb obu klas
      - $P(\omega_1)$, $P(\omega_2)$ (może być jednakowe $P(\omega_1)=P(\omega_2)$)
      - $P(\omega_1) + P(\omega_2) = 1$ (prawdopodobieństwo wszystkich zdarzeń)

- Reguła decyzyjna przy informacji apriorycznej
  - Wybierz klasę ω_1 jeżeli $P(\omega_1) > P(\omega_2)$, w przeciwnym przypadku wybierz ω_2
- Informacja w postaci rozkładów warunkowych
- $P(x \mid \omega_1)$ i $P(x \mid \omega_2)$ opisują prawdopodobieństwa wyników pomiaru jasności w populacjach sea and salmon

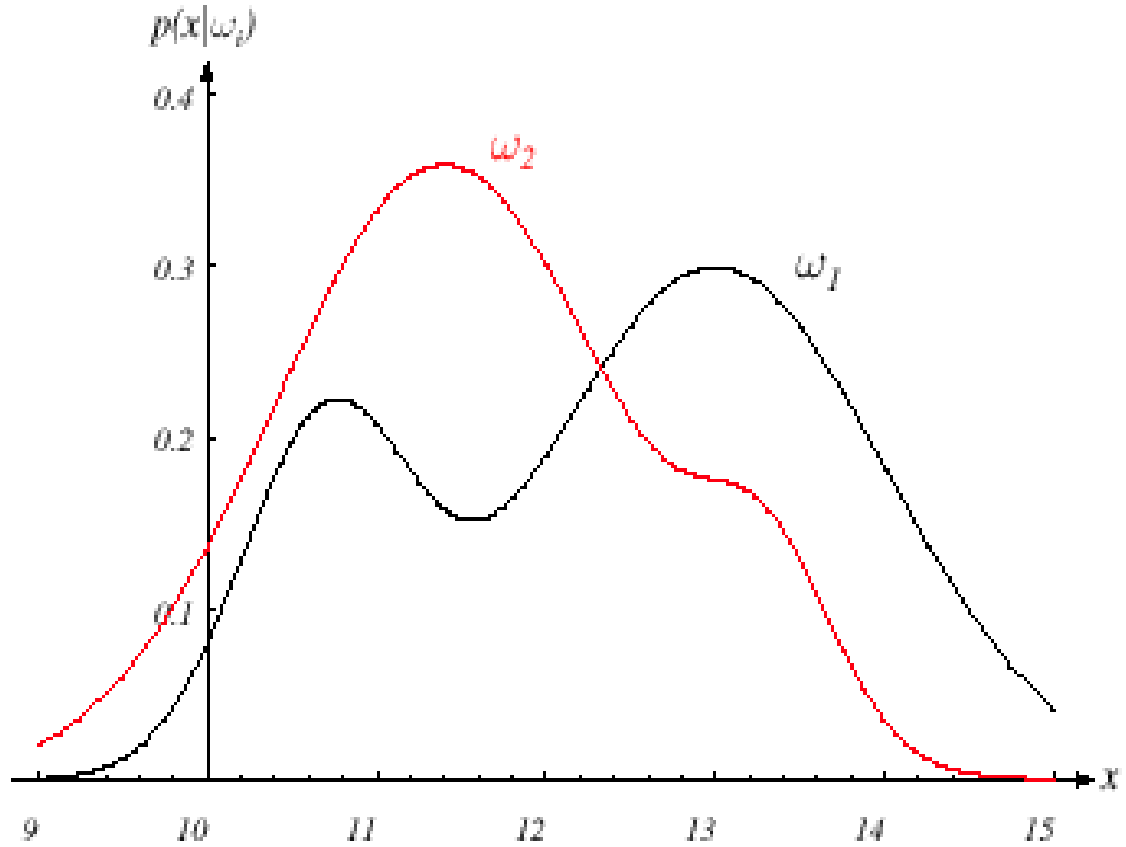


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Prawdopodobieństwa: a'posteriori, warunkowe, całkowite
  - $P(\omega_j \mid x) = P(x \mid \omega_j)\, P(\omega_j) / P(x)$
  - W przypadku dwóch klas

$$P(x) = \sum_{j=1}^{j=2} P(x \mid \omega_j) P(\omega_j)$$

  - P.a'posteriori = (P.warunkowe * P.a'priori) /P.całkowite

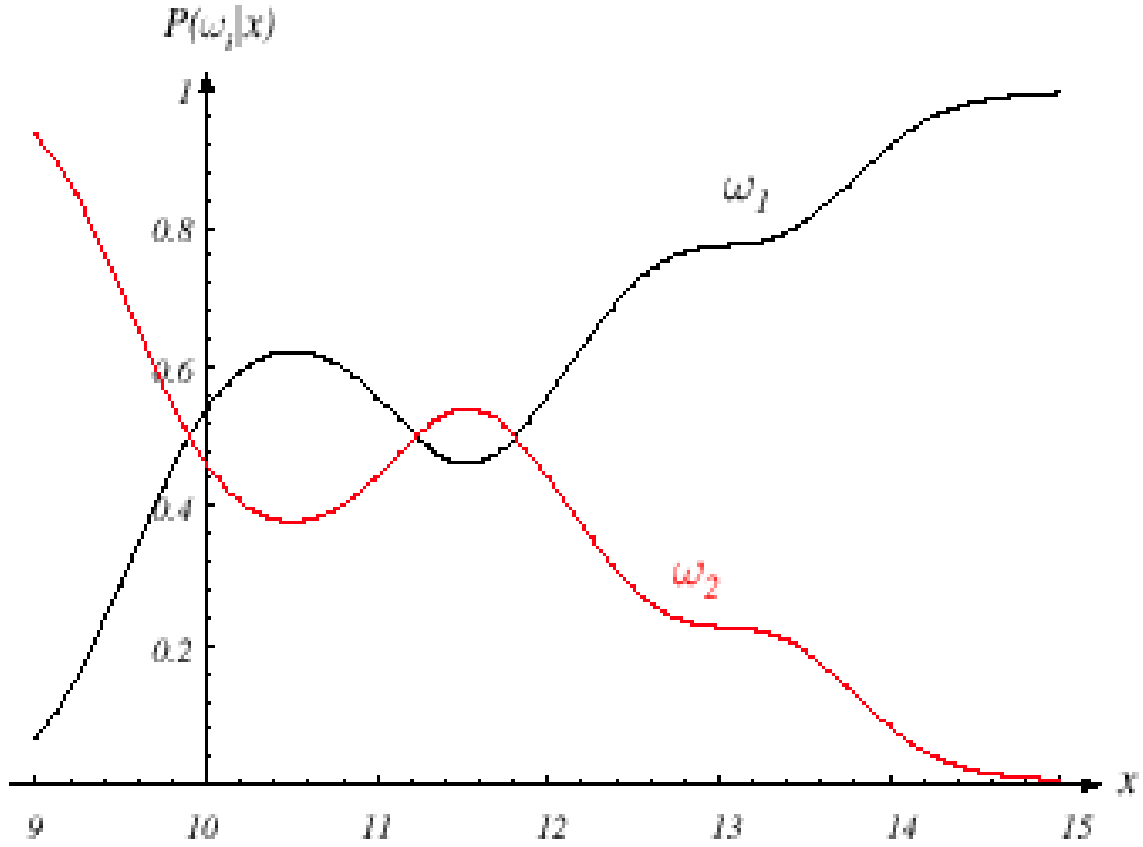


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.

