CC-226 Aula 05 - Teoria da Decisão Bayesiana

Carlos Henrique Q. Forster - Instituto Tecnológico de Aeronáutica

2008

Classificador Bayesiano

Considerando M classes $C_1 \dots C_M$.

N observações $\mathbf{x}_j$.

L atributos $\mathbf{x}_j = (x_{1j}, \dots, x_{kj}, \dots, x_{Lj})$.

Considere um padrão representado pelo vetor de características $\mathbf{x}$.

M probabilidades condicionais $P(C_i|\mathbf{x}), i = 1 \dots M$ (probabilidades posteriores).

Escolhemos para um dado x a classe C_i mais provável.

O problema agora é estimar as funções de densidade de probabilidades (dado um conjunto de observações).

Caso para 2 classes

Supor $P(C_1)$ e $P(C_2)$ as probabilidades a priori conhecidas de duas classes.

É fácil de estimar a partir de observações:

Se $|C_1 \cup C_2| = N$, $|C_1| = N_1$ e $|C_2| = N_2$, estima-se

$$P(C_1) = \frac{N_1}{N}$$

$$P(C_2) = \frac{N_2}{N}$$

Supor conhecida a pdf $p(\mathbf{x}|C_i), i = 1, 2$

Esta é a função de verossimilhança de C_i com respeito a $\mathbf{x}$.

Aplicando a regra de Bayes:

$$P(C_i|\mathbf{x}) = \frac{p(\mathbf{x}|C_i)P(C_i)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|C_i)P(C_i)}{p(\mathbf{x}|C_1)P(C_1) + p(\mathbf{x}|C_2)P(C_2)}$$

Regra de classificação de Bayes:

Se $P(C_1|\mathbf{x}) > P(C_2|\mathbf{x})$, $\mathbf{x}$ é classificado como C_1.

Se $P(C_1|\mathbf{x}) < P(C_2|\mathbf{x})$, $\mathbf{x}$ é classificado como C_2.

No caso de igualdade, o resultado é arbitrado.

É equivalente à comparação de $p(x|C_1)P(C_1) < / > p(x|C_2)P(C_2)$.

Ou então, se $P(C_1) = P(C_2)$, $p(x|C_1) < / > p(x|C_2)$ (razão de verossimilhanças).

Visualizando a classificação em 1D

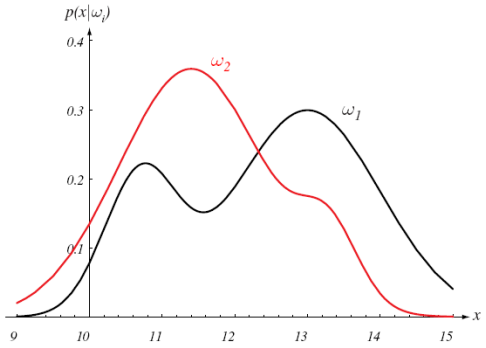


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two 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.

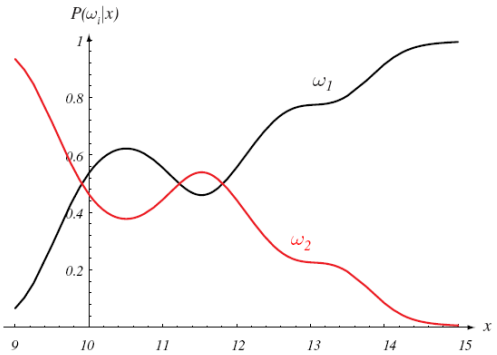


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 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*. Copyright © 2001 by John Wiley & Sons, Inc.

