Approaches for Solving decision problems

- Decision Rule or Classification Strategy attempts to integrate all available problem information, such as measurements and a priori probabilities
- Decision rules may be formulated by i)Inference problem
  Solution: Converting an a priori class probability $P(C_i)$ and class-conditional densities $p(\mathbf{x}|\mathbf{C_k})$ into a measurement-conditioned (a posteriori) probability $P(C_i|x)$ or model the joint distribution of $p(\mathbf{x}, \mathbf{C_k})$ directly and obtain the prosterior probabilities
  ii) Use discriminative models(that compute the posterior class probabilities) and use decision theory
  iii) Use discriminant function $f(x)$ which maps each input $\mathbf{x}$ onto a class label
- By formulating a measure of expected classification error or risk, and choosing a decision rule that minimizes this measure

Comparision

- Inference approach: Demanding computationally as it involves finding $P(\mathbf{x}, \mathbf{C_k})$
  Large training set required
  Advantage is it allows marginal density to be used from total probability$p(\mathbf{x}) = \sum_{\mathbf{k}} \mathbf{p}(\mathbf{x}|\mathbf{C_k})\mathbf{p}(\mathbf{C_k})$ which may be used for novelty detection or outlier detection
- Discriminant Models: Preferred when only classification decision is to be made
  No need of large data and resources
- Discriminant function:Simpler approach
  Use training data to find a discriminant function $f9\mathbf{x})$ that maps each $\mathbf{x}$ directly onto a class label

- Example 1: No measurement, c=2
- A two -class problem where $P(C_1) = 0.7$ and $P(C_2) = 0.3$ with no measurement
- No features as there are no mesurements

- Probability of error :

$$P(\text{error}) = P(\text{choose } C_2|C_1)P(C_1) + P(\text{choose } C_1|C_2)P(C_2)$$

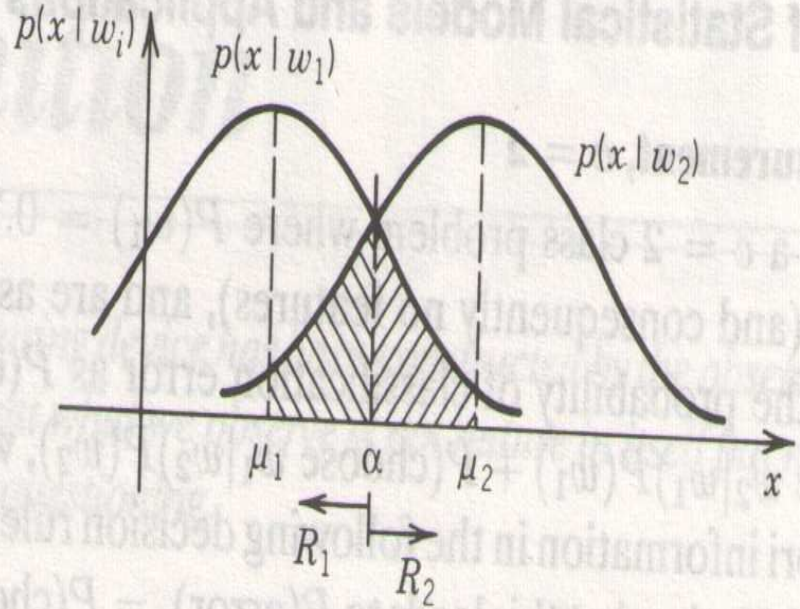
- Rule: Always Choose C_1, since $P(C_1) > P(C_2)$
- Error: $P(\textit{choose } C_1|C_2)P(C_2) = 1 * 0.3 = 0.3$
- Example 2: Single measurement, c=2, Gaussian Densities, equal a priori probabilities
- Consider the case where d=1,c=2, $P(C_1) = P(C_2)$ and

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

- As the variance and a priori probabilities for both the classes are same, class-means provide specific-information

$$P(C_i|x) = p(x|C_i)[P(C_i)/p(x)]$$

- This is shown by the crosshatched regions in Figure 1(a). The exercises explo further. ■



(a)

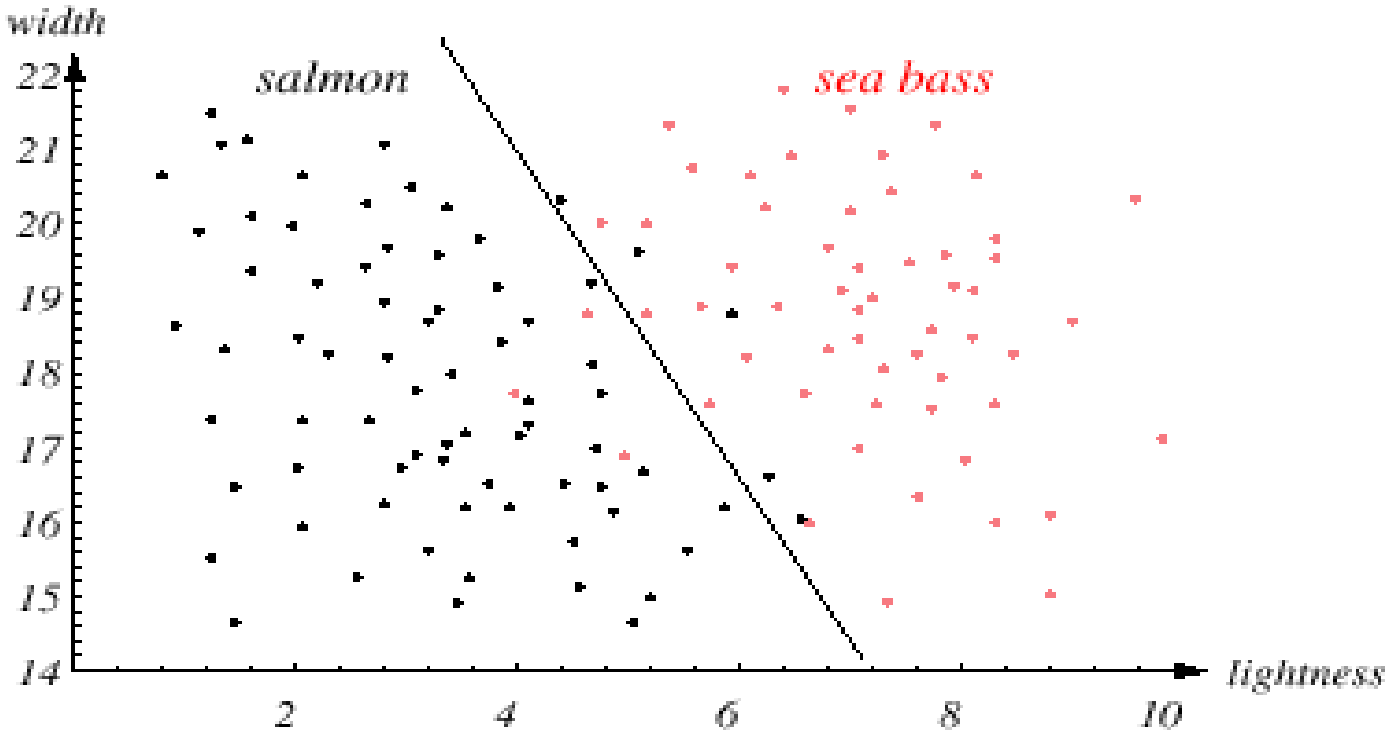
Figure 1(a): Densities for

Error assessment

- $$P(\text{error}) = P(\underline{x} \text{ is assigned to the wrong class})$$

- For a c=2 class problem,

$$\begin{aligned} P(\text{error}) &= P(\text{choose } C_1 \text{ and } x \text{ actually from } C_2) \\ &\quad + P(\text{choose } C_2 \text{ and } x \text{ actually from } C_1) \\ &= P(\text{error}|C_1)P(C_1) + P(\text{error}|C_2)P(C_2) \\ &= P(\underline{x} \in R_2|C_1)P(C_1) + P(\underline{x} \in R_1|C_2)P(C_2) \\ &= P(x < \alpha|C_2)P(C_2) + P(x < \alpha|C_1)P(C_1) \\ &= P(C_2)\int_{-\infty}^{\alpha} p(\zeta|C_2)d\zeta + P(C_1)\int_{\alpha}^{\infty} p(\zeta|C_1)d\zeta \end{aligned}$$

- The above classifier can also be looked at as a minimum-distance classifier as the decision rule for $P(C_1) = P(C_2)$ can be looked at as: Assign x to R_i, where x is closest to μ_i
- Example 3: Extended d=2, c=2 Classifier design
- 

Bayesian Decision theory

- When we have a overlapping feature space, perfect classification is not possible
- classification error indicates the likelihood that an incorrect classification or decision occurs
- Density functions characterize random vectors
- Using Bayes theory, a priori estimate of the probability of a certain class is converted into the a posterior , or measurement conditioned, probablity of a state of nature

Bayesian Decision Theory

- Statistical approach to the problem of pattern classification
- The decision problem is posed in probabilistic terms assuming that all the relevant probability values are known
- Let C define the state of nature with $C = C_1$ for class 1 and $C = C_2$ for class 2
- Let $P(C_1)$ and $P(C_2)$denote the prior probabilities
- Let $p(x|C)$ denote the class-conditional pdf

- The joint pdf of finding a pattern that is in category C_j and has feature value x is

$$p(C_j, x) = P(C_j|x)p(x) = p(x|C_j)P(C_j)$$

- Bayes' formula

$$P(C_j|x) = \frac{p(x|C_j)P(C_j)}{p(x)}$$

- Where

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

- In words,

$$posterior = \frac{likelihood * prior}{evidence}$$

- By observing the value of x we can convert the prior probability $P(C_j)$ to the a posterior probability$P(C_j|x)$- the probability of the state of the nature being C_j given that the feature value x has been measured
- $p(x)$ is the evidence can be viewed as a scale factor
- Both a prior information and measurement related information are combined in decision procedure

Normal Density

- Analytical tractability
- Appropriate model for cantinuous valued, randomly corrupted versions of signal
- Continuous univariate normal or Gaussinan density

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-1/2(\frac{(x-\mu)}{\sigma})^2)$$

- The expected value of x and its variance is

$$\mu = \varepsilon[x] = \int_{-\infty}^{\infty} xp(x)dx$$

$$\sigma^2 = \varepsilon[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx$$

- Specified completely by mean and variance

$$p(x) \approx N(\mu, \sigma^2)$$

- Entropy of a distribution is given by

$$H(p(x)) = -\int p(x) \ln(p(x)) dx$$

- Nats or bits (log2)
- Entropy measures the fundamental uncertainity in the values of the points selected randomly from a distribution
- Normal distribution has maximum entropy of all the distributions
- Central limit theorem: The aggregate of the sum of a large number of small independent random disturbances will lead to a Gaussian distribution

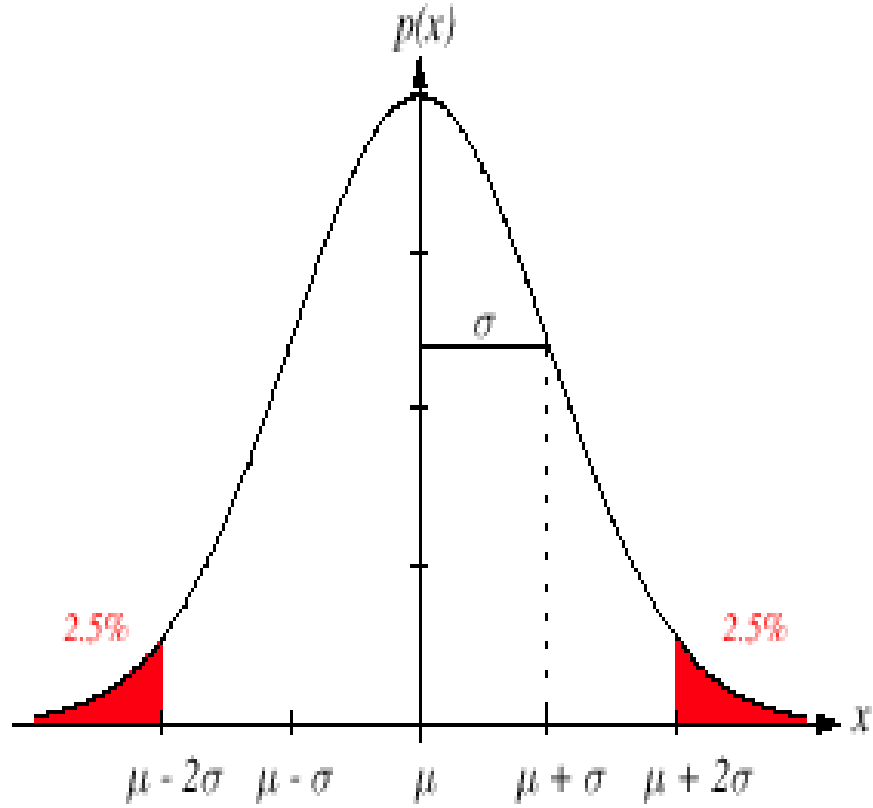


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John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