- Reguła decyzyjna z informacją aposterioryczną

  X reprezentuje pomiar, dla którego:

  Jeżeli $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ ⇒ to klasa = ω_1
  Jeżeli $P(\omega_1 \mid x) < P(\omega_2 \mid x)$ ⇒ to klasa = ω_2

  Zatem, dla konkretnego pomiaru x, prawdopodobieństwo błędnej decyzji jest następujące:

  $P(\text{błąd} \mid x) = P(\omega_1 \mid x)$ *dla decyzji* ω_2
  $P(\text{błąd} \mid x) = P(\omega_2 \mid x)$ *dla decyzji* ω_1

- Minimalizacja prawdopodobieństwa błędnej decyzji przy regule decyzyjnej:

  Wybierz klasę ω_1 *jeżeli* $P(\omega_1 \mid x) > P(\omega_2 \mid x)$;
  w przeciwnym przypadku wybierz ω_2

  prowadzi do prawdopodobieństwa błędnej decyzji o postaci:

  $P(\text{błąd} \mid x) = \min [P(\omega_1 \mid x), P(\omega_2 \mid x)]$

  (optymalna decyzja Bayesa)

Bayesowska teoria decyzji – Cechy ciągłe

- Uogólnienie dotychczasowych rozważań
  - Pomiar wielu cech
  - Uwzględnienie wielu klas obiektów
  - Zdefiniowanie funkcji strat, która jest ogólniejsza od prawdopodobieństwa błędnej decyzji.

- Możliwe jest dopuszczenie decyzji odrzucenia obiektu

- Funkcja strat definiuje koszty związane z podjęciem błędnej decyzji

$\{\omega_1, \omega_2, \ldots, \omega_c\}$ – zbiór klas

$\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ – zbiór decyzji

$\lambda(\alpha_i \mid \omega_j)$ – funkcja strat związanych z podjęciem decyzji α_i dla obiektu z klasy ω_j

Ryzyko całkowite *R*

całka po wszystkich x $\underbrace{R(\alpha_i \mid x)}$

Ryzyko warunkowe

Minimalizacja R $\Leftrightarrow$ Minimalizacja $R(\alpha_i \mid x)$ *dla* $i = 1,\ldots, a$

$$R(\alpha_i \mid x) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

Dla i = 1,…,a

Podejmij taką decyzję α_i , dla której $R(\alpha_i \mid x)$ przyjmuje wartość najmniejszą

w takim przypadku R również przyjmuje wartość najmniejszą i jest nazywane

Ryzykiem bayesowskim

Bayesowska reguła decyzyjna prowadzi do najmniejszej możliwej wartości R

- Rozpoznawanie zero-jedynkowe

α_1 : *wybierz* ω_1

α_2 : *wybierz* ω_2

$\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$

to strata związana z decyzją α_i dla obiektu z klasy ω_j

Ryzyko warunkowe:

$$R(\alpha_1 \mid x) = \lambda_{11}P(\omega_1 \mid x) + \lambda_{12}P(\omega_2 \mid x)$$
$$R(\alpha_2 \mid x) = \lambda_{21}P(\omega_1 \mid x) + \lambda_{22}P(\omega_2 \mid x)$$

Reguła decyzyjna jest następująca:

Jeżeli $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$

Podejmowana jest decyzja α_1: "wybierz ω_1"

Równoważnie:

Wybierz ω_1 jeżeli:

$$(\lambda_{21} - \lambda_{11})\, P(x \mid \omega_1)\, P(\omega_1) > (\lambda_{12} - \lambda_{22})\, P(x \mid \omega_2)\, P(\omega_2)$$

w przeciwnym przypadku wybierz ω_2

Wskaźnik wiarygodności:

Poprzednio przestawiona reguła decyzyjna jest równoważna następującej:

Jeżeli $$\frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

to podejmij decyzję α_1 („wybierz ω_1”)

w przeciwnym przypadku podejmij decyzję α_2 („wybierz ω_2”)

Własność decyzji optymalnych

"Jeżeli wskaźnik wiarygodności przekracza wartość progową niezależnie od wartości cech obiektu x, możliwe jest podejmowanie optymalnych decyzji"

Ćwiczenie

Podejmij optymalną decyzję w następującym zadaniu:

$\Omega = \{\omega_1, \omega_2\}$

$P(x \mid \omega_1)$ → N(2, 0.5) (rozkład normalny)

$P(x \mid \omega_2)$ → N(1.5, 0.2)

$P(\omega_1) = 2/3$

$P(\omega_2) = 1/3$

$$\lambda = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

Chapter 2 (Part 2): Bayesowska teoria decyzji (Sections 2.3-2.5)

- Optymalny klasyfikator Bayesa
- Klasyfikatory, funkcje dyskryminujące i obszary decyzyjne
- Rozkład normalny

Optymalny klasyfikator Bayesa

- Decyzje dotyczące numeru klasy obiektu
  Jeżeli podjęto decyzję α_i dla obiektu klasy ω_j to:
  decyzja jest poprawna jeśli $i = j$, a błędna jeśli $i \neq j$

- Poszukujemy reguły decyzyjnej minimalizującej *prawdopodobieństwo błędnej klasyfikacji*

- Zero-jedynkowa funkcja strat:

$$\lambda(\alpha_i, \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1, \ldots, c$$

Ryzyko warunkowe przyjmuje postać:

$$R(\alpha_i \mid x) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

$$= \sum_{\substack{j=1 \\ j \neq i}}^{c} P(\omega_j \mid x) = 1 - P(\omega_i \mid x)$$

"Ryzyko związane z zerojedynkową funkcją strat jest średnim błędem klasyfikacji"

- Minimalizacja ryzyka sprowadza się do maksymalizacji $P(\omega_i \mid x)$

  (ponieważ $R(\alpha_i \mid x) = 1 - P(\omega_i \mid x)$)

- Wybierz ω_i *jeżeli* $P(\omega_i \mid x) > P(\omega_j \mid x) \;\; \forall j \neq i$

- Obszary decyzyjne :

$$\text{Niech } \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}.\frac{P(\omega_2)}{P(\omega_1)}=\theta_\lambda \text{ to wybierz } \omega_1 \text{ gdy } \frac{P(x\,|\,\omega_1)}{P(x\,|\,\omega_2)}>\theta_\lambda$$

- Jeśli λ jest zero-jedynkową funkcją strat:

$$\lambda=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}$$

$$\text{to } \theta_\lambda=\frac{P(\omega_2)}{P(\omega_1)}=\theta_a$$

$$\text{dla } \lambda=\begin{pmatrix}0 & 2\\ 1 & 0\end{pmatrix} \text{ mamy } \theta_\lambda=\frac{2P(\omega_2)}{P(\omega_1)}=\theta_b$$