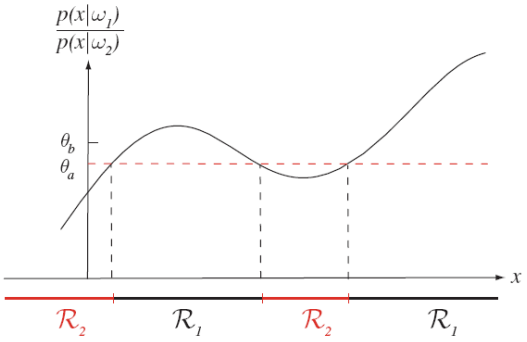


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse,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

No caso de uma única feição x:
x_0 é um limiar particionando o espaço de características em duas regiões R_1 e R_2.
A regra de decisão de Bayes decide C_1 para $x_1 \in R_1$ e C_2 para $x_2 \in R_2$.
Há probabilidade de decidir incorretamente (área hachurada).

$$P_e = \int_{R_1} p(x|C_2)dx + \int_{R_2} p(x|C_1)dx$$

Teorema

O classificador bayesiano é ótimo com respeito a minimizar a probabilidade de erro de classificação.

$$P_e = P(x \in R_2, C_1) + P(x \in R_1, C_2)$$

$$P_e = P(x \in R_2|C_1)P(C_1) + P(x \in R_1|C_2)P(C_2)$$

Pela regra de Bayes:

$$P_e = \int_{R_2} P(C_1|x)p(x)dx + \int_{R_1} P(C_2|x)p(x)dx$$

Escolhemos por hipótese R_1 e R_2 tais que

$$R_1 : P(C_1|x) > P(C_2|x)$$

$$R_2 : P(C_2|x) > P(C_1|x)$$

$$P(C_1) = \int_{R_1} P(C_1|x)p(x)dx + \int_{R_2} P(C_1|x)p(x)$$

porque $R_1 \cup R_2$ cobre todo o espaço $R_1 \cap R_2 = \emptyset$.

Substituindo no P_e:

$$P_e = P(C_1) - \int_{R_1} (P(C_1|x) - P(C_2|x))p(x)dx$$

É mínima se R_1 é tal que $P(C_1|x) > P(C_2|x)$.

Conversamente, a afirmação é válida para R_2.

A extensão para múltiplas classes é trivial.

$\diamond$

Minimização do risco médio

Riscos diferentes para falsos-positivos e falsos-negativos:

- Diagnóstico classifica uma pessoa sã como necessitando de intervenção cirúrgica *versus* classificar pessoa doente como sã.
- Fornecer empréstimo para uma pessoa com grande risco de estar aplicando um golpe *versus* deixar de fornecer empréstimo a uma pessoa idônea.
- Permitir acesso a um impostor *versus* impedir acesso de uma pessoa cadastrada.
- Colocar "corbinas" nas latas de salmão *versus* colocar salmão nas latas de corbina.

Classificador com $R_j, j = 1 \ldots M$ regiões correpondentes a decisões pelas classes C_j.

Se $x \in C_k$, mas está na região $R_i, i \neq k$ (misclassification), a perda será

$$\lambda_{ki}$$

(era para ser k, mas classifiquei como i).

Com os valores de perda, montamos uma matriz de perda L.

Risco ou perda associado à classe C_k:

$$r_k = \sum_{i=1}^{M} \lambda_{ki} \int_{R_i} p(x|C_k)dx$$

Risco médio

$$r = \sum_{k=1}^{M} r_k P(C_k)$$

$$= \sum_{i=1}^{M} \int_{R_i} \left(\sum_{k=1}^{m} \lambda_{ki} p(x|C_k) P(C_k) \right) dx$$

Minimizar o risco médio corresponde a escolher regiões R_i particionando o espaço de forma que

$$x \in R_i \text{ se } l_i = \sum_{k=1}^{M} \lambda_{ki} p(x|C_k) P(C_k) < l_j, \forall j \neq i$$

A matriz L que contém zeros na diagonal principal e uns nas demais células corresponde a minimizar a probabilidade de erro da classificação.

