

Bayesian Decision Theory

CS536, Spring 2013

Based on slides from DHS

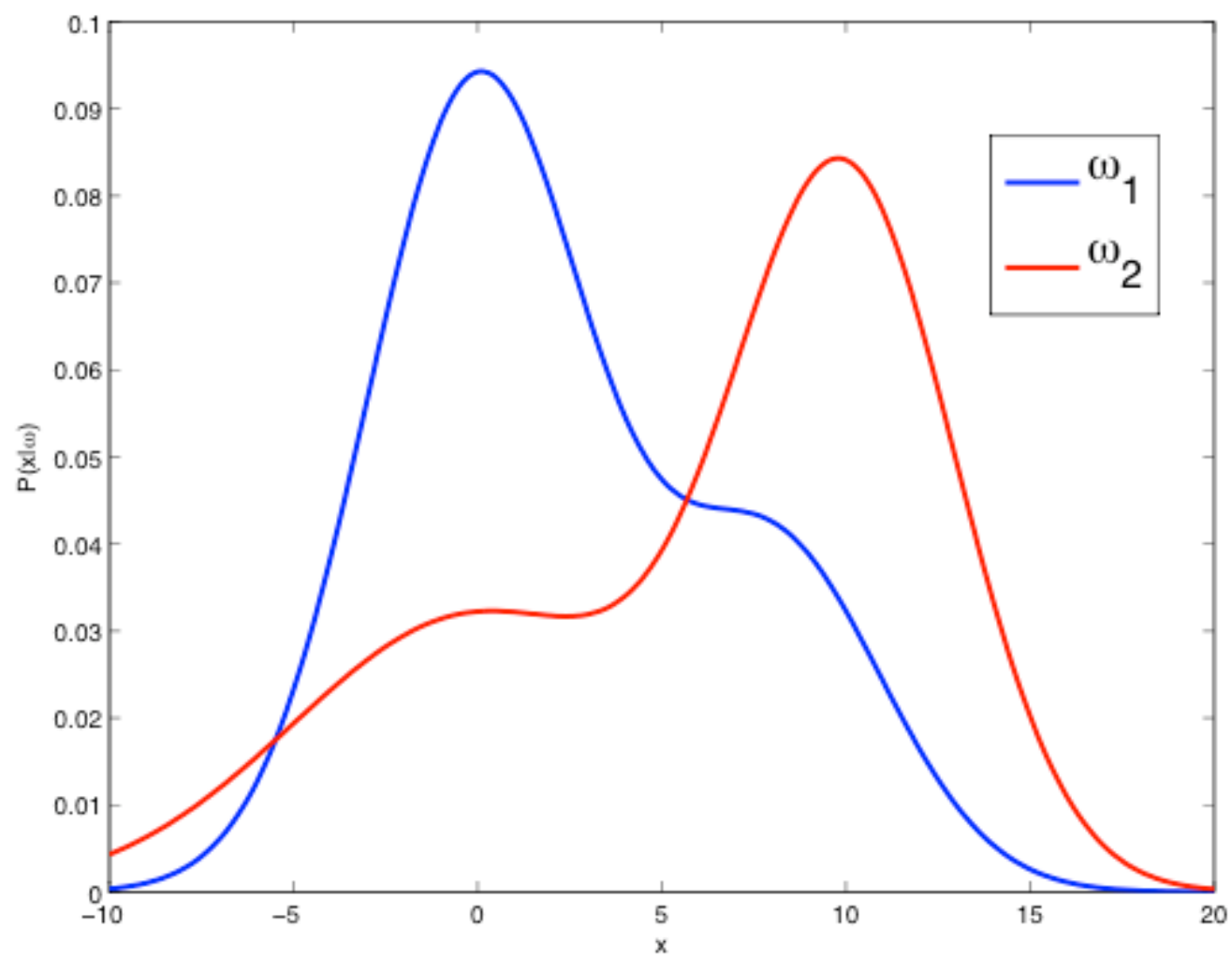
Today's Topics

- Bayesian decision theory
- Optimal Bayesian decision
- Loss minimization
- Discriminant function
- Bayesian decision theory for normal density

Introduction

- The family car example (with engine power as feature)
 - State of nature, prior
 - State of nature is a random variable
 - The chance of family car or other car is equiprobable
 - $P(\omega_1) = P(\omega_2)$ (uniform priors)
 - $P(\omega_1) + P(\omega_2) = 1$ (exclusivity and exhaustively)

- Decision rule with only the prior information
 - Decide ω_1 if $P(\omega_1) > P(\omega_2)$ otherwise decide ω_2
- Use of the class –conditional information
- $P(x | \omega_1)$ and $P(x | \omega_2)$ describe the difference in engine power between populations of family & non-family cars