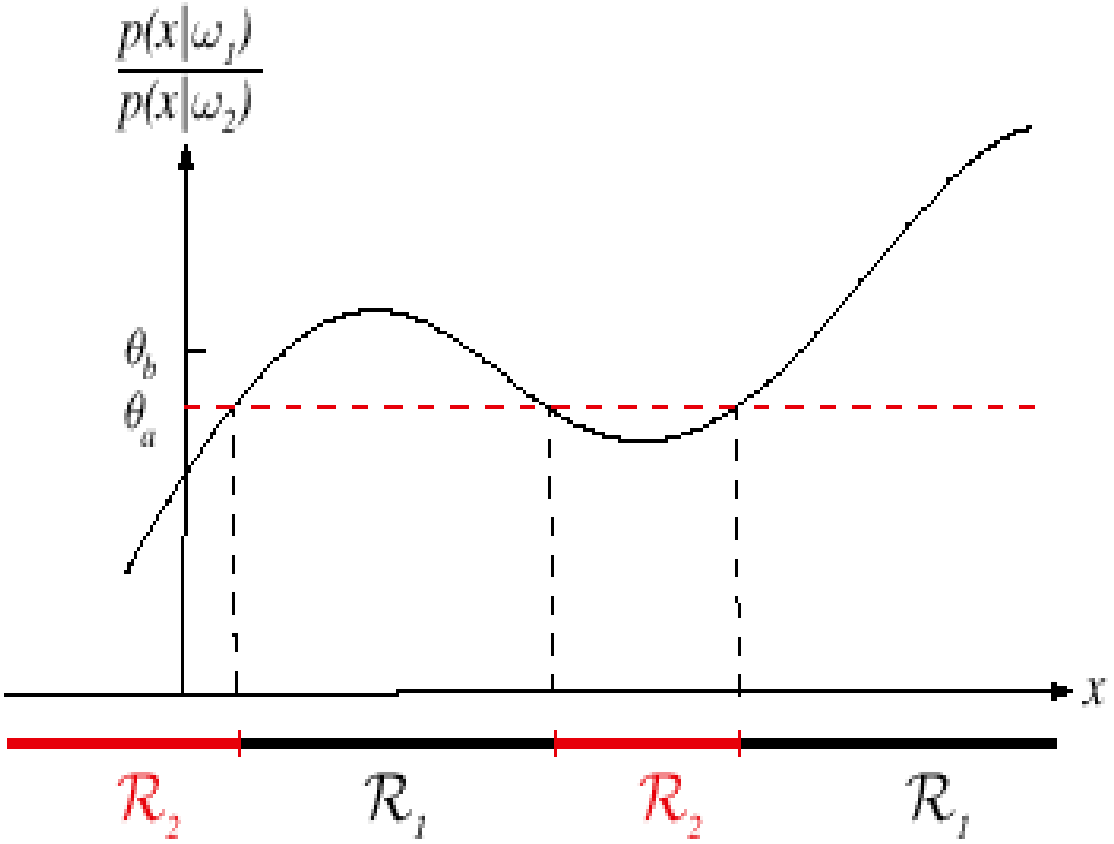


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, 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.

Klasyfikatory, funkcje dyskryminujące i obszary decyzyjne

- Przypadek wielu klas
  - Zbiór funkcji dyskryminujących: $g_i(x)$, $i = 1,\ldots,c$
  - Klasyfikator przyporządkowuje wektorowi cech x klasę ω_i jeżeli:

$$g_i(x) > g_j(x) \quad \forall j \neq i$$

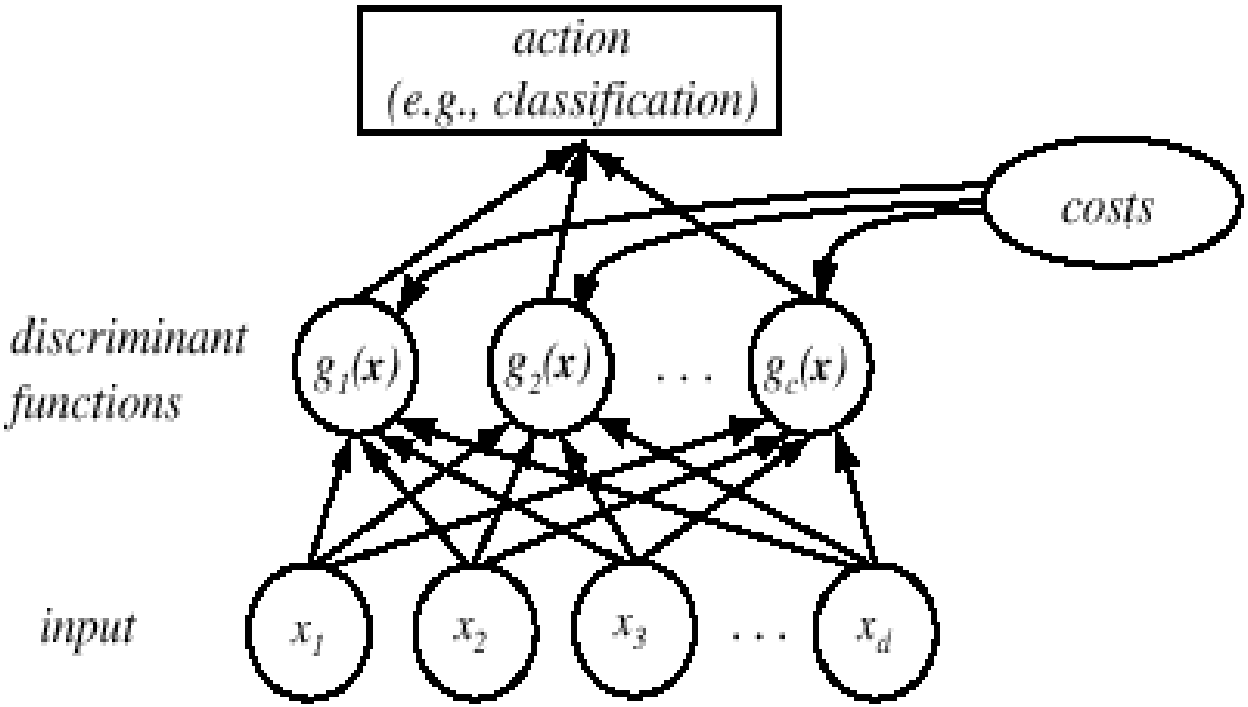


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction 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.

- *Niech* $g_i(x) = - R(\alpha_i \mid x)$

  (maksimum funkcji dyskryminującej odpowiada minimum funkcji ryzyka)

- W celu minimalizacji prawdopodobieństwa błędnej decyzji bierzemy

$$g_i(x) = P(\omega_i \mid x)$$

(maksimum funkcji dyskryminującej odpowiada maksimum rozkładu a'posteriori)

$$g_i(x) \equiv P(x \mid \omega_i)\, P(\omega_i)$$

$$g_i(x) = \ln P(x \mid \omega_i) + \ln P(\omega_i)$$

(ln: logarytm naturalny)

- Przestrzeń cech zostaje podzielona na c regionów decyzyjnych:

$$Jeżeli\ g_i(x) > g_j(x)\ \ \forall j \neq i\ to\ x\ leży\ w\ \mathcal{R}_i$$

($\mathcal{R}_i$ oznacza obszar decyzyjny klasy o numerze *i*)

- Przypadek dychotonomiczny (dwie klasy)
  - Klasyfikator ma do dyspozycji dwie funkcje g_1 i g_2