Multivariate density

- The multivariate normal density in d dimensions is given by

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

- Notation:

$$p(\mathbf{x}) \approx N(\mu, \mathbf{\Sigma})$$

$$\mu \equiv \varepsilon[\mathbf{x}] = \int \mathbf{x}\mathbf{p}(\mathbf{x})\mathbf{dx}$$

$$\mathbf{\Sigma}\varepsilon[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^{\mathbf{t}}] = \int (\mathbf{x}-\mu)(\mathbf{x}-\mu)^{\mathbf{t}}\mathbf{p}(\mathbf{x})\mathbf{dx}$$

- Where $\mu_i = \varepsilon[x_i]$ and $\sigma_{ij} = \varepsilon[(x_i - \mu_i)(x_j - \mu_j)]$

Multivariate density

- Covariance matrix is always symmetric and positive semidefinite
- Diagonal elements are variances and off-diagonal elements are covariances
- If all the off-diagonal elements are zero, multivariate pdf reduces to the product of the univariate densities

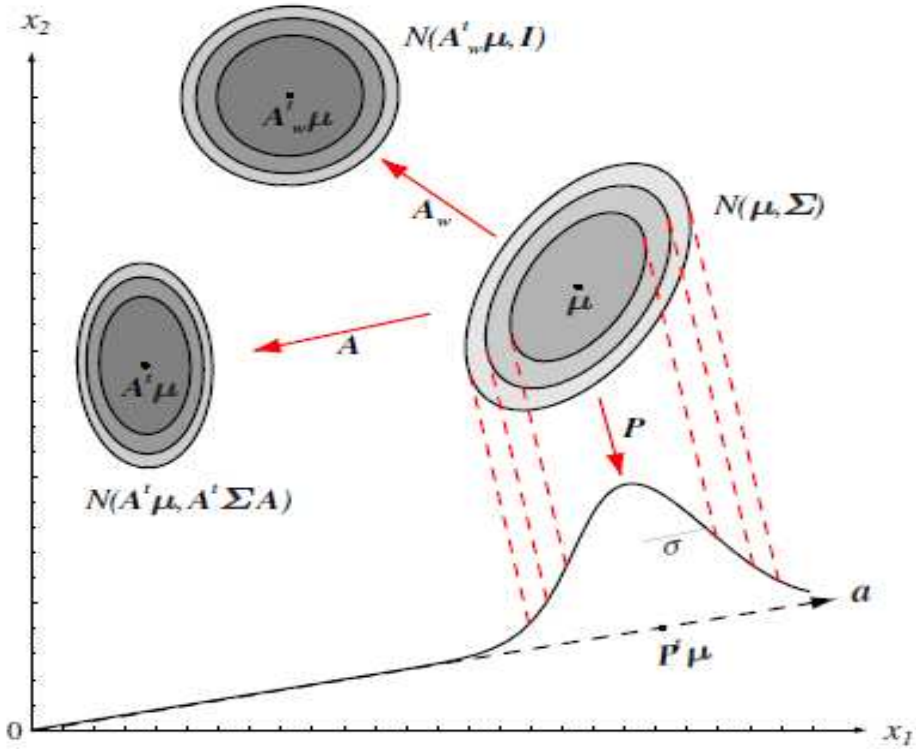


FIGURE 2.8. The action of a linear transformation on the feature space will convert an arbitrary normal distribution into another normal distribution. One transformation, $\mathbf{A}$, takes the source distribution into distribution $N(\mathbf{A}^t\boldsymbol{\mu}, \mathbf{A}^t\boldsymbol{\Sigma}\mathbf{A})$. 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.

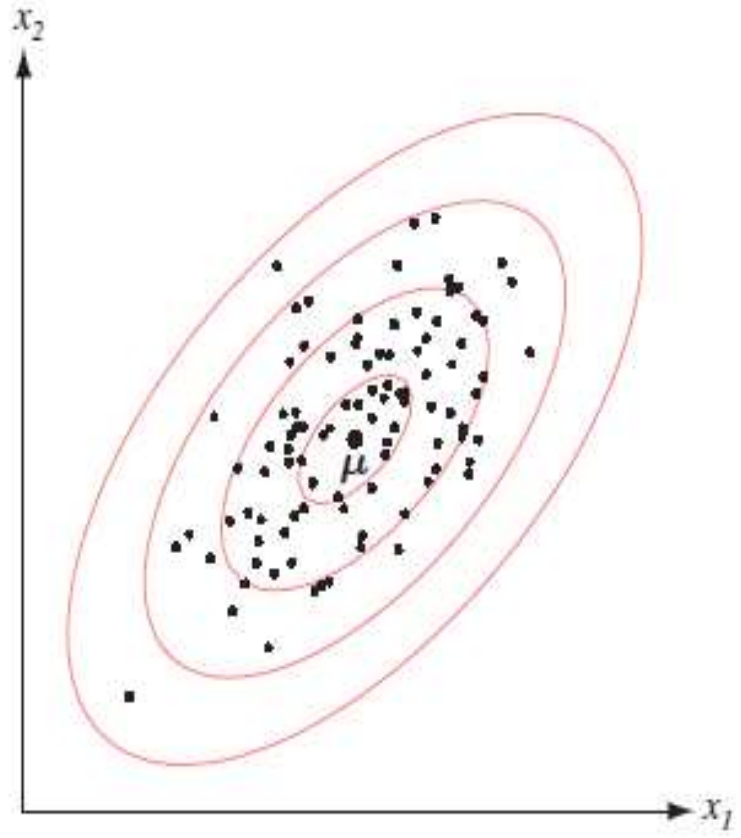


FIGURE 2.9. Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean μ. The ellipses show lines of equal probability density of the Gaussian. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Discriminant functions

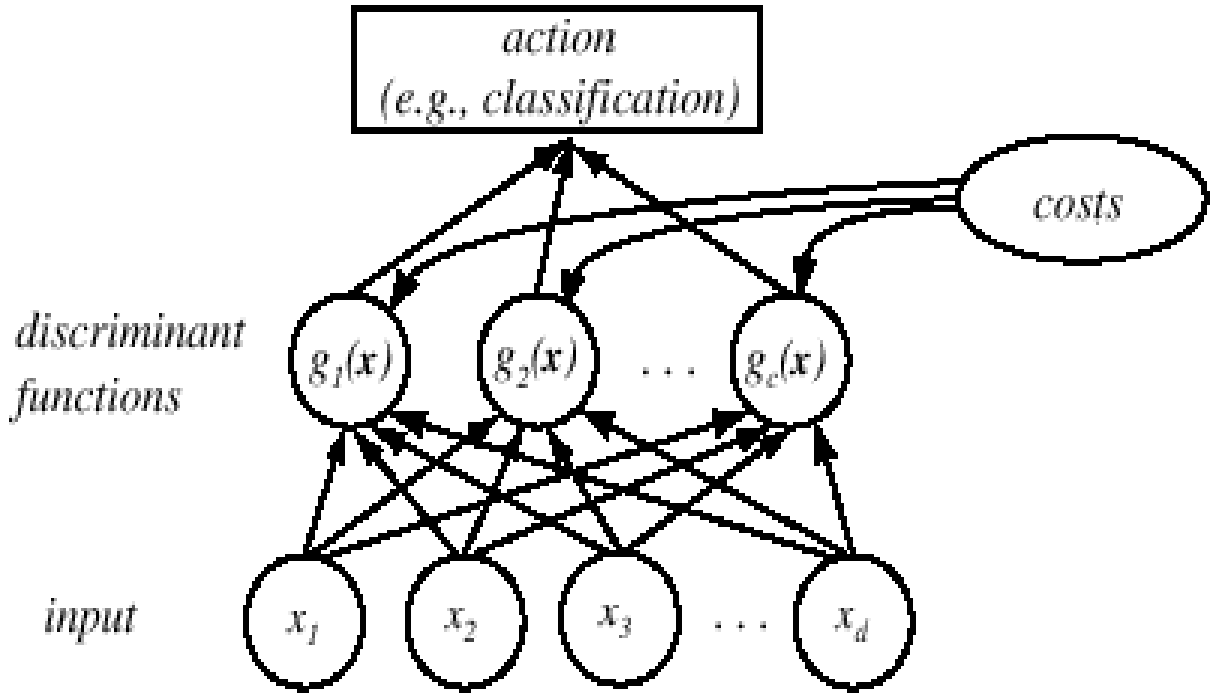


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction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

- Define a discriminant function for the ith class as

$$g_i(\underline{x}) = P(C_i|\underline{x})$$

- The discriminant function $g_i(x), i = 1, 2, ...c$ assign a feature vector to class C_i if

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

- The classifier may be viewed as a network or a machine that computes c discriminant functions and selects the category corresponding to the largest discriminant(minimum conditional risk)
- Any monotonically increasing function of $g_i(\underline{x})$ is a valid discriminant function

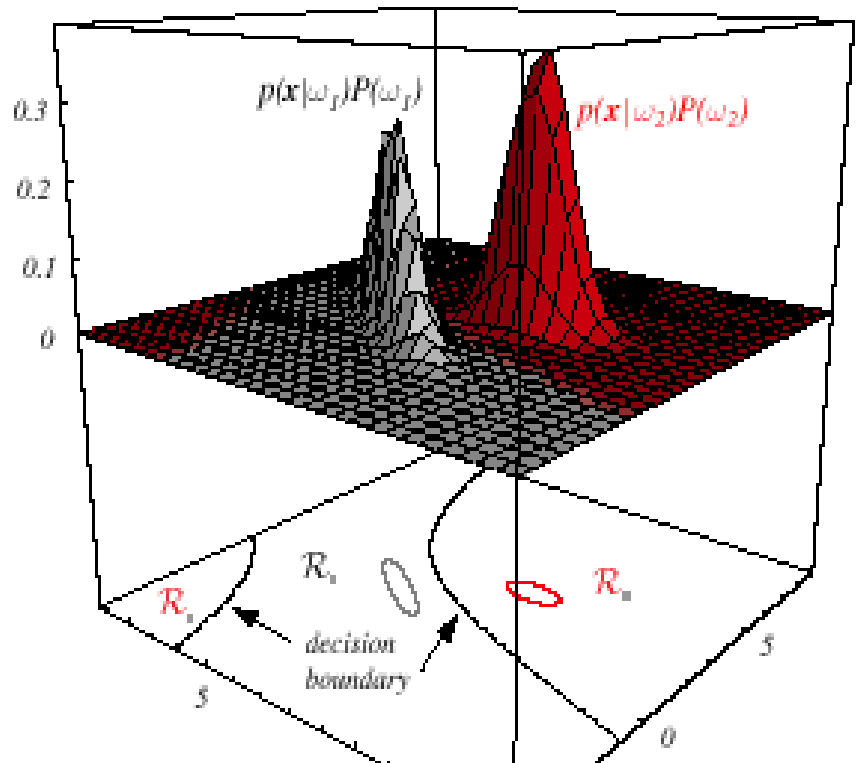
- For minimum error-rate calculations, all the following choices gives identical results

$$g_i(\mathbf{x}) = \mathbf{P(C_i|x)} = \frac{\mathbf{p(x|C_i)}}{\sum_{\mathbf{j=1}}^{\mathbf{c}} \mathbf{p(x|C_j)P(C_j)}}$$

$$g_i(\mathbf{x}) = \mathbf{p(x|C_i)P(C_i)}$$

$$g_i(\mathbf{x}) = \mathbf{lnp(x|C_i|) + lnP(C_i)}$$

- The Effect of any decision rule is to partition the feature



Two-class classification

- $g_1(\vec{x})$ and $g_2(\vec{x})$ are the discriminant functions
- Decision rule: Assign $\vec{x}$ to y_1 if

$$g_1(\vec{x}) > g_2(\vec{x})$$
$$\Rightarrow g_1(\vec{x}) - g_2(\vec{x}) > 0$$
$$p(C_1|\vec{x}) > p(C_2|\vec{x})$$