No caso de duas classes

$$l_1 = \lambda_{11} p(x|C_1) P(C_1) + \lambda_{21} p(x|C_2) P(C_2)$$

$$l_2 = \lambda_{12} p(x|C_1) P(C_1) + \lambda_{22} p(x|C_2) P(C_2)$$

Associamos x a C_1 se $l_1 < l_2$, ou seja,

$$(\lambda_{21} - \lambda_{22}) p(x|C_2) P(C_2) < (\lambda_{12} - \lambda_{11}) p(x|C_1) P(C_1)$$

Naturalmente, de forma geral,

$$\lambda_{ij} > \lambda_{ii}$$

.

A regra de decisão fica da forma:

$$x \in C_1 \text{ se } l_{12} \equiv \frac{p(x|C_1)}{p(x|C_2)} > \frac{P(C_2)}{P(C_1)} \frac{\lambda_{21} - \lambda_{22}}{\lambda_{12} - \lambda_{11}}$$

l_{12} é uma razão de verossimilhanças.

No caso que

$$L = \begin{bmatrix} 0 & \lambda_{12} \\ \lambda_{21} & 0 \end{bmatrix}$$

Associar a C_2 se $p(x|C_2) > p(x|C_1) \frac{\lambda_{12}}{\lambda_{21}}$ supondo $P(C_1) = P(C_2)$.

Exemplo

(2.1 Theodoridis)

Única feição x com pdfs gaussianas com $\sigma^2 = 1/2$ e médias 0 e 1 respectivamente.

$$p(x|C_1) = \frac{1}{\sqrt{\pi}} \exp(-x^2)$$

$$p(x|C_2) = \frac{1}{\sqrt{\pi}} \exp(-(x-1)^2)$$

Se $P(C_1) = P(C_2) = 1/2$, calcular o limiar x_0 para a

a) probabilidade de erro mínima

b) risco mínimo para matriz de perda

$$L = \begin{bmatrix} 0 & 0,5 \\ 1 & 0 \end{bmatrix}$$

Solução

a)

$$exp(-x^2) = exp(-(x-1)^2)$$

$$-x^2 = -(x-1)^2$$

$$x_0 = 1/2$$

b)

$$exp(-x^2) = 2exp(-(x-1)^2)$$

$$x_0 = \frac{1 - \ln 2}{2}$$

Superfícies de Decisão

Se duas regiões R_i e R_j de um problema de lcassificação são vizinhas, são separadas pela superfície definida pela equação

$$P(C_i|x) - P(C_j|x) = 0$$

(no caso de minimização do erro de classificação)

De um lado da superfície, o valor é positivo. Do outro, é negativo.

Definimos $g_i(x) = f(P(C_i|x))$, onde

f é monotonicamente crescente,

$g_i(x)$ é função discriminante e

seu teste de decisão decorrente é:

$$\text{Classificar } x \in C_i \text{ se } g_i(x) > g_j(x), \forall j \neq i$$

Superfícies de decisão para regiões contíguas são descritas por

$$g_{ij}(x) = g_i(x) - g_j(x) = 0 \text{ com } i, j = 1 \ldots M, i \neq j$$

Desvantagens da classificação Bayesiana

Torna-se difícil no caso de pdfs complicadas com estimação difícil.

Função de discriminação complica com a quantidade de classes.

No caso de dependência, o custo computacional é elevado para muitos atributos.

Classificação Bayesiana para Distribuições Normais

Caso particular de distribuições normais permite tratabilidade computacional.

Modelos complexos em que uma variável aleatória corresponde à soma de várias variáveis aleatórias podem ser modelados por uma distribuição normal.

$$p(x|C_i) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right)$$

onde $i = 1 \dots M$

$\mu_i = E[x]$ é o valor médio de x da classe C_i e Σ é a matriz covariância.

$$\Sigma_i = E[(x-\mu_i)(x-\mu_i)^T]$$

$|\Sigma|$ é seu determinante.