Niech $g(x) \equiv g_1(x) - g_2(x)$

Wybierz ω_1 *jeżeli* $g(x) > 0$; w przec. przypadku wybierz ω_2

- Wyznaczanie funkcji g(x)

$$g(x) = P(\omega_1 \mid x) - P(\omega_2 \mid x)$$

lub

$$g(x) = \ln \frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

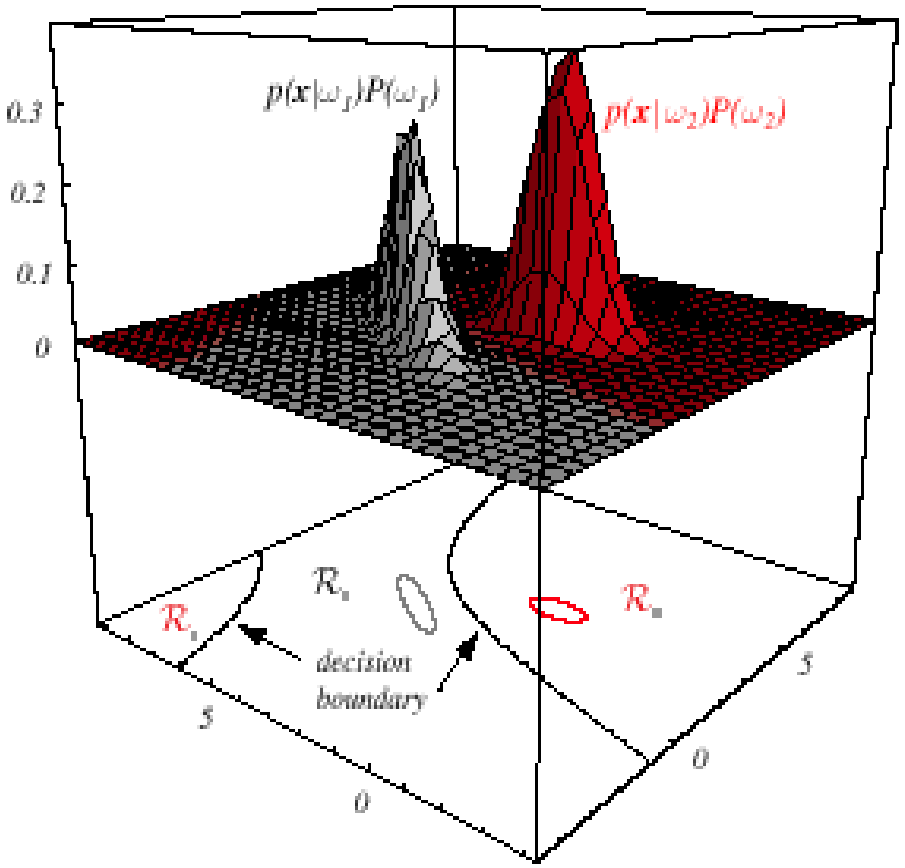


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution.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

Rozkład normalny

- Rozkład jednowymiarowy
  - Dogodny w obliczeniach analitycznych
  - Ciągły
  - Wiele procesów jest asymptotycznie normalnymi
  - Znaki pisane odręcznie, sygnały mowy można rozpatrywać jako idealne wzorce zakłócone przez proces losowy (centralne twierdzenie graniczne)

$$P(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right],$$

gdzie:

μ = średnia(lub wartość oczekiwana) zmiennej x

σ^2 = wariancja (kwadrat odchylenia standardowego)

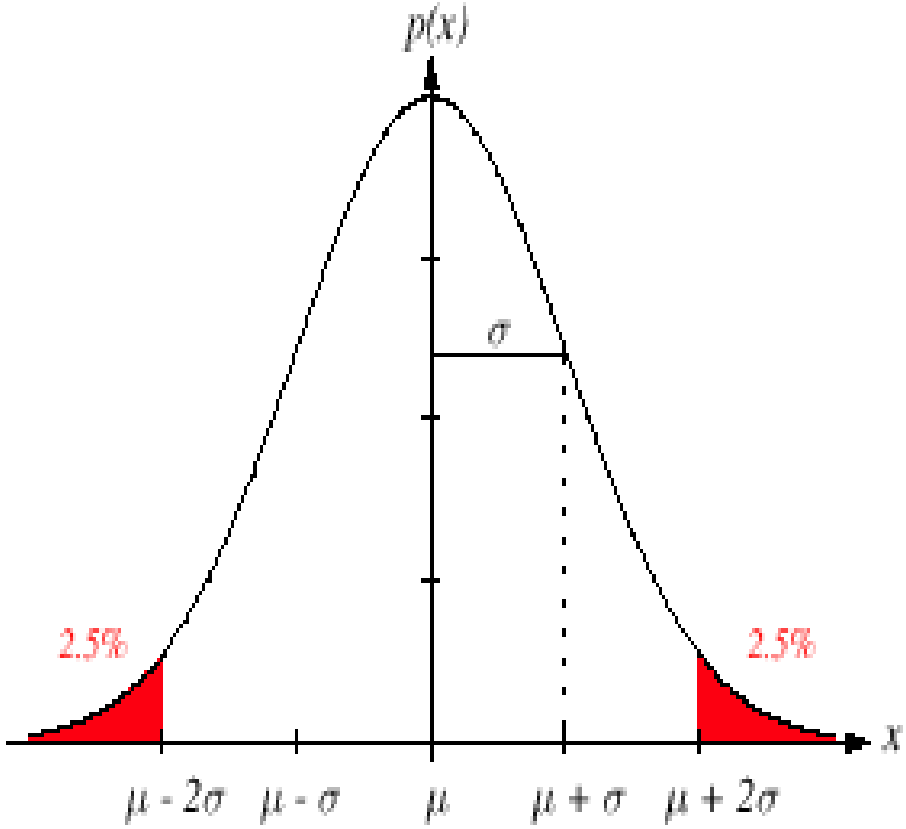


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Rozkład wielowymiarowy
  - d – wymiarowy rozkład normalny:

$$P(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]$$

gdzie:

$x = (x_1, x_2, \ldots, x_d)^T$ (T oznacza transpozycję)

$\mu = (\mu_1, \mu_2, \ldots, \mu_d)^T$ – wektor średnich

$\Sigma = d*d$ – macierz kowariancji

$|\Sigma|$ i Σ^{-1} oznaczają wyznacznik i odwrotność macierzy

Chapter 2 (part 3)
Bayesowska teoria decyzji
(Sections 2-6,2-9)

- Funkcje dyskryminujące dla rozkładu normalnego
- Bayesowska teoria decyzji – cechy dyskretne

Funkcje dyskryminujące dla rozkładu normalnego

- Optymalny klasyfikator Bayesa wiąże się z funkcjami dyskryminującymi o postaci:

  $g_i(x) = \ln P(x \mid \omega_i) + \ln P(\omega_i)$

- W przypadku wielowymiarowym:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

- Przypadek $\Sigma_i = \sigma^2.I$ (I oznacza macierz jednostkową)

$$g_i(x) = w_i^T x + w_{i0} \text{ (liniowa funkcja dyskrymin.)}$$

gdzie

$$w_i = \frac{\mu_i}{\sigma^2};\ w_{i0} = -\frac{1}{2\sigma^2}\mu_i^T \mu_i + \ln P(\omega_i)$$

(ω_{i0} jest progiem dla ***i*** - tej klasy!)