- Let us use Single discriminant function: $g(\mathbf{x}) = \mathbf{g_1}(\mathbf{x}) - \mathbf{g_2}(\mathbf{x})$
- Decide C_1 if $g(\mathbf{x}) > \mathbf{0}$ otherwise decide C_2

Case 1:Isotropic covariance matrix

- Same variance in all directions and statistically independent features
- Geometrically this corresponds to a situation in which samples fall in equal sized hyper-spherical clusters
- Covariance matrix has only the spread information

$$\Sigma_i = \sigma^2 I, i = 1, 2$$

-

$$Class1 : g_1(\vec{x}) = \ln p(\vec{x}|C_1)P(C_1)$$
$$\text{Class 2:} g_2(\vec{x}) = \ln p(\vec{x}|C_2)P(C_2)$$
$$g(\vec{x}) = g_1(\vec{x}) - g_2(\vec{x}) = 0$$

- Assuming Normal densities:

$$p(\vec{x}|C_1) \approx N(\vec{\mu_1}, \Sigma_1^2)$$
$$p(\vec{x}|C_2) \approx N(\vec{\mu_2}, \Sigma_2^2)$$

- The class-conditional density functions

$$p(\vec{x}|C_1) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp \frac{-(\vec{x} - \mu_1)^T(\vec{x} - \mu 1)}{2\sigma^2}$$

- Since

$$K\vec{a}^t\vec{a} = K(\parallel \vec{a} \parallel)^2$$

$$p(\vec{x}|C_1) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp \frac{-(||\vec{x} - \mu_1||^2)}{2\sigma^2}$$

$$g_1(\vec{x}) = -\frac{1}{2}\frac{(\parallel \vec{x} - \vec{\mu_1} \parallel^2)}{\sigma^2} - \ln(2\pi)^{d/2} - \ln|\Sigma|^{\frac{1}{2}} + \ln P(C_1)$$

$$g_2(\vec{x}) = -\frac{1}{2}\frac{(\parallel \vec{x} - \vec{\mu_2} \parallel^2)}{\sigma^2} - \ln(2\pi)^{d/2} - \ln|\Sigma|^{\frac{1}{2}} + \ln P(C_2)$$

- Ignoring constant terms as they donot have any roles decision making

$$g_1(\vec{x}) = -\frac{(\parallel \vec{x} - \vec{\mu_1} \parallel)^2}{2\sigma^2} + \ln P(C_1)$$

$$g_2(\vec{x}) = -\frac{(\parallel \vec{x} - \vec{\mu_2} \parallel)^2}{2\sigma^2} + \ln P(C_2)$$

- As $\| \vec{a} - \vec{b} \|^2 = \| a \|^2 + \| b \|^2 - 2\vec{a}^t \vec{b}$

$$g_1(\vec{x}) = -\frac{1}{2\sigma^2}[\| \vec{x} \|^2 + \| \mu_1 \|^2 - 2\vec{x}^t \vec{\mu_1}] + \ln P(C_1)$$

$$g_2(\vec{x}) = -\frac{1}{2\sigma^2}[\| \vec{x} \|^2 + \| \mu_2 \|^2 - 2\vec{x}^t \vec{\mu_2}] + \ln P(C_2)$$

- Neglecting terms which are not class-specific

$$\| x \|^2 = \vec{x}^t \vec{x}$$

$$\vec{\omega}^t \vec{x} = \vec{x}^t \vec{\omega} = \text{scalar}$$

-

$$g_1(\vec{x}) = -\frac{1}{2\sigma^2}[\vec{\mu_1}^t\vec{\mu_1} - 2\vec{\mu_1}^t\vec{x}] + \ln P(C_1)$$

$$g_2(\vec{x}) = -\frac{1}{2\sigma^2}[\vec{\mu_2}^t\vec{\mu_2} - 2\vec{\mu_2}^t\vec{x}] + \ln P(C_2)$$

- Writing $g_i(\vec{x}) = \vec{w_1}^t\vec{x} + w_{i0}$
- Where

$$\vec{w} = \frac{1}{\sigma^2}\vec{\mu_i}$$

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

Distance function for decision surfaces

- Discriminant function

$$g(\vec{x}) = g_1(\vec{x}) - g_2(\vec{x}) = 0$$

$$\vec{w_1}^t\vec{x} + w_{10} - (\vec{w_2}^t\vec{x} + w_{20}) = 0$$

$$(\vec{w_1}^t - \vec{w_2}^t)\vec{x} + (w_{10} - w_{20}) = 0$$

$$\text{linear surface } \vec{w}^t\vec{x} + w_0 = 0$$

$$\vec{w} = (\vec{w_1} - \vec{w_2}) = \frac{1}{\sigma^2}(\vec{\mu_1} - \vec{\mu_2})$$

$$w_0 = w_{10} - w_{20} = -\frac{1}{2 * \sigma^2}(\vec{\mu_1}^t\vec{\mu_1} - \vec{\mu_2}^t\vec{\mu_2}) + \ln\frac{P(C_1)}{P(C_2)}$$

$$= -\frac{1}{2 * \sigma^2} \parallel \vec{\mu_1} \parallel^2 - \parallel \vec{\mu_2} \parallel^2 + \ln\frac{P(C_1)}{P(C_2)}$$

$$\parallel \vec{a} \parallel^2 - \parallel \vec{b} \parallel^2 = (\vec{a} + \vec{b})^t(\vec{a} - \vec{b}) = (\vec{a} - \vec{b})^t(\vec{a} + \vec{b})$$

- Writing

$$g(\vec{x}) = \frac{1}{\sigma^2}(\vec{\mu_1} - \vec{\mu_2})^t\vec{x} - \frac{1}{2 * \sigma^2}(\vec{\mu_1} - \vec{\mu_2})^t(\vec{\mu_1} + \vec{\mu_2}) + \ln\frac{P(C_1)}{P(C_2)}$$

$$= \frac{1}{\sigma^2}(\vec{\mu_1} - \vec{\mu_2})^t(\vec{x} - \frac{1}{2}(\vec{\mu_1} + \vec{\mu_2}) + \frac{\sigma^2(\vec{\mu_1} - \vec{\mu_2})}{\| \vec{\mu_1} - \vec{\mu_2} \|^2}\ln\frac{P(C_1)}{P(C_2)})$$

- This can be expressed as $\vec{w}^t(\vec{x} - \vec{x_0}) = 0$

$$g(\vec{x}) = \frac{1}{\sigma^2}(\vec{\mu_1} - \vec{\mu_2})^t(\vec{x} - (\frac{1}{2}(\mu_1 + \mu_2) - \frac{\sigma^2(\vec{\mu_1} - \vec{\mu_2})}{\| \mu_1 - \mu_2 \|^2}\ln\frac{P(C_1)}{P(C_2)}))$$

- This is a hyperplane that passes through $\vec{x_0}$
- When $P(C_1) = P(C_2)$, then $\vec{x_0}$ is exactly midway between μ_1 and μ_2 perpendicular to the line between means

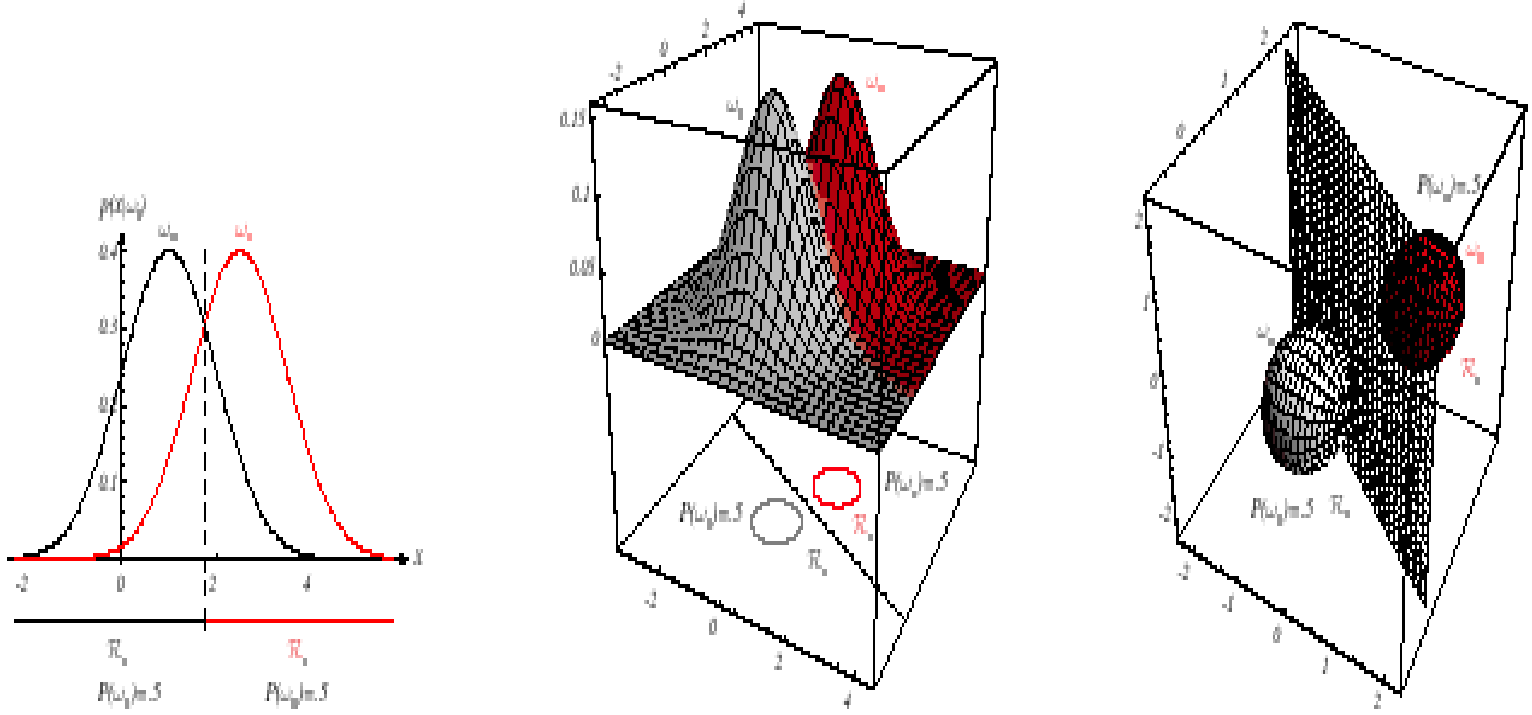
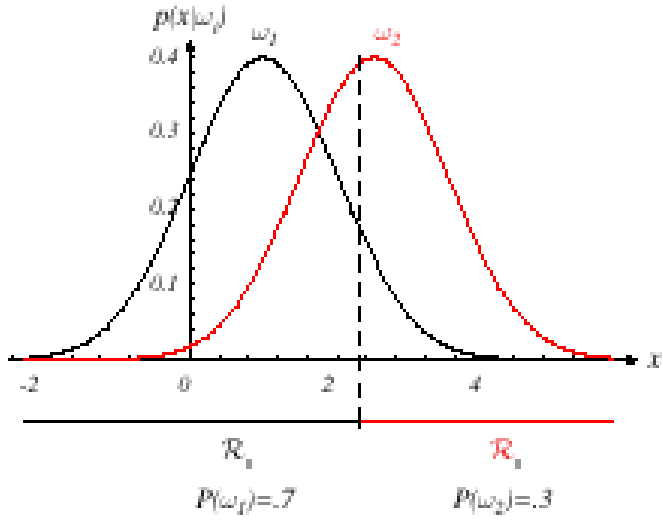
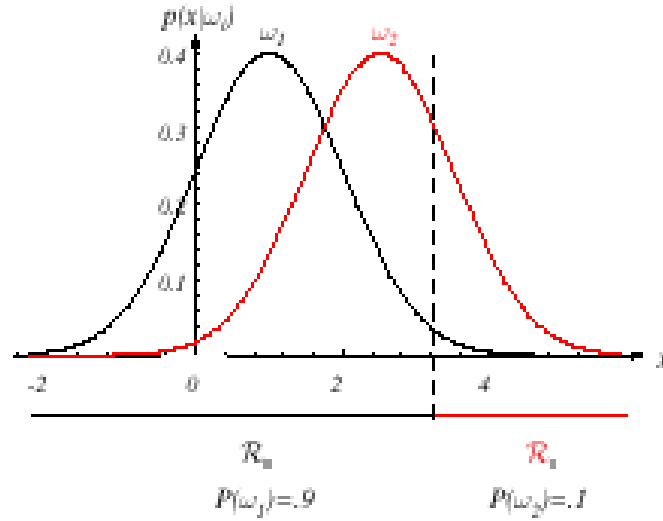
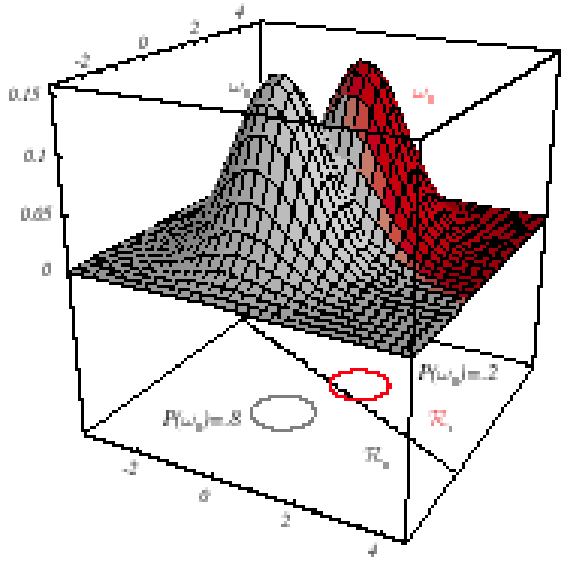
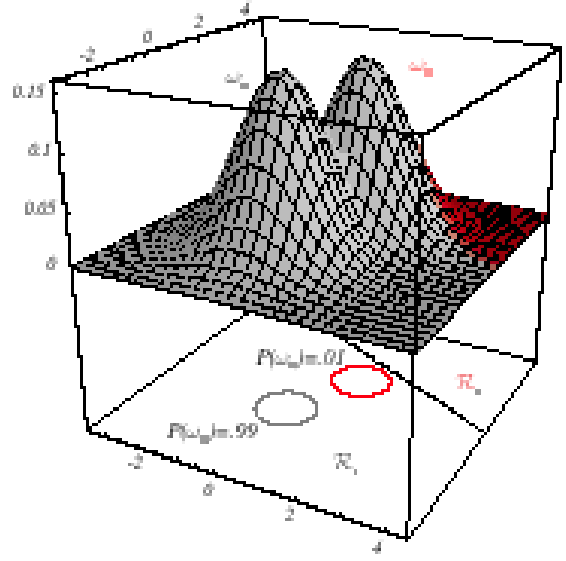


