

Bayesian Decision Theory

Chapter 2

(Jan 11, 18, 23, 25)

- Bayes decision theory is a fundamental statistical approach to pattern classification
- Assumption: decision problem posed in probabilistic terms and relevant probability values are known

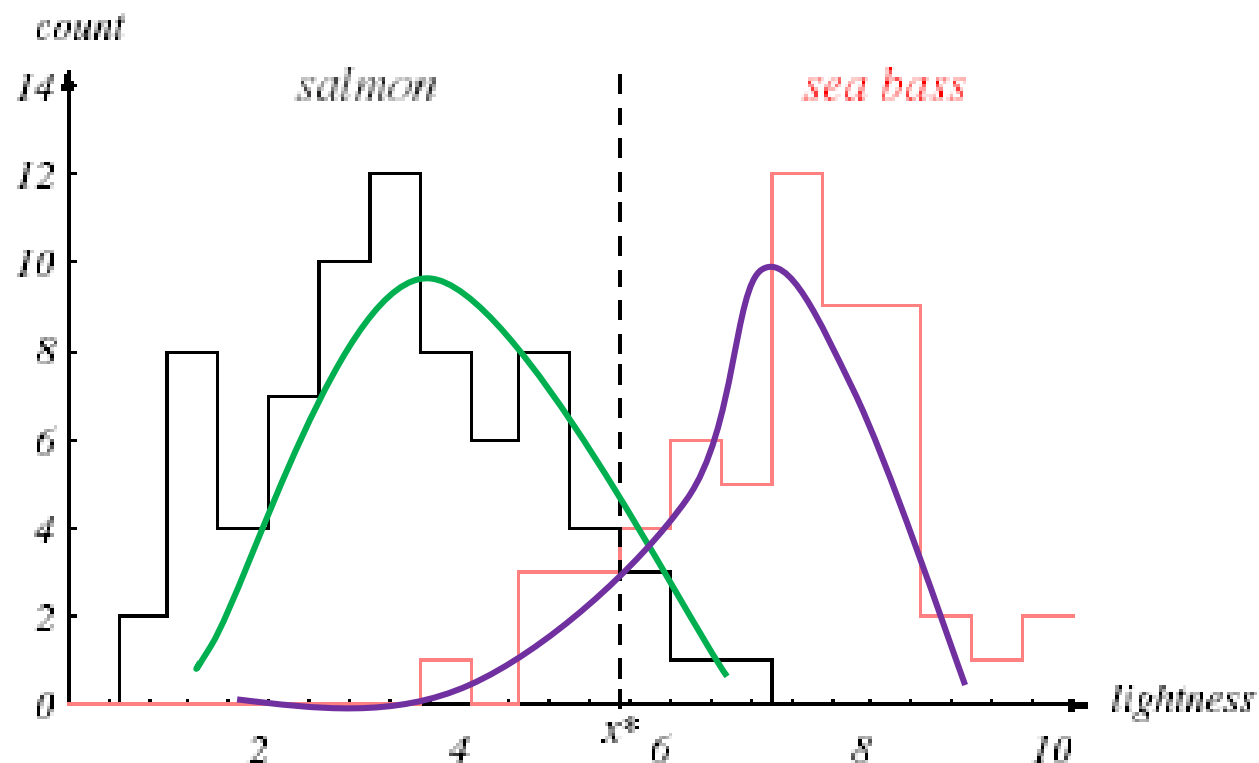


FIGURE 1.3. Histograms for the lightness feature for the two categories. No single threshold value x^* (decision boundary) will serve to unambiguously discriminate between the two categories; using lightness alone, we will have some errors. The value x^* marked will lead to the smallest number of errors, on average. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

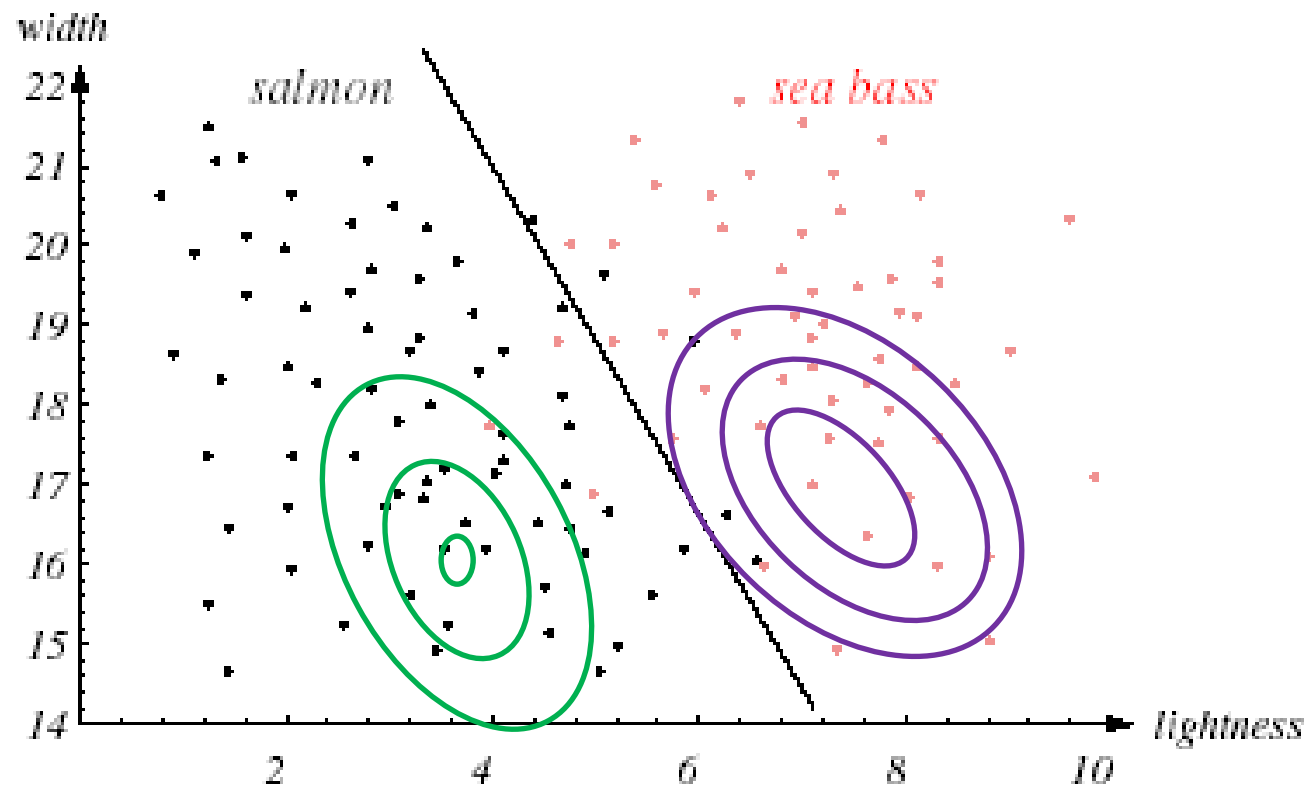
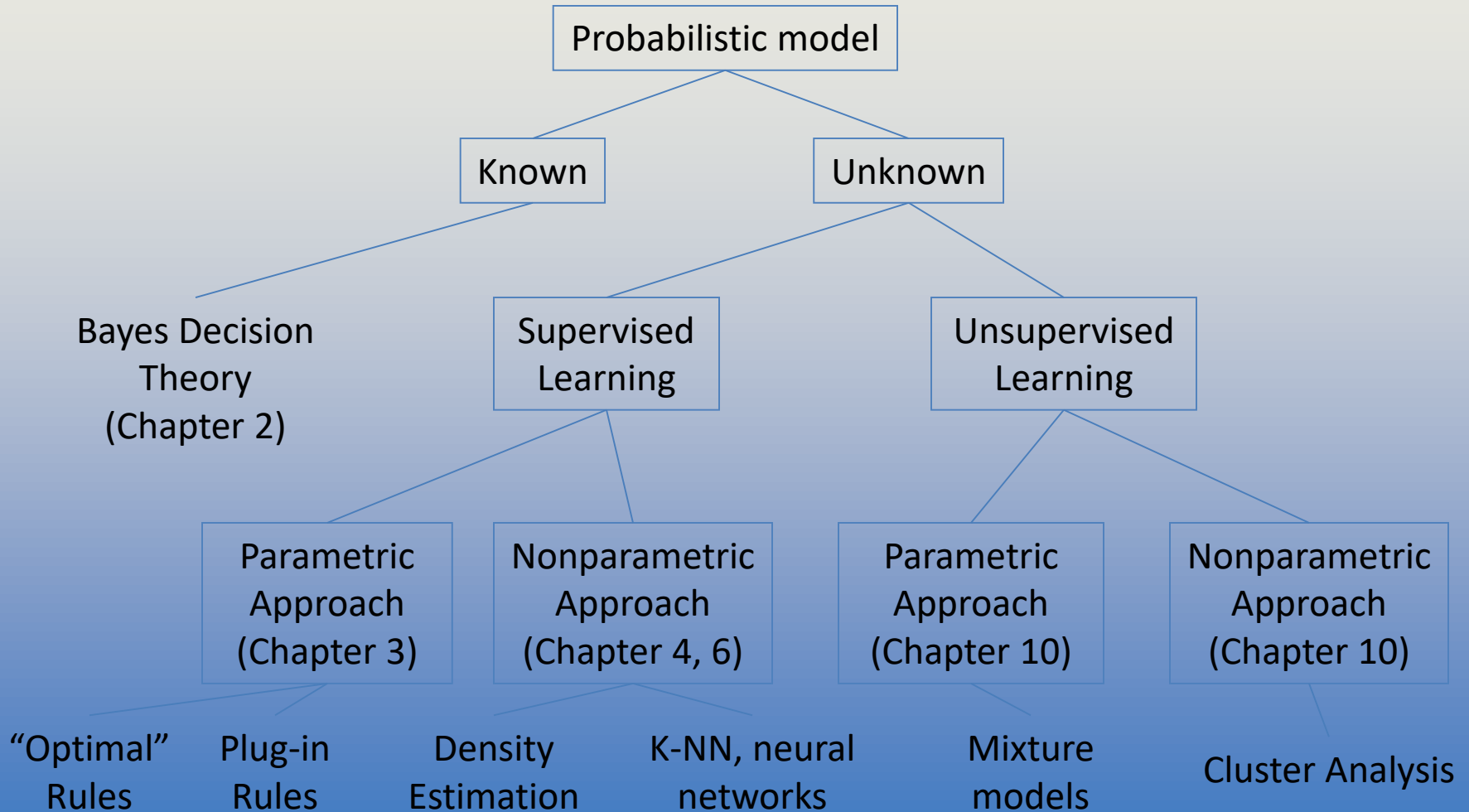


FIGURE 1.4. The two features of lightness and width for sea bass and salmon. The dark line could serve as a decision boundary of our classifier. Overall classification error on the data shown is lower than if we use only one feature as in Fig. 1.3, but there will still be some errors. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Decision Making



Sea bass v. Salmon Classification

- Each fish appearing on the conveyor belt is either sea bass or salmon; two “states of nature”
- Let ω denote the state of nature: ω_1 = sea bass and ω_2 = salmon; ω is a random variable that must be described probabilistically
- *a priori* (prior) probability: $P(\omega_1)$ and $P(\omega_2)$; $P(\omega_1)$ is the probability next fish observed is a sea bass
- *If no other types of fish are present then*
 - $P(\omega_1) + P(\omega_2) = 1$ (exclusivity and exhaustivity)
- $P(\omega_1) = P(\omega_2)$ (uniform priors)
- Prior prob. reflects our prior knowledge about how likely we are to observe a sea bass or salmon; prior prob. may depend on time of the year or the fishing area!

- **Case 1:** Suppose we are asked to make a decision without observing the fish. We only have prior information
- Bayes decision rule given only prior information
 - Decide ω_1 if $P(\omega_1) > P(\omega_2)$, otherwise decide ω_2
- *Error rate* = $\text{Min} \{P(\omega_1) , P(\omega_2)\}$
- Suppose now we are allowed to measure a feature on the state of nature - say the fish lightness value
- Define class-conditional probability density function (pdf) of feature x ; x is a r.v.
- $P(x | \omega_i)$ is the prob. of x given class ω_i , $i = 1, 2$. $P(x | \omega_i) \geq 0$ and area under the pdf is 1.

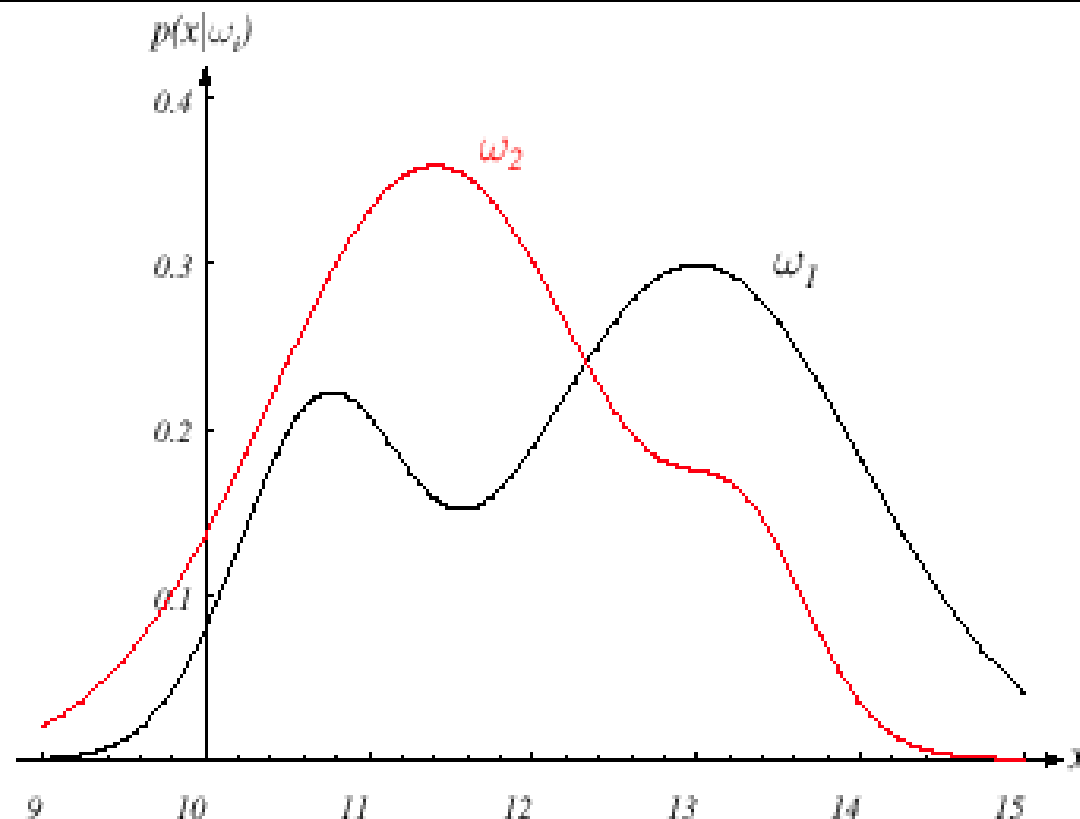


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i . If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two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

Less the densities overlap, better the feature

- **Case 2:** Suppose we only have class-conditional densities and no prior information
- Maximum likelihood decision rule
 - Assign input pattern x to class ω_1 if
$$P(x | \omega_1) > P(x | \omega_2), \text{ otherwise } \omega_2$$
- $P(x | \omega_1)$ is also the likelihood of class ω_1 given the feature value x
- **Case 3:** We have both prior densities and class-conditional densities
- How does the feature x influence our attitude (prior) concerning the true state of nature?
- Bayes decision rule