- Klasyfikator korzystający z liniowych funkcji dyskryminujących nazywany jest “klasyfikatorem liniowym”

- Powierzchnie decyzyjne klasyfikatora liniowego są fragmentami hiperpłaszczyzn zdefiniowanych jako:

$$g_i(x) = g_j(x)$$

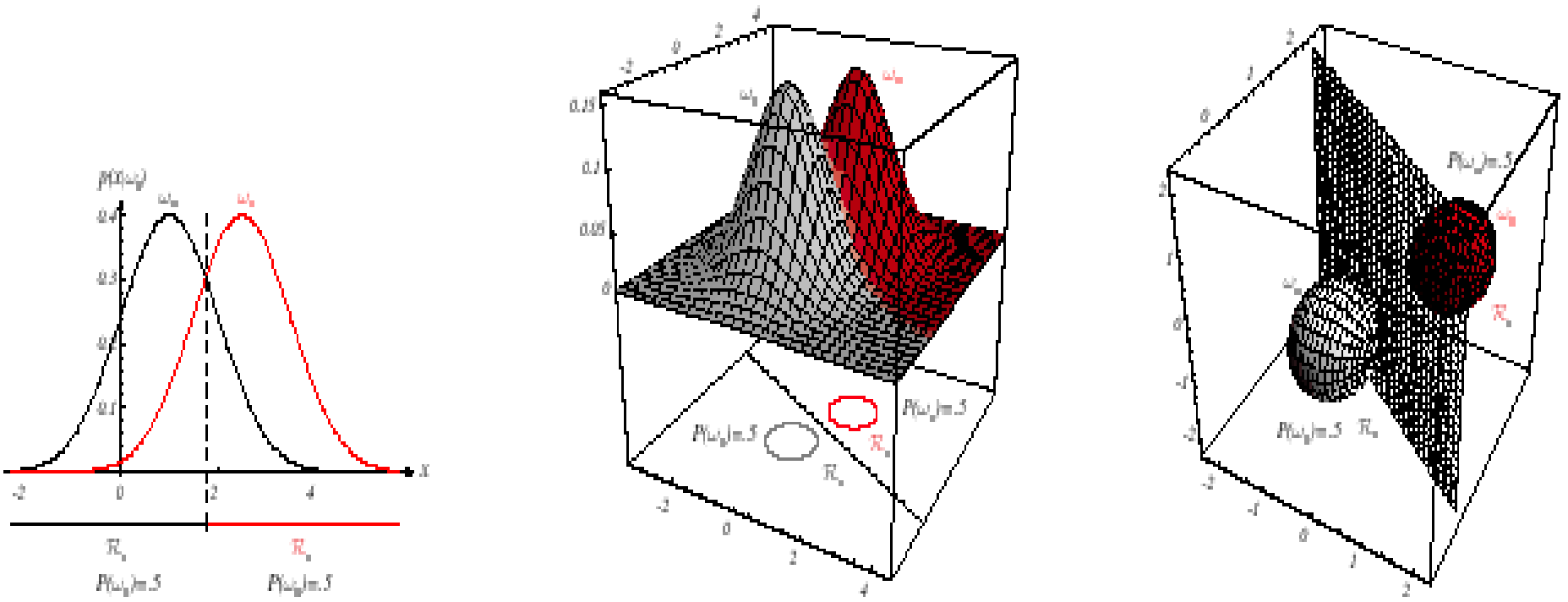


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-, 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.

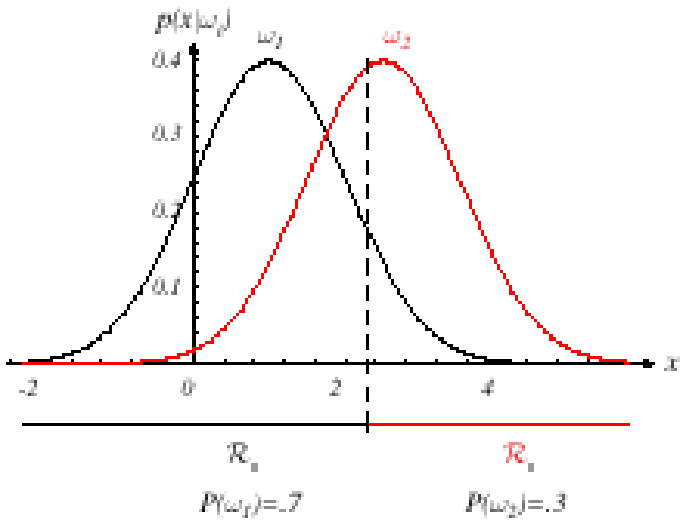
- Hiperpłaszczyzna rozdzielająca $\mathcal{R}_i$ *oraz* $\mathcal{R}_j$

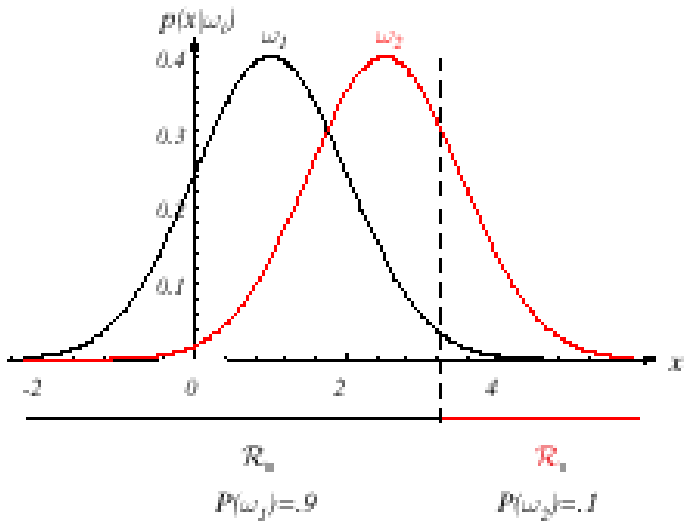
$$w^t\left(x - x_0\right) = 0, \quad gdzie\ w = \mu_i - \mu_j$$

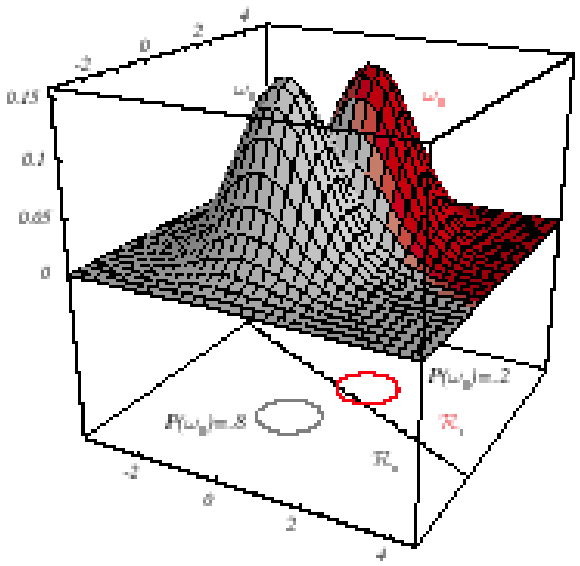
$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\left\|\mu_i - \mu_j\right\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)$$