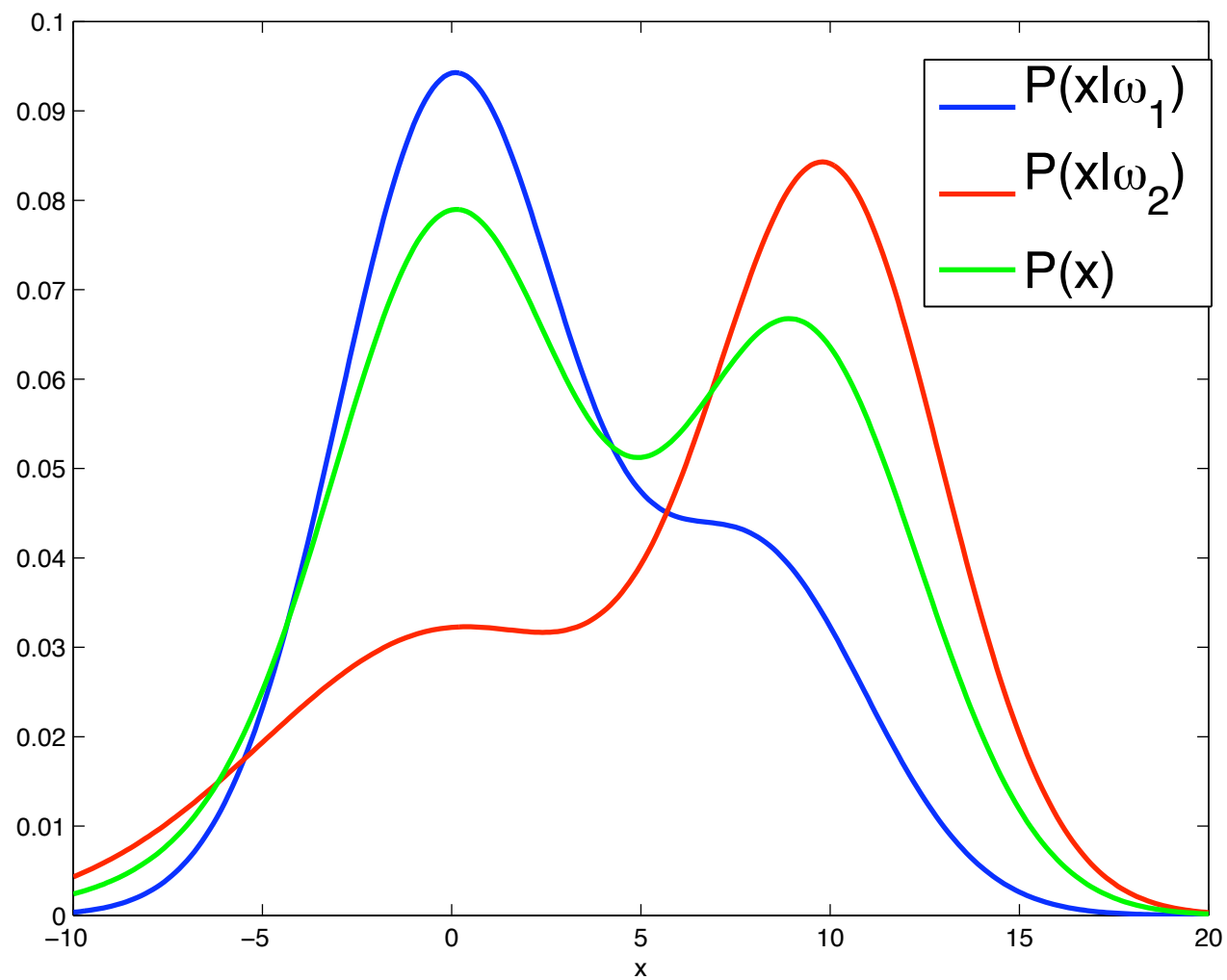
- Posterior, likelihood, evidence

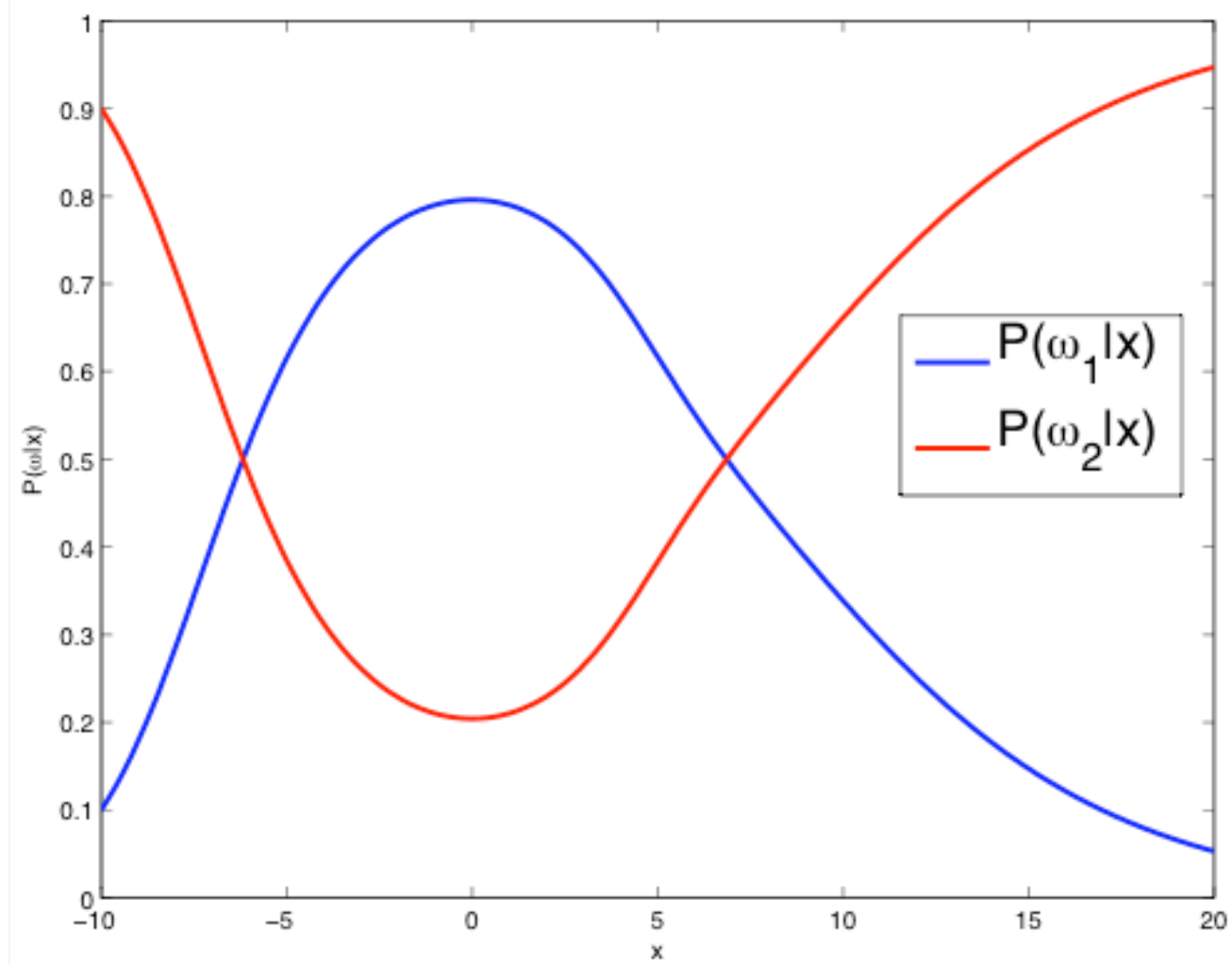
$$P(\omega_j | x) = \frac{P(x | \omega_j)P(\omega_j)}{P(x)}$$

where in case of two categories



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

$$\textit{posterior} = \frac{\textit{likelihood} \cdot \textit{prior}}{\textit{evidence}}$$





– $P(\omega_1) = 2/3$, $P(\omega_2) = 1/3$

- Decision given the posterior probabilities
- X is an observation for which:
 - if $P(\omega_1 | x) > P(\omega_2 | x)$  True state of nature = ω_1
 - if $P(\omega_1 | x) < P(\omega_2 | x)$  True state of nature = ω_2

- Therefore:

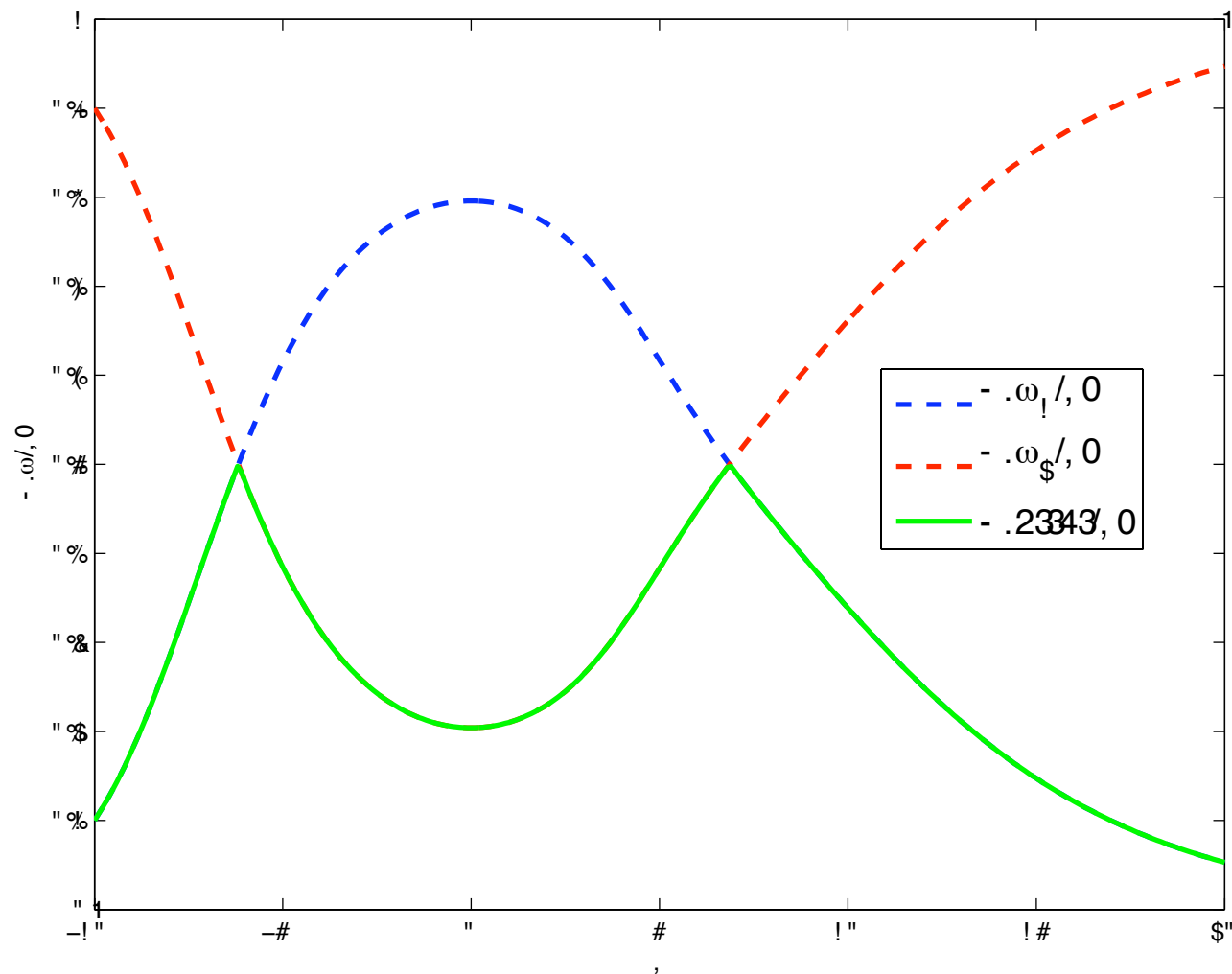
whenever we observe a particular x , the probability of error is :

- $P(\text{error} | x) = P(\omega_1 | x)$ if we decide ω_2
- $P(\text{error} | x) = P(\omega_2 | x)$ if we decide ω_1

- Minimizing the probability of error
- Decide ω_1 if $P(\omega_1 | x) > P(\omega_2 | x)$;
otherwise decide ω_2
- Therefore:

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

(Bayes decision)



Bayesian Decision Theory – Continuous Features

- Generalization of the preceding ideas
 - Use of more than one feature
 - Use more than two states of nature
 - Allowing actions and not only decide on the state of nature
 - Introduce a loss of function which is more general than the probability of error

- Allowing actions other than classification primarily allows the possibility of rejection
- Refusing to make a decision in close or bad cases!
- The loss function states how costly each action taken is

- Let $\{\omega_1, \omega_2, \dots, \omega_c\}$ be the set of c states of nature (or “categories”)
- Let $\{\alpha_1, \alpha_2, \dots, \alpha_a\}$ be the set of possible actions
- Let $\lambda(\alpha_i \mid \omega_j)$ be the loss incurred for taking action α_i when the state of nature is ω_j

- Overall risk
- $R = \text{Sum of all } R(\alpha_i | x) \text{ for } i = 1, \dots, a$


$\underbrace{\hspace{10em}}$
Conditional risk

- Minimizing $R \iff$ Minimizing $R(\alpha_i | x)$ for $i = 1, \dots, a$

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

for $i = 1, \dots, a$

- Select the action a_i for which $R(a_i | x)$ is minimum

 R is minimum and R in this case is called the Bayes risk = best performance that can be achieved!