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-,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


p(x|ω_j)
ω_1
ω_2
0.4
0.3
0.2
0.1
x
-2
0
2
4
R_1
R_2
P(ω_1)=.7
P(ω_2)=.3


p(x|ω_j)
ω_1
ω_2
0.4
0.3
0.2
0.1
x
-2
0
2
4
R_1
R_2
P(ω_1)=.9
P(ω_2)=.1


0.15
0.1
0.05
0
P(ω_2)=.2
P(ω_1)=.8
R_1
R_2


0.15
0.1
0.05
0
P(ω_2)=.01
P(ω_1)=.99
R_1
R_2

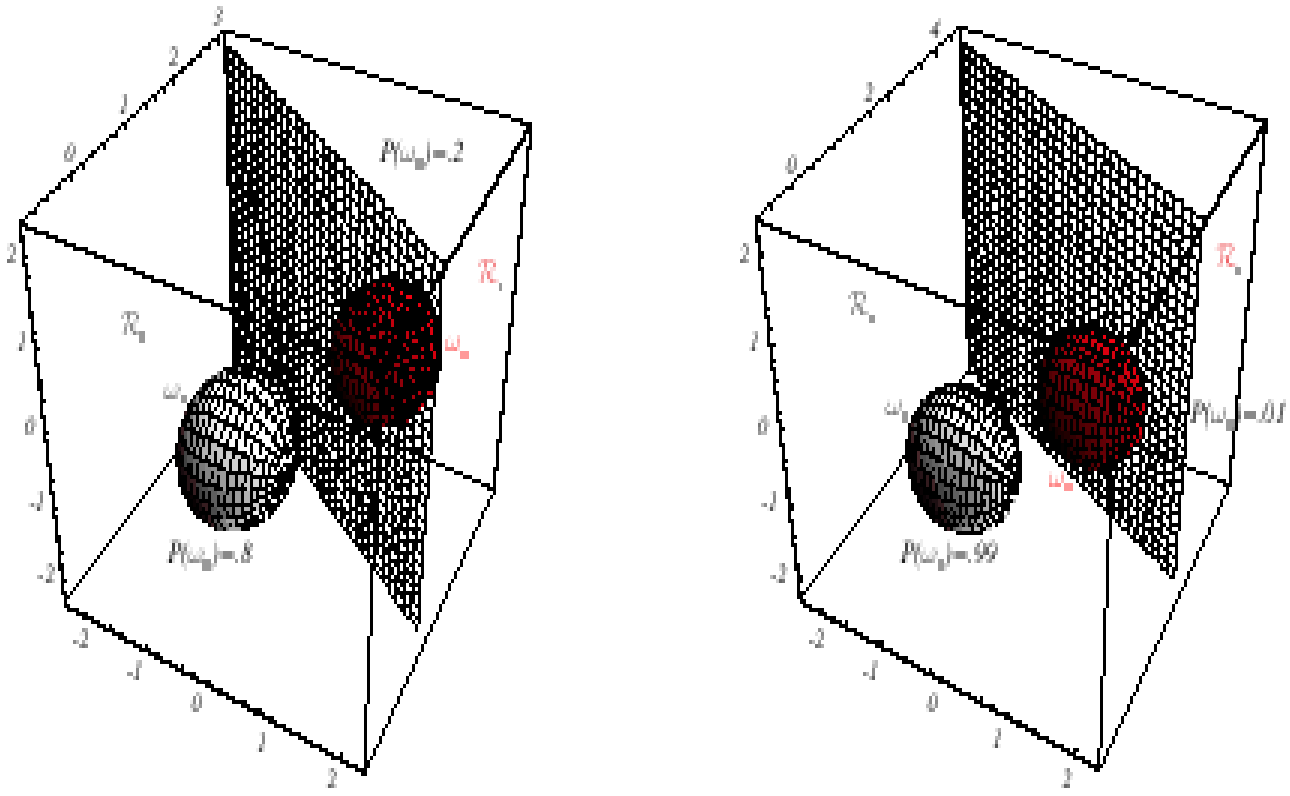


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

- Discriminant function: Two class problem:

$$w_2 x_2 + w_1 x_1 + w_0 = 0$$
$$\text{or} \vec{w}^t \vec{x} + w_0 = 0$$

- Where

$$\vec{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$
$$x_2 = \frac{-w_1}{w_2} x_1 - \frac{w_0}{w_2}$$

- Which is similar to a $y = mx + c$
- Eqn of a plane:

$$w_3 x_3 + w_2 x_2 + w_1 x_1 + w_0 = 0$$

- Eqn of a hyperplane:

$$w_4 x_4 + w_3 x_3 + w_2 x_2 + w_1 x_1 + w_0 = 0$$

Case 2:Identical Covariances

- This corresponds to a situation in which the samples fall in hyperellipsoidal clusters of equal size and shape, the cluster for the i^{th}class being centered about the mean vector

$$\begin{aligned}
\Sigma_1 &= \Sigma_2 = \Sigma \\
g_i(\vec{x}) &= -\frac{1}{2}(\vec{x}-\mu_i)^t\Sigma^{-1}(\vec{x}-\mu_i) \\
&- \ln(2\pi)^{d/2} - \ln|\Sigma|^{1/2} + \ln P(C_i) \\
&\text{Neglecting Constants} \\
g_i(\vec{x}) &= -\frac{1}{2}(\vec{x}-\mu_i)^t\Sigma^{-1}(\vec{x}-\mu_i) \\
&+ \ln P(C_i) \\
&= -\frac{1}{2}(\vec{x}^t\Sigma^{-1}\vec{x} + \vec{x}^t\Sigma^{-1}\vec{\mu_i} \\
&-\vec{\mu_i}^t\Sigma^{-1}\vec{x} + \vec{\mu_i}^t\Sigma^{-1}\vec{\mu_i})
\end{aligned}$$

Case 2:Identical Covariances

- Ignoring class independent terms and as

$$\vec{a}^t M \vec{b} = \vec{b}^t M \vec{a}$$

$$g_i(\vec{x}) = -\frac{1}{2}(-2\vec{\mu_i}^t \Sigma^{-1} \vec{x} + \vec{\mu_i}^t \Sigma^{-1} \vec{\mu_i}) + \ln P(C_i)$$

$$= (\Sigma^{-1}\vec{\mu_i})^t \vec{x} - \frac{1}{2}\vec{\mu_i}^t \Sigma^{-1} \vec{\mu_i} + \ln P(C_i)$$

$$= \vec{w_i}^t \vec{x} + w_{i0}$$

$$g(\vec{x}) = \vec{w}^t(\vec{x} - \vec{x_0}) = 0 = g_1(\vec{x}) - g_2(\vec{x})$$

$$\vec{w_i} = \Sigma^{-1}\vec{\mu_i}, w_{i0} = -\frac{1}{2}\vec{\mu_i}^t \Sigma^{-1} \vec{\mu_i} + \ln P(C_i)$$

- Simplifying

$$g(\vec{x}) = \vec{w}^t \vec{x} + w_0 = 0$$
$$\vec{w} = \Sigma^{-1}(\vec{\mu_1} - \vec{\mu_2})$$
$$w_0 = \frac{-1}{2}(\vec{\mu_1}^t \Sigma^{-1} \vec{\mu_1} - \vec{\mu_2}^t \Sigma^{-1} \vec{\mu_2}) + \ln \frac{P(C_1)}{P(C_2)}$$

- Adding $\vec{\mu_1}^t \Sigma^{-1} \vec{\mu_2}$ and subtracting $\vec{\mu_2}^t \Sigma^{-1} \vec{\mu_1}$

$$
\begin{aligned}
w_0 &= -1/2(\vec{\mu_1}^t \Sigma^{-1} \vec{\mu_1} + \vec{\mu_1}^t \Sigma^{-1} \vec{\mu_2}) \\
&\quad + \ln \frac{P(C_1)}{P(C_2)} \\
&= -1/2(\vec{\mu_1}^t \Sigma^{-1} (\vec{\mu_1} + \vec{\mu_2}) \\
&\quad - \vec{\mu_2}^t \Sigma^{-1} (\vec{\mu_1} + \vec{\mu_2})) + \ln \frac{P(C_1)}{P(C_2)} \\
&= -1/2(\vec{\mu_1} + \vec{\mu_2}) + \ln \frac{P(C_1)}{P(C_2)} \frac{(\vec{\mu_1} - \vec{\mu_2})}{(\vec{\mu_1} - \vec{\mu_2})^t \Sigma^{-1} (\vec{\mu_1} - \vec{\mu_2})}
\end{aligned}
$$