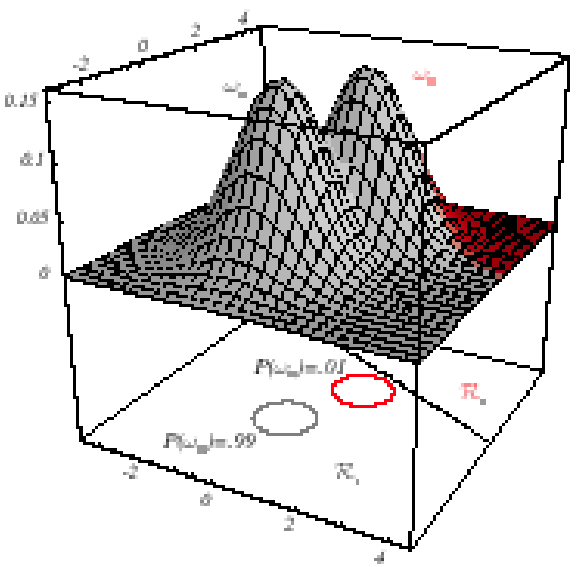
jest ortogonalna do prostej łączącej średnie

Jeżeli $P(\omega_i) = P(\omega_j)$ to $w_{i0} = \frac{1}{2}(\mu_i + \mu_j)$


p(x|ωj)
ω1
ω2
0.4
0.3
0.2
0.1
x
-2
0
2
4
P(ω1)=.7
P(ω2)=.3


p(x|ωj)
ω1
ω2
0.4
0.3
0.2
0.1
x
-2
0
2
4
P(ω1)=.9
P(ω2)=.1


0.15
0.1
0.05
0
P(ω2)=.2
P(ω1)=.8


0.15
0.1
0.05
0
P(ω2)=.01
P(ω1)=.99

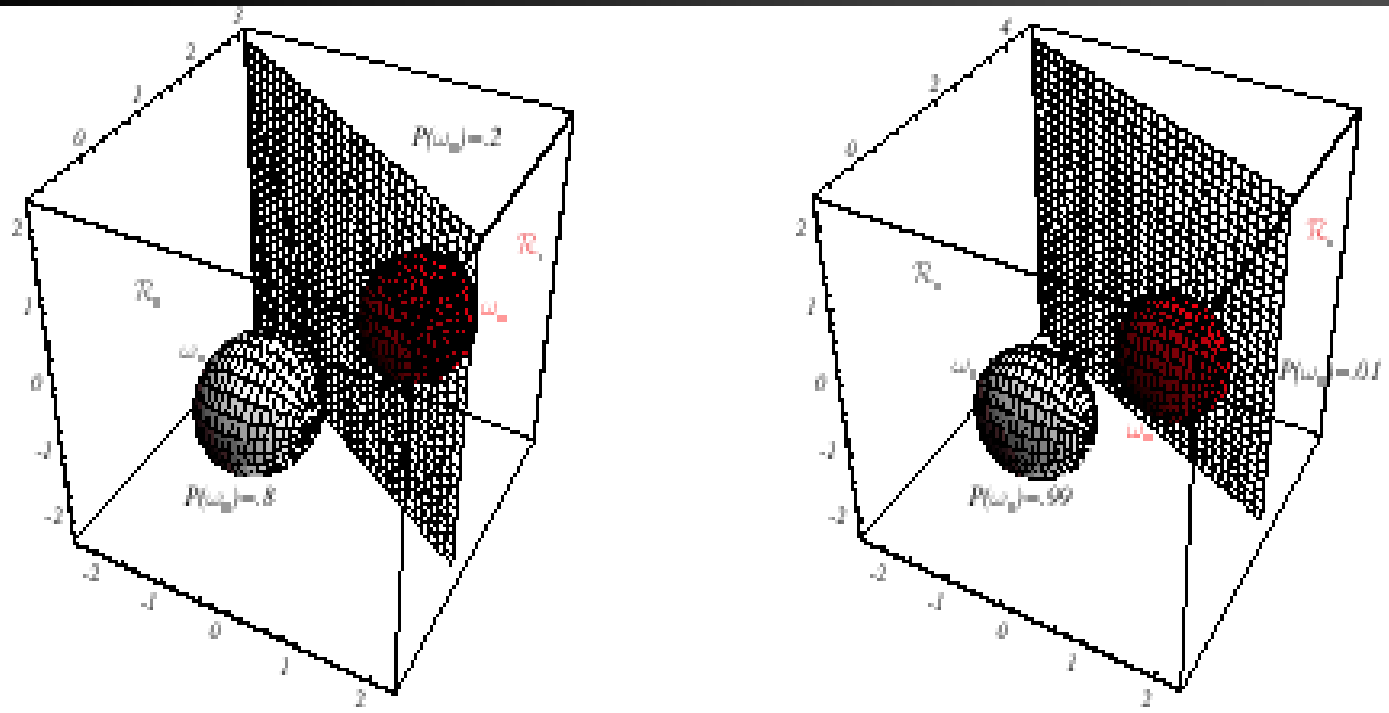


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

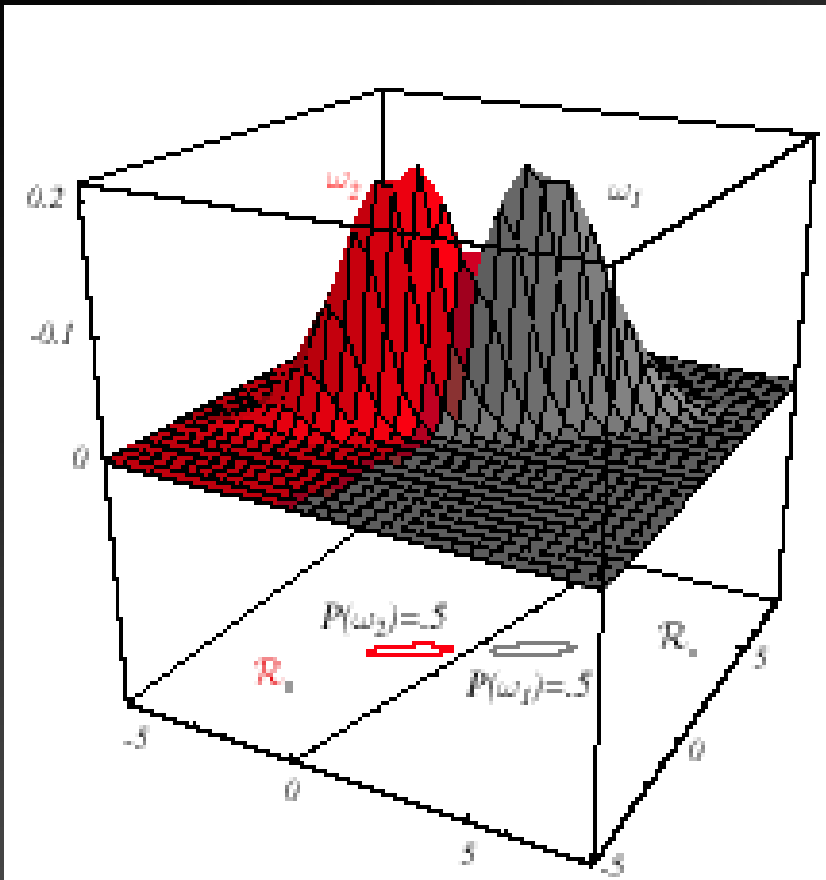
- Przypadek $\Sigma_i = \Sigma$ (kowariancje wszystkich klas są identyczne, z góry ustalone)