- Posteriori prob. is a function of likelihood & prior
 - Joint density: $P(\omega_j, x) = P(\omega_j | x)p(x) = p(x | \omega_j) P(\omega_j)$
 - *Bayes rule*

$$P(\omega_j | x) = \{p(x | \omega_j) \cdot P(\omega_j)\} / p(x), j = 1, 2$$

where

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

- Posterior = (Likelihood x Prior) / Evidence
- Evidence $P(x)$ can be viewed as a scale factor that guarantees that the posterior probabilities sum to 1
- $P(x)$ is also called the unconditional density of feature x

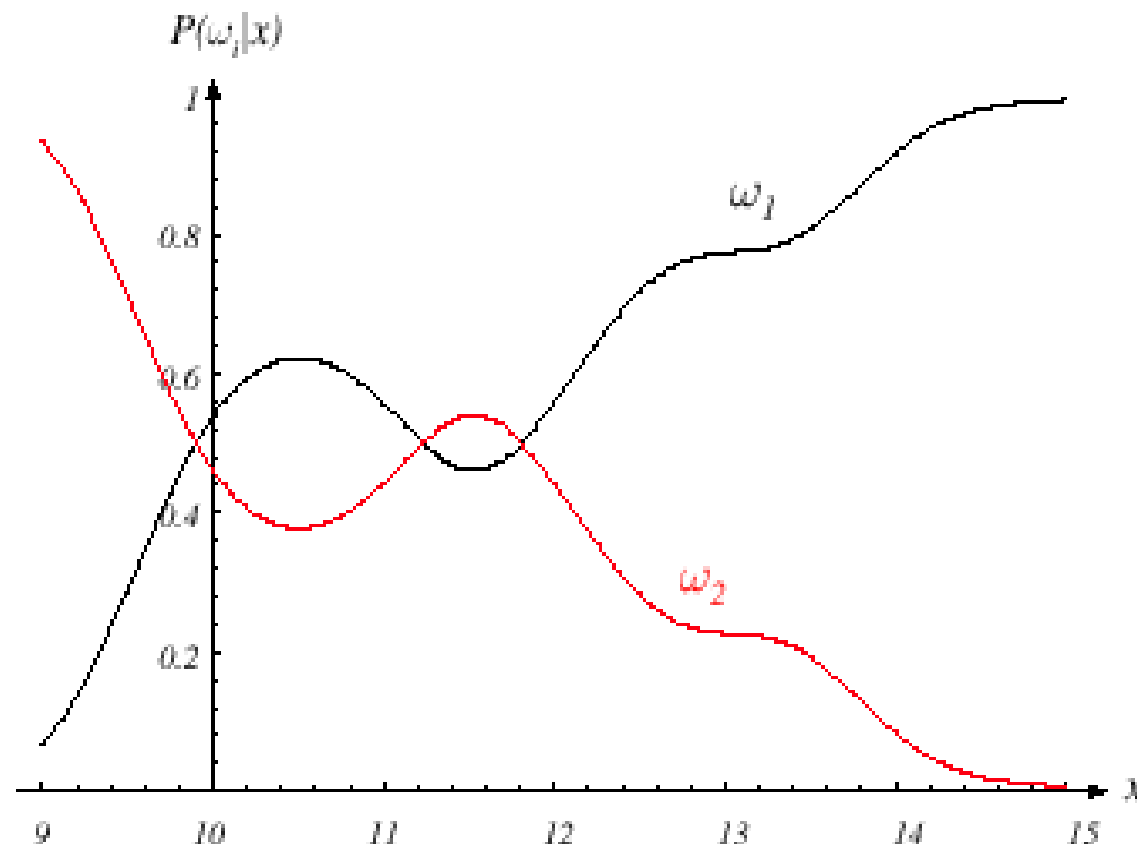


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly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

- $P(\omega_1 | x)$ is the probability of the state of nature being ω_1 given that feature value x has been observed
- Decision based on the posterior probabilities is called the “**Optimal**” Bayes Decision rule. What does optimal mean?

For a given observation (feature value) X :

if $P(\omega_1 | x) > P(\omega_2 | x)$  decide ω_1

if $P(\omega_1 | x) < P(\omega_2 | x)$  decide ω_2

To justify the above rule, calculate the probability of error:

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

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

- So, for a given x , we can minimize the prob. of error by deciding ω_1 if

$$P(\omega_1 | x) > P(\omega_2 | x);$$

otherwise decide ω_2

Therefore:

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

- For each observation x , Bayes decision rule minimizes the probability of error
- Unconditional error: $P(\text{error})$ obtained by *integration over all possible observed x w.r.t. $p(x)$*

- Optimal Bayes decision rule

Decide ω_1 if $P(\omega_1 | \mathbf{x}) > P(\omega_2 | \mathbf{x})$; otherwise decide ω_2

- Special cases:

(i) $P(\omega_1) = P(\omega_2)$; Decide ω_1 if

$p(\mathbf{x} | \omega_1) > p(\mathbf{x} | \omega_2)$, otherwise ω_2

(ii) $p(\mathbf{x} | \omega_1) = p(\mathbf{x} | \omega_2)$; Decide ω_1 if

$P(\omega_1) > P(\omega_2)$, otherwise ω_2

Bayesian Decision Theory – Continuous Features

- Generalization of the preceding formulation
 - Use of more than one feature (d features)
 - Use of more than two states of nature (c classes)
 - Allowing actions other than deciding on the state of nature
 - Introduce a “loss function”; minimizing the “risk” is more general than minimizing the probability of error

- Allowing actions other than classification primarily allows the possibility of “rejection”
- **Rejection:** Input pattern is rejected when it is difficult to decide between two classes or the pattern is too noisy!
- The loss function specifies the cost of each action

- Let $\{\omega_1, \omega_2, \dots, \omega_c\}$ be the set of c states of nature (or “categories” or “classes”)
- Let $\{\alpha_1, \alpha_2, \dots, \alpha_a\}$ be the set of a possible actions that can be taken for an input pattern x
- Let $\lambda(\alpha_i | \omega_j)$ be the loss incurred for taking action α_i when the true state of nature is ω_j
- Decision rule: $\alpha(x)$ specifies which action to take for every possible observation x

Conditional Risk

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

For a given x , suppose we take the action α_i

- If the true state is ω_j , we will incur the loss $\lambda(\alpha_i / \omega_j)$
- $P(\omega_j / x)$ is the prob. that the true state is ω_j
- But, any one of the C states is possible for given x

Overall risk

$R = \text{Expected value of } R(\alpha_i / x) \text{ w.r.t. } p(x)$

Conditional risk

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

Select the action α_i for which $R(\alpha_i | x)$ is minimum

- This action minimizes the overall risk
- The resulting risk is called the **Bayes risk**
- It is the best classification performance that can be achieved given the priors, class-conditional densities and the loss function!

- Two-category classification

α_1 : *decide* ω_1

α_2 : *decide* ω_2

$\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$; loss incurred in deciding α_i when the true state of nature is ω_j

Conditional risk:

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

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

Bayes decision rule is stated as:

$$\text{if } R(\alpha_1 | \mathbf{x}) < R(\alpha_2 | \mathbf{x})$$

Take action α_1 : “decide ω_1 ”

This rule is equivalent to: decide ω_1 if:

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

decide ω_2 otherwise

In terms of the Likelihood Ratio (LR), 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)

The “threshold” term on the right hand side now involves the prior and the loss function

Interpretation of the Bayes decision rule:

If the **likelihood ratio** of class ω_1 and class ω_2 exceeds a threshold value (independent of the input pattern x), the **optimal action is: decide ω_1**

Maximum likelihood decision rule is a special case of minimum risk decision rule:

- Threshold value = 1
- 0-1 loss function
- Equal class prior probability

Bayesian Decision Theory

(Sections 2.3-2.5)

- Minimum Error Rate Classification
- Classifiers, Discriminant Functions and Decision Surfaces
- Multivariate Normal (Gaussian) Density

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* or the *error rate*

- Zero-one (0-1) loss function: no loss for correct decision and a unit loss for incorrect decision

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

The conditional risk can now be simplified as:

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

"The risk corresponding to the 0-1 loss function is the average probability of error"

- Minimizing the risk under 0-1 loss function requires maximizing the posterior probability $P(\omega_i / x)$ since

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

- For Minimum error rate
 - Decide ω_i if $P(\omega_i / x) > P(\omega_j / x) \forall j \neq i$

- Decision boundaries and decision regions

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

- If λ is the 0-1 loss function then the threshold involves only the priors:

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

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

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

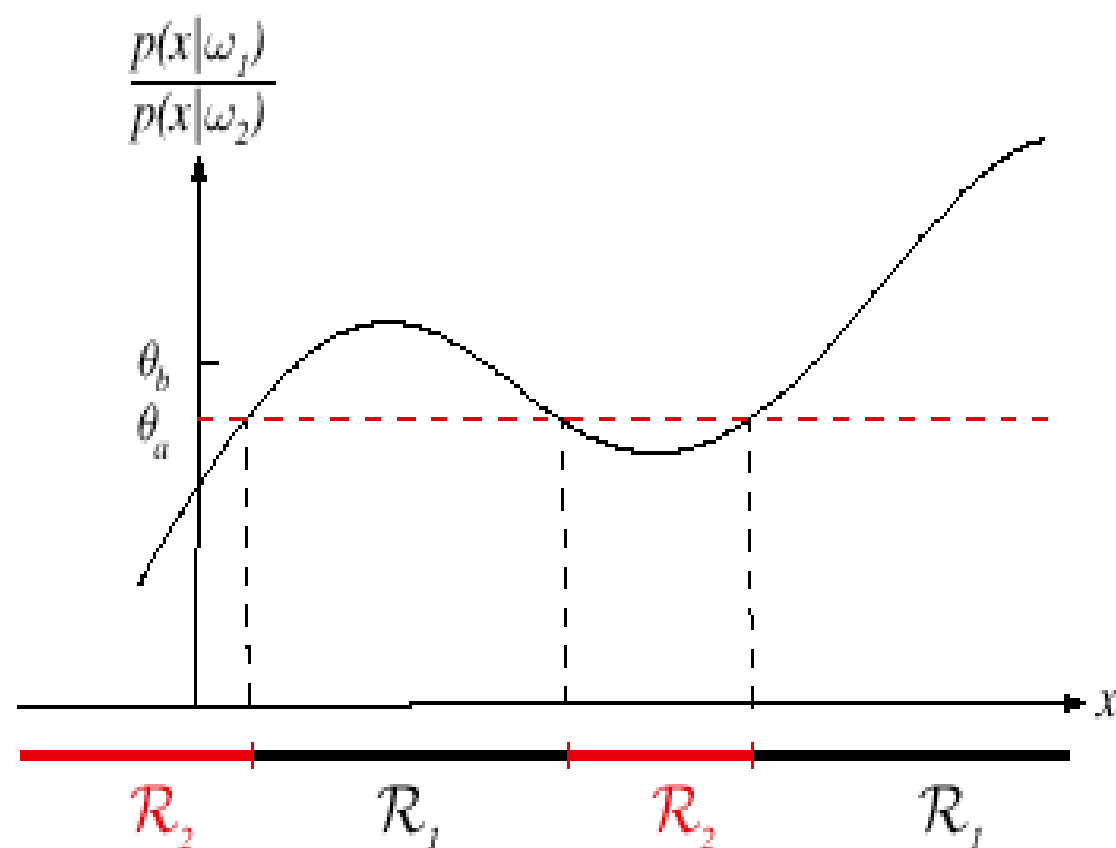


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a . If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse,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

Classifiers, Discriminant Functions and Decision Surfaces

- Many different ways to represent classifiers or decision rules;
- One of the most useful is in terms of “discriminant functions”
- The multi-category case
 - Set of discriminant functions $g_i(x)$, $i = 1, \dots, c$
 - Classifier assigns a feature vector x to class ω_i if:
$$g_i(x) > g_j(x) \quad \forall j \neq i$$

Network Representation of a Classifier

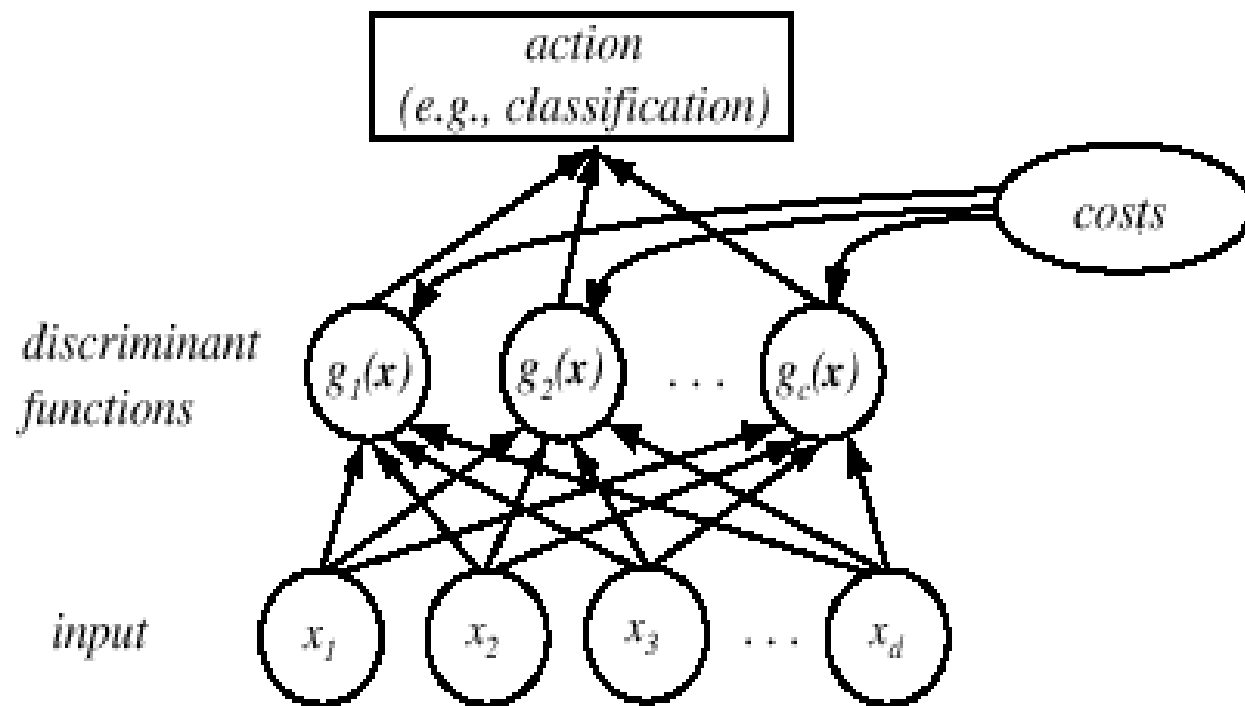


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction 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 classifier can be represented in this way, but the choice of discriminant function is not unique
- $g_i(x) = -R(\alpha_i / x)$
(max. discriminant corresponds to minimum risk)

- For the minimum error rate, we take

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

(max. discrimination corresponds to max. posterior!)

$$g_i(x) \equiv P(x / \omega_i) P(\omega_i)$$

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

(ln: natural log)