- This term is squared mahanalobis distance of two distributions with same covariances and means $\vec{\mu_1}$ and $\vec{\mu_2}$

$$g(\vec{x}) = \vec{w}^t(\vec{x} - \vec{x_0}) \Rightarrow$$

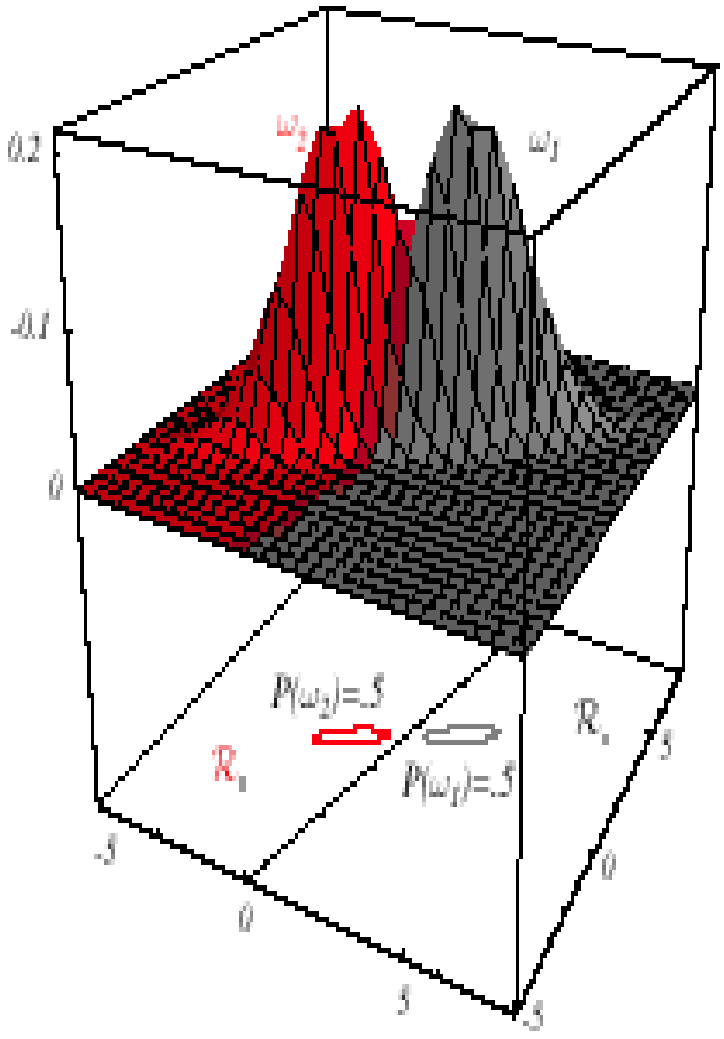
$$\vec{x_0} = \frac{1}{2}(\vec{\mu_1} + \vec{\mu_2}) - \frac{(\vec{\mu_1} - \vec{\mu_2})}{(\vec{\mu_1} - \vec{\mu_2})^t \Sigma^{-1} (\vec{\mu_1} - \vec{\mu_2})} \ln \frac{P(C_1)}{P(C_2)}$$

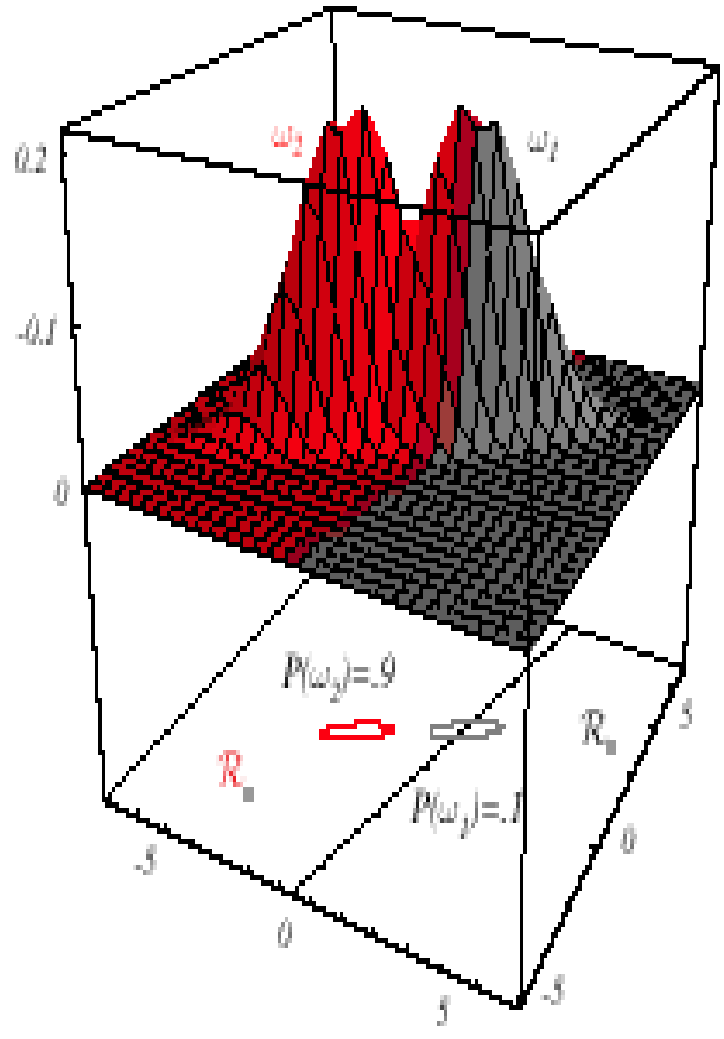
- Since

$$\vec{w_0} = -\vec{w}^t \vec{x_0}$$

- and

$$\vec{w}^t = (\Sigma^{-1}(\vec{\mu_1} - \vec{\mu_2}))^t$$


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


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

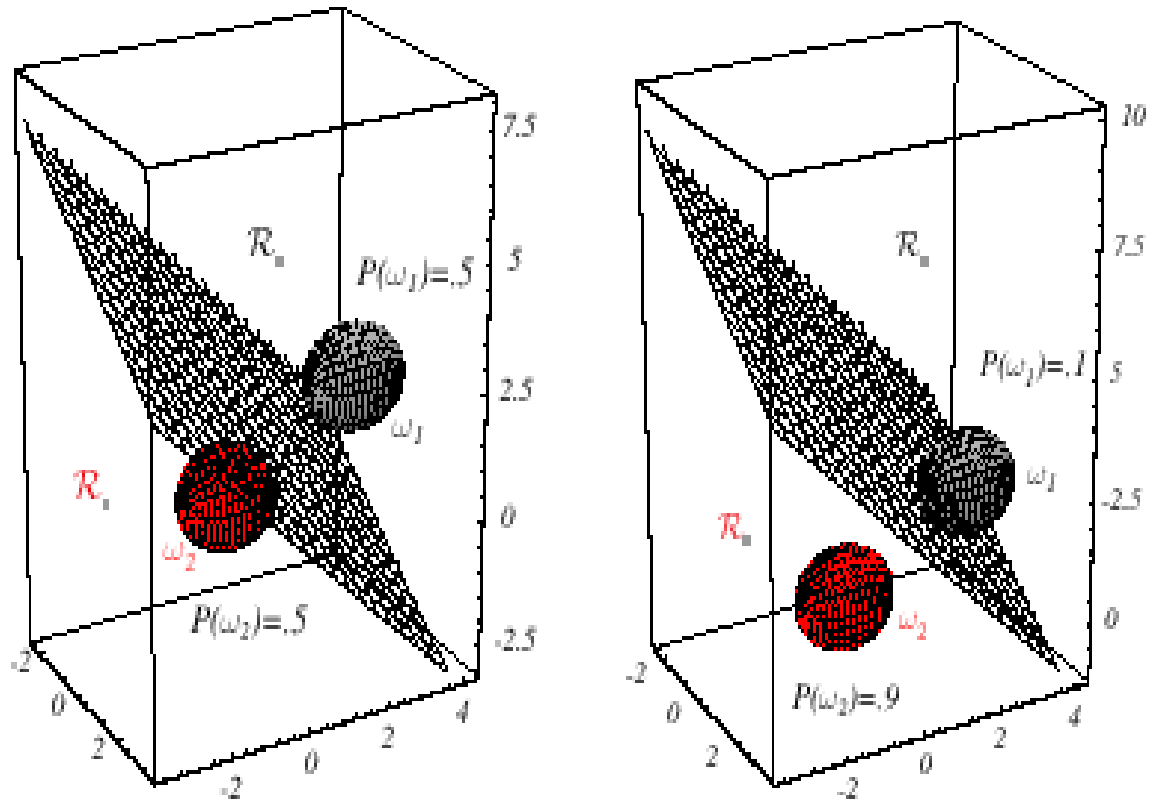


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda,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

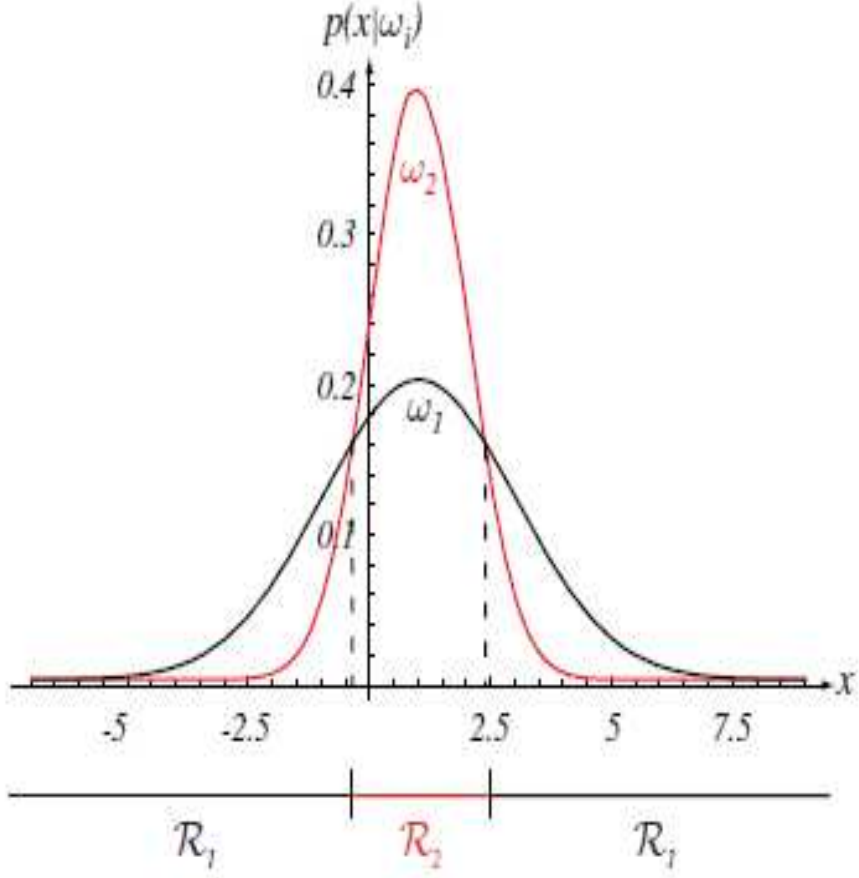


FIGURE 2.13. Non-simply connected decision regions can arise in one dimensions for Gaussians having unequal variance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Arbitrary covariance matrix

- Covariance matrices are different for each category

$$p(\vec{x}|C_1) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} exp[\frac{-1}{2}(\vec{x}-\vec{\mu_i})^t\Sigma_i^{-1}(\vec{x}-\vec{\mu_i})]$$

$$g_i(\vec{x}) = -\frac{1}{2}[(\vec{x}-\vec{\mu_i})^t\Sigma_i^{-1}(\vec{x}-\vec{\mu_i})] + \ln P(C_i) - \frac{1}{2}\ln|\Sigma_i|$$

$$= -\frac{1}{2}[\vec{x}^t\Sigma_i^{-1}\vec{x} - 2\vec{\mu_i}^t\Sigma_i^{-1}\vec{x} + \vec{\mu_i}^t\Sigma^{-1}\vec{\mu_i}] + \ln P(C_i) - \frac{1}{2}\ln|\Sigma_i|$$

$$g(\vec{x}) = g_1(\vec{x}) - g_2(\vec{x}) = 0$$

- Simplifying

$$g(\vec{x}) = -\frac{1}{2}[\vec{x}^t(\Sigma_1^{-1} - \Sigma_2^{-1})\vec{x}] + [\Sigma_1^{-1}\vec{\mu}_1 - \Sigma_2^{-1}\vec{\mu}_2]^t\vec{x}$$
$$-\frac{1}{2}\vec{\mu_1}^t\Sigma_1^{-1}\vec{\mu_1} + \frac{1}{2}\vec{\mu_2}^t\Sigma_2^{-1}\vec{\mu_2}$$
$$-\frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|} + \ln\frac{P(C_1)}{P(C_2)}$$