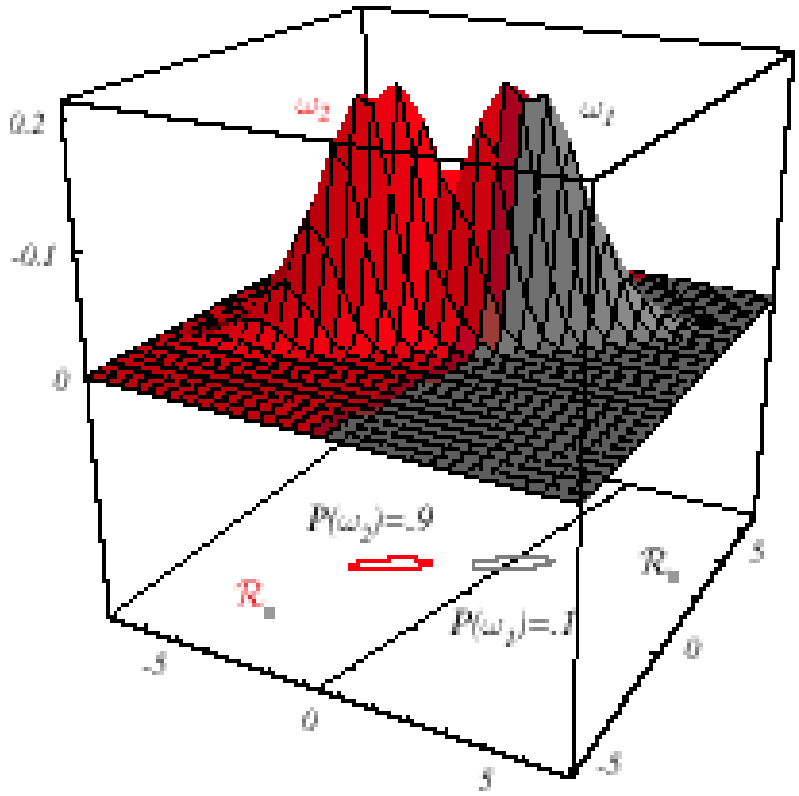
  - Hiperpłaszczyzna rozdzielająca $\mathcal{R}_i$ i $\mathcal{R}_j$

$$w^t\left(x - x_0\right) = 0, \quad gdzie\ w = \Sigma^{-1}\left(\mu_i - \mu_j\right)$$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\mu_i - \mu_j)^T \Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$$

(hiperpłaszczyzna rozdzielająca $\mathcal{R}_i$ i $\mathcal{R}_j$ na ogół nie jest ortogonalna do prostej łączącej średnie)


0.2
-0.1
0
ω_2
ω_1
$P(\omega_2)=.5$
$P(\omega_1)=.5$
$\mathcal{R}_1$
$\mathcal{R}_2$
-5
0
5
-5
0
5


0.2
-0.1
0
ω_2
ω_1
$P(\omega_2)=.9$
$P(\omega_1)=.1$
$\mathcal{R}_1$
$\mathcal{R}_2$
-5
0
5
-5
0
5

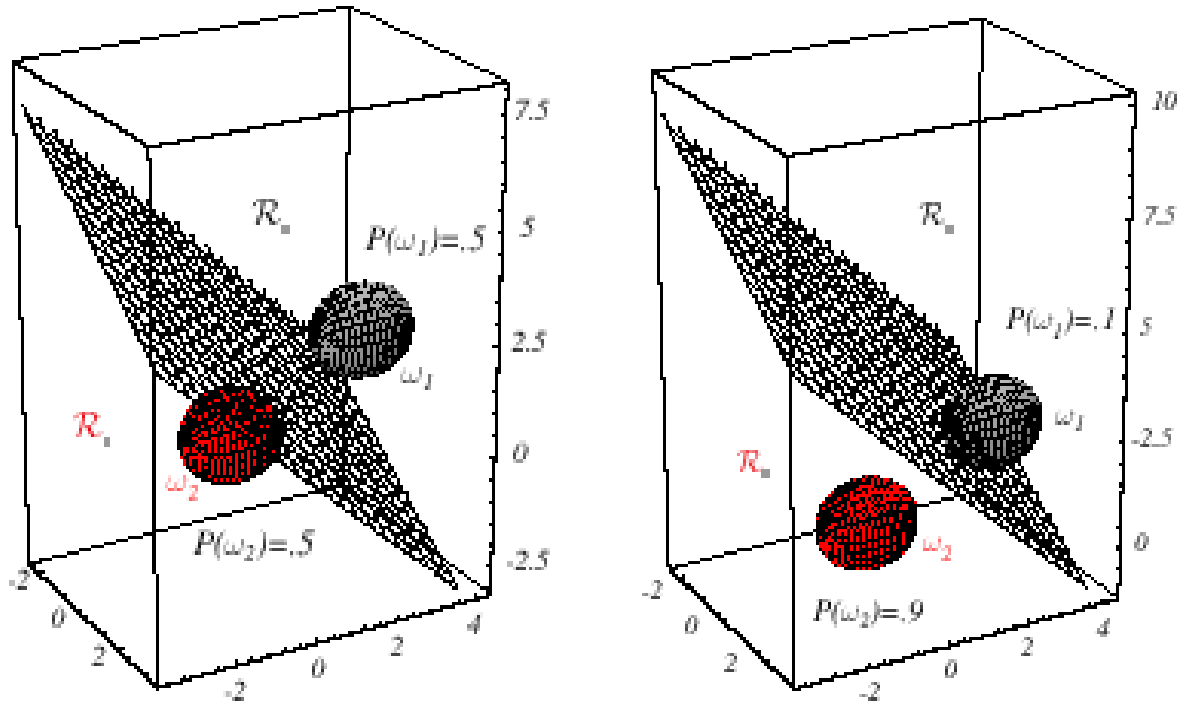


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda,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