$$g_i(x) = \ln\left(p(x|C_i)P(C_i)\right) = \ln p(x|C_i) + \ln P(C_i)$$

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) + \ln P(C_i) + K_i$$

Obtemos a forma quadrática da superfície de decisão

$$g_i(x) = -\frac{1}{2}x^T\Sigma_i^{-1}x + \frac{1}{2}x^T\Sigma_i^{-1}\mu_i - \frac{1}{2}\mu_i^T\Sigma_i^{-1}\mu_i + \frac{1}{2}\mu_i^T\Sigma_i^{-1}x + \ln P(C_i) + K_i$$

Exemplo

$$l = 2$$

e

$$\Sigma_i = \begin{bmatrix} \sigma_i^2 & 0 \\ 0 & \sigma_i^2 \end{bmatrix}$$

$$g_i(X) = \frac{1}{2\sigma_i^2}({x_1}^2 + {x_2}^2) + \frac{1}{\sigma_i^2}(\mu_{i1}x_1 + \mu_{i2}x_2) - \frac{1}{2\sigma_i^2}(\mu_{i1}^2 + \mu_{i2}^2) + \ln P(C_i) + K_i$$

Logo $g_i(x) - g_j(x) = 0$ são superfícies de separação que são quádricas: elipsóides, parábolas, hipérboles, pares de linhas.

Para $l > 2$, são hiperquádricas.

Gráficos para diferentes Σ

Hiperplanos decisórios

Supondo que todas as matrizes de covariância são iguais

$$\Sigma_i = \Sigma$$

O termo quadrático se cancela em comparações.

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

Onde

$$w_i = \Sigma^{-1}\mu_i$$

e

$$w_{i0} = \ln P(C_i) - \frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i$$

$g_i(x) - g_j(x) = 0$ define um hiperplano de separação.

Caso da matriz de covariância diagonal

$$E[(x_i - \mu_i)(x_j - \mu_j)] = \sigma^2 \delta_{ij}$$

ou

$$\Sigma = \sigma^2 I$$

Os atributos são icorrelatos mutuamente e de mesma variância

$$g_i(x) = \frac{1}{\sigma^2}\mu_i^T x + w_{i0}$$

$$g_{ij}(x) = g_i(x) - g_j(x) = w^T(x - x_0) = 0$$

com $w = \mu_i - \mu_j$
e

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

onde $\|x\| = \sqrt{{x_1}^2 + {x_2}^2 + \ldots + {x_l}^2}$ é a norma euclidiana de x
Hiperplano passando por x_0

Caso da matriz de covariância não-diagonal

$$g_{ij}(x) = w^T(x - x_0) = 0$$

onde $w = \Sigma^{-1}(\mu_i - \mu_j)$

e

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \ln\left(\frac{P(C_i)}{P(C_j)}\right)\frac{\mu_i - \mu_j}{\|\mu_i - \mu_j\|^2_{\Sigma^{-1}}}$$

onde

$$\|x\|_{\Sigma^{-1}} \equiv (x^T\Sigma^{-1}x)^{1/2}$$

é a norma Σ^{-1} de x.

O plano não é mais ortogonal a $\mu_i - \mu_j$, mas a $\Sigma^{-1}(\mu_i - \mu_j)$.

Classificadores de mínima distância

Supomos classes equiprováveis com mesma matriz de covariância

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^T\Sigma^{-1}(x - \mu_i)$$

Foram desconsideradas as constantes desnecessárias à comparação das funções discriminantes.

Para $\Sigma = \sigma^2 I$, máximo $g_i(x)$ implica mínima distância euclidiana

$$d_{\mathcal{E}} = \|x - \mu_i\|$$

As superfícies de nível são circunferências ou cascas hiper-esféricas.

Para Σ não-diagonal, minimização da norma Σ^{-1} para maximizar $g_i(x)$ minimiza-se a distância de Mahalanobis

$$d_m = \left((x - \mu_i)^T\Sigma^{-1}(x - \mu_i)\right)^{-1/2}$$

Propriedade: A matriz de covariância é simétrica.