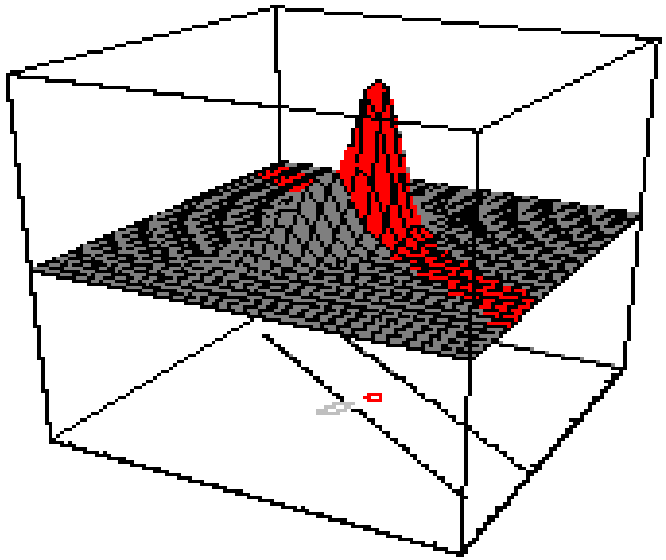
- This can be written as

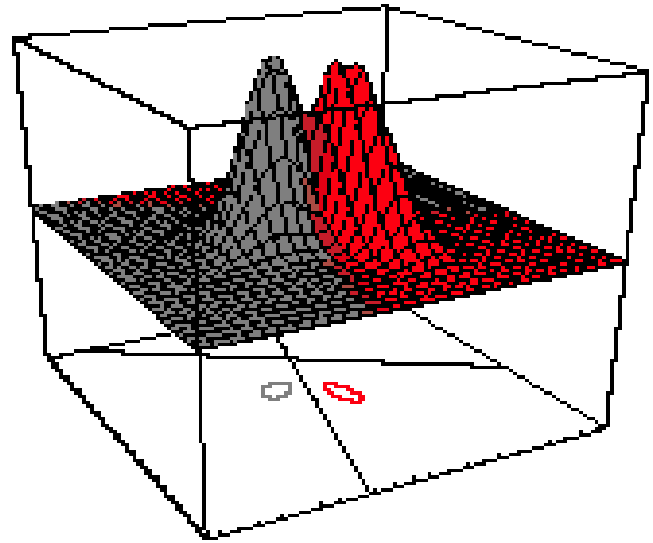
$$g(\vec{x}) = \vec{x}^t W\vec{x} + \vec{W}^t\vec{x} + w_0 = 0$$

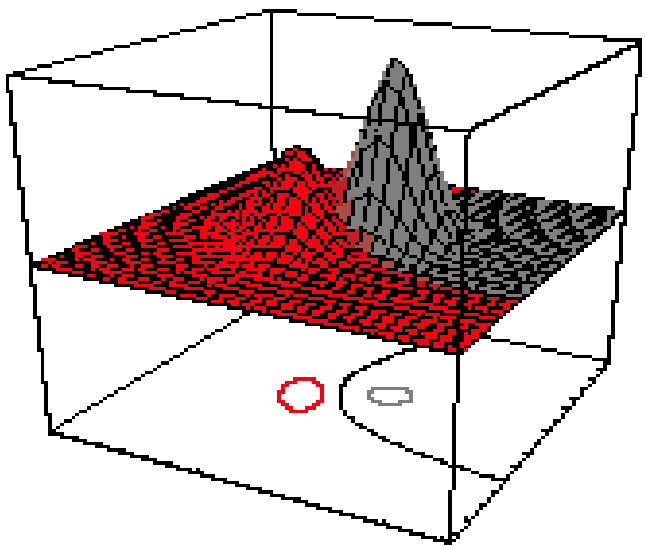
- This is the equation for the decision surface i.e., hyperquadratic
- The shape of the surface depends on the matrix W
- IF W is positive definite, then surface is hyperellipsoid, whose axes are in the directions of the eigenvectors of W
- If $W = kI$ where $k > 0$, then its a hypersphere
- If W is positive semi-definite, then surface is hyperellipsoidal cylinder
- Otherwise the decision surface is referred to as hyperhyperboloid

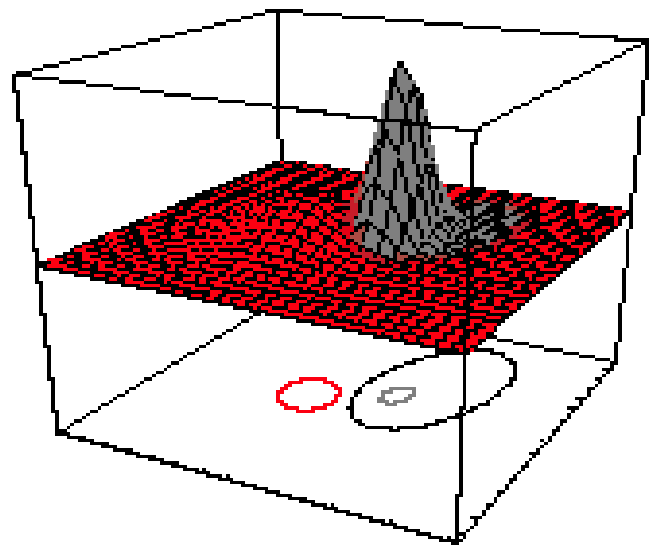
- Circle:$(x-h)^2+(y-k)^2=r^2$:d-dim- hypersphere
- Ellipse: $\frac{(x-h)^2}{a^2}+\frac{(y-k)^2}{b^2}=c$:hyperellipse
- Parabola: $y^2=4ax$:hyperparabola
- Hyperbola:d-dim-hyperhyperboloid

Arbitrary covariance matrix









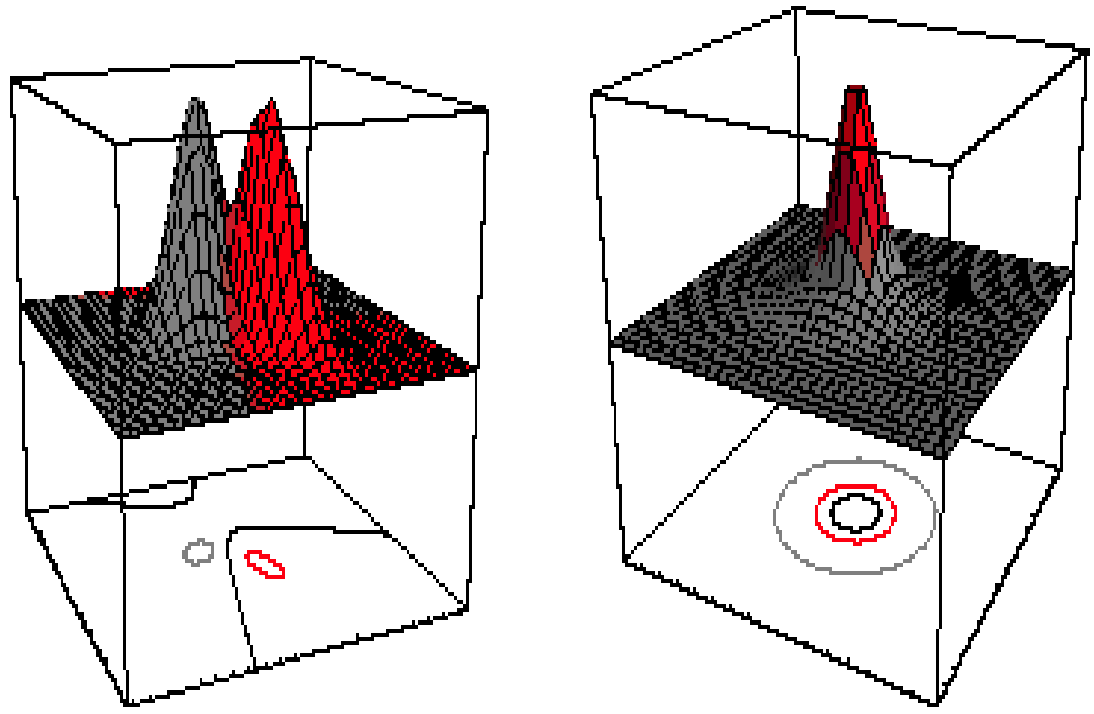
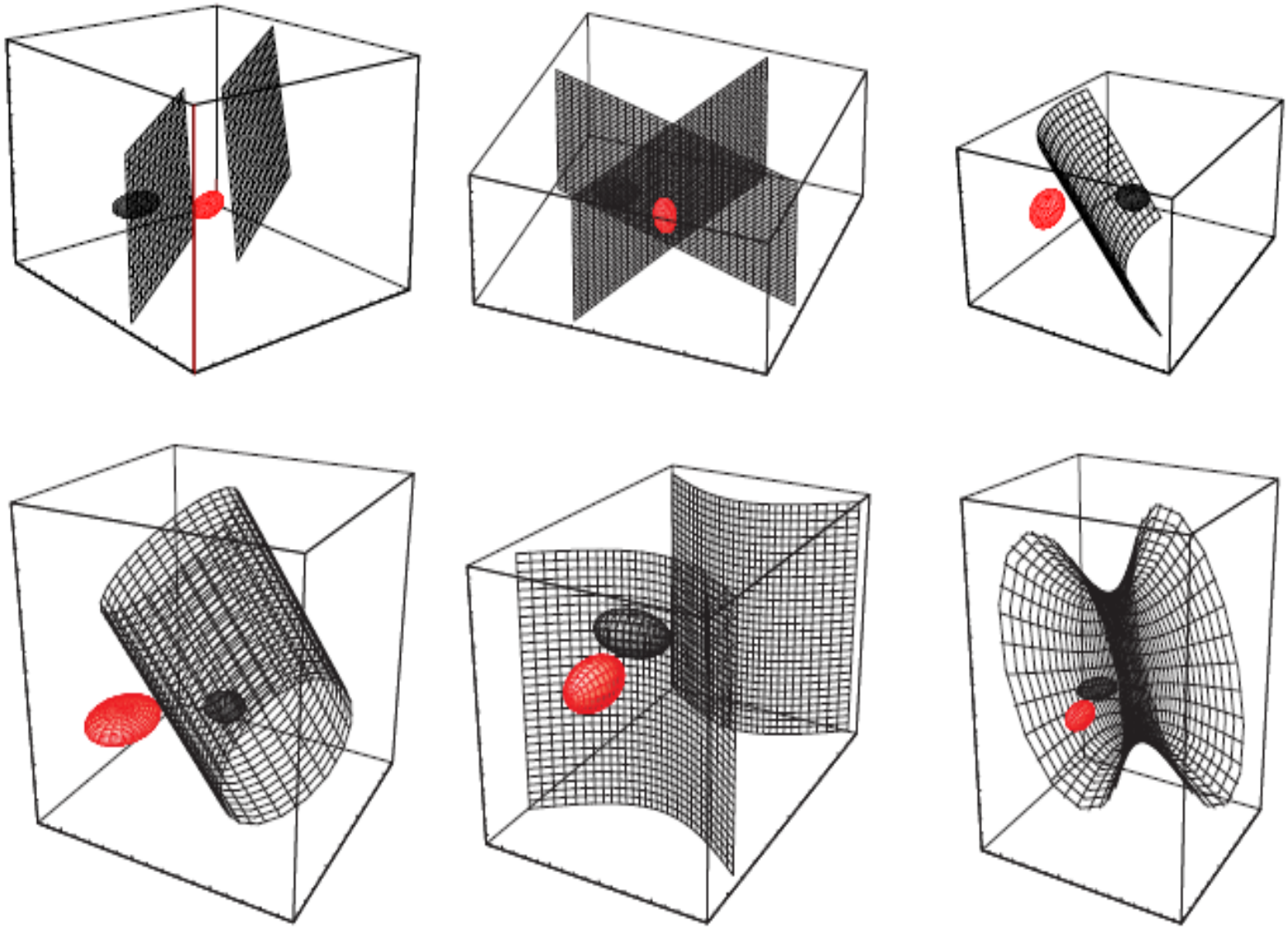


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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