- A decision rule partitions the feature space into c decision regions

if $g_i(x) > g_j(x) \forall j \neq i$ then x is in R_i

In region R_i input pattern x is assigned to class ω_i

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

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

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

- A “dichotomizer” computes a single discriminant function $g(x)$ and classifies x according to whether $g(x)$ is positive or not
- Computation of $g(x) = g_1(x) - g_2(x)$

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

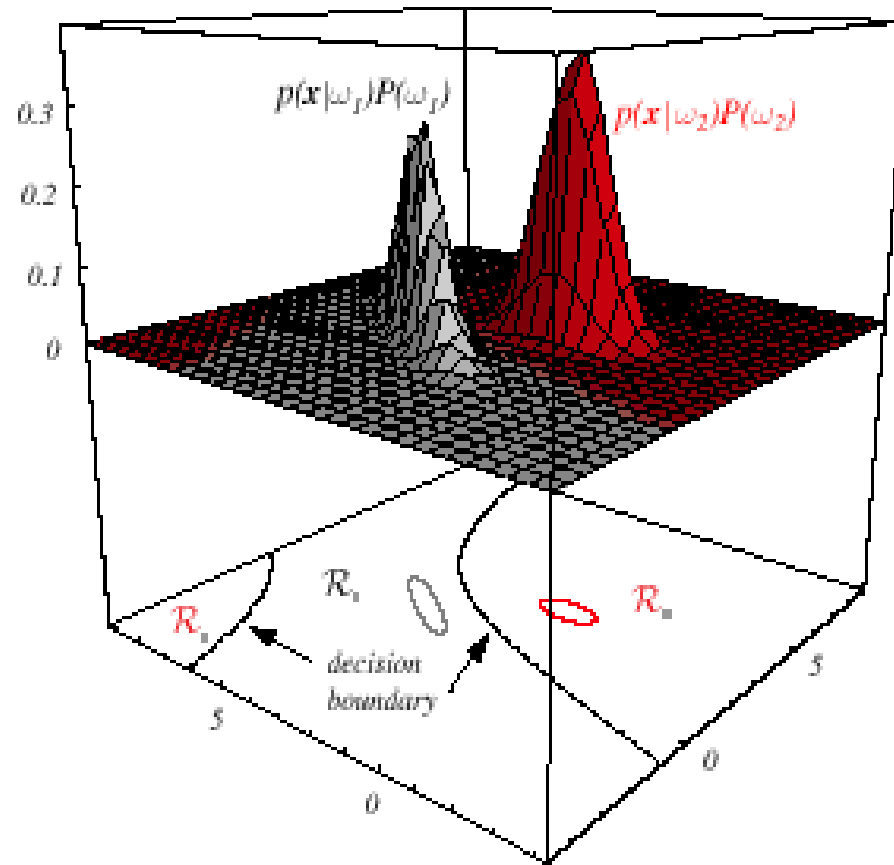


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region \mathcal{R}_2 is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution.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

The Normal Density

- Univariate density: $N(\mu, \sigma^2)$
 - Normal density is analytically tractable
 - Continuous density with two parameters (mean, variance)
 - A number of processes are asymptotically Gaussian (CLT)
 - Patterns (e.g., handwritten characters, speech signals) can be viewed as randomly corrupted (noisy) versions of a single typical or prototype pattern

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

where:

μ = mean (or expected value) of x

σ^2 = variance (or expected squared deviation) of x

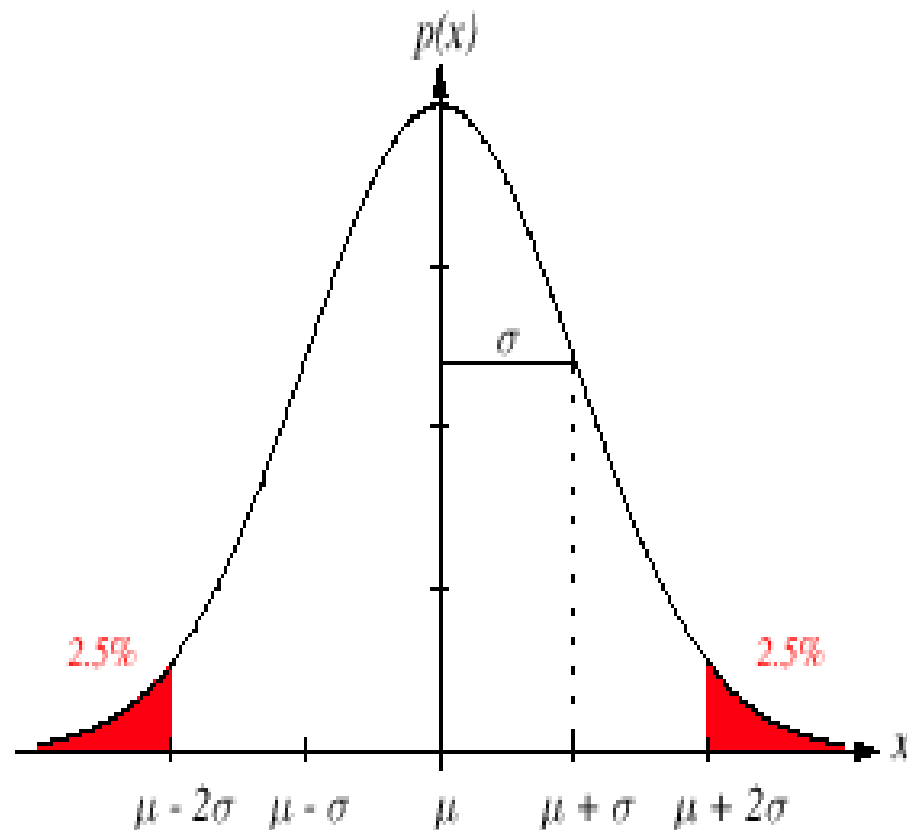


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John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

- Multivariate density: $N(\mu, \Sigma)$

- Multivariate normal density in d dimensions:

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

where:

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

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

$\Sigma = d \times d$ covariance matrix

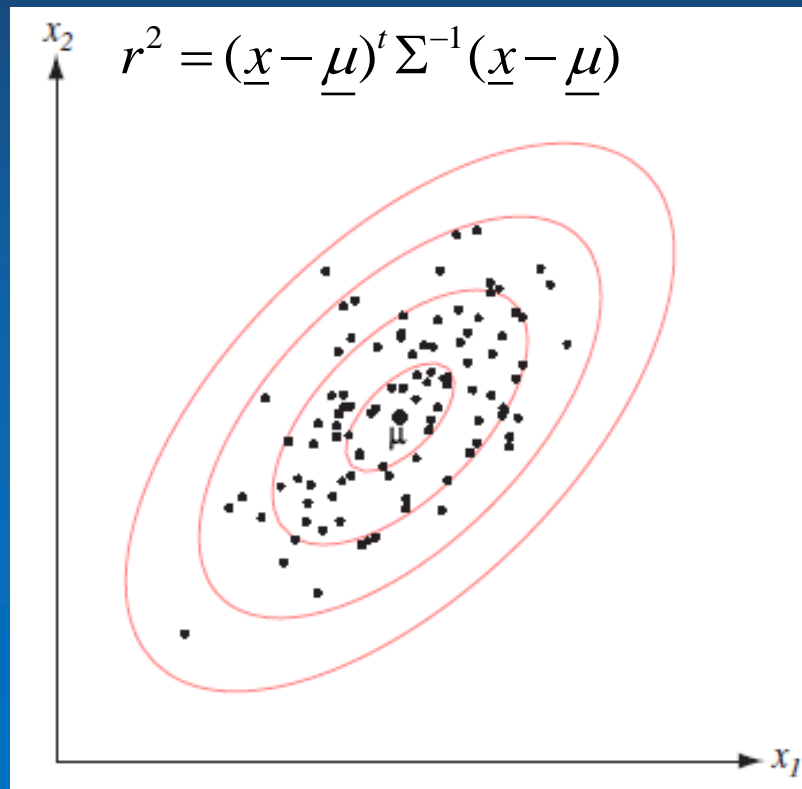
$|\Sigma|$ and Σ^{-1} are determinant and inverse of Σ , respectively

- Covariance matrix is symmetric and positive semidefinite; we assume Σ is positive definite so the determinant of Σ is strictly positive
- Multivariate normal density is completely specified by $[d + d(d+1)/2]$ parameters
- If variables x_1 and x_2 are "statistically independent" then the covariance of x_1 and x_2 is zero.

Multivariate Normal density

Samples drawn from a normal population tend to fall in a single cloud or cluster; cluster center is determined by the mean vector and shape by the covariance matrix

The loci of points of constant density are hyperellipsoids whose principal axes are the eigenvectors of Σ

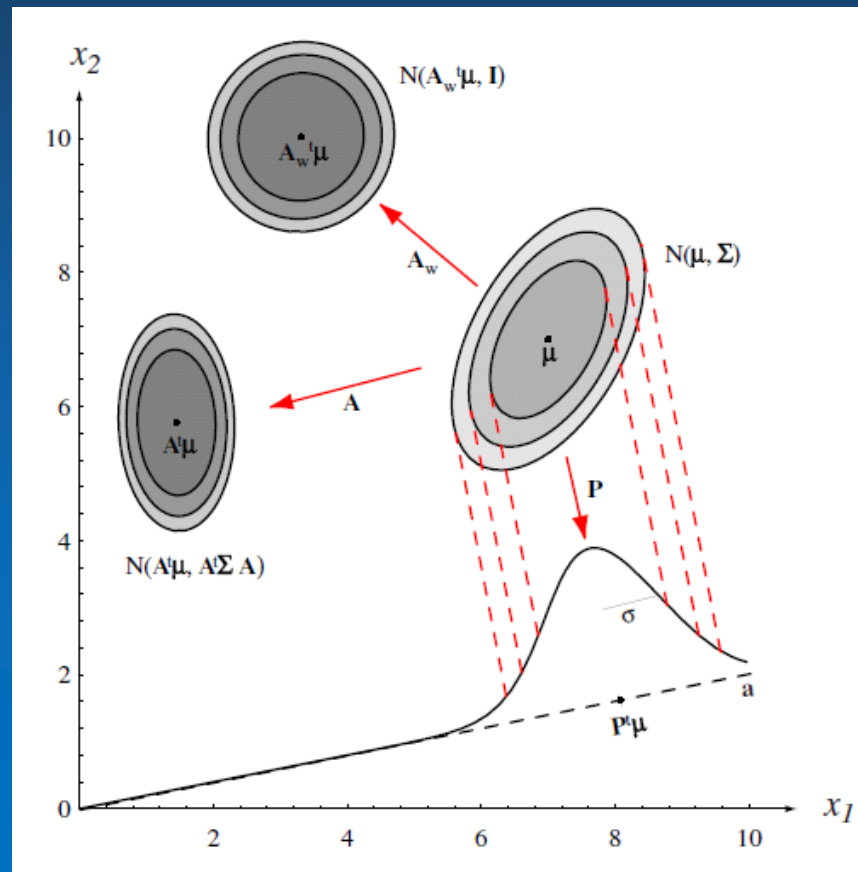


squared Mahalanobis distance from \mathbf{x} to $\boldsymbol{\mu}$

Transformation of Normal Variables

Linear combinations of jointly normal random variables have normal distribution

Linear transformation can convert an arbitrary multivariate normal distribution into a spherical one ("Whitening")



Bayesian Decision Theory

(Sections 2.6 to 2.9)

- Discriminant Functions for the Normal Density
- Bayes Decision Theory – Discrete Features

Discriminant Functions for the Normal Density

- The minimum error-rate classification can be achieved by the discriminant functions

$$g_i(x) = \ln P(x | \omega_i) + \ln P(\omega_i), \quad i = 1, 2, \dots, c$$

- In case of multivariate normal densities

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

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

Features are statistically independent and each feature has the same variance irrespective of the class

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

where :

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

(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 (boundaries) for a linear machine are pieces of **hyperplanes** defined by the linear equations:

$$g_i(\mathbf{x}) = g_j(\mathbf{x})$$

- The hyperplane separating \mathcal{R}_i and \mathcal{R}_j

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

is orthogonal to the line linking the means!