- Two-category classification
 - α_1 : deciding ω_1
 - α_2 : deciding ω_2
 - $\lambda_{ij} = \lambda(\alpha_i | \omega_j)$
- loss incurred for deciding ω_i when the true state of nature is ω_j
- Conditional risk:
 - $R(\alpha_1 | x) = \lambda_{11}P(\omega_1 | x) + \lambda_{12}P(\omega_2 | x)$
 - $R(\alpha_2 | x) = \lambda_{21}P(\omega_1 | x) + \lambda_{22}P(\omega_2 | x)$

- Our rule is the following:

if $R(\alpha_1 | x) < R(\alpha_2 | x)$
action α_1 : “decide ω_1 ” is taken

- This results in the equivalent rule :

– decide ω_1 if:

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

– and decide ω_2 otherwise

- Likelihood ratio:
- The preceding rule is equivalent to the following rule:

$$\text{if } \frac{P(x | \omega_1)}{P(x | \omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

- Then take action α_1 (decide ω_1)
- Otherwise take action α_2 (decide ω_2)

- Optimal decision property

“If the likelihood ratio exceeds a threshold value independent of the input pattern x , we can take optimal actions”

Exercise

- Select the optimal decision where:
 - $= \{\omega_1, \omega_2\}$
- $P(x | \omega_1) \longrightarrow N(2, 0.5)$
- $P(x | \omega_2) \longrightarrow N(1.5, 0.2)$
- $P(\omega_1) = 2/3$
- $P(\omega_2) = 1/3$

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

Minimum-Error-Rate Classification

- Actions are decisions on classes
If action α_i is taken and the true state of nature is ω_j then:
the decision is correct if $i = j$ and in error if $i \neq j$
- Seek a decision rule that minimizes the *probability of error*
which is the *error rate*

- Introduction of the zero-one loss function:

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

Therefore, the conditional risk is:

$$\begin{aligned} R(\alpha_i | x) &= \sum_{j=1}^{j=c} \lambda(\alpha_i | \omega_j) P(\omega_j | x) \\ &= \sum_{j \neq i} P(\omega_j | x) = 1 - P(\omega_i | x) \end{aligned}$$

“The risk corresponding to this loss function is the average probability error”

- Minimize the risk requires maximize $P(\omega_i | x)$
(since $R(\alpha_i | x) = 1 - P(\omega_i | x)$)
- For Minimum error rate
 - Decide ω_i if $P(\omega_i | x) > P(\omega_j | x)$ for all $j \neq i$

- Regions of decision and zero-one loss function, therefore:

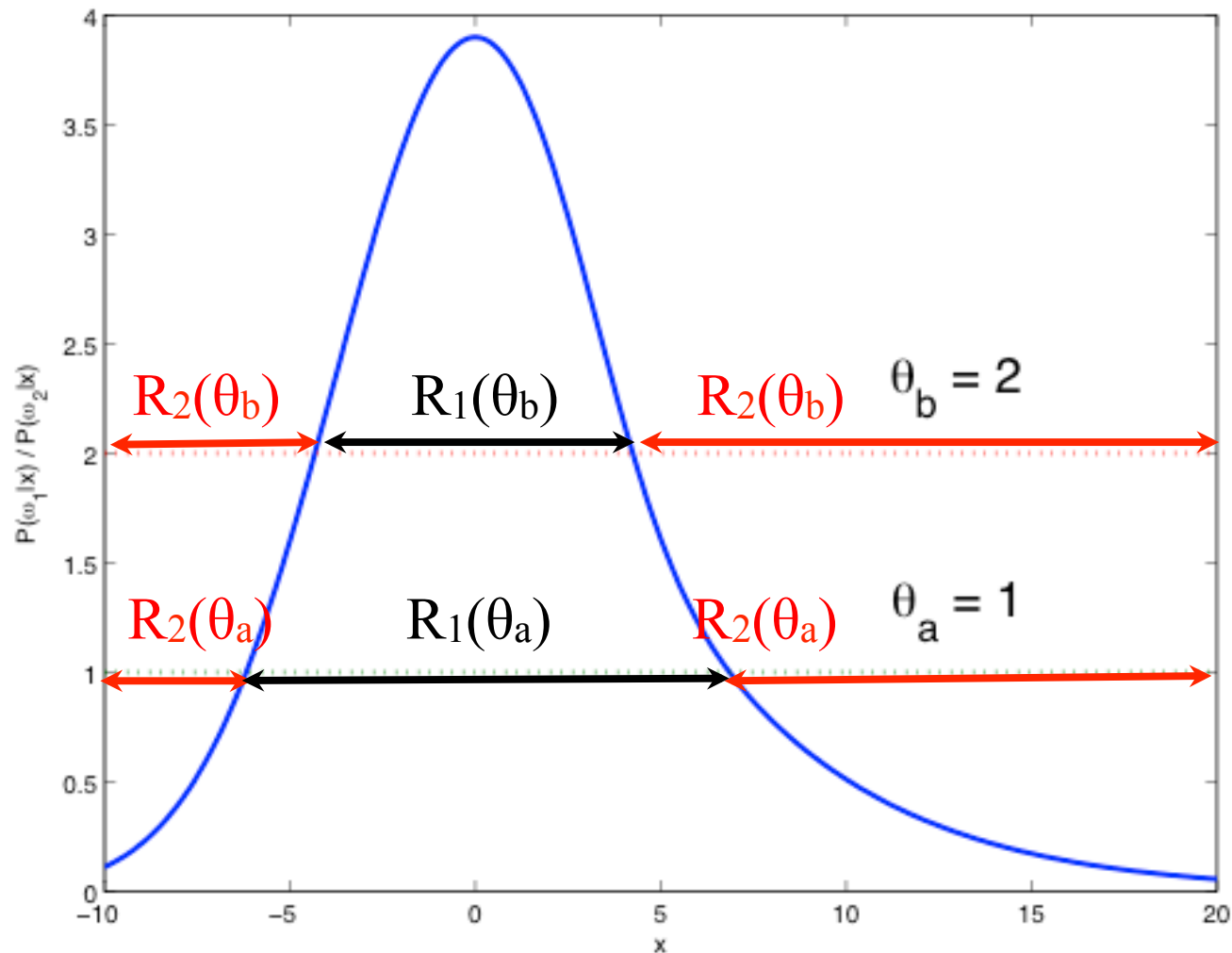
$$\text{Let } \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)} = \theta_\lambda \text{ then decide } \omega_1 \text{ if: } \frac{P(x | \omega_1)}{P(x | \omega_2)} > \theta_\lambda$$

- If λ is the zero-one loss function which means:

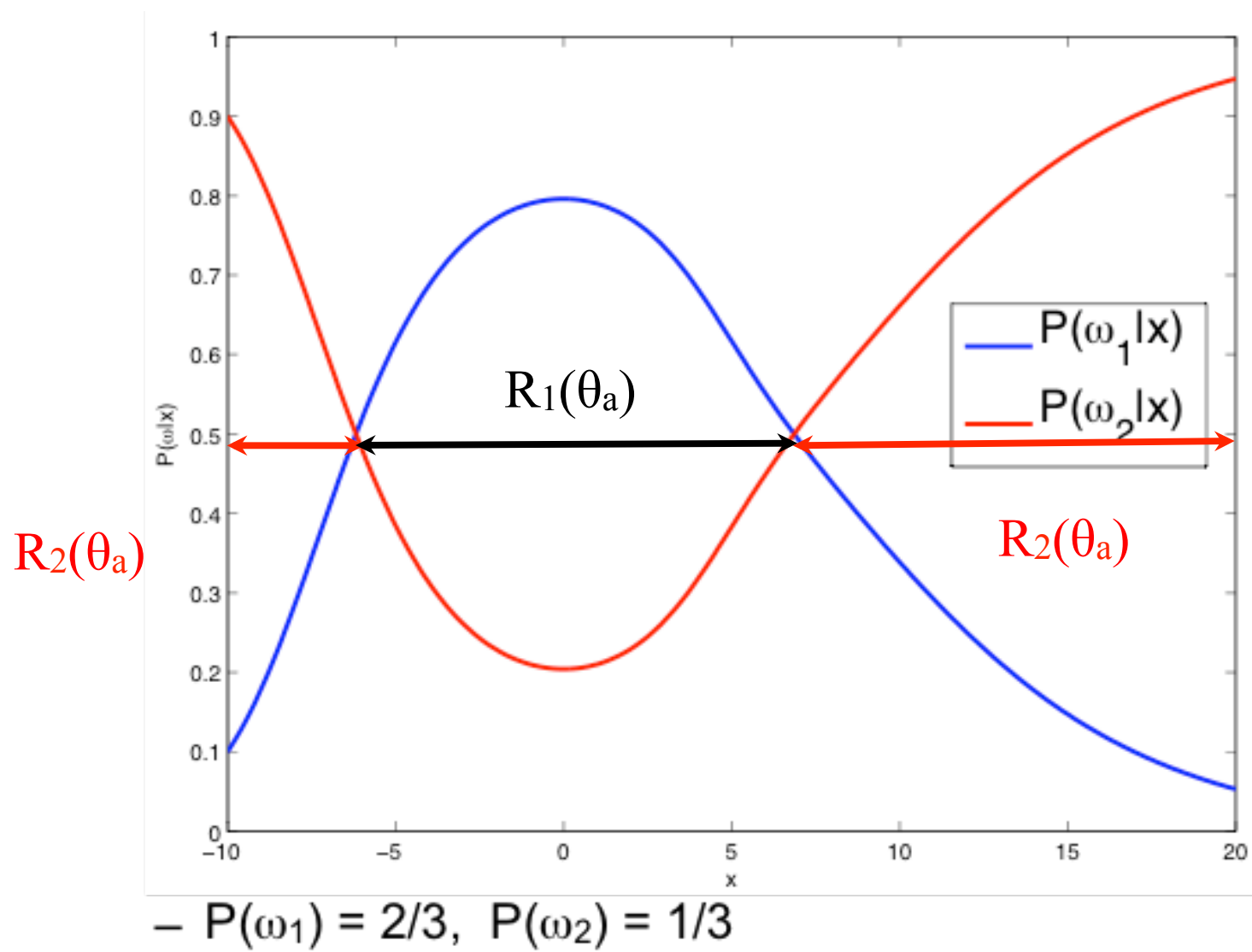
$$\lambda = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

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

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



If we employ 0-1 loss, the threshold function is θ_a . If missclassification of ω_2 is penalized 2x as much as that of ω_1 , the threshold increases to θ_b and the decision region R_1 becomes smaller.