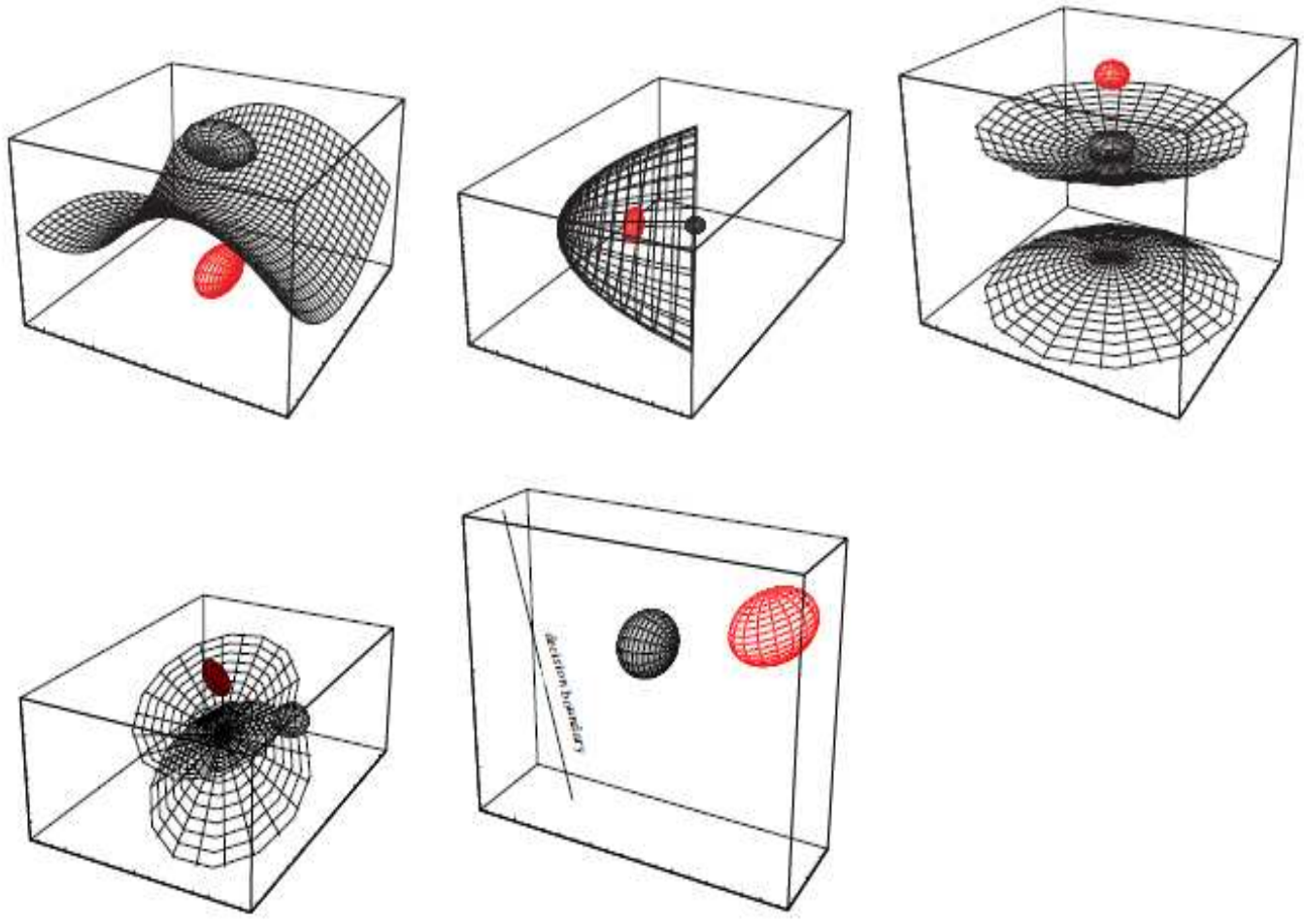


FIGURE 2.15. Arbitrary three-dimensional Gaussian distributions yield Bayes decision boundaries that are two-dimensional hyperquadrics. There are even degenerate cases in which the decision boundary is a line. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright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

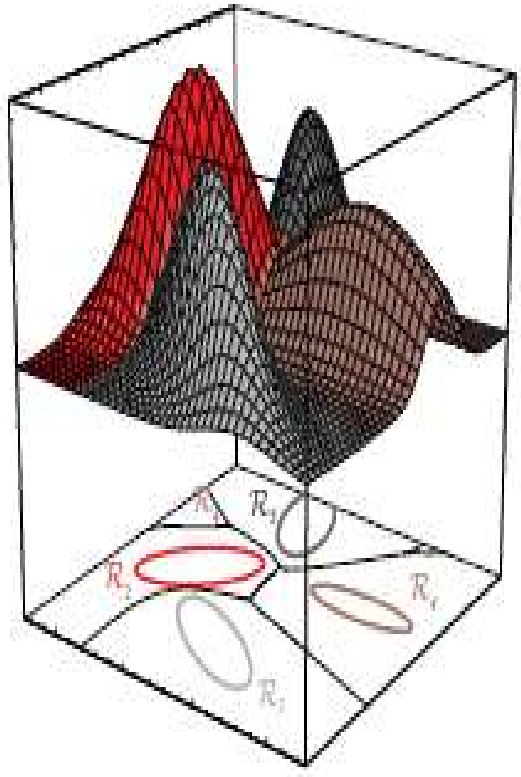


FIGURE 2.16. The decision regions for four normal distributions. Even with such a low number of categories, the shapes of the boundary regions can be rather complex. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Measures of classification performance

- As features are common to more than one class, regions of support overlap resulting in classification error
- The classification error for a two-class problem can be written as

$$P(error|\vec{x}) = \begin{cases} P(\omega_1|\vec{x}) \text{if we decide} \omega_2 \\ P(\omega_2|\vec{x}) \text{if we decide} \omega_1 \end{cases}$$

- Expressing Conditional error: i)Joint occurrence of a random variable $\vec{X}$ and an event A_i

$$P_{\vec{X}A}(x, A_i) = P(\vec{X} \leq \vec{x} \cap \text{event} A_i \text{occurs}$$

- ii)the Joint occurrence of two random variables :

$$P_{\vec{X}\omega}(\vec{x}, \omega_i) = P(\vec{X} \leq \vec{x} \cap \omega = \omega_i \cup)$$

- Error formulation for case i)

$$P(error) = \int P(error, \vec{x}) d\vec{x}$$

- Error formulation for case ii)

$$P(error) = \sum_{i=1}^{c} P(error \cap \omega_i) = \sum_{i=1}^{c} P(error|\omega_i) P(\omega_i)$$

- Both formulations are useful in error analysis

$$P(error, \vec{x}) = P(error|\vec{x}) p(\vec{x})$$

$$P(error) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} .... \int_{-\infty}^{\infty} P(error|\vec{x}) p(\vec{x}) d\vec{x}$$

- To minimize error, Choose ω_i(over all $\vec{x}$ such that $P(error|\vec{x})$ is smallest

General measures of CLassification Risk

- Case 1: The incoming signal is a valid signal and correctly classify it (detection)
- Case 2: The incoming signal is a valid signal and incorrectly classify it as noise(a miss)
- Case 3: The incoming signal is noise and incorrectly classify it as a valid signal (false alarm)
- Case 4: The incoming signal is noise and correctly classify it
- Formulate a loss function, cost function or risk function λ_{ij}, as
- λ_{ij} is the cost or risk of choosing class ω_i when class ω_j is the true class

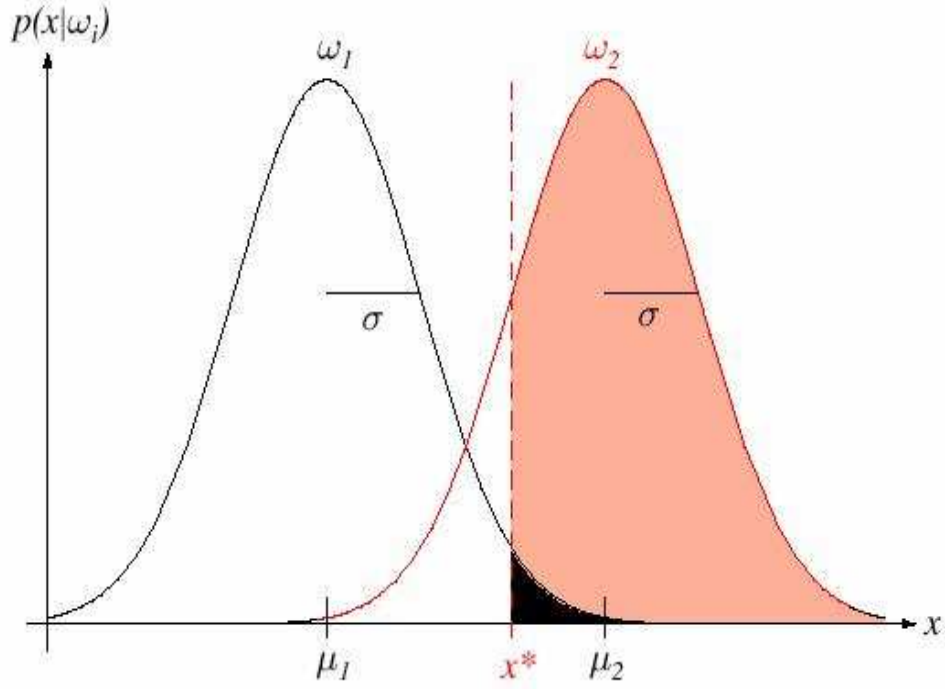


FIGURE 2.19. During any instant when no external pulse is present, the probability density for an internal signal is normal, that is, $p(x|\omega_1) \sim N(\mu_1, \sigma^2)$; when the external signal is present, the density is $p(x|\omega_2) \sim N(\mu_2, \sigma^2)$. Any decision threshold x^* will determine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

- As λ_{11} and λ_{22} are the costs(rewards) for correct decision
- and λ_{12} and λ_{21} are the costs for classification error for a two-class problem
- A decision rule or classification is a mapping of the observed feature vector,$\vec{x}$ into a *alpha*$_i$ through a decision rule

$$\alpha(\vec{x}) = \{\alpha_1, \alpha_2 ... \alpha_c\}$$

- Overall risk measure for a two-class problem is

$$R = \lambda_{11} P(\alpha_1|\omega_1) P(\omega_1) + \lambda_{21} P(\alpha_2|\omega_1) P(\omega_1) + \lambda_{12} P(\alpha_1|\omega_2) P(\omega_2) -$$

- A measure of conditional risk associated with a two-class rpoblem is

$$R\left[\alpha(\vec{x}) \rightarrow \alpha_1\right] = R(\alpha_1|\vec{x}) = \lambda_{11}P(\omega_1|\vec{x}) + \lambda_{12}P(\omega_2|\vec{x})$$
$$R\left[\alpha(\vec{x}) \rightarrow \alpha_2\right] = R(\alpha_2|\vec{x}) = \lambda_{21}P(\omega_1|\vec{x}) + \lambda_{22}P(\omega_2|\vec{x})$$

- The expected risk for a c class decision problem can be written using total probability theorem as

$$R\left[\alpha(\vec{x})\right] = \int R\left[\alpha(\vec{x})|\vec{x}\right] p(\vec{x}) d\vec{x}$$

- Minimizing conditional risk minimizes the expected risk
- Lower bound on $R\left[\alpha(\vec{x})\right]$ is referred to as the Bayes risk

Minimizing Bayes risk

- For a two-class problem, decision rule is formulated as

$$R(\alpha_1|\vec{x}) \gtrless_{\alpha_1}^{\alpha_2} R(\alpha_2|\vec{x})$$

- Substituting for R

$$\lambda_{11}P(\omega_1|\vec{x}) + \lambda_{12}P(\omega_2|\vec{x}) \gtrless_{\alpha_1}^{\alpha_2} \lambda_{21}P(\omega_1|\vec{x}) + \lambda_{22}P(\omega_2|\vec{x})$$
$$(\lambda_{11} - \lambda_{21})p(\vec{x}|\omega_1)P(\omega_1) \gtrless_{\alpha_1}^{\alpha_2} (\lambda_{22} - \lambda_{12})p(\vec{x}|\omega_2)P(\omega_2)$$

- As $\lambda_{11} = \lambda_{22} = 0$ there is no cost or risk associated with correct classification

$$\frac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)} \gtrless_{\alpha_1}^{\alpha_2} \frac{(\lambda_{22} - \lambda_{12})}{(\lambda_{11} - \lambda_{21})} \frac{P(\omega_2)}{P(\omega_1}$$

- Similar to LRT
- For $\lambda_{11} = \lambda_{22} = 0$ and $\lambda_{12} = \lambda_{21} = 1$ this reduces to intuitive classification strategy