Decomposição em auto-valores:

$$\Sigma = \Phi\Lambda\Phi^T$$

onde

$$\Phi^T = \Phi^{-1}$$

ou seja, Φ é matriz ortogonal, e Λ é uma matriz diagonal cujos elementos são autovalores λ_i de Σ e, se

$$\Phi = [\mathbf{v_1\_v_2} \ldots \mathbf{v_l}]$$

então os vetores $\mathbf{v}_i$ são autovalores correspondentes a λ_i.

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_l \end{bmatrix}$$

Substituindo na expressão da distância de Mahalanobis:

$$(x-\mu_i)^T \Phi \Lambda^{-1} \Phi^T (x-\mu_i) = c^2$$

Fazendo que

$$x' = {v_k}^T x, k = 1 \ldots l$$

Temos

$$\frac{(x'_1 - \mu'_{i1})^2}{\lambda_1} + \cdots + \frac{(x'_l - \mu'_{il})^2}{\lambda_l} = c^2$$

Assim, as superfícies de nível formadas são hiper-elipsóides.

Exemplo 2.2 Theodoridis

Classificação 2 classes e 2 dimensões (atributos)

Dadas distribuições normais com a mesma matriz de covariância

$$\Sigma = \begin{bmatrix} 1,1 & 0,3 \\ 0,3 & 1,9 \end{bmatrix}$$

e os vetores média são $\mu_1 = [00]^T$ e $\mu_2 = [33]^T$ respectivamente.

a) classificar o vetor $[1,02,2]^T$ de acordo com o classificador Bayesiano.

b) computar os eixos principais da elipse centrada em $[00]^T$ que corresponde à distância de Mahalanobis constante $d_m = \sqrt{2,952}$ ao centro.

Solução

a) Basta computar a distância de Mahalanobis a $[1,02,2]^T$ aos vetores de média

$${d_m}^2(\mu_1, x) = (x-\mu_1)^T \Sigma^{-1} (x-\mu_1)$$

$$= [1,02,2] \begin{bmatrix} 0,95 & -0,15 \\ -0,15 & 0,55 \end{bmatrix} \begin{bmatrix} 1,0 \\ 2,2 \end{bmatrix} = 2,952$$

$${d_m}^2(\mu_2, x) = (x-\mu_2)^T \Sigma^{-1} (x-\mu_2)$$

$$= [-2,0-0,8] \begin{bmatrix} 0,95 & -0,15 \\ -0,15 & 0,55 \end{bmatrix} \begin{bmatrix} -2,0 \\ -0,8 \end{bmatrix} = 3,672$$

Assim, associa-se o ponto dado à classe C_1 correspondente à menor distância de Mahalanobis.

b) Obter os autovalores de Σ

$$\det\left(\begin{bmatrix} 1,1-\lambda & 0,3 \\ 0,3 & 1,9-\lambda \end{bmatrix} \right) = \lambda^2 - 3\lambda + 2 = 0$$

$$\lambda_1 = 1 \text{ e } \lambda_2 = 2$$

Substituindo em $(\Sigma - \lambda I)v = 0$

Encontrado o espaço nulo, isto é, determinar vetores normalizados que formam uma base para a solução da equação.

$$v_1 = \begin{bmatrix} \frac{3}{\sqrt{10}} \\ \frac{-1}{\sqrt{10}} \end{bmatrix}$$

$$v_1 = \begin{bmatrix} \frac{1}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} \end{bmatrix}$$

Os eixos principais são paralelos a v_1 e v_2 respectivamente com comprimentos 3,436 e 4,859.