Classifiers, Discriminant Functions and Decision Surfaces

- The multi-category case
 - Set of discriminant functions $g_i(x)$, $i = 1, \dots, c$
 - The classifier assigns a feature vector x to class ω_i if:

$$g_i(x) > g_j(x) \text{ for all } j \neq i$$

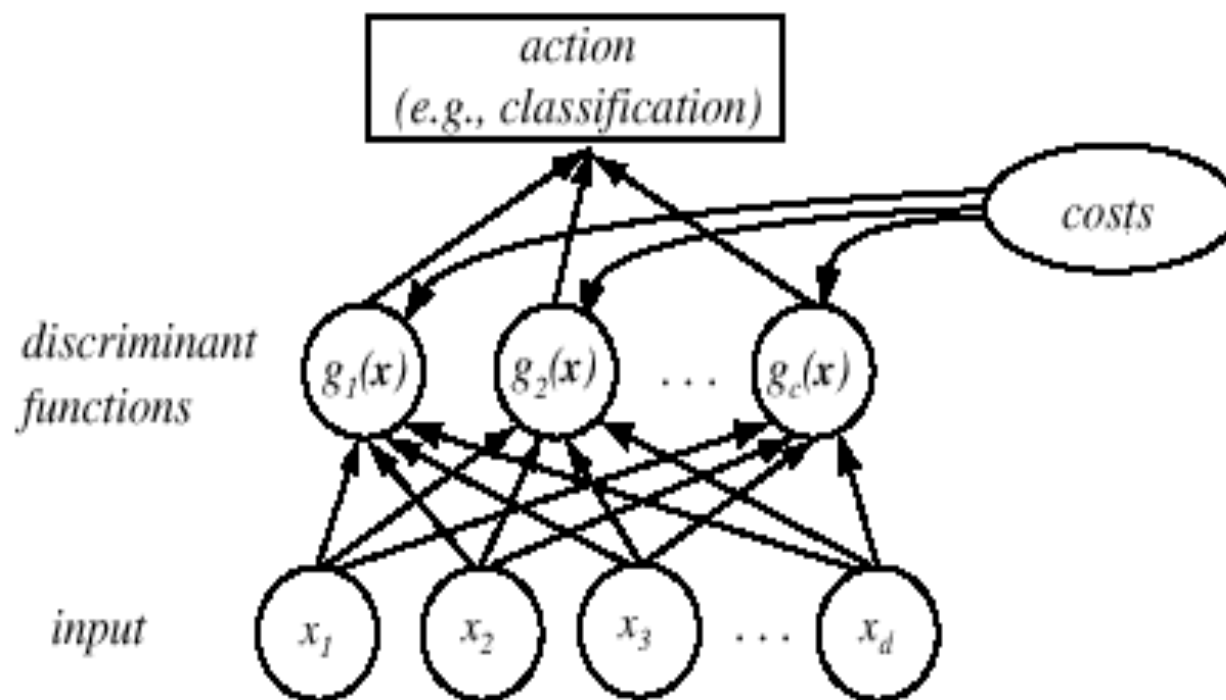


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction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

- Let $g_i(x) = -R(\alpha_i | x)$
(max. discriminant corresponds to min. risk!)

- For the minimum error rate, we take
$$g_i(x) = P(\omega_i | x)$$

(max. discrimination corresponds to max. posterior!)

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

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

(ln: natural logarithm!)

- Feature space divided into c decision regions

if $g_i(x) > g_j(x)$ for all $j \neq i$ then x is in \mathcal{R}_i

(\mathcal{R}_i means assign x to ω_i)

- The two-category case
 - A classifier is a “dichotomizer” that has two discriminant functions g_1 and g_2

Let $g(x) = g_1(x) - g_2(x)$

Decide ω_1 if $g(x) > 0$; Otherwise decide ω_2

- The computation of $g(x)$

$$\begin{aligned} g(x) &= P(\omega_1 | x) - P(\omega_2 | x) \\ &= \ln \frac{P(x | \omega_1)}{P(x | \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)} \end{aligned}$$

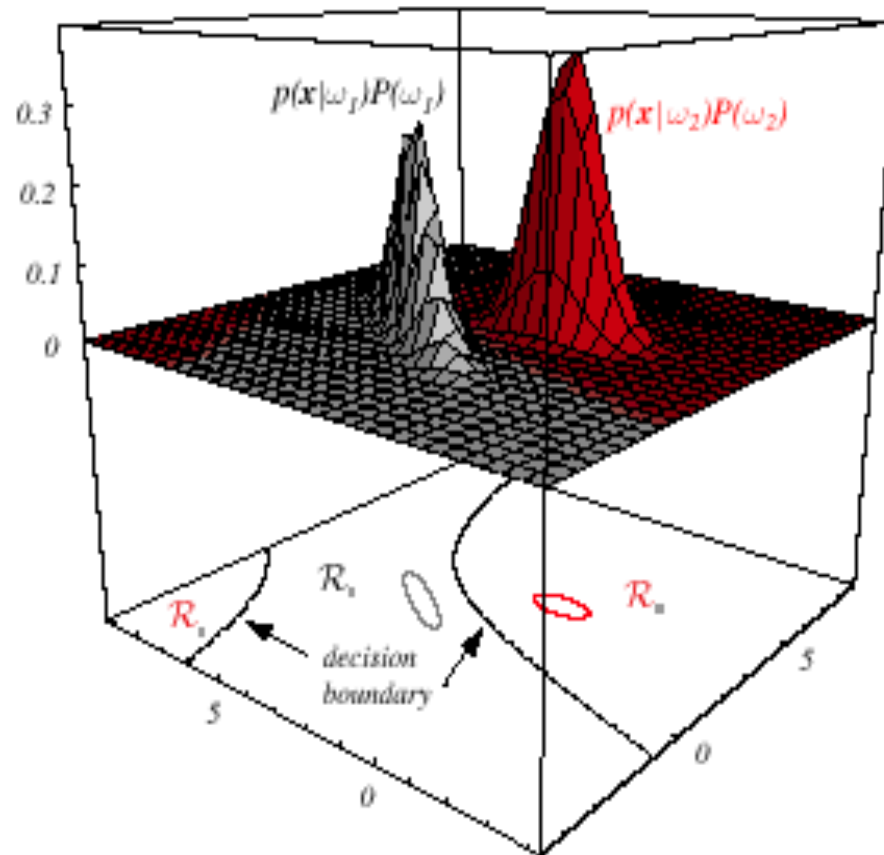


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region \mathcal{R}_2 is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution.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

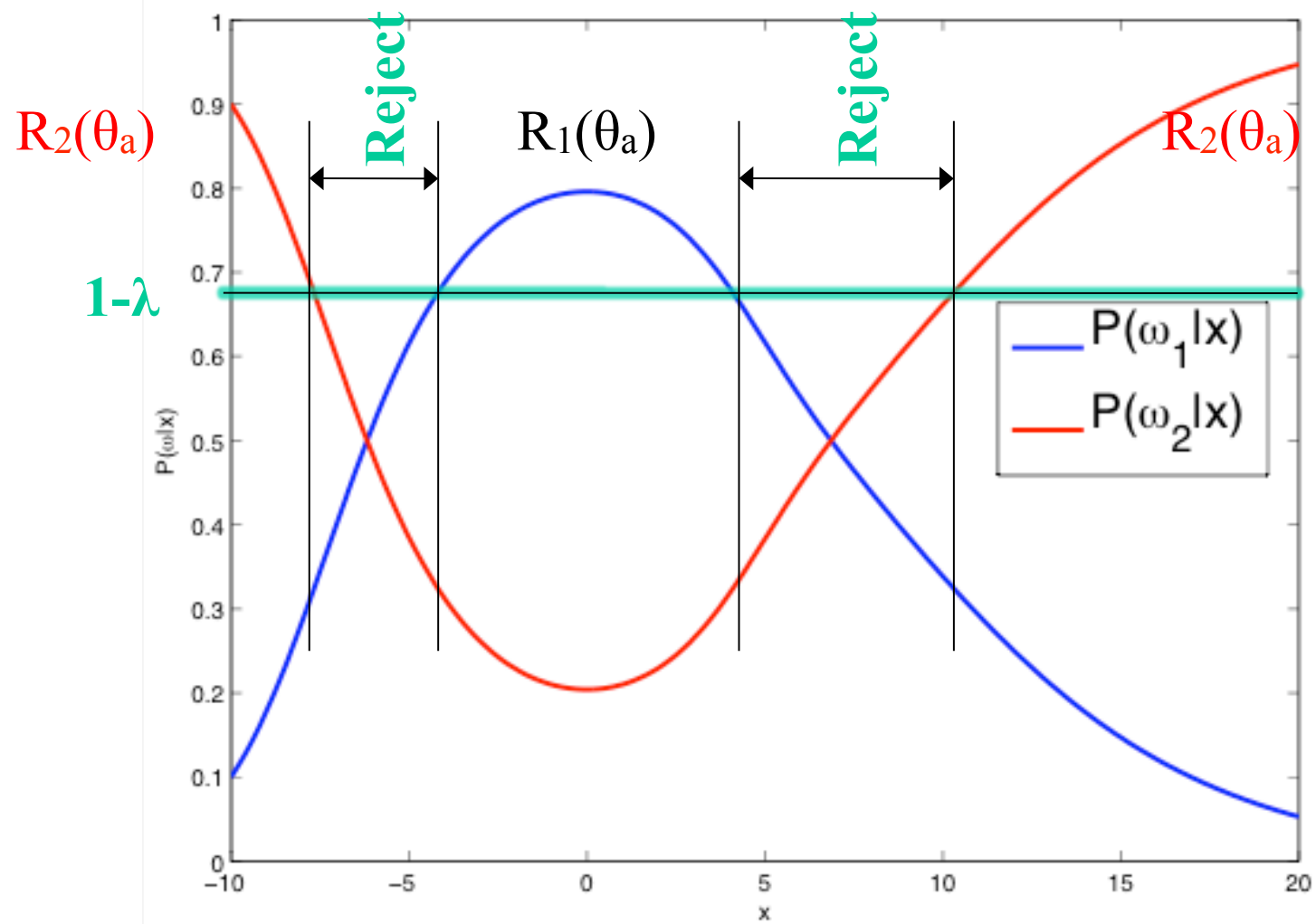
Decision with Reject Option

$$\lambda_{ik} = \begin{cases} 0 & \text{if } i = k \\ \lambda & \text{if } i = K + 1, \quad 0 < \lambda < 1 \\ 1 & \text{otherwise} \end{cases}$$

$$R(\alpha_{K+1} | \mathbf{x}) = \sum_{k=1}^K \lambda P(\omega_k | \mathbf{x}) = \lambda$$

$$R(\alpha_i | \mathbf{x}) = \sum_{k \neq i} P(\omega_k | \mathbf{x}) = 1 - P(\omega_i | \mathbf{x})$$

choose ω_i if $P(\omega_i | \mathbf{x}) > P(\omega_k | \mathbf{x}) \quad \forall k \neq i$ and $P(\omega_i | \mathbf{x}) > 1 - \lambda$
reject otherwise