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

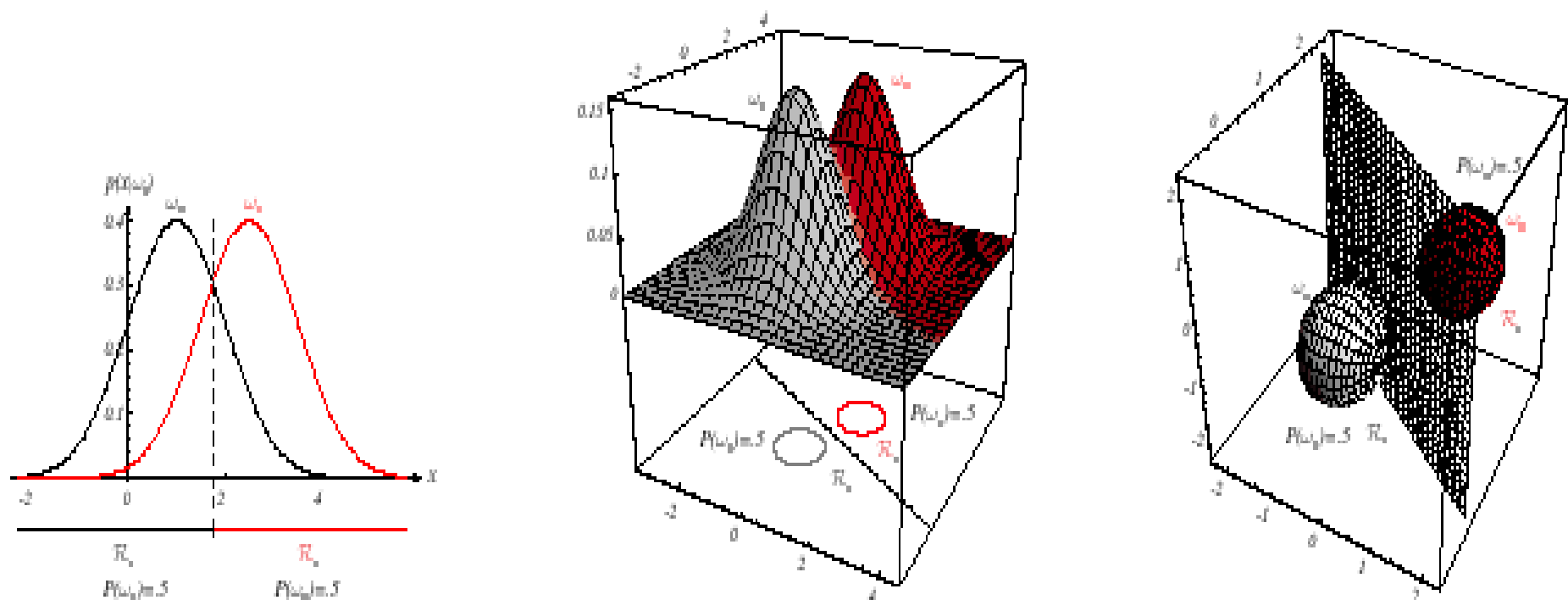
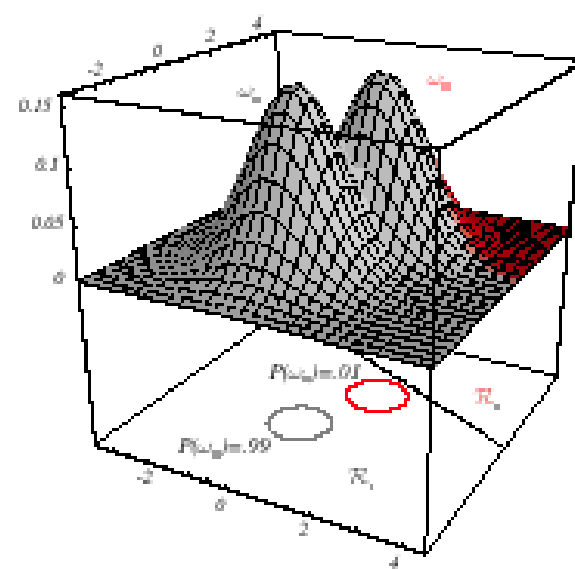
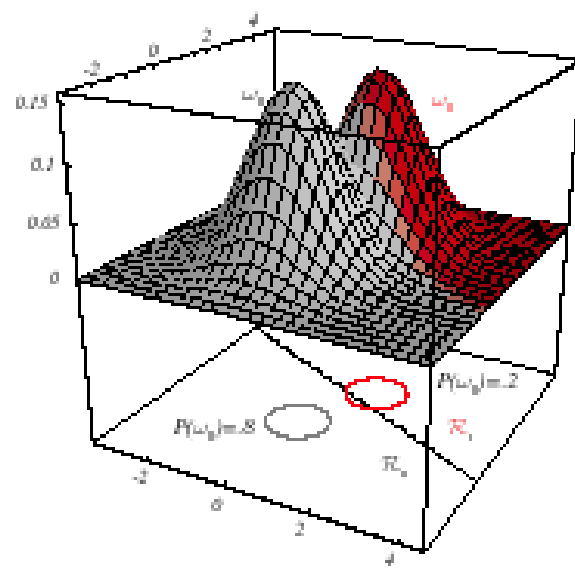
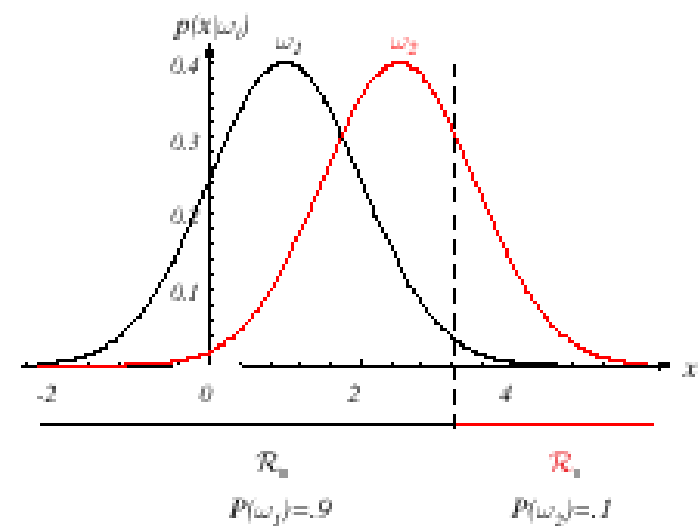
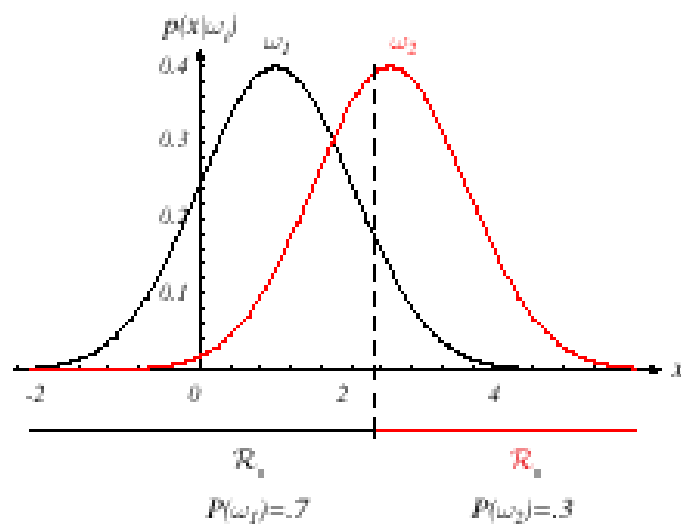


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d - 1$ dimensions, perpendicular to the line separating the means. In these one-, two-,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



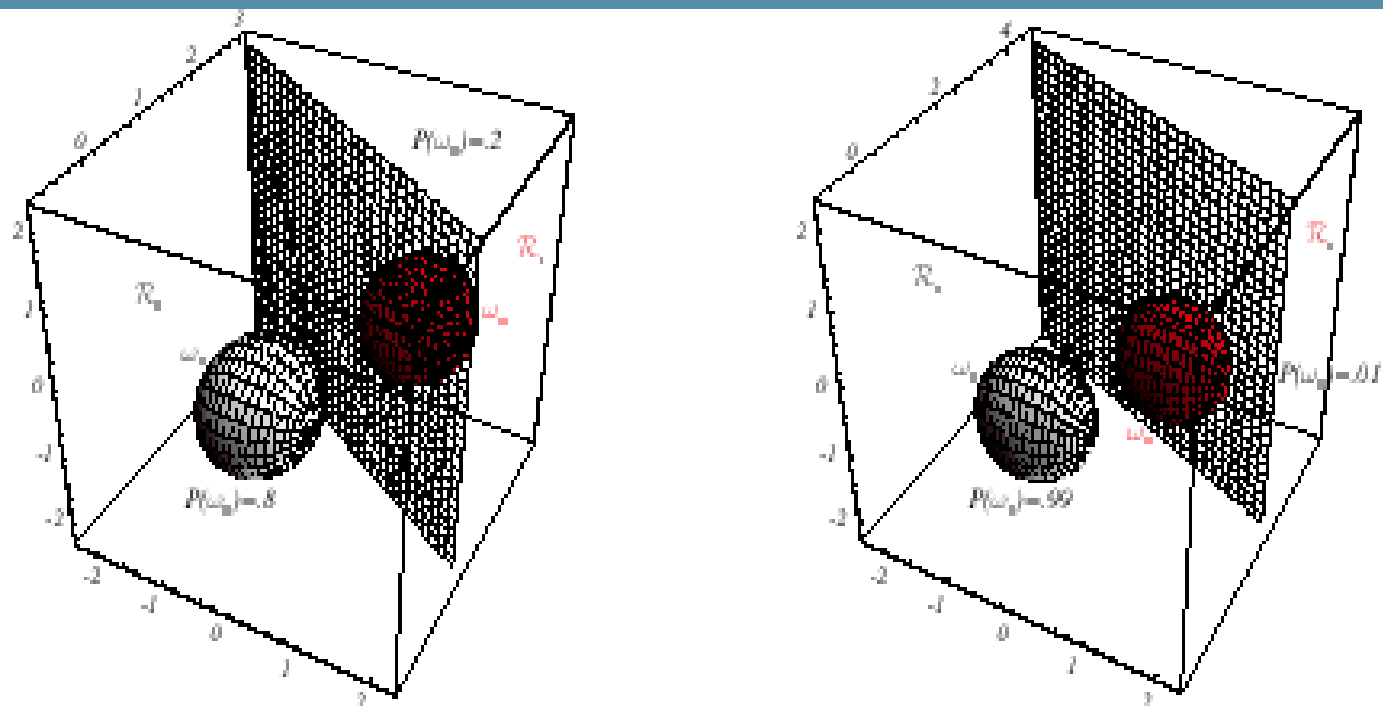
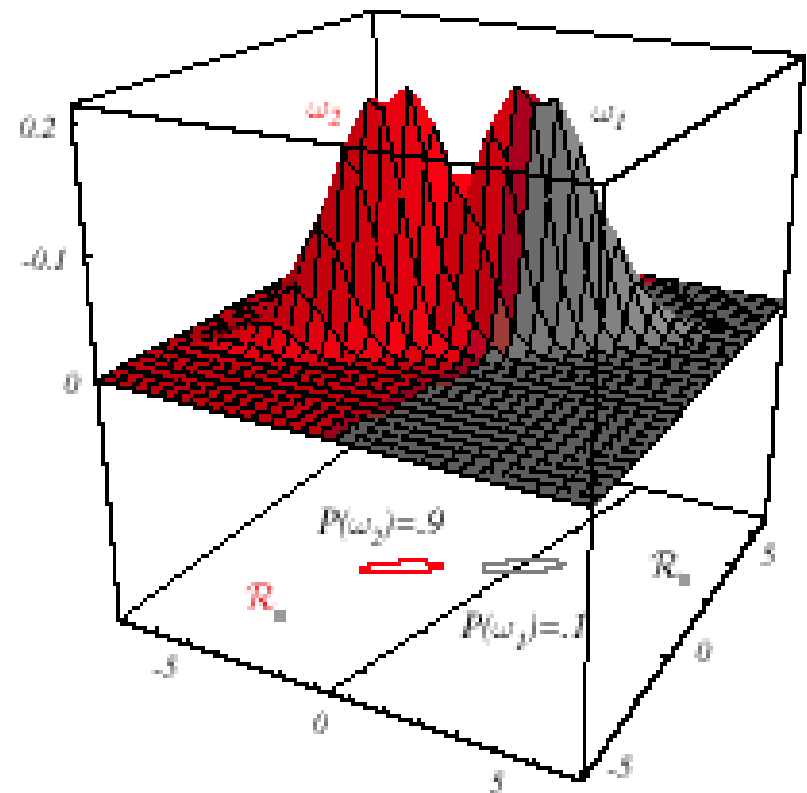
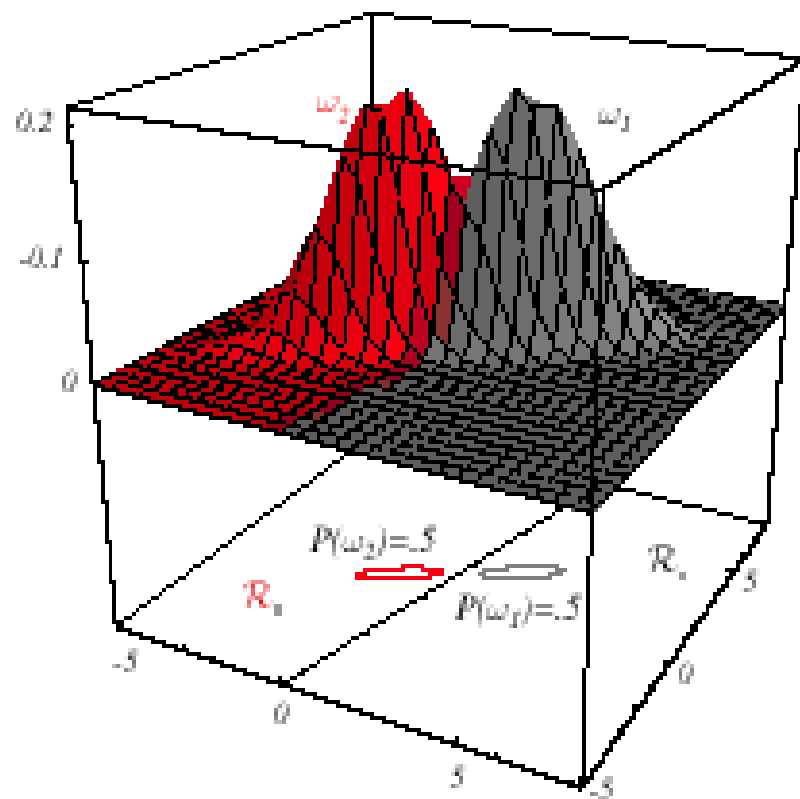


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

- **Case 2:** $\Sigma_i = \Sigma$ (covariance matrices of all classes are identical, but otherwise arbitrary!)
- Hyperplane separating \mathcal{R}_i and \mathcal{R}_j

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

- The hyperplane separating \mathcal{R}_i and \mathcal{R}_j is generally not orthogonal to the line between the means!
- To classify a feature vector x , measure the squared Mahalanobis distance from x to each of the c means; assign x to the category of the nearest mean



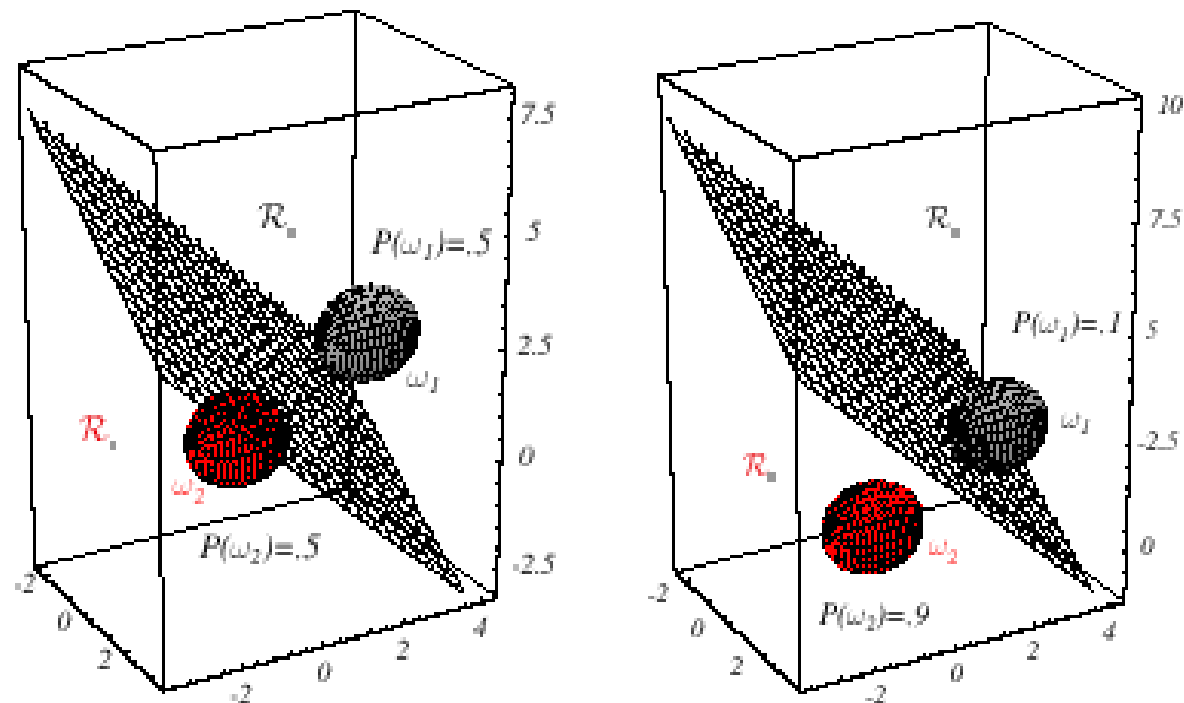


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda,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

- Case 3: $\Sigma_i = \text{arbitrary}$
 - 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 :

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

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

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

In the 2-category case, the decision surfaces are **hyperquadrics** that can assume any of the general forms: hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, hyperhyperboloids)