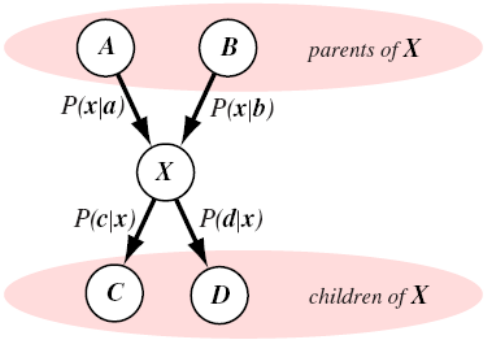


FIGURE 2.25. A portion of a belief network, consisting of a node **X**, having variable values $(x_1, x_2, \ldots)$, its parents (**A** and **B**), and its children (**C** and **D**). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

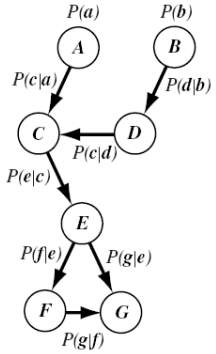


FIGURE 2.24. A belief network consists of nodes (labeled with uppercase bold letters) and their associated discrete states (in lowercase). Thus node **A** has states a_1, $a_2, \ldots$, denoted simply **a**; node **B** has states b_1, $b_2, \ldots$, denoted **b**, and so forth. The links between nodes represent conditional probabilities. For example, $P(\mathbf{c}|\mathbf{a})$ can be described by a matrix whose entries are $P(c_i|a_j)$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

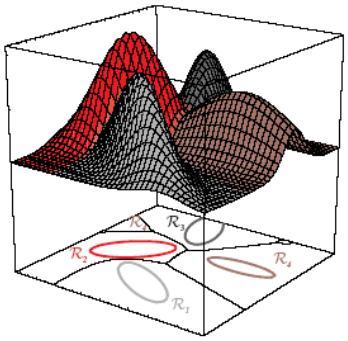


FIGURE 2.16. The decision regions for four normal distributions. Even with such a low number of categories, the shapes of the boundary regions can be rather complex. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

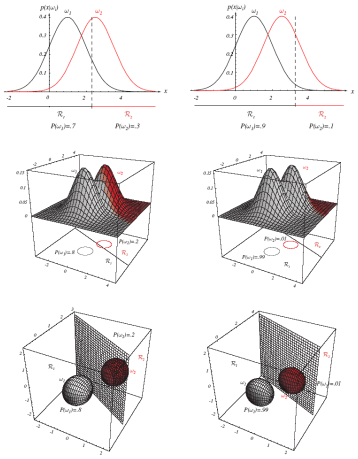


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

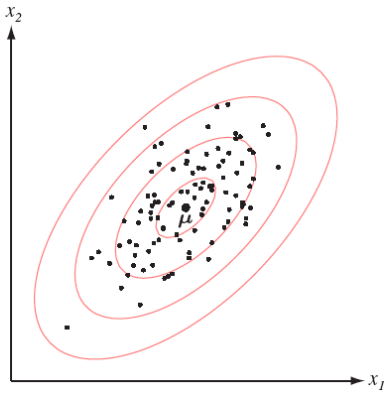


FIGURE 2.9. Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean $\boldsymbol{\mu}$. The ellipses show lines of equal probability density of the Gaussian. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

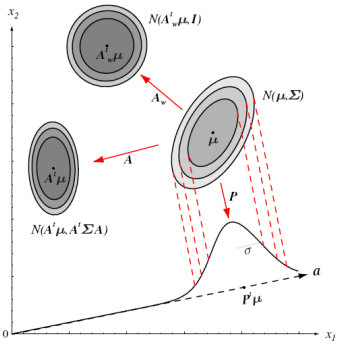


FIGURE 2.8. The action of a linear transformation on the feature space will convert an arbitrary normal distribution into another normal distribution. One transformation, $\mathbf{A}$, takes the source distribution into distribution $N(\mathbf{A}^t\boldsymbol{\mu}, \mathbf{A}^t\boldsymbol{\Sigma}\mathbf{A})$. Another linear transformation—a projection $\mathbf{P}$ onto a line defined by vector $\mathbf{a}$—leads to $N(\mu, \sigma^2)$ measured along that line. While the transforms yield distributions in a different space, we show them superimposed on the original x_1x_2-space. A whitening transform, $\mathbf{A}_w$, leads to a circularly symmetric Gaussian, here shown displaced. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