– $P(\omega_1) = 2/3$, $P(\omega_2) = 1/3$

Utility Theory

- Prob of state k given evidence \mathbf{x} : $P(\omega_k | \mathbf{x})$
- Utility of α_i when state is k : U_{ik}
- Expected utility:

$$EU(\alpha_i | \mathbf{x}) = \sum_k U_{ik} P(\omega_k | \mathbf{x})$$

Choose α_i if $EU(\alpha_i | \mathbf{x}) = \max_j EU(\alpha_j | \mathbf{x})$

Value of Information

- Expected utility using \mathbf{x} only

$$EU(\mathbf{x}) = \max_i \sum_k U_{ik} P(\omega_k | \mathbf{x})$$

- Expected utility using \mathbf{x} and new feature z

$$EU(\mathbf{x}, z) = \max_i \sum_k U_{ik} P(\omega_k | \mathbf{x}, z)$$

- z is useful if $EU(\mathbf{x}, z) > EU(\mathbf{x})$

Overall Risk

- What one really wants to minimize is overall risk

$$R = \int R(\alpha(x) | x) p(x) dx$$

- Note that overall risk will be minimized if the risk for any argument is minimized! So we know how to minimize overall risk.
- For the special case of 0-1 loss we have the total Bayesian error

$$P(error) = \int P(error | x) p(x) dx$$

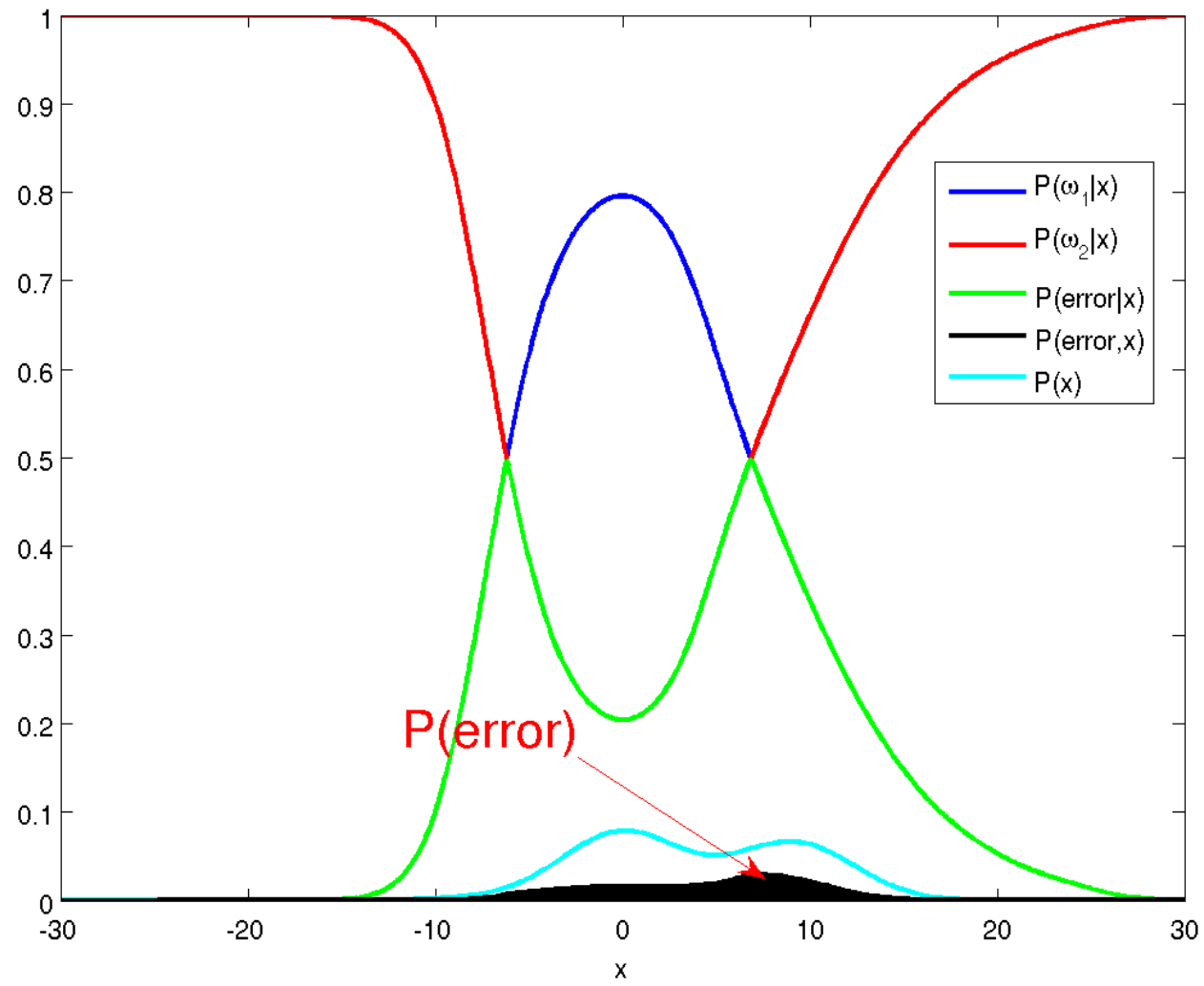
Bayesian Error

- Another way to write the total Bayesian error (binary classes)

$$\begin{aligned}P(error) &= \int P(error | x) p(x) dx \\&= \int \min[P(\omega_1 | x), P(\omega_2 | x)] p(x) dx \\&= \int_{R_1} P(\omega_2 | x) p(x) dx + \int_{R_2} P(\omega_1 | x) p(x) dx\end{aligned}$$

- Note: we cannot typically compute the true Bayesian error unless we know the distribution of the data. We will use the approximation

$$\hat{R} = \sum_{x \in D} R(\alpha(x) | x)$$



Other Risk Minimization Options

- There are other ways to minimize risk.
- Q: What if prior probabilities are not known?
A: Minimize risk over all possible priors.

- Minimax risk

$$\begin{aligned} R &= \int_{R_1} [P(x | \omega_1)p(\omega_1) + \lambda_{12}P(x | \omega_2)p(\omega_2)]dx + \\ &\quad \int_{R_2} [\lambda_{21}P(x | \omega_1)p(\omega_1) + \lambda_{22}P(x | \omega_2)p(\omega_2)]dx \\ &= \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{R_1} P(x | \omega_2)dx + \\ &\quad p(\omega_1) \left[\lambda_{11} - \lambda_{22} + (\lambda_{21} - \lambda_{11}) \int_{R_2} P(x | \omega_1)dx - (\lambda_{12} - \lambda_{22}) \int_{R_1} P(x | \omega_2)dx \right] \end{aligned}$$

The Normal Density

- Univariate density
 - Density which is analytically tractable
 - Continuous density
 - A lot of processes are asymptotically Gaussian
 - Handwritten characters, speech sounds are ideal or prototype corrupted by random process (central limit theorem)

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

Where:

μ = mean (or expected value) of x

σ^2 = expected squared deviation or variance

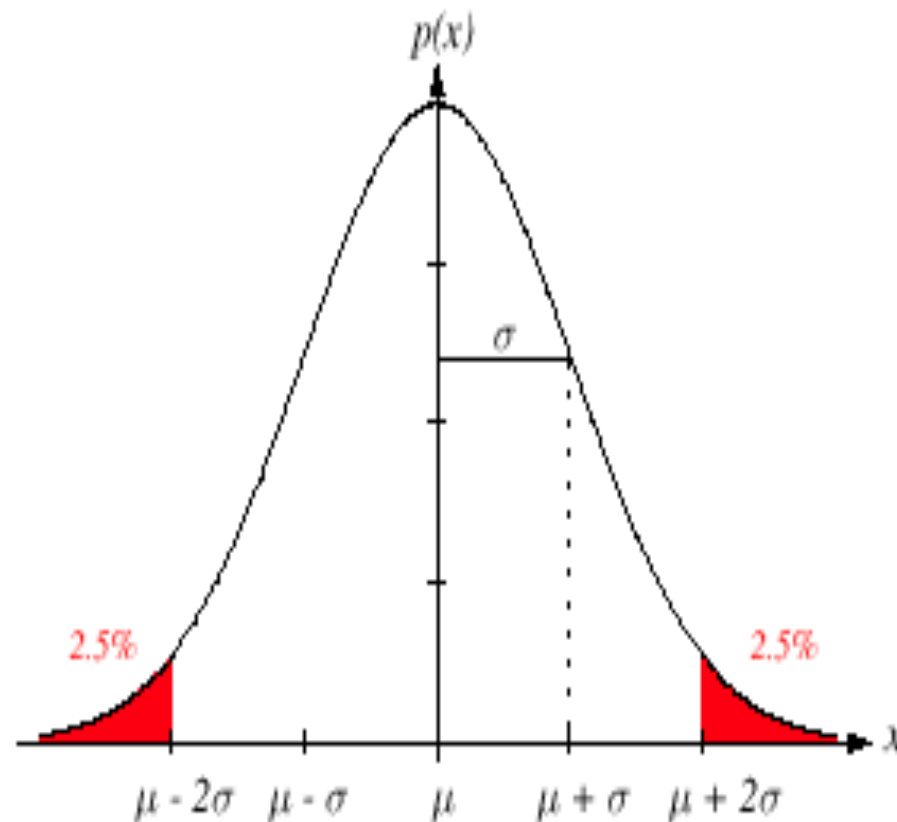


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John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