Discriminant Functions for 1D Gaussian

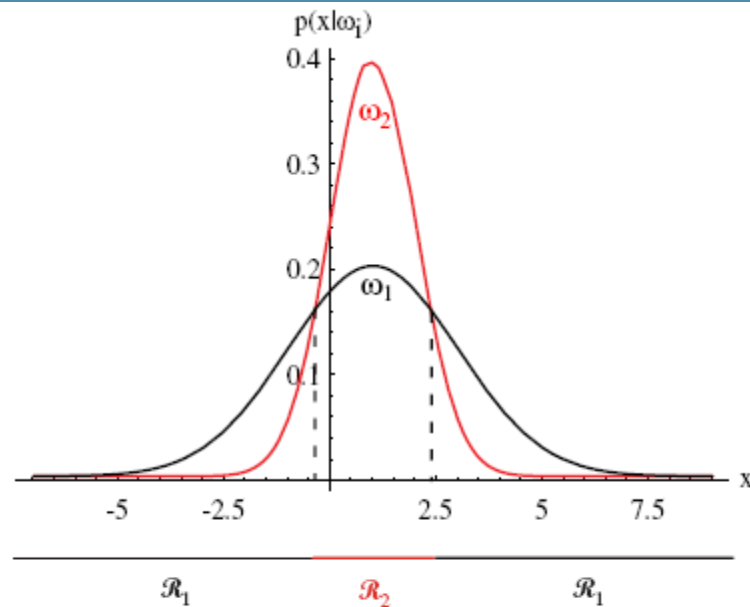
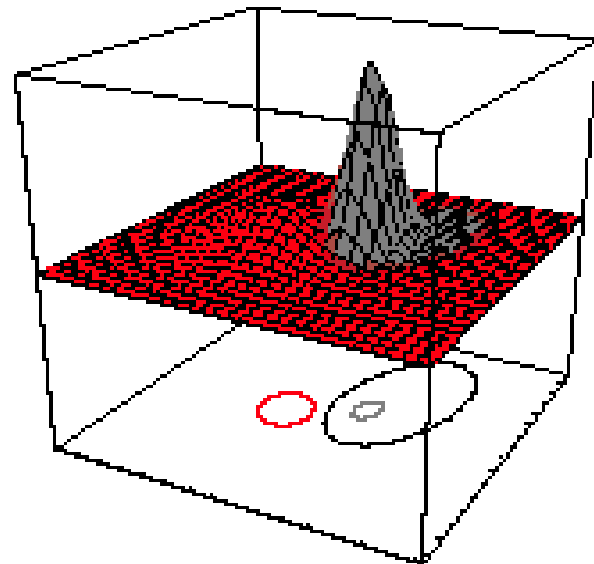
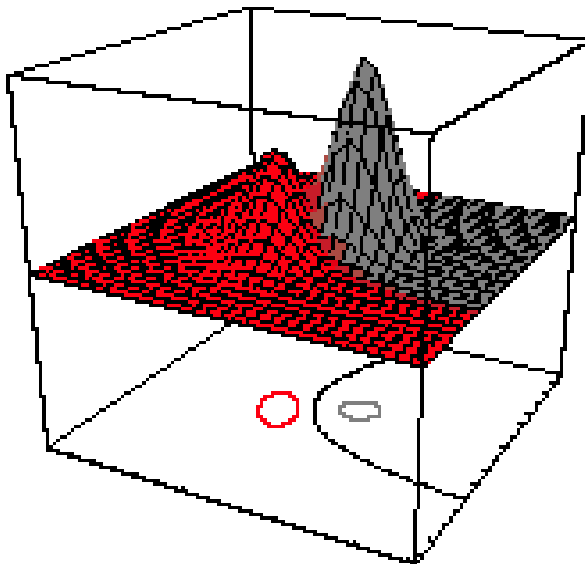
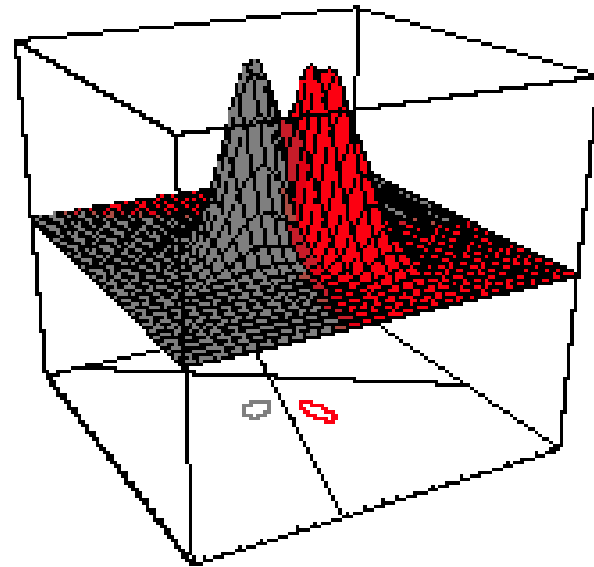
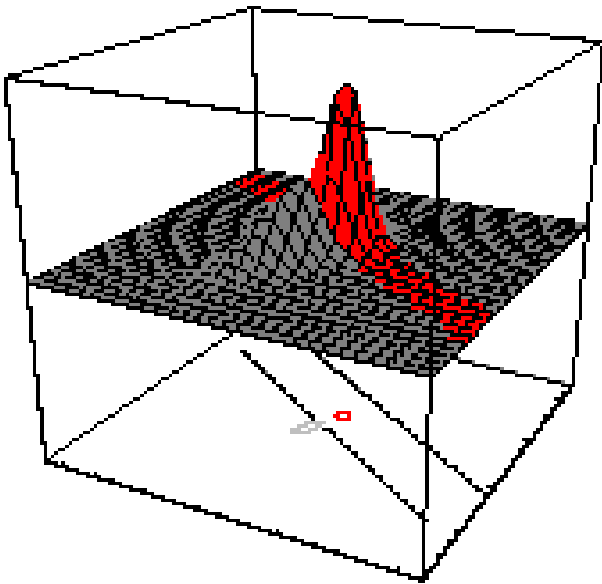


Figure 2.13: Non-simply connected decision regions can arise in one dimensions for Gaussians having unequal variance.

Discriminant Functions for the Normal Density



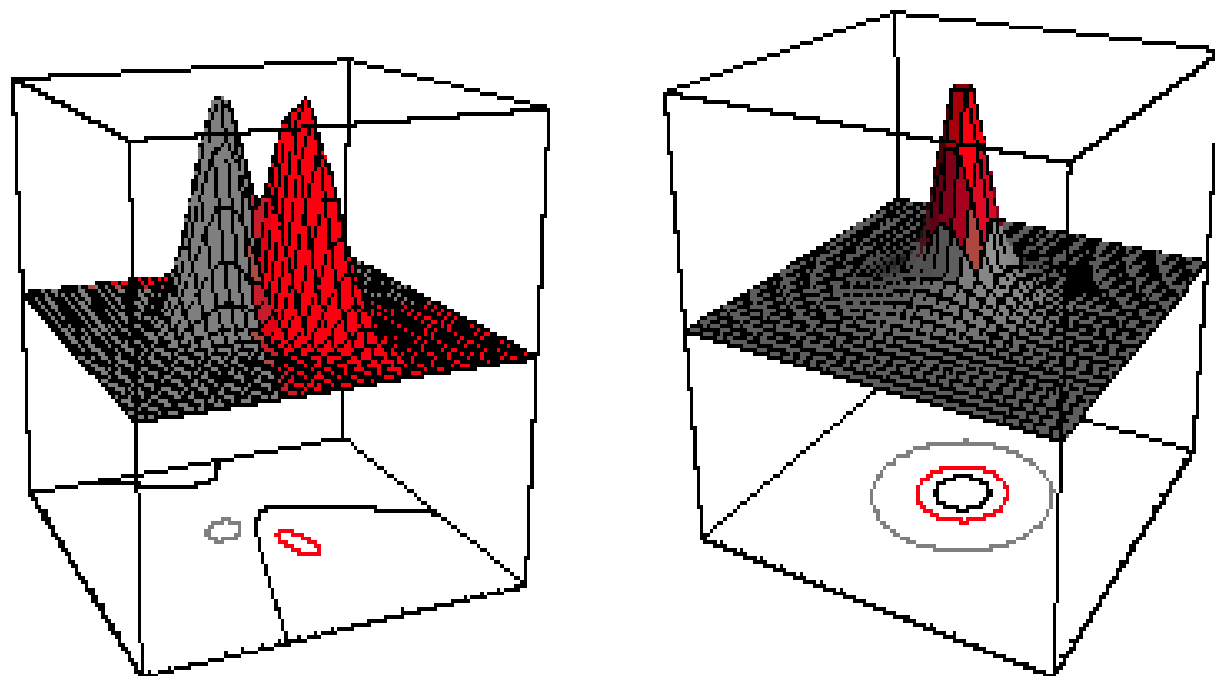


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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

Discriminant Functions for the Normal Density

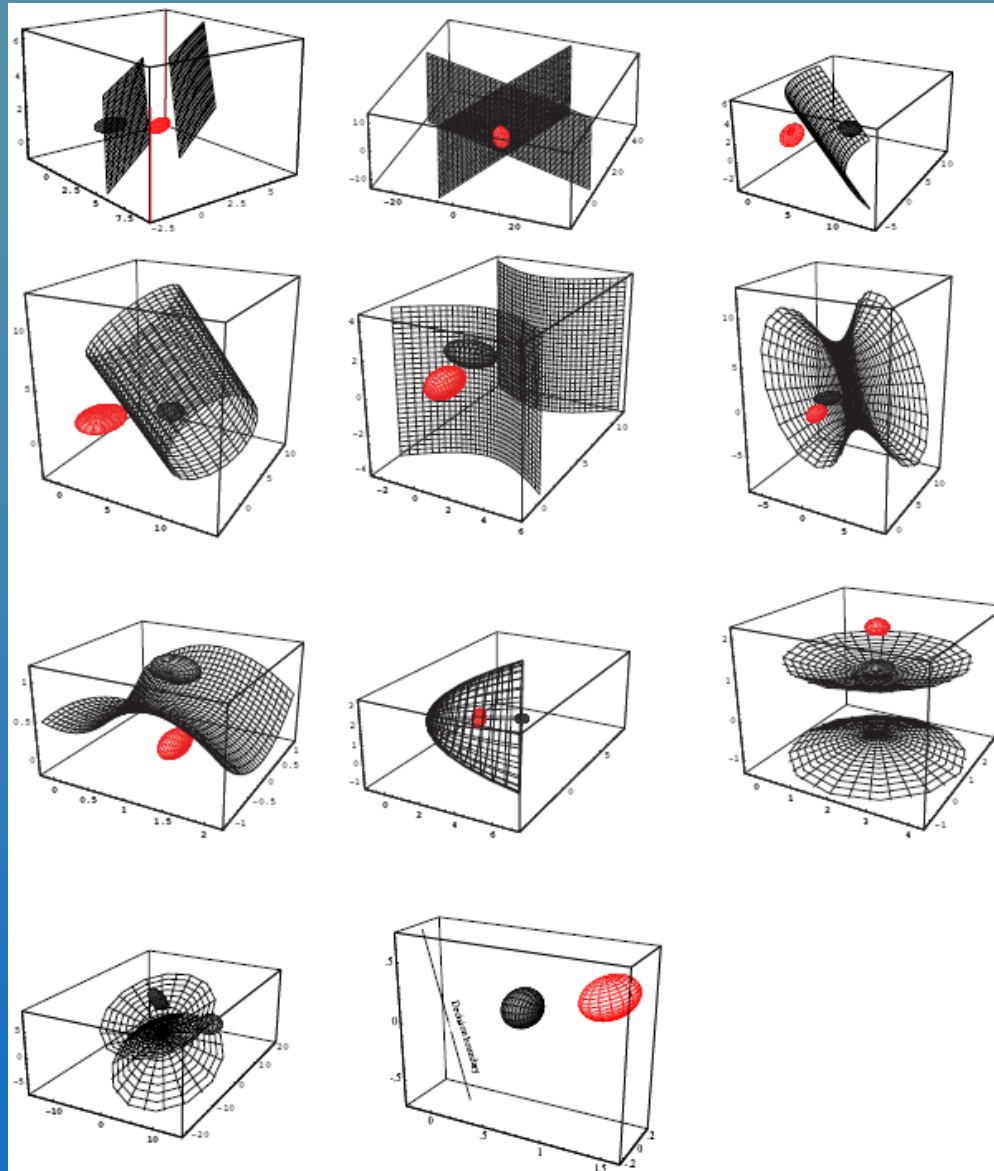


Figure 2.15: Arbitrary three-dimensional Gaussian distributions yield Bayes decision boundaries that are two-dimensional hyperquadrics. There are even degenerate cases in which the decision boundary is a line.

Discriminant Functions for the Normal Density

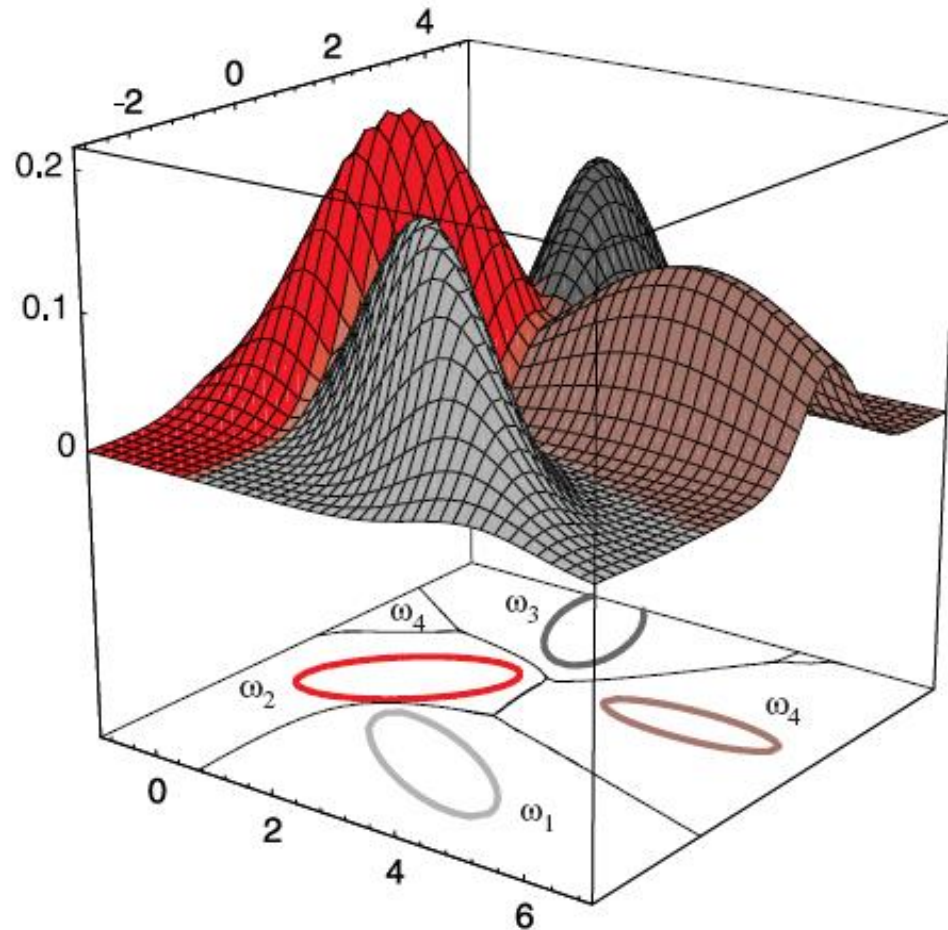
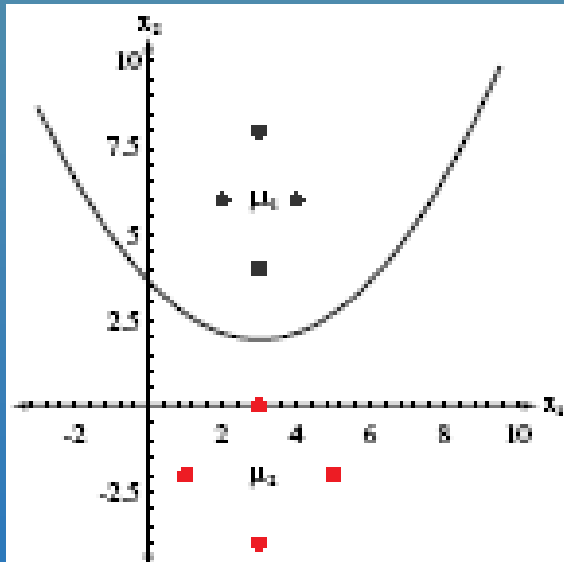


Figure 2.16: The decision regions for four normal distributions. Even with such a low number of categories, the shapes of the boundary regions can be rather complex.