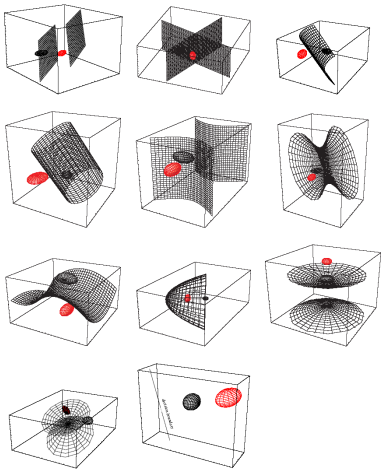


FIGURE 2.15. Arbitrary three-dimensional Gaussian distributions yield Bayes decision boundaries that are two-dimensional hyperquadrics. There are even degenerate cases in which the decision boundary is a line. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright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

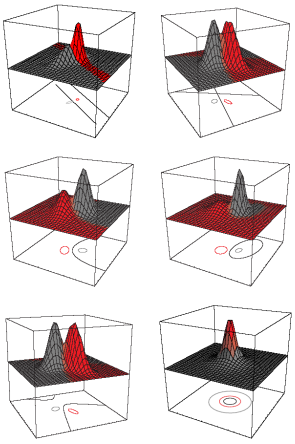


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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

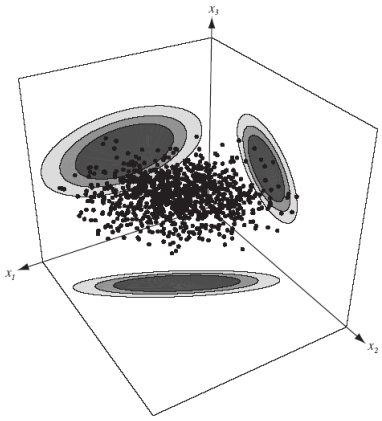


FIGURE 2.23. A three-dimensional distribution which obeys $p(x_1, x_3) = p(x_1)p(x_3)$; thus here x_1 and x_3 are statistically independent but the other feature pairs are not. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

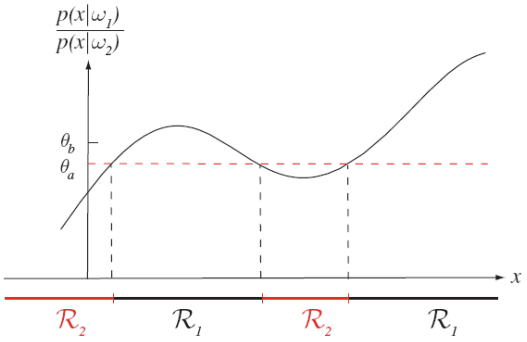


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse,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

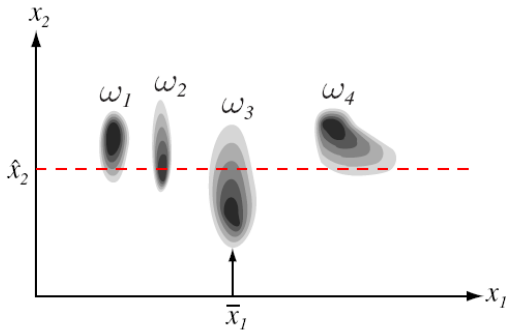


FIGURE 2.22. Four categories have equal priors and the class-conditional distributions shown. If a test point is presented in which one feature is missing (here, x_1) and the other is measured to have value $\hat{x}_2$ (red dashed line), we want our classifier to classify the pattern as category ω_2, 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.