- Multivariate density

- Multivariate normal density in d dimensions is:

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

where:

$\mathbf{x} = (x_1, x_2, \dots, x_d)^t$ (t stands for the transpose vector form)

$\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_d)^t$ mean vector

$\Sigma = d \times d$ covariance matrix

$|\Sigma|$ and Σ^{-1} are determinant and inverse respectively

- Mahalanobis distance

- The term in the exponent of the multivariate normal density is a distance

$$d_{Mahalanobis}^2(x, \mu) = (x - \mu)^t \Sigma^{-1} (x - \mu)$$

$$P(x) = \frac{1}{Z} \exp \left[-\frac{1}{2} d_{Mahalanobis}^2(x, \mu) \right]$$



Discriminant Functions for the Normal Density

- We saw that the minimum error-rate classification can be achieved by the discriminant function
- $g_i(x) = \ln P(x | \omega_i) + \ln P(\omega_i)$
- Case of multivariate normal

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

Three cases

- We will consider three cases:
 - Spherical, shared covariance
 - Arbitrary, shared covariance
 - Arbitrary covariances

Case 1: Spherical, Shared Covariance

- Case $\Sigma_i = \sigma^2.I$ (I stands for the identity matrix)

$$g_i(x) = w_i^t x + w_{i0} \text{ (linear discriminant function)}$$

where :

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

(ω_{i0} is called the threshold for the i th category!)

- A classifier that uses linear discriminant functions is called “a linear machine”
- The decision surfaces for a linear machine are pieces of hyperplanes defined by:

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

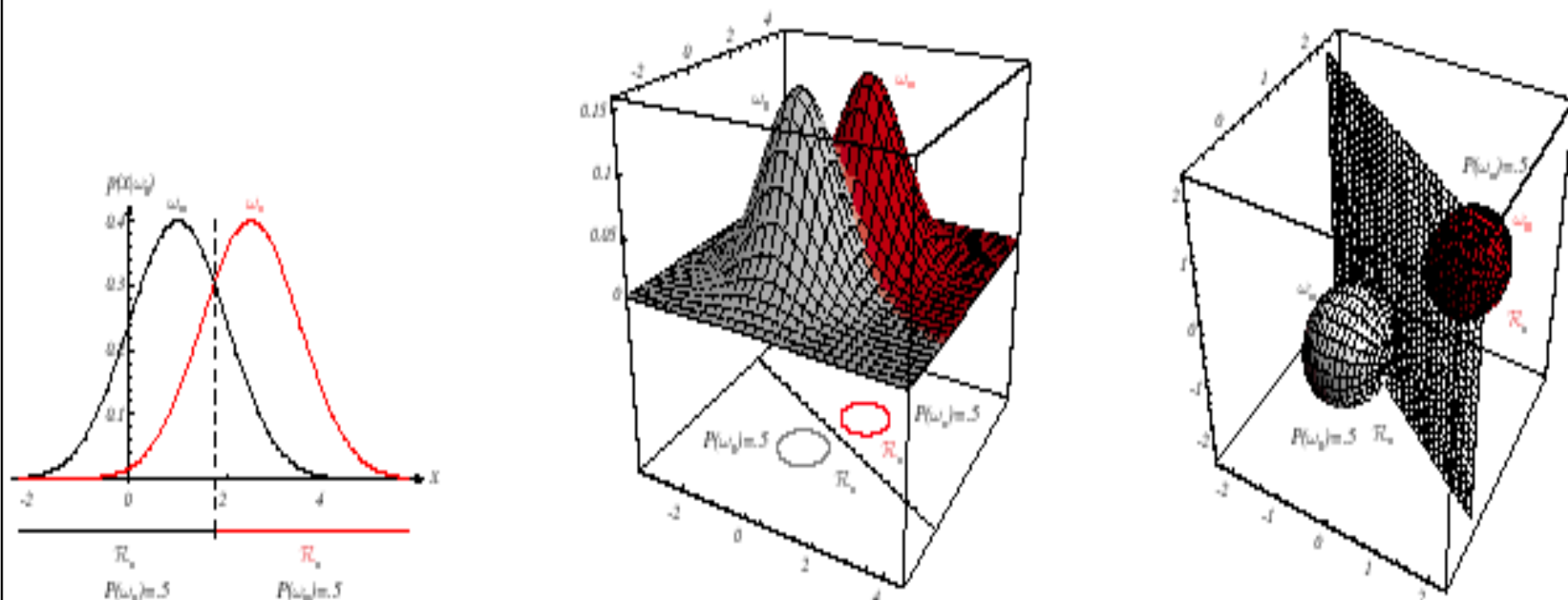


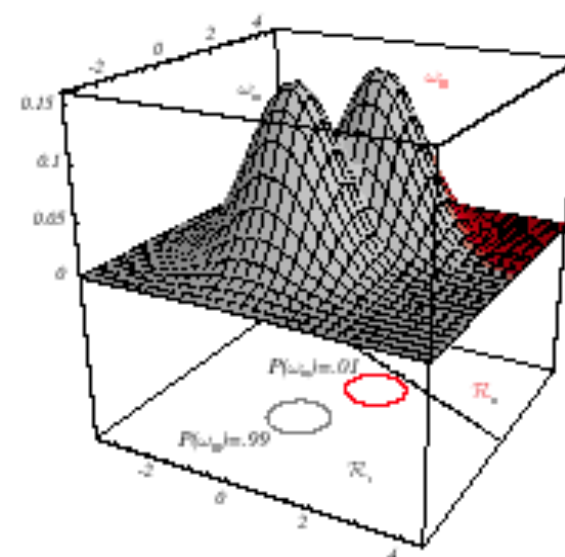
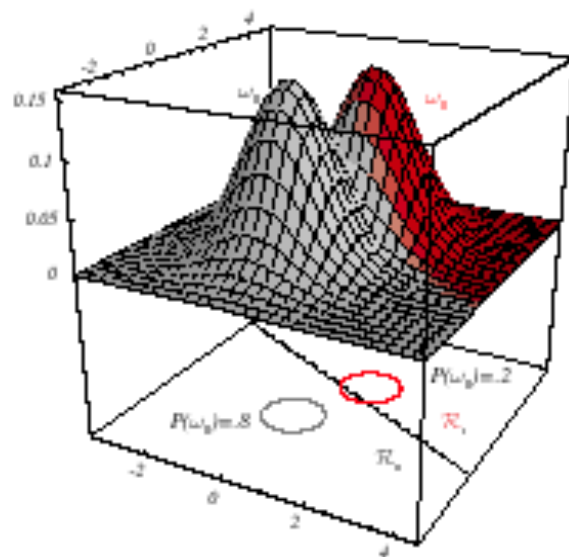
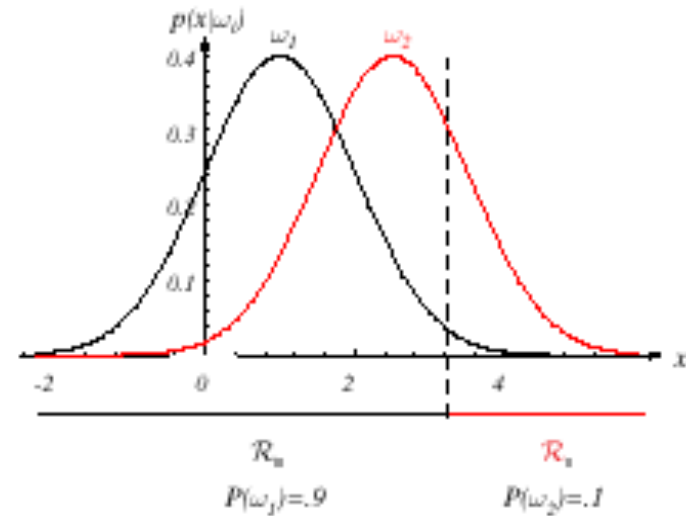
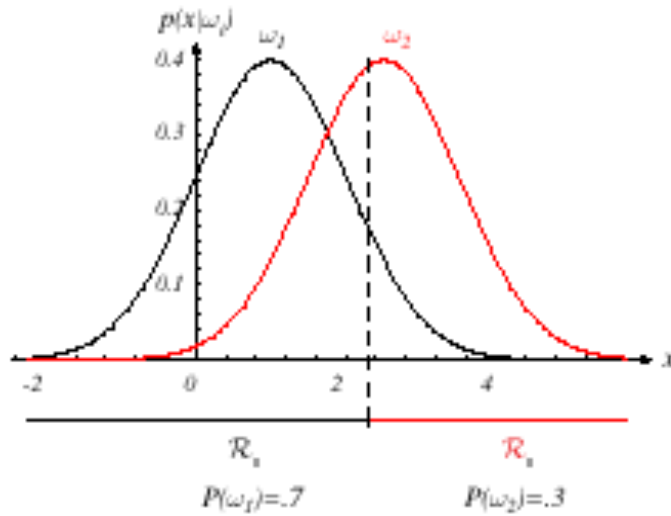
FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d - 1$ dimensions, perpendicular to the line separating the means. In these one-, two-,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