Decision Regions for Two-Dimensional Gaussian Data



$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

$$\Sigma_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \quad \text{and} \quad \Sigma_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

$$P(\omega_1) = P(\omega_2) = 0.5$$

$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2$$

Error Probabilities and Integrals

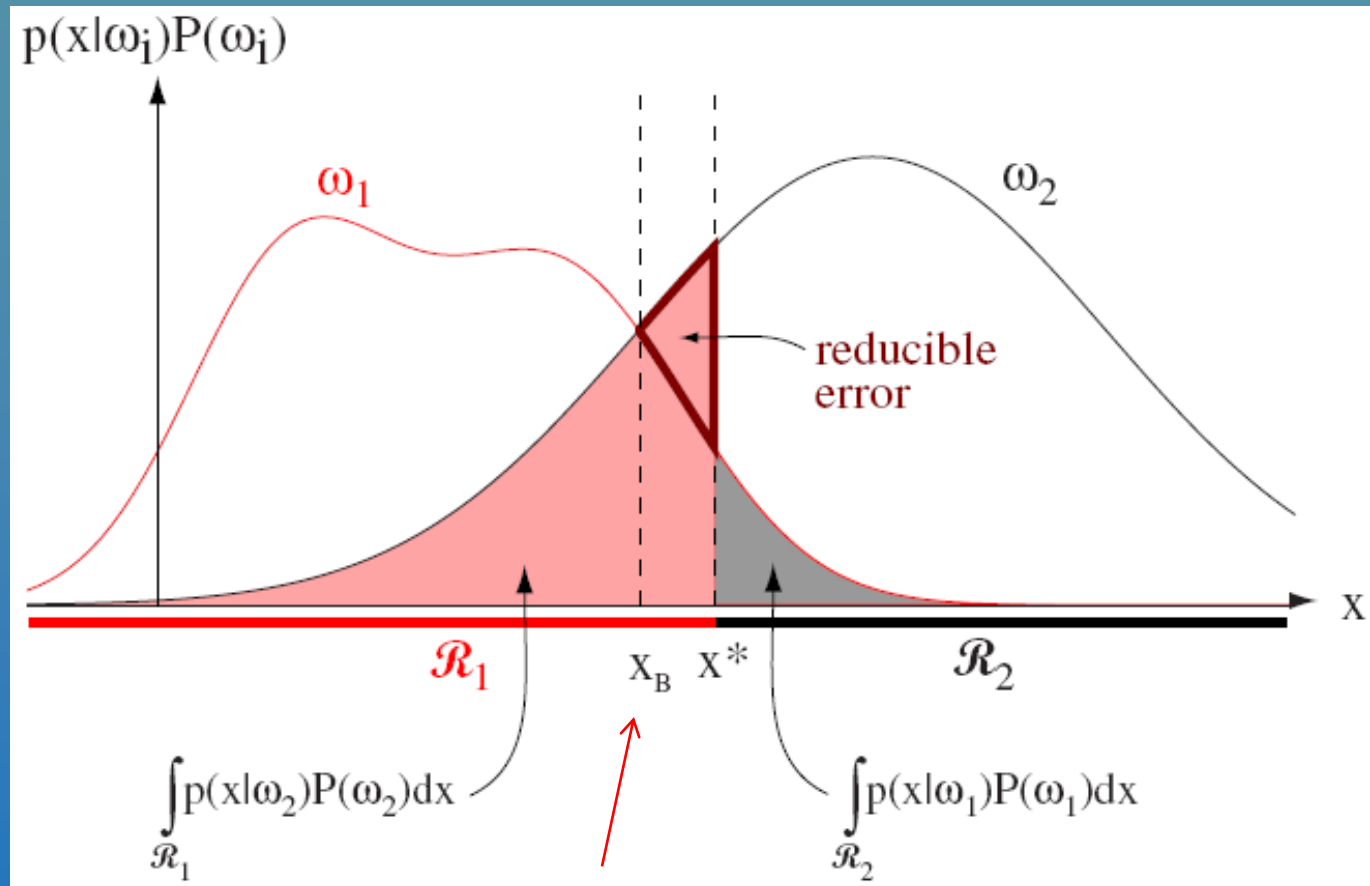
- 2-class problem
 - There are two types of errors

$$\begin{aligned}P(\text{error}) &= P(\mathbf{x} \in \mathcal{R}_2, \omega_1) + P(\mathbf{x} \in \mathcal{R}_1, \omega_2) \\&= P(\mathbf{x} \in \mathcal{R}_2 | \omega_1)P(\omega_1) + P(\mathbf{x} \in \mathcal{R}_1 | \omega_2)P(\omega_2) \\&= \int_{\mathcal{R}_2} p(\mathbf{x} | \omega_1)P(\omega_1) d\mathbf{x} + \int_{\mathcal{R}_1} p(\mathbf{x} | \omega_2)P(\omega_2) d\mathbf{x}.\end{aligned}$$

- Multi-class problem
 - Simpler to compute the prob. of being correct (more ways to be wrong than to be right)

$$\begin{aligned}P(\text{correct}) &= \sum_{i=1}^c P(\mathbf{x} \in \mathcal{R}_i, \omega_i) \\&= \sum_{i=1}^c P(\mathbf{x} \in \mathcal{R}_i | \omega_i)P(\omega_i) \\&= \sum_{i=1}^c \int_{\mathcal{R}_i} p(\mathbf{x} | \omega_i)P(\omega_i) d\mathbf{x}.\end{aligned}$$

Error Probabilities and Integrals



Bayes optimal decision boundary in 1-D case

Figure 2.17: Components of the probability of error for equal priors and (non-optimal) 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. 68. 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.

Error Rate of Linear Discriminant Function (LDF)

- Assume a 2-class problem

$$p(\underline{x}|\omega_1) \sim N(\underline{\mu}_1, \Sigma), \quad p(\underline{x}|\omega_2) \sim N(\underline{\mu}_2, \Sigma)$$

$$g_i(\underline{x}) = \log[P(\underline{x}|\omega_i)] = -\frac{1}{2}(\underline{x} - \underline{\mu}_i)^t \Sigma^{-1}(\underline{x} - \underline{\mu}_i) + \log[P(\omega_i)]$$

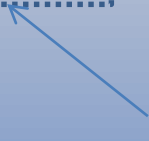
- Due to the symmetry of the problem (identical Σ), the two types of errors are identical
- Decide $\underline{x} \in \omega_1$ if $g_1(\underline{x}) > g_2(\underline{x})$ or

$$-\frac{1}{2}(\underline{x} - \underline{\mu}_1)^t \Sigma^{-1}(\underline{x} - \underline{\mu}_1) + \log[P(\omega_1)] > -\frac{1}{2}(\underline{x} - \underline{\mu}_2)^t \Sigma^{-1}(\underline{x} - \underline{\mu}_2) + \log[P(\omega_2)]$$

or $(\underline{\mu}_2 - \underline{\mu}_1)^t \Sigma^{-1} \underline{x} + \frac{1}{2}(\underline{\mu}_1^t \Sigma^{-1} \underline{\mu}_1 - \underline{\mu}_2^t \Sigma^{-1} \underline{\mu}_2) < \log[P(\omega_1) / P(\omega_2)]$

Error Rate of LDF

- Let $h(\underline{x}) = (\underline{\mu}_2 - \underline{\mu}_1)^t \Sigma^{-1} \underline{x} + \frac{1}{2} (\underline{\mu}_1^t \Sigma^{-1} \underline{\mu}_1 - \underline{\mu}_2^t \Sigma^{-1} \underline{\mu}_2)$
- Compute expected values & variances of $h(\underline{x})$ when $\underline{x} \in \omega_1$ & $\underline{x} \in \omega_2$

$$\begin{aligned}\eta_1 &= E[h(\underline{x}) | \underline{x} \in \omega_1] = (\underline{\mu}_2 - \underline{\mu}_1)^t \Sigma^{-1} \boxed{E[\underline{x} | \omega_1]} + \frac{1}{2} (\underline{\mu}_1^t \Sigma^{-1} \underline{\mu}_1 - \underline{\mu}_2^t \Sigma^{-1} \underline{\mu}_2) \\ &= -\frac{1}{2} (\underline{\mu}_2 - \underline{\mu}_1)^t \Sigma^{-1} (\underline{\mu}_2 - \underline{\mu}_1) \\ &= -\eta\end{aligned}$$


where $(\underline{\mu}_2 - \underline{\mu}_1)^t \Sigma^{-1} (\underline{\mu}_2 - \underline{\mu}_1)$
= squared Mahalanobis distance between
 $\underline{\mu}_1$ & $\underline{\mu}_2$

Error Rate of LDF

- Similarly
$$\eta_2 = +\frac{1}{2}(\underline{\mu}_2 - \underline{\mu}_1)^t \Sigma^{-1}(\underline{\mu}_2 - \underline{\mu}_1)$$
$$= +\eta$$

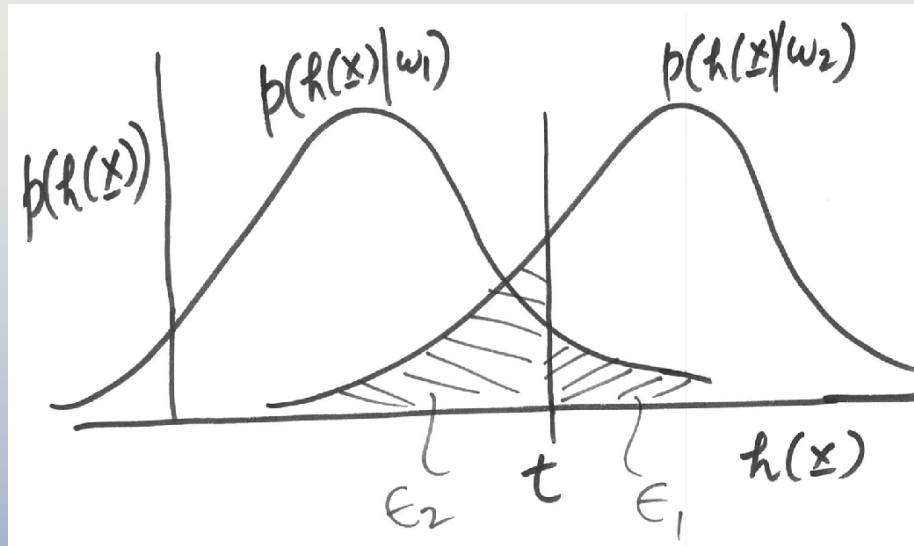
$$\begin{aligned}\sigma_1^2 &= E\left[\left(h(\underline{x}) - \eta_1\right)^2 \mid \omega_1\right] = E\left[(\underline{\mu}_2 - \underline{\mu}_1)^t \Sigma^{-1}(\underline{x} - \underline{\mu}_1) \mid \underline{x} \in \omega_1\right] \\ &= (\underline{\mu}_2 - \underline{\mu}_1)^t \Sigma^{-1}(\underline{\mu}_2 - \underline{\mu}_1) \\ &= 2\eta\end{aligned}$$

$$\sigma_2^2 = 2\eta$$

$$p\left(h(\underline{x}) \mid \underline{x} \in \omega_1\right) \sim N(-\eta, 2\eta)$$

$$p\left(h(\underline{x}) \mid \underline{x} \in \omega_2\right) \sim N(+\eta, 2\eta)$$

Error Rate of LDF



$$\varepsilon_1 = P(g_1(\underline{x}) < g_2(\underline{x}) | \underline{x} \in \omega_1) = \int_t^{\infty} P(h(\underline{x}) | \omega_1) dh \quad h(\underline{x}) \sim \frac{1}{\sqrt{2\Pi \cdot 2\phi}} e^{-\frac{1}{2}(\cdot)}$$

$$= \int_{\frac{n+t}{\sqrt{2\eta}}}^{\infty} \frac{1}{\sqrt{2\Pi}} e^{-\frac{1}{2}\xi^2} d\xi$$

$$= \frac{1}{2} - \frac{1}{2} \operatorname{erf}\left(\frac{\eta+t}{\sqrt{4\eta}}\right)$$

Error Rate of LDF

$$t = \log \left[\frac{P(\omega_1)}{P(\omega_2)} \right]$$

$$\text{erf}(r) = \frac{2}{\sqrt{\Pi}} \int_0^r e^{-x^2} dx$$

$$\varepsilon_2 = \frac{1}{2} - \frac{1}{2} \text{erf} \left(\frac{\eta - t}{\sqrt{4\eta}} \right)$$

Total probability of error

$$P_e = P(\omega_1)\varepsilon_1 + P(\omega_2)\varepsilon_2$$

Error Rate of LDF

$$P(\omega_1) = P(\omega_2) = \frac{1}{2} \Rightarrow t = 0$$

$$\varepsilon_1 = \varepsilon_2 = \frac{1}{2} - \frac{1}{2} \operatorname{erf} \left(\frac{\eta}{\sqrt{4\eta}} \right) = \frac{1}{2} - \frac{1}{2} \operatorname{erf} \left(\frac{(\underline{\mu}_1 - \underline{\mu}_2)^t \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2)}{2\sqrt{2}} \right)$$

(i) No Class Separation

$$(\underline{\mu}_1 - \underline{\mu}_2)^t \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2) = 0$$

$$\Rightarrow \varepsilon_1 = \varepsilon_2 = \frac{1}{2}$$

(ii) Perfect Class Separation

$$(\underline{\mu}_1 - \underline{\mu}_2)^t \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2) \rightarrow \infty$$

$$\Rightarrow \varepsilon_1 = \varepsilon_2 \rightarrow 0 \quad (\operatorname{erf} \rightarrow 1)$$

Mahalanobis distance is a good measure of separation between classes

Error Bounds for Normal Densities

- The exact calculation of the error for the general Gaussian case (case 3) is extremely difficult
- However, in the 2-category case the general error can be approximated analytically to give us an upper bound on the error

Chernoff Bound

- To derive a bound for the error, we need the following inequality

$$\min[a, b] \leq a^\beta b^{1-\beta} \quad \text{for } a, b \geq 0 \text{ and } 0 \leq \beta \leq 1.$$

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}, \quad P(error|x) = \min [P(\omega_1|x), P(\omega_2|x)].$$

$$P(error) \leq P^\beta(\omega_1)P^{1-\beta}(\omega_2) \int p^\beta(\mathbf{x}|\omega_1)p^{1-\beta}(\mathbf{x}|\omega_2) d\mathbf{x} \quad \text{for } 0 \leq \beta \leq 1.$$