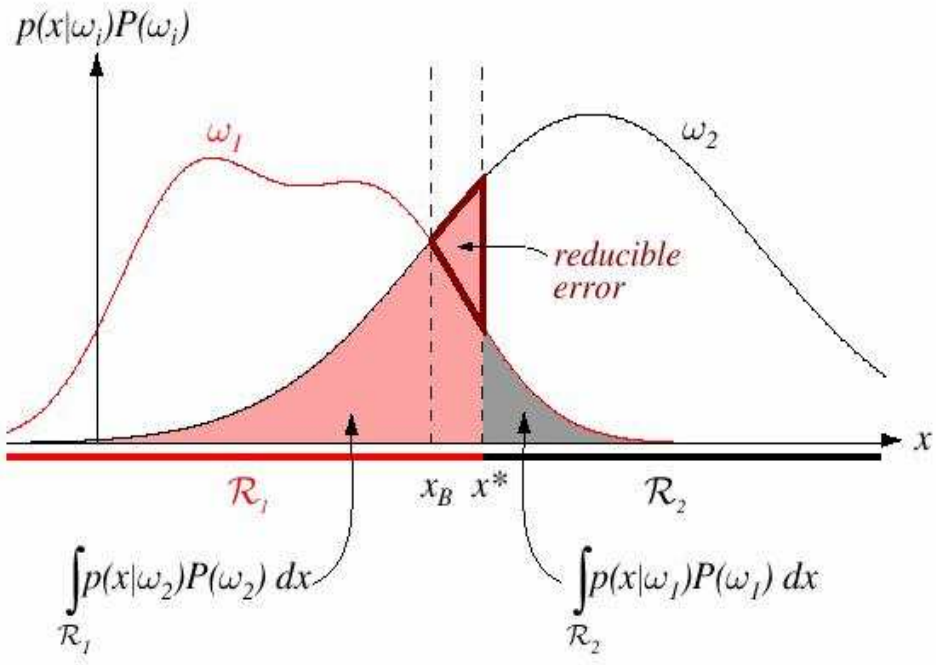


FIGURE 2.17. Components of the probability of error for equal priors and (nonoptimal) decision point x^*. The pink area corresponds to the probability of errors for deciding ω_1 when the state of nature is in fact ω_2; the gray area represents the converse, as given in Eq. 70. If the decision boundary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

C-class classification

- The loss function for c classes

$$\lambda_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

- Assuming all errors equally costly
- The conditional risk of decision α_i is based on generalizing

$$R\left[\alpha(\vec{x}) \rightarrow \alpha_i\right] = \sum_{j=1}^{c} \lambda_{ij} P(\omega_j|\vec{x})$$

$$= \sum_{j \neq i} P(\omega_j|\vec{x}) = 1 - P(\omega_i|\vec{x})$$

- To minimize the conditional risk, the decision rule is to choose α_i that maximizes $P(\omega_i|\vec{x})$, that is ω_i for which $P(\omega_i|\vec{x})$ is largest
- This is MAP classifier which may be formulated as

$$P(\omega_i|\vec{x}) >^{\alpha_i} P(\omega_j|\vec{x}) \; \forall j \neq i$$

or

$$R(\alpha_i|\vec{x}) <^{\alpha_i} R(\alpha_j|\vec{x}) \; \forall i \neq j$$

Error Bounds

- Discontinuous nature of the decision regions makes evaluation of integral for Gaussina case to be very difficult
- Error integral can be approximated to give an upper bound on error
- Chernoff Bound:

$$P(error) = P^{\beta}(\omega_1)P(1-\beta)(\omega_2)\int p^{\beta}(\vec{x}|\omega_1)p^{1-\beta}(\vec{x}|\omega_2)d\vec{x}\, 0 \leq \beta \leq$$

- This integral is over all feature space-not corresponding to decision boundaries
- IF the conditional probabilities are normal, then

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

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

- Chernoff bound depends on β and the bound is loose for exterme values of β(i.e., $\beta \to 1$ and $\beta \to 0$
- Bhattacharya Bound:BOund is tighter for intermediate values
- Bhattacharya bound: Slightly less tighter bound

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

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

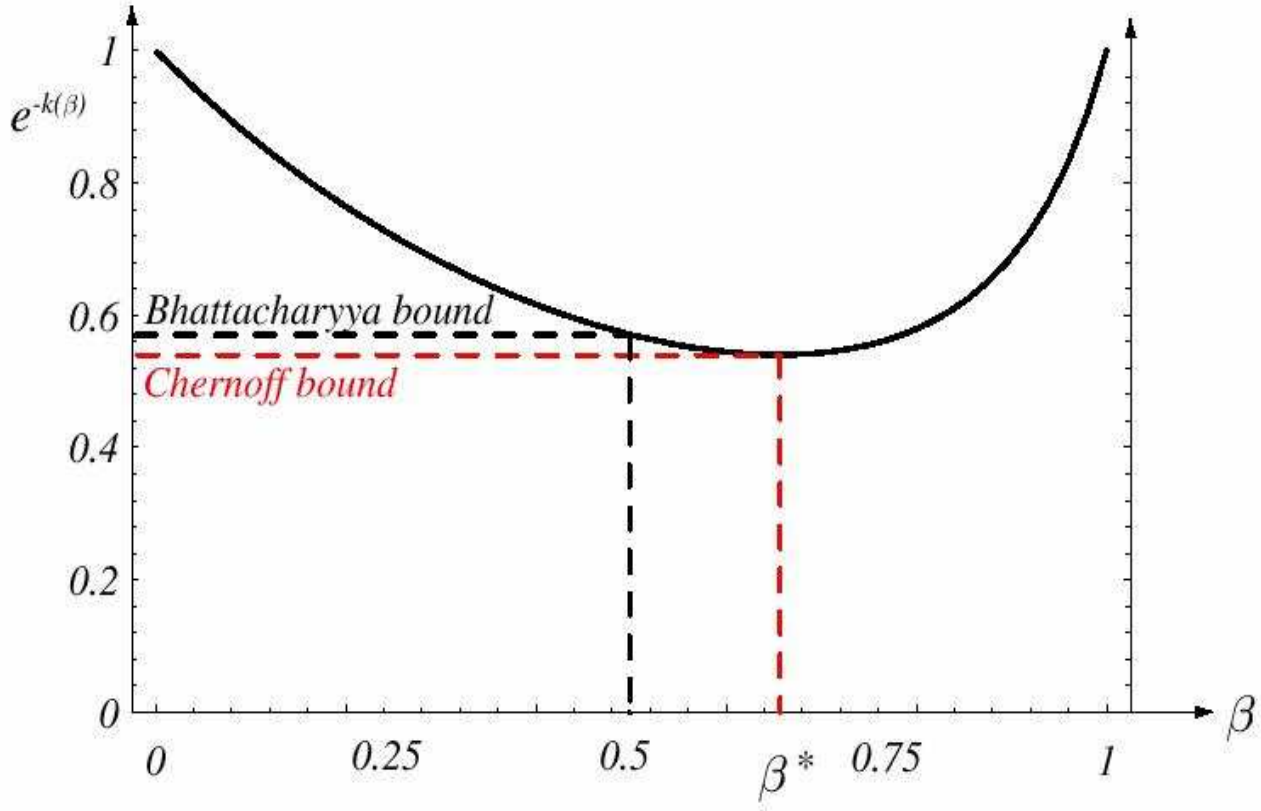


FIGURE 2.18. The Chernoff error bound is never looser than the Bhattacharyya bound. For this example, the Chernoff bound happens to be at $\beta^* = 0.66$, and is slightly tighter than the Bhattacharyya bound ($\beta = 0.5$). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

GMM

- Given a c-class problem, decision surface is linear or quadratic depending on the covariance
  or interrelationship among classes
- Discriminant function for a decision surface

$$g_{ij}(\vec{x}) = g_i(\vec{x}) - g_j(\vec{x})$$

- The distributions can be Gaussian or Poisson, Rayleigh distribution or Erlang...
- Even if the data has a multimodal distribution, then it can be approximated
  as a combination of more than one Gaussian distribution

$$p(\vec{x}|\omega_i) = \sum_{k=1}^{k_i} \omega_k N(\tilde{\mu_{ik}}, \Sigma_{ik})$$

- For d-dimensional data, we require d parameters for mean and $d(d+1)/2$ for Covariance matrix as it is symmetric
- For k no. of d-dim distributions, $(d + \frac{d(d+1)}{2}) * k + k$

Discrete features

- Components of $\vec{x}$(feature vector) in many practical applications are discrete-valued
- They may be binary, ternary or higher-integer valued, so that $\vec{x}$ can assume only one of m discrete values

$$\vec{x} = [x_1 x_2 .... x_d]^t$$
$$x_j \in (0, 1)$$

- For biased coins:$P(H) \neq P(T)$: feature takes binary value
- Presence or absence of a event
- Bayes formula in discrete feature case involves probabilities rather than probability densities
- Two class: $P(\vec{x}|\omega_1), P(\vec{x}|\omega_2)$

Discrete features

- Compute their prior probability and

$$P(C_1|\vec{x}) = \frac{P(\vec{x}|\omega_1)P(C_1)}{P(\vec{x}}$$
$$P(C_2|\vec{x}) = \frac{P(\vec{x}|C_2)P(C_2)}{P(\vec{x})}$$

- Gaussian or normal density assumption donot hold good as the probabilities are discrete

$$P(H|C_1) = p; P(T|C_2) = 1 - p$$
$$P(x|C_1) = p^x(1-p)^{1-x}$$

- $P(x|C_1)$ is the Probability mass function

Independent Binary features

- Consider a two-category problem with binary conditionally independent features
- By assuming conditional independence

$$p(a, b|c) = p(a|c)p(b|c)$$

- Conditionally independent though $p(a, b) \neq p(a)p(b)$

$$P(\vec{x}|C_1) = P(x_1|C_1).P(x_2|C_1)....P(x_d|C_1)$$
$$= \prod_{j=1}^{d} P(x_j|C_1)$$

- (Bernoulli distribution:

$$P(x_j|C_1) = p_j^{x_j}(1-p_j)^{1-x_j}$$

$$P(\vec{x}|C_1) = \prod_{j=1}^{d} p_j^{x_j}(1-p_j)^{(1-x_j)}$$

$$P(\vec{x}|C_2) = \prod_{j=1}^{d} q_j^{x_j}(1-q_j)^{(1-x_j)}$$

$$P(C_1|\vec{x}) = \frac{P(\vec{x}|C_1)P(C_1)}{P(\vec{x})}$$

$$g_1(\vec{x}) = \ln \frac{P(\vec{x}|C_1).P(C_1)}{P(\vec{x})}$$

$$g_2(\vec{x}) = \ln \frac{P(\vec{x}|C_2).P(C_2)}{P(\vec{x})}$$

- $$g(\vec{x}) = g_1(\vec{x}) - g_2(\vec{x})$$
$$= \ln \prod_{j=1}^{d} p_j^{x_j} (1-p_j)^{(1-x_j)} P(C_1)$$
$$- \ln \prod_{j=1}^{d} q_j^{x_j} (1-q_j)^{(1-x_j)} P(C_q)$$

- By expansion, for independent binary features
$$g(\vec{x}) = \sum_{j=1}^{d} \left[x_j \ln \frac{p_j}{q_j} + (1-x_j) \ln \frac{1-p_j}{1-q_j} \right] + \ln \frac{P(C_1)}{P(C_2)}$$
$$g(\vec{x}) = \sum_{j=1}^{d} w_j x_j + w_0 = \frac{\vec{w}^t \vec{x} + w_0}{d}$$
$$\Rightarrow w_j = \ln \frac{p_j/(1-p_j)}{q_j/(1-q_j)}; w_0 = \sum_{j=1}^{d} \ln \frac{(1-p_j)}{1-q_j)} + \ln \frac{P(C_1)}{P(C_2)}$$

- If $p_j = q_j$, then behaviour of features is same for both classes($\omega_j = 0$): Non discriminative
- If $p_j > q_j$, $\Rightarrow w_j = +ve$ More contribution or
- If $p_j < q_j$, $\Rightarrow w_j = -ve$ Less contribution
- Special class of Naive-Bayes classifier where statistically independent features are assumed
- Covariance matrices must be same for linear decision surface
- Problem