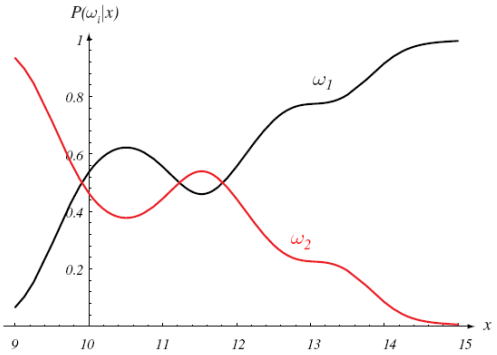


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 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*. Copyright © 2001 by John Wiley & Sons, Inc.

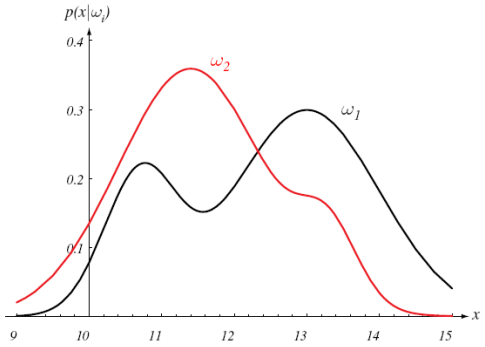


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two 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.

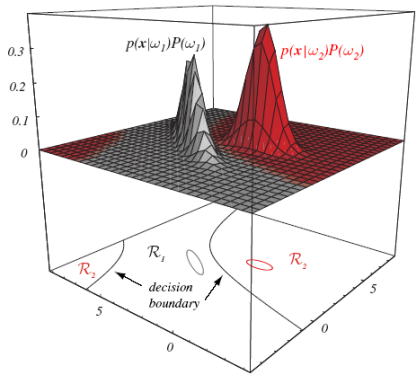


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution.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

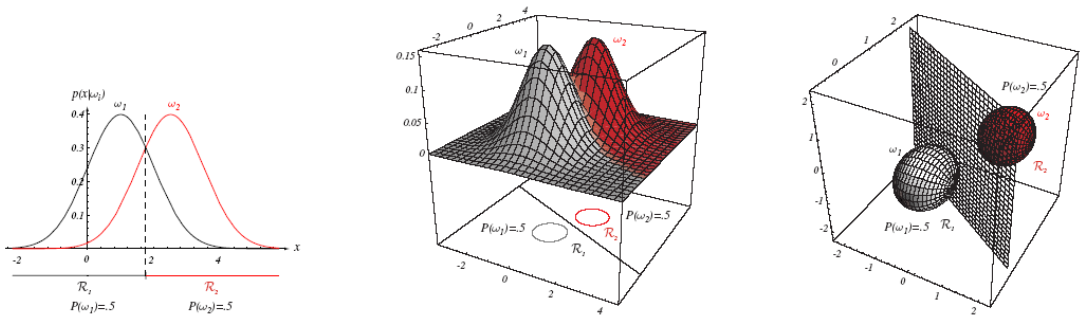
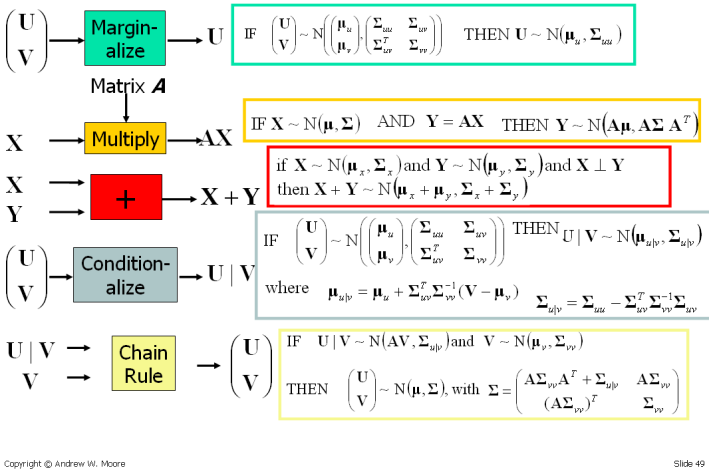


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-,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

Available Gaussian tools



 Slide 49