The hyperplane separating \mathcal{R}_i and \mathcal{R}_j passes through

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

and is always orthogonal to the line linking the means!

if $P(\omega_i) = P(\omega_j)$ then $x_0 = \frac{1}{2}(\mu_i + \mu_j)$



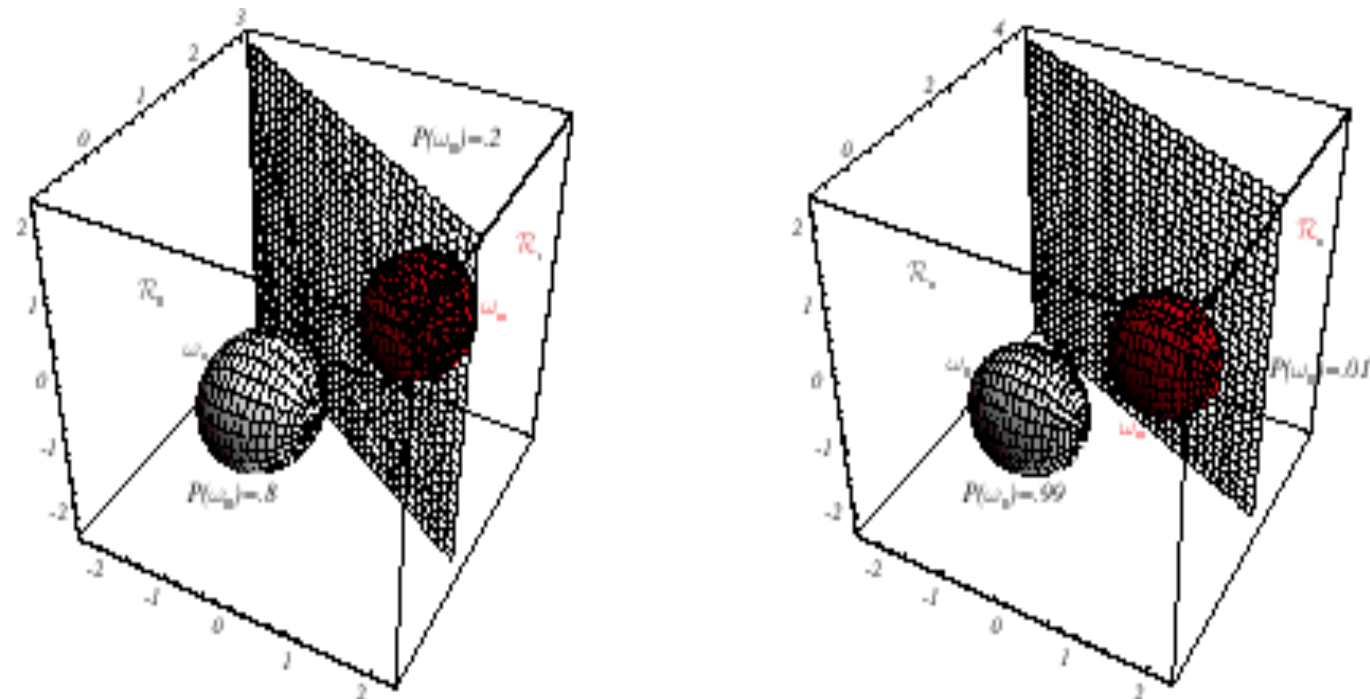


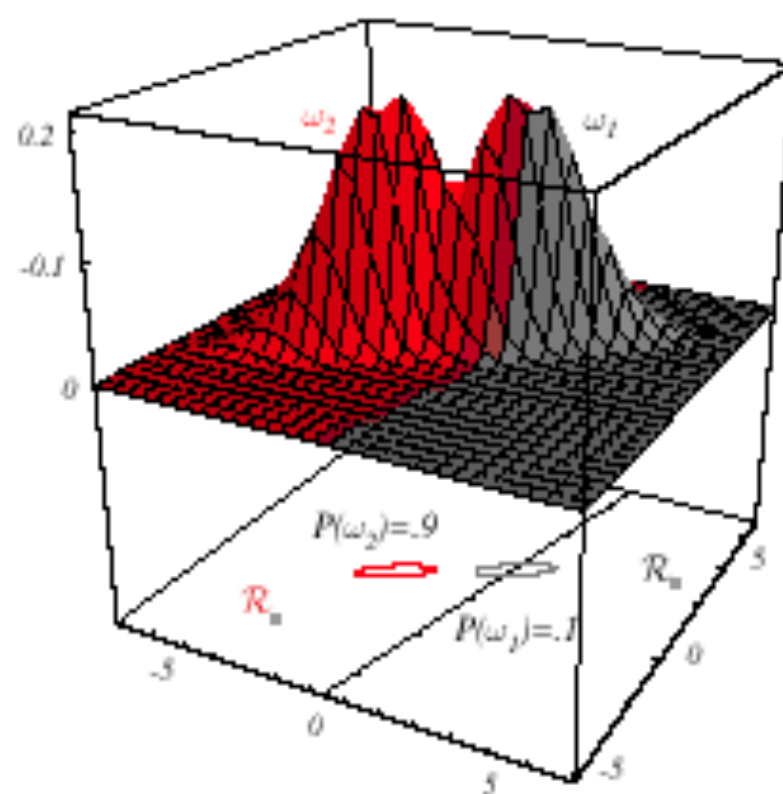
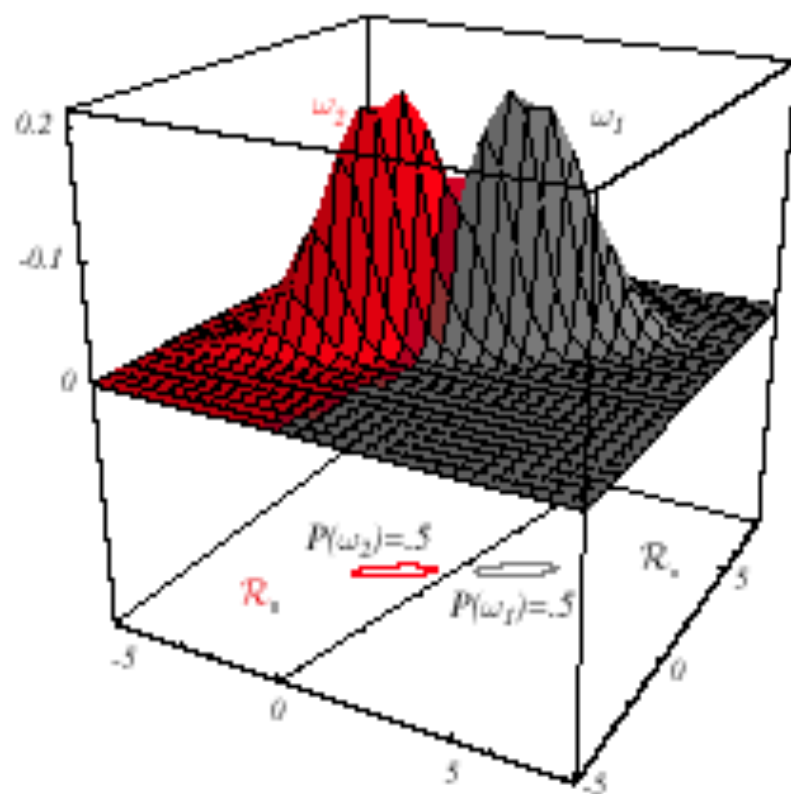
FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

Case 2: Arbitrary Shared Covariances

- **Case $\Sigma_i = \Sigma$** (covariance of all classes are identical but arbitrary!)
- Hyperplane separating \mathcal{R}_i and \mathcal{R}_j passes through

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

(the hyperplane separating \mathcal{R}_i and \mathcal{R}_j is generally not orthogonal to the line between the means!)



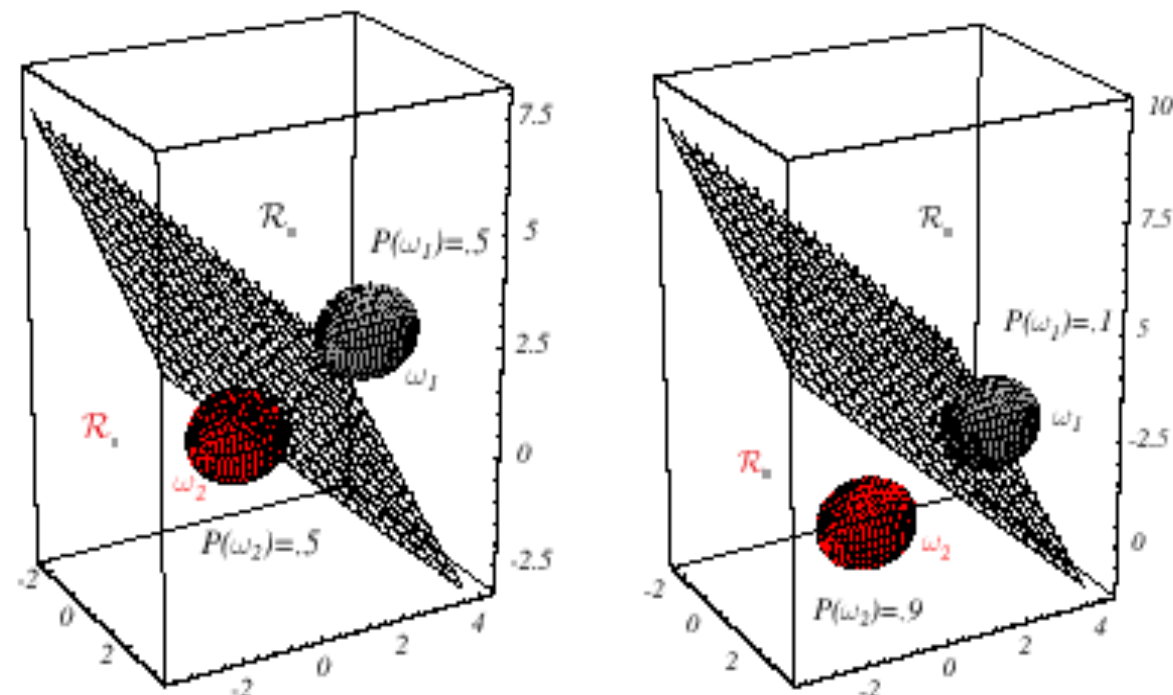


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda,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

