Bayes Minimum Error Classifier – 2 Classes

2-Class

$p(\underline{x}|S_1)P(S_1) \underset{<}{>} p(\underline{x}|S_2)P(S_2)$

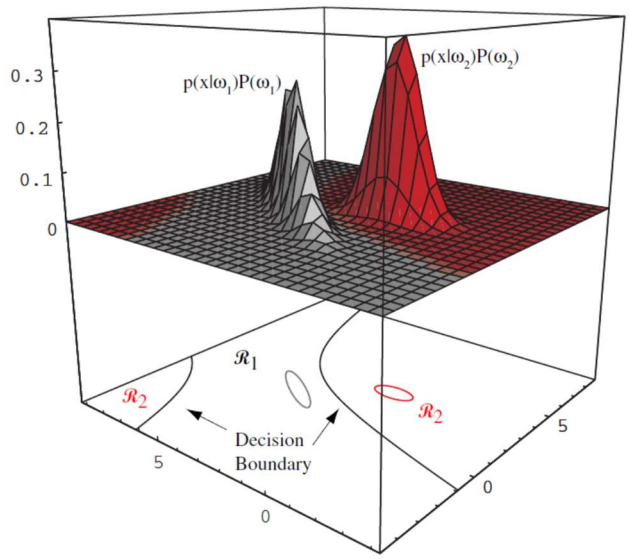


Figure 2.6: In this two-dimensional two-category classifier, the probability densities are Gaussian (with $1/e$ ellipses shown), the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected.

Likelihood ratio

$$l(\underline{x}) = \frac{p(\underline{x}|S_1)}{p(\underline{x}|S2)} \underset{<}{>} \frac{P(S_2)}{P(S_1)} = T$$

Log likelihood ratio $h(\underline{x}) = -\ln[\, l(\underline{x}) \,]$

$$h(\underline{x}) \underset{<}{>} -\ln T$$

Bayes Minimum Error Classifier – Multiple Classes

Bayes Minimum Error – Multiple Classes

$P(S_i|\underline{x}) > P(S_j|\underline{x})$ for all $j \neq i \Rightarrow \underline{x} \in S_i$

$p(\underline{x}|S_i)P(S_i) > p(\underline{x}|S_j)P(S_j)$ for all $j \neq i \Rightarrow \underline{x} \in S_i$

For Multiclass: if $p(\underline{x}|S_i)P(S_i) > p(\underline{x}|S_j)P(S_j)$ $\forall j \neq i$ then $\underline{x} \in S_i$

A Set of Discriminant functions: $g_i(\underline{x}) = p(\underline{x}|S_i)P(S_i)$

Bayes Classifiers with Normal Density Functions

Again, use a set of discriminant functions $g_i(\underline{x})$
i.e., $g_i(\underline{x}) > g_j(\underline{x})$ for all $j \neq i$.

Express $g_i(\underline{x})$ in terms of probabilites

$g_i(\underline{x}) = p(S_i | \underline{x})$

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

$g_i(\underline{x}) = \ln [p(\underline{x} | S_i)] + \ln [P(S_i)]$

Now Let's consider Gaussian (normal) density functions

Normal Density

- So far, general forms of density functions are considered
- Most widely studied density functions are the multivariate normal or Gaussian density
- Why? analytical tractability, most appropriate model

Univariate Density

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

μ=expected value of x, average or mean

σ=standard deviation

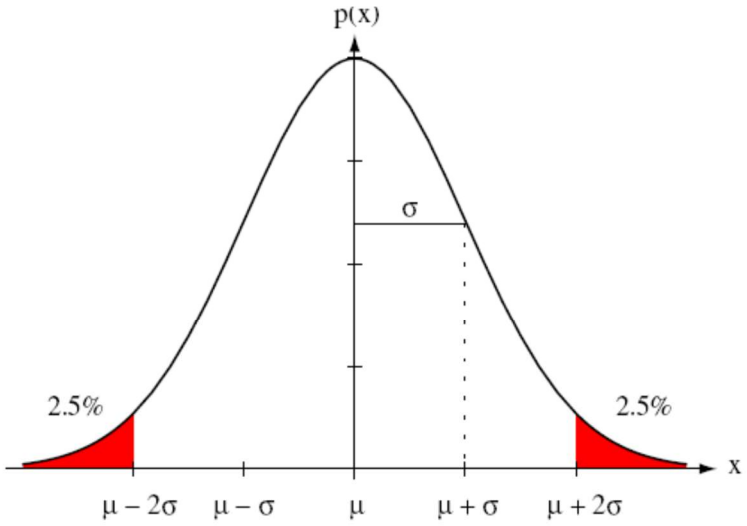


Figure 2.7: A univariate normal distribution has roughly 95% of its area in the range $|x-\mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$.