Assume conditional prob. are normal $\int p^\beta(\mathbf{x}|\omega_1)p^{1-\beta}(\mathbf{x}|\omega_2) d\mathbf{x} = e^{-k(\beta)}$

where

$$k(\beta) = \frac{\beta(1-\beta)}{2}(\mu_2 - \mu_1)^t [\beta \Sigma_1 + (1-\beta)\Sigma_2]^{-1}(\mu_2 - \mu_1) + \frac{1}{2} \ln \frac{|\beta \Sigma_1 + (1-\beta)\Sigma_2|}{|\Sigma_1|^\beta |\Sigma_2|^{1-\beta}}.$$

Chernoff Bound

Chernoff bound for $P(\text{error})$ is found by determining the value of β that minimizes $\exp(-k(\beta))$

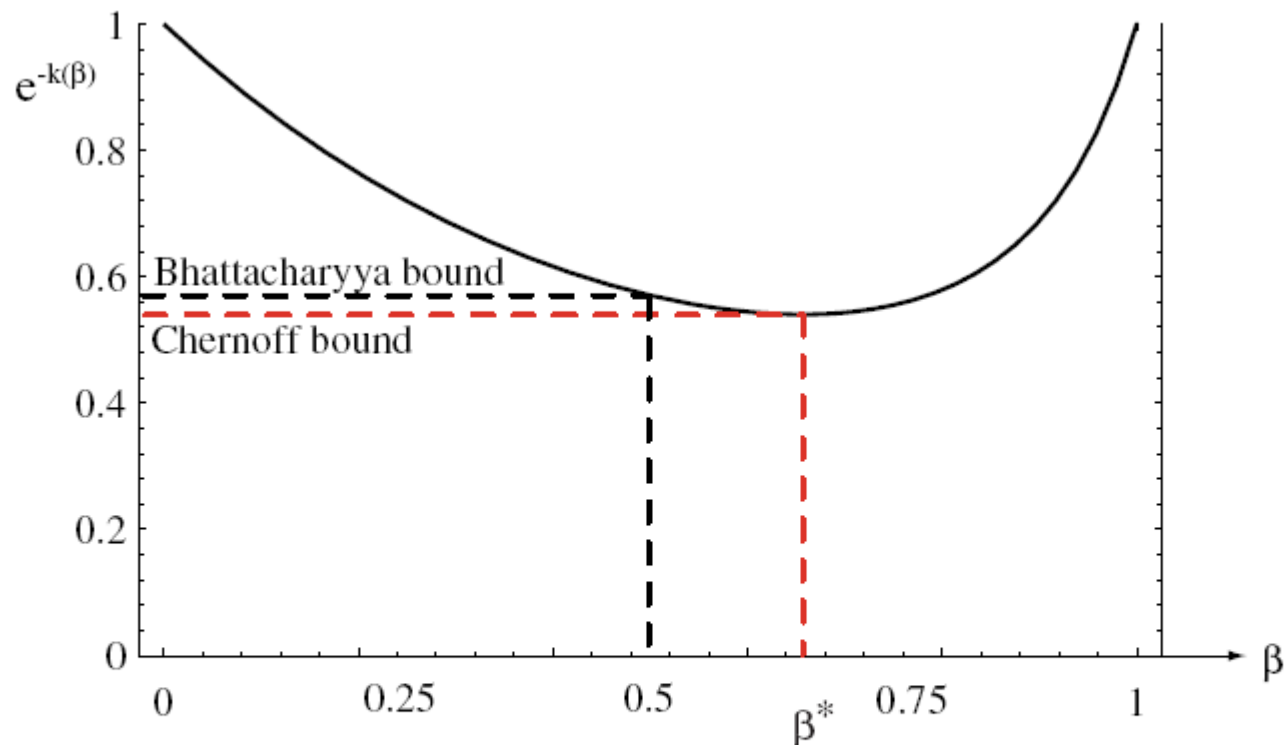


Figure 2.18: The Chernoff error bound is never looser than the Bhattacharyya bound. For this example, the Chernoff bound happens to be at $\beta^* = 0.66$, and is slightly tighter than the Bhattacharyya bound ($\beta = 0.5$).

Error Bounds for Normal Densities

- Bhattacharyya Bound
 - Assume $\beta = 1/2$
 - computationally simpler
 - slightly less tight bound
- Now, Eq. (73) has the form

$$\begin{aligned} P(\text{error}) &\leq \sqrt{P(\omega_1)P(\omega_2)} \int \sqrt{p(\mathbf{x}|\omega_1)p(\mathbf{x}|\omega_2)} d\mathbf{x} \\ &= \sqrt{P(\omega_1)P(\omega_2)} e^{-k(1/2)}, \end{aligned}$$

$$\begin{aligned} k(1/2) &= 1/8(\mu_2 - \mu_1)^t \left[\frac{\Sigma_1 + \Sigma_2}{2} \right]^{-1} (\mu_2 - \mu_1) + \\ &\quad \frac{1}{2} \ln \frac{\left| \frac{\Sigma_1 + \Sigma_2}{2} \right|}{\sqrt{|\Sigma_1| |\Sigma_2|}}. \end{aligned}$$

When the two covariance matrices are equal, $k(1/2)$ is the same as the Mahalanobis distance between the two means

Error Bounds for Gaussian Distributions

Chernoff Bound

$$P(\text{error}) \leq P^\beta(\omega_1)P^{1-\beta}(\omega_1) \int p^\beta(x|\omega_1)p^{1-\beta}(x|\omega_2)dx \quad 0 \leq \beta \leq 1$$

$$\int p^\beta(x|\omega_1)p^{1-\beta}(x|\omega_2)dx = e^{-k(\beta)}$$

$$k(\beta) = \frac{\beta(1-\beta)}{2} (\mu_2 - \mu_1)^t [\beta\Sigma_1 + (1-\beta)\Sigma_2]^{-1} (\mu_2 - \mu_1) + \frac{1}{2} \ln \frac{\beta\Sigma_1 + (1-\beta)\Sigma_2}{|\Sigma_1|^\beta |\Sigma_2|^{1-\beta}}$$

Best Chernoff error bound is 0.008190

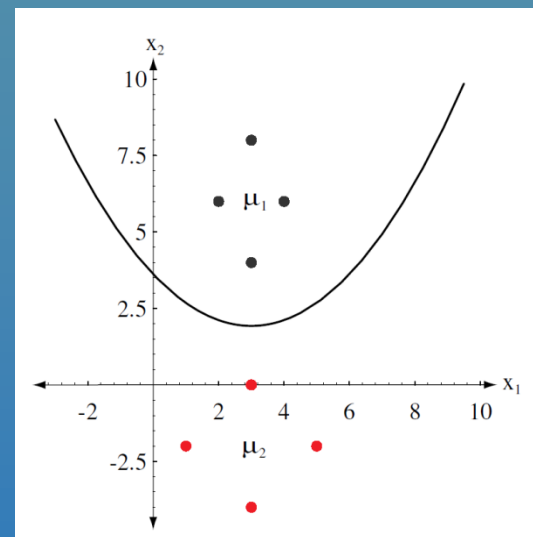
Bhattacharya Bound ($\beta=1/2$)

$$P(\text{error}) \leq \sqrt{P(\omega_1)P(\omega_2)} \int \sqrt{P(x|\omega_1)P(x|\omega_2)}dx = \sqrt{P(\omega_1)P(\omega_2)} e^{-k(1/2)}$$

$$k(1/2) = 1/8 (\mu_2 - \mu_1)^t \left[\frac{\Sigma_1 + \Sigma_2}{2} \right]^{-1} (\mu_2 - \mu_1) + \frac{1}{2} \ln \frac{\Sigma_1 + \Sigma_2}{2\sqrt{|\Sigma_1||\Sigma_2|}}$$

Bhattacharya error bound is 0.008191

True error using numerical integration = 0.0021



2-category, 2D data

Signal Detection Theory

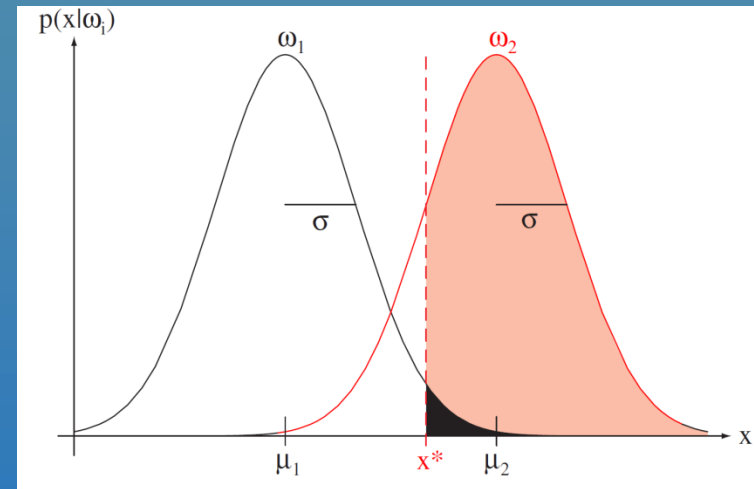
We are interested in detecting a single weak pulse, e.g. radar reflection; the internal signal (x) in detector has mean m_1 (m_2) when pulse is absent (present)

$$p(x | \omega_1) \sim N(\mu_1, \sigma^2)$$

$$p(x | \omega_2) \sim N(\mu_2, \sigma^2)$$

The detector uses a threshold x^* to determine the presence of pulse

Discriminability: ease of determining whether the pulse is present or not



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

For given threshold, define *hit*, *false alarm*, *miss* and *correct rejection*

$$P(x > x^* | x \in \omega_2): \textit{hit}$$

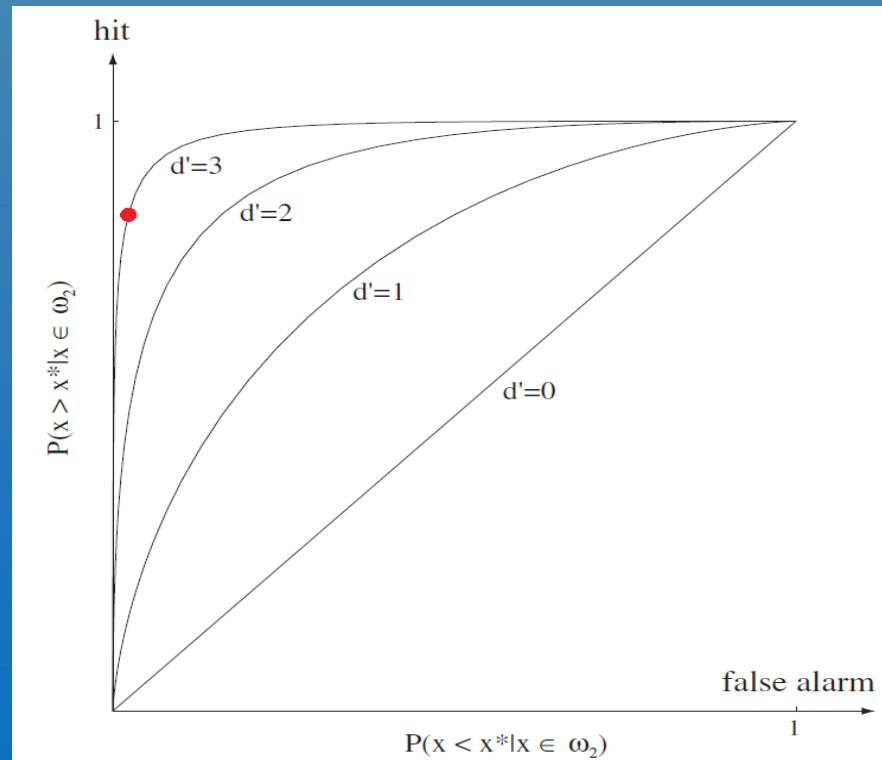
$$P(x > x^* | x \in \omega_1): \textit{false alarm}$$

$$P(x < x^* | x \in \omega_2): \textit{miss}$$

$$P(x < x^* | x \in \omega_1): \textit{correct rejection}$$

Receiver Operating Characteristic (ROC)

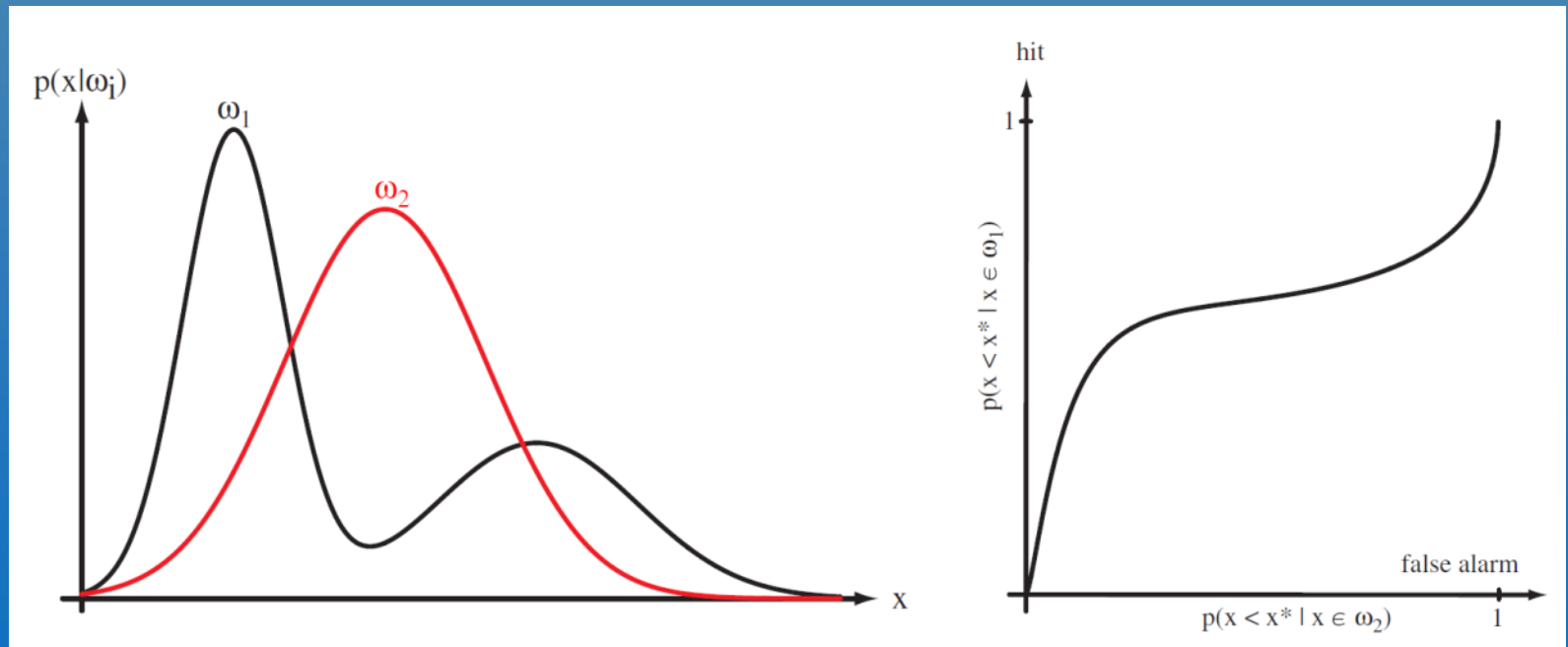
- Experimentally compute *hit* and *false alarm* rates for fixed x^*
- Changing x^* will change the *hit* and *false alarm* rates
- A plot of hit and false alarm rates is called the ROC curve



Performance shown at different operating points

Operating Characteristic

- In practice, distributions may not be Gaussian and will be multidimensional; ROC curve can still be plotted
- Vary a single control parameter for the decision rule and plot the resulting *hit* and *false alarm* rates