- Przypadek Σ_i = z góry ustalona
  - Macierze kowariancji dla każdej klasy są inne

$$g_i(x) = x^T W_i x + w_i^T x + w_{i0}$$

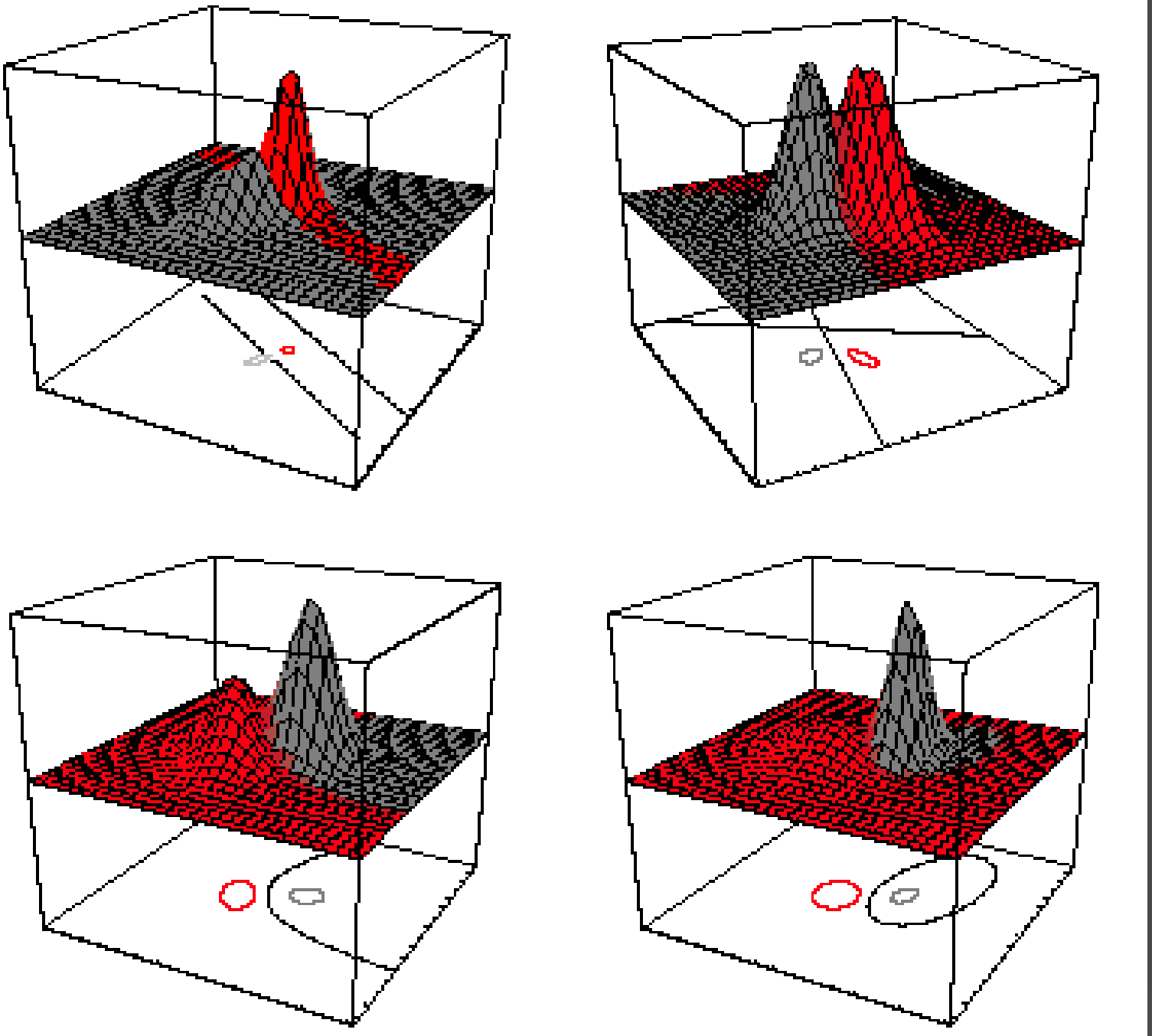
gdzie:

$$\mathrm{W_i} = -\frac{1}{2}\Sigma_i^{-1}$$

$$\mathrm{w_i} = \Sigma_i^{-1}\mu_i$$

$$\mathrm{w_{i0}} = -\frac{1}{2}\mu_i^T\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln\left|\Sigma_i\right| + \ln P(\omega_i)$$

(Hiperkwadratyki: hiperpowierzchnie, pary hiperpowierzchni, hipersfery, hiperelipsoidy, hiperparaboloidy, hiperhiperboloidy)



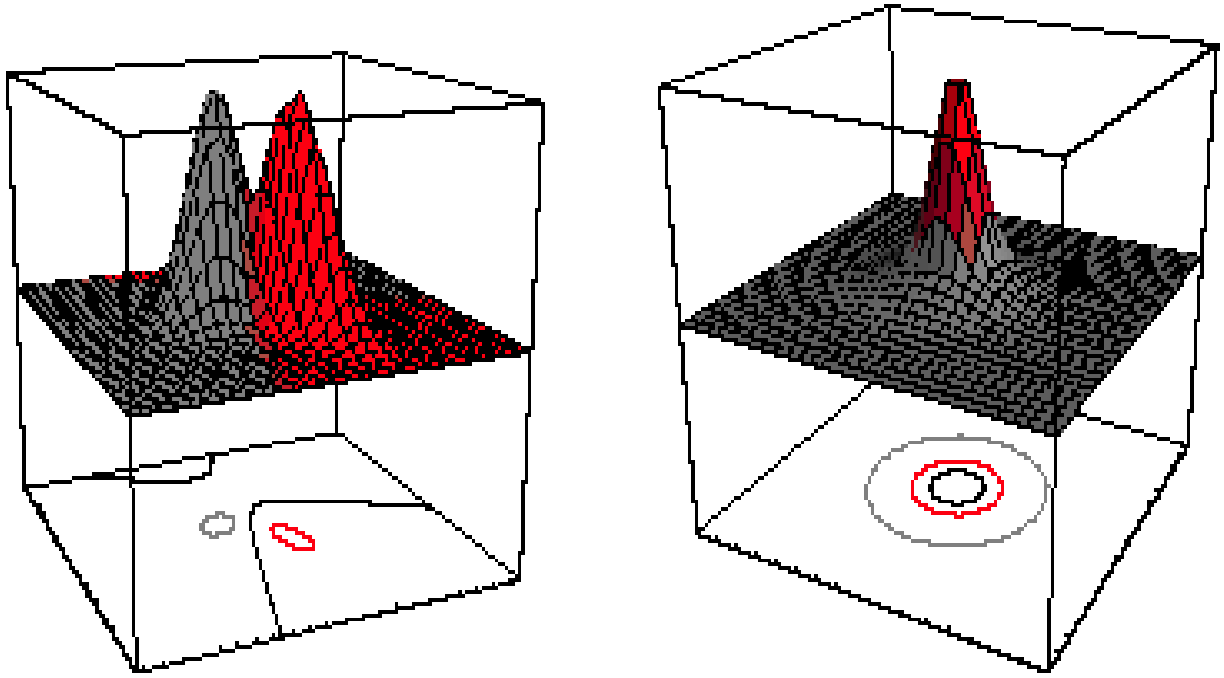


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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

Bayesowska teoria decyzji – cechy dyskretne

- Współrzędne wektora cech x są binarne lub całkowitoliczbowe, x może przyjąć jedynie m różnych wartości

$$v_1, v_2, \ldots, v_m$$

- Przypadek niezależnych cech binarnych w zadaniu klasyfikacji z dwiema klasami

  Niech $x = [x_1, x_2, \ldots, x_d]^T$, przy czym x_i ma wartość 0 lub 1, z prawdopodobieństwami:

$$p_i = P(x_i = 1 \mid \omega_1)$$
$$q_i = P(x_i = 1 \mid \omega_2)$$

- Funkcje dyskryminujące mają postać:

$$g(x) = \sum_{i=1}^{d} w_i x_i + w_0$$

gdzie:

$$w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad i = 1,\ldots,d$$

oraz:

$$w_0 = \sum_{i=1}^{d} \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

wybierz ω_1 gdy $g(x) > 0$ albo ω_2 gdy $g(x) \leq 0$