Case 3: Arbitrary Covariances

- Case $\Sigma_i = \text{arbitrary}$
 - The covariance matrices are different for each category

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

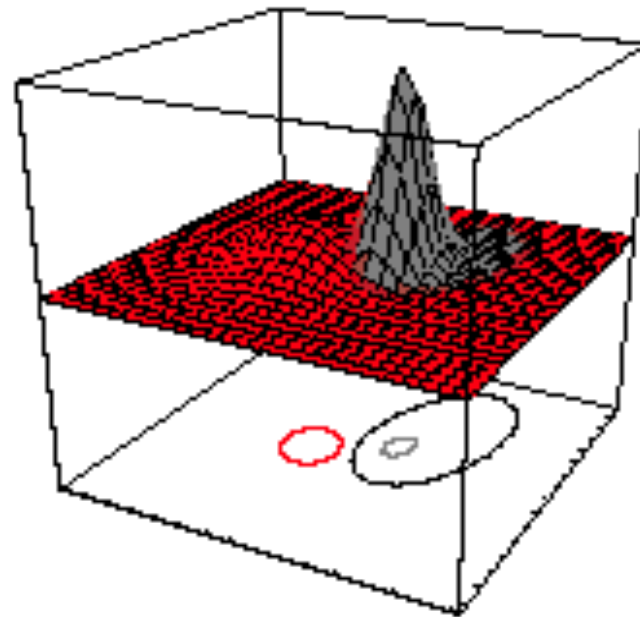
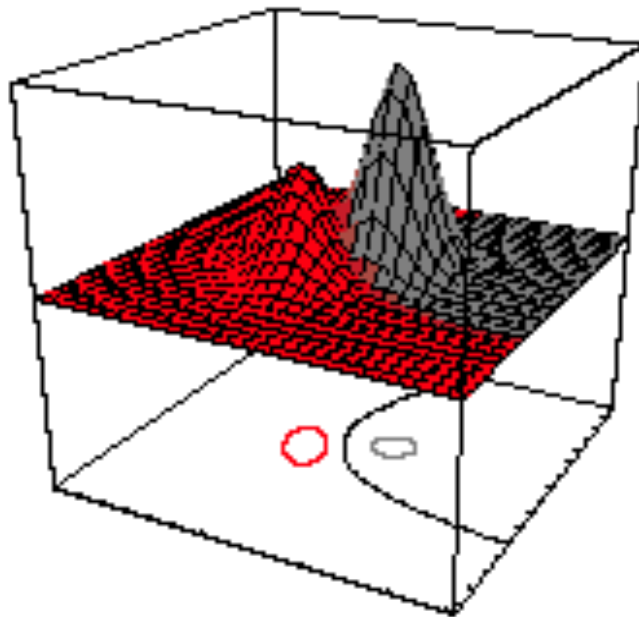
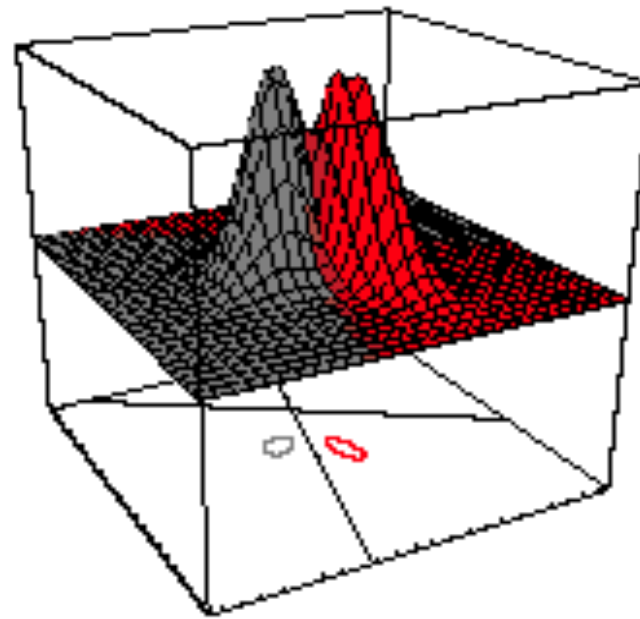
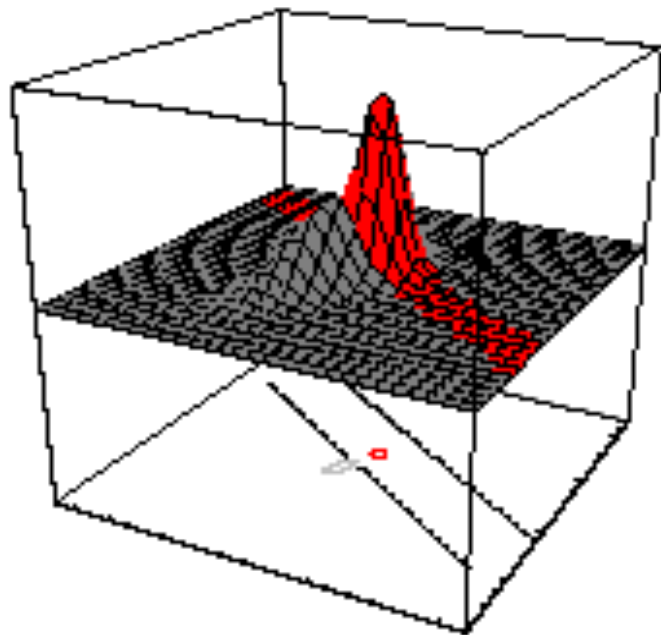
where :

$$\mathbf{W}_i = -\frac{1}{2} \Sigma_i^{-1}$$

$$\mathbf{w}_i = \Sigma_i^{-1} \boldsymbol{\mu}_i$$

$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^t \Sigma_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i)$$

(Hyperquadrics which are: hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, hyperhyperboloids)



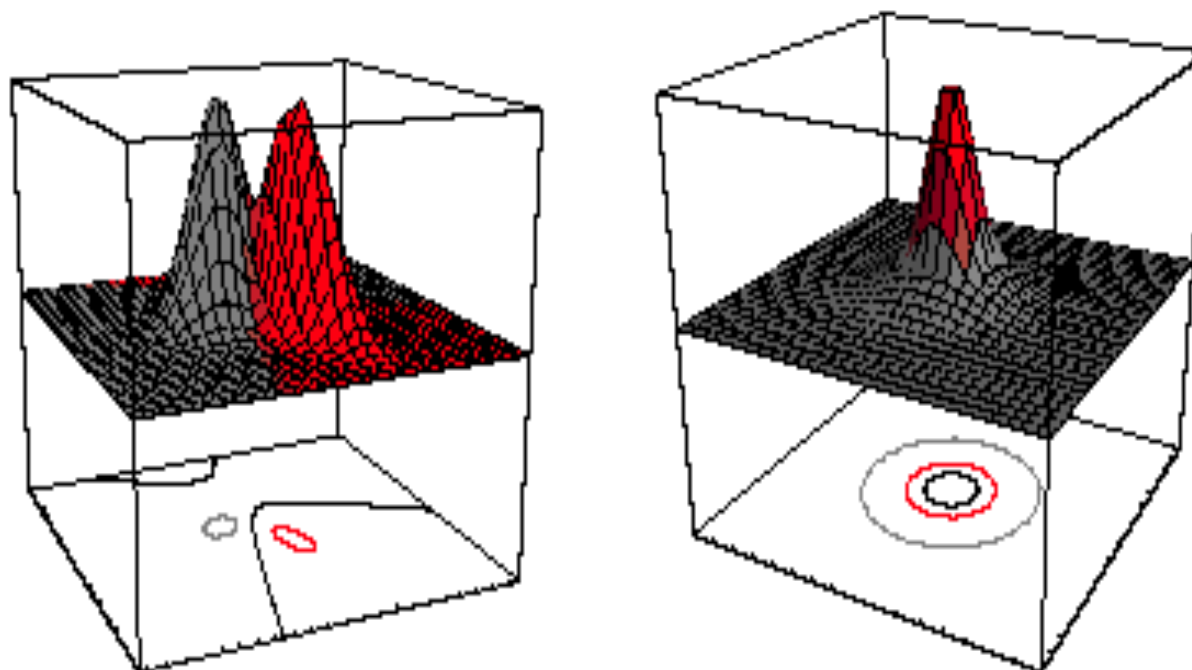


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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

Bayes Decision Theory – Discrete Features

- Components of x are binary or integer valued, x can take only one of m discrete values
- v_1, v_2, \dots, v_m
- Case of independent binary features in 2 category problem
- Let $x = [x_1, x_2, \dots, x_d]^t$ where each x_i is either 0 or 1, with probabilities:

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

- The discriminant function in this case is:

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

where :

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

and :

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

decide ω_1 if $g(x) > 0$ and ω_2 if $g(x) \leq 0$