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 for 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 :

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

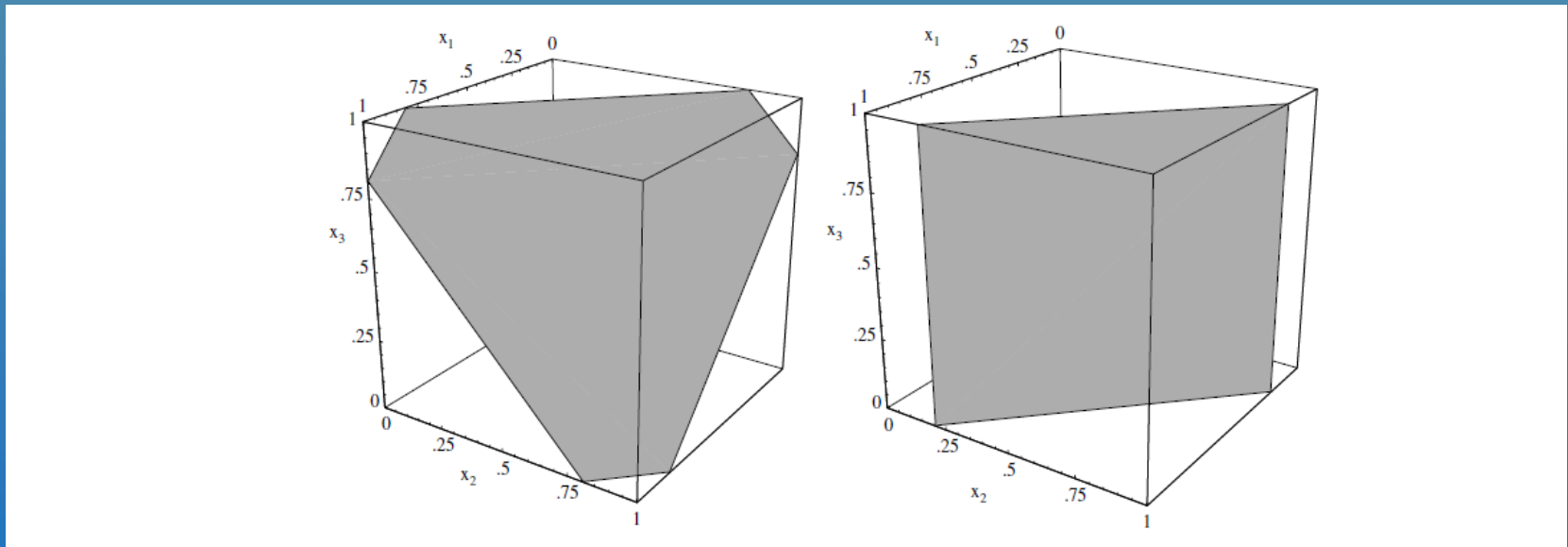
and :

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

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

Bayesian Decision for 3-dim Binary Data

- A 2-class problem; 3 independent binary features; class priors are equal; $p_i = 0.8$ and $q_i = 0.5$, $i = 1, 2, 3$
- $w_i = 1.3863$; $w_0 = 1.2$
- Decision surface $g(x) = 0$ is shown below



Left figure: $p_i = 0.8$ and $q_i = 0.5$. Right figure: $p_3 = q_3$ (feature 3 does not provide any discriminatory information) so the decision surface is parallel to x_3 axis

Neyman-Pearson Rule

I. Neyman-Pearson Rule

Let $\Omega = \{\omega_0, \omega_1\}$, $A = \{a_0, a_1\}$. This rule is related to the following hypothesis testing problem.

H_0 (null hypothesis) : \mathbf{x} comes from a population governed by $p(\mathbf{x} \mid \omega_0)$; ω_0 is the state of nature.

H_1 (alternative hypothesis) : the state of nature is ω_1 .

Neyman-Pearson Rule

The N-P decision rule is also called a 'test' which divides the *feature space* (X) into two regions : critical region $C_\delta = \{\mathbf{x} \mid \delta(\mathbf{x}) = a_1\}$ and its complement C_δ^* . A point in C_δ leads to the acceptance of the alternative hypothesis.

There are two types of errors defined below.

1. False alarm : decide H_1 when H_0 is true; also called type I error, denoted by α .
2. False dismissal : decide H_0 when H_1 is true; also called type II error, denoted by β .

This terminology comes from the field of communication theory where we want to detect a message in the presence of noise.

H_0 : a received signal is noise alone.

H_1 : a received signal is message plus noise.

Neyman-Pearson Rule

The false alarm probability $= \int_{C_\delta} p(\mathbf{x} | \omega_0) d\mathbf{x}$

The false dismissal probability $= 1 - \int_{C_\delta} p(\mathbf{x} | \omega_1) d\mathbf{x}$

These error probabilities can be related to the risk function when a 0-1 loss function is used. That is,

$$L(\omega_k, \alpha_i) = \begin{cases} 0 & \text{if } i = k \\ 1 & \text{if } i \neq k \end{cases}$$

$$\text{Since } R_\delta(\omega_0) = \int L(\omega_0, \delta(\mathbf{x})) p(\mathbf{x} | \omega_0) d\mathbf{x}$$

$$\text{Hence } R_\delta(\omega_0) = \int_{C_\delta} p(\mathbf{x} | \omega_0) d\mathbf{x} \text{ (false alarm probability)}$$

$$R_\delta(\omega_1) = \int L(\omega_1, \delta(\mathbf{x})) p(\mathbf{x} | \omega_1) d\mathbf{x} = \int_{C_\delta^*} p(\mathbf{x} | \omega_1) d\mathbf{x}$$

$$R_\delta(\omega_1) = 1 - \int_{C_\delta} p(\mathbf{x} | \omega_1) d\mathbf{x} \text{ (false dismissal probability)}$$

Assuming that the probability of error of type $I(\alpha)$ is given, then the Neyman-Pearson classifier will minimize β for a fixed α , where

α is called the level or size of the test.

$(1-\beta)$ is called the power of the test.

Neyman-Pearson Rule

The Neyman-Pearson classifier that minimizes β for a given value of α is a likelihood ratio test with the threshold k_α ,

$$Pr \left\{ \frac{p(\mathbf{x} | \omega_1)}{p(\mathbf{x} | \omega_0)} \geq k_\alpha \mid \mathbf{x} \in \omega_0 \right\} = \alpha$$

The critical region for the N-P rule is

$$C_\alpha = \left[\mathbf{x} \mid \frac{p(\mathbf{x} | \omega_1)}{p(\mathbf{x} | \omega_0)} \geq k_\alpha \right].$$

If $k_\alpha = 1$, then the N-P rule is a maximum likelihood rule. The expression of the rule can be simplified by choosing a monotone increasing function g such that $\delta_{NP}(\mathbf{x}) = a_1$ if $g\left[\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_0)}\right] \geq g(k_\alpha)$.

J. Neyman-Pearson Lemma

Consider the class of all decision rules having size α . No rule in this class has power larger than the N-P rule at level α . In other words, for a 0-1 loss function, the N-P rule at level α minimizes $R_\delta(\omega_1)$ among all rules for which $R_\delta(\omega_0) \leq \alpha$.

Neyman-Pearson Rule

Example 2.4

Let the two class-conditional densities be univariate Gaussian. We will now determine the critical region and the power of the N-P rule for a given α .

$$p(x | \omega_0) \sim N(\mu_0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_0)^2}{2\sigma^2}}$$

$$p(x | \omega_1) \sim N(\mu_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}$$

$$\text{Thus, } \log \left[\frac{p(x|\omega_1)}{p(x|\omega_0)} \right] = \frac{1}{\sigma^2} \left[(\mu_1 - \mu_0)x - \frac{1}{2}(\mu_1^2 - \mu_0^2) \right],$$

$$\begin{aligned} C_\alpha &= \left[x \mid \log \left[\frac{p(x|\omega_1)}{p(x|\omega_0)} \right] \geq \log k_\alpha \right] \\ &= \left[x \mid x \geq x_0, \text{ where } x_0 = \frac{\sigma^2 \log k_\alpha}{\mu_1 - \mu_0} + \frac{1}{2}(\mu_1 + \mu_0) \right], \end{aligned}$$

$$\begin{aligned} \alpha &= \int_{C_\alpha} p(x|\omega_0) dx = \int_{x_0}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_0)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{\pi}} \int_{\frac{x_0-\mu_0}{\sqrt{2}\sigma}}^{\infty} e^{-t^2} dt \\ &= \frac{-1}{\sqrt{\pi}} \int_0^{\frac{x_0-\mu_0}{\sqrt{2}\sigma}} e^{-t^2} dt + \frac{1}{\sqrt{\pi}} \int_0^{\infty} e^{-t^2} dt. \end{aligned}$$

Neyman-Pearson Rule

Therefore, $\alpha = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{x_0 - \mu_0}{\sqrt{2}\sigma}\right)$.

The threshold value x_0 can be written as

$$x_0 = \mu_0 + \sqrt{2}\sigma\text{erf}^{-1}(1 - 2\alpha).$$

The power of the test becomes

$$1 - \beta = \int_{x_0}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} dx = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{x_0 - \mu_1}{\sqrt{2}\sigma}\right).$$

The decision boundary is shown in Figure 2.2.

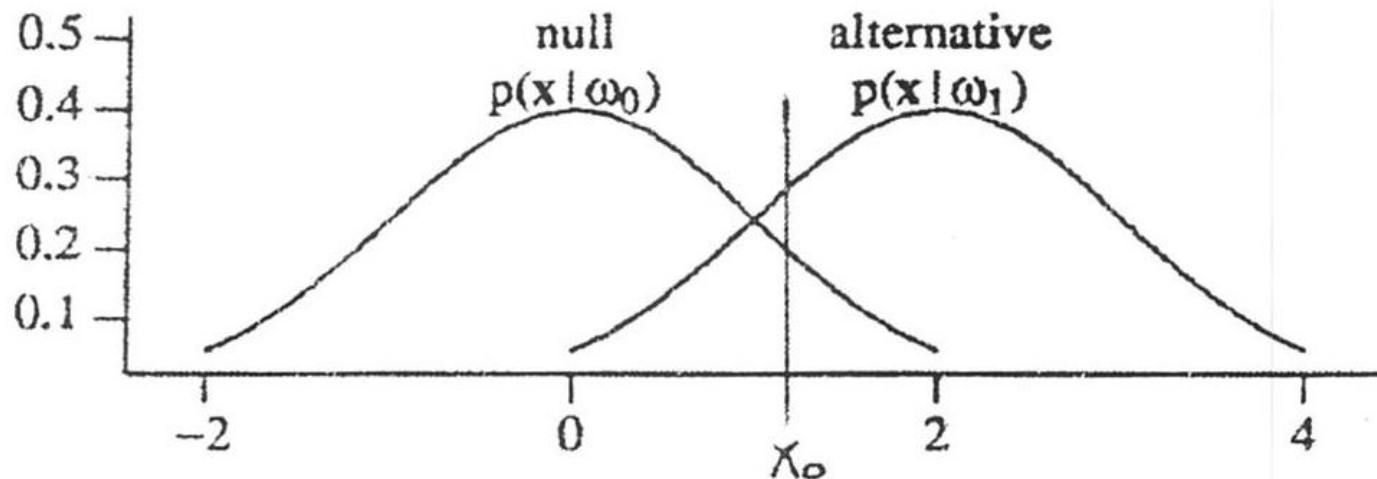


Figure 2.2: N-P rule

Missing Feature Values

- $n \times d$ pattern matrix; $n \times n$ (dis)similarity matrix
- Suppose it is not possible to measure a certain feature for a given pattern
- Possible solutions:
 - Reject the pattern
 - Approximate the missing value
 - Replace missing value by mean for that feature
 - Marginalize over distribution of the missing feature

$$\mathbf{x} = [x_g, x_b]$$

$$\begin{aligned} P(\omega_i | x_g) &= \frac{p(\omega_i, x_g)}{p(x_g)} = \frac{\int p(\omega_i, x_g, x_b) dx_b}{p(x_g)} = \frac{\int p(\omega_i, x_g, x_b) dx_b}{p(x_g)} \\ &= \frac{\int p(\omega_i | x_g, x_b) p(x_g, x_b) dx_b}{p(x_g)} = \frac{\int g_i(x) p(x) dx_b}{\int p(x) dx_b} \end{aligned}$$

Handling Missing Feature value

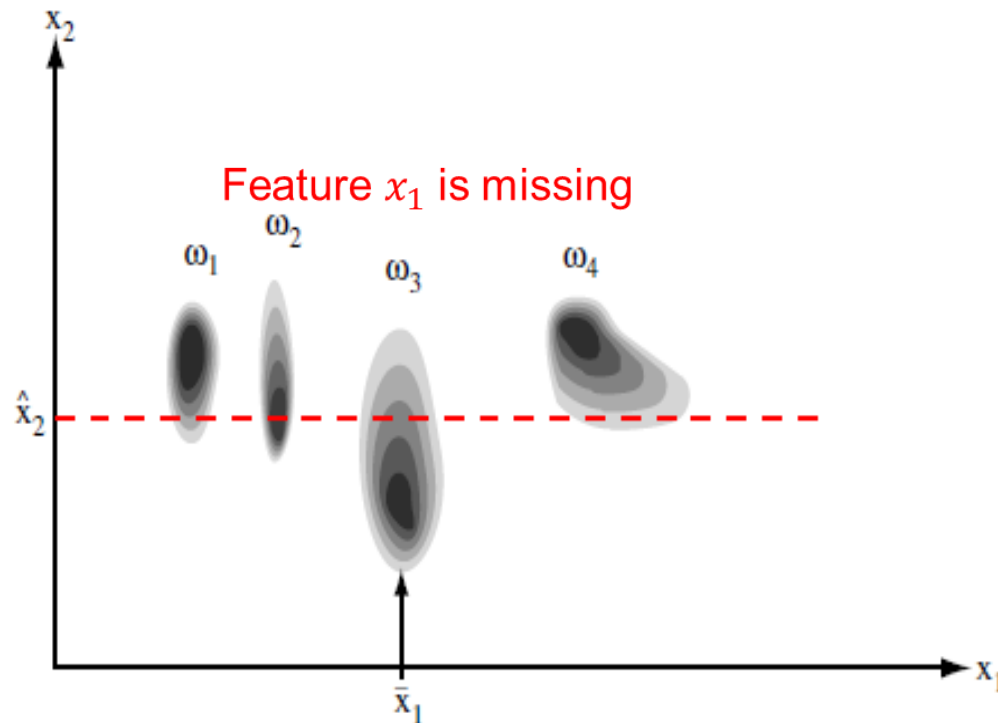


Figure 2.22: Four categories have equal priors and the class-conditional distributions shown. If a test point is presented in which one feature is missing (here, x_1) and the other is measured to have value \hat{x}_2 (red dashed line), we want our classifier to classify the pattern as category ω_2 , because $p(\hat{x}_2|\omega_2)$ is the largest of the four likelihoods.

Other Topics

- Compound Bayes Decision Theory & Context
 - Consecutive states of nature may be dependent; state of next fish may depend on state of the previous fish
 - Exploit such statistical dependence to gain improved performance (use of context)
 - Compound decision vs. sequential compound decision
 - Markov dependence
- Sequential Decision Making
 - Feature measurement process is sequential
 - Feature measurement cost
 - Minimize a combination of feature measurement cost and the classification accuracy

Context in Text Recognition

A T
7 B 9 C A T
C E
D
Q V E S T