Multivariate Density

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

μ=mean vector

Σ=covariance matrix

Whitening Transformation

Transformation of an arbitrary multivariate normal distribution into a spherical one

That is one having a covariance matrix proportional to the identity matrix, I

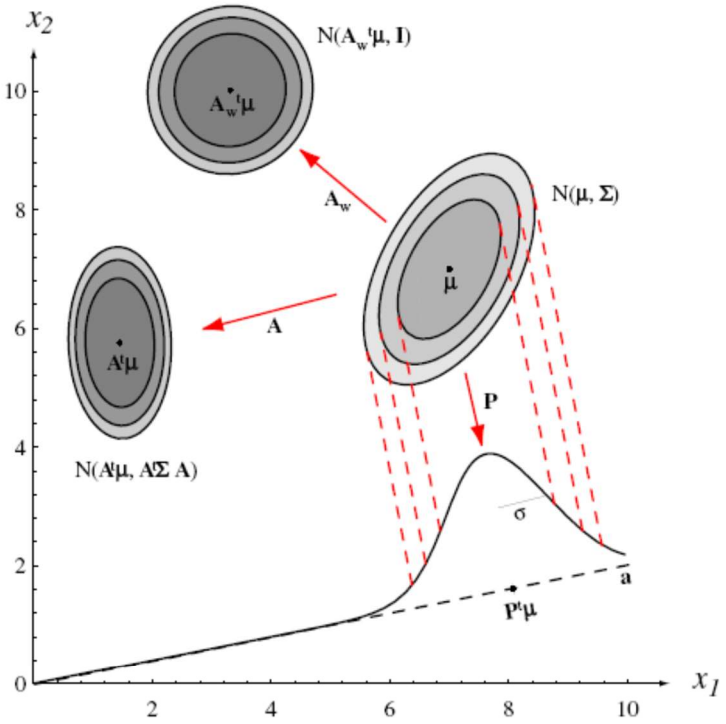


Figure 2.8: The action of a linear transformation on the feature space will convert an arbitrary normal distribution into another normal distribution. One transformation, $\mathbf{A}$, takes the source distribution into distribution $N(\mathbf{A}^t\mu, \mathbf{A}^t\mathbf{\Sigma}\mathbf{A})$. Another linear transformation — a projection $\mathbf{P}$ onto line $\mathbf{a}$ — leads to $N(\mu, \sigma^2)$ measured along $\mathbf{a}$. While the transforms yield distributions in a different space, we show them superimposed on the original $x_1 - x_2$ space. A whitening transform leads to a circularly symmetric Gaussian, here shown displaced.

Mahalanobis Distance = d_M

$d_M^2(\underline{x},\underline{m})=(\underline{x}-\underline{m})^T\boldsymbol{\Sigma}^{-1}(\underline{x}-\underline{m})$

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

The contours of constant density are hyperellipsoids of constant Mahalanobis distance to $\boldsymbol{\mu}$ in Fig. 2.9

$$\boldsymbol{\Sigma}=E\{(\underline{x}-\underline{m})\,(\underline{x}-\underline{m})^T\}=\begin{bmatrix}\sigma_{11}^2 & \sigma_{12}^2 & & \\ \sigma_{21}^2 & & & \\ & & & \\ & & & \sigma_{NN}^2\end{bmatrix}$$

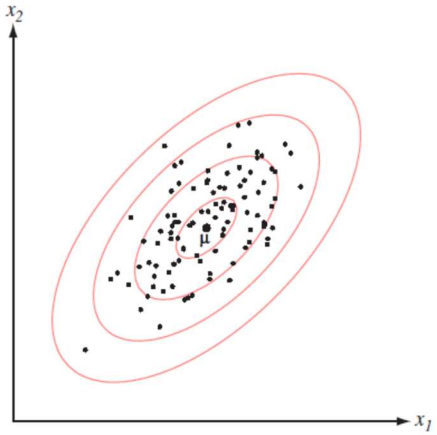
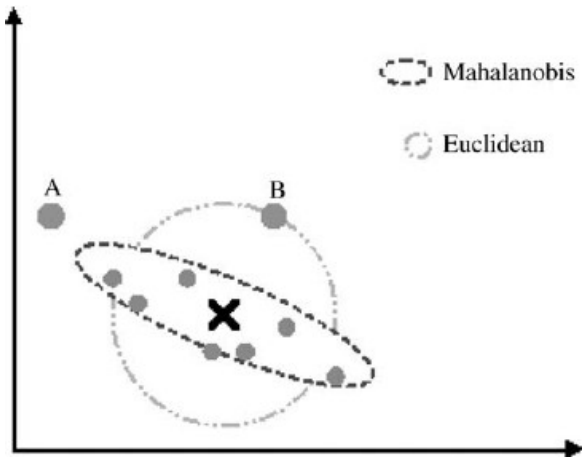


Figure 2.9: Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean μ. The red ellipses show lines of equal probability density of the Gaussian.



Discriminant Functions with Normal Density

The minimum error rate classification can be done using the discriminant functions:

$g_i(\underline{x})=\ln [p(\underline{x}|S_i)] + \ln [P(S_i)]$

If $p(\underline{x}|S_i)=N(\mu_i,\sum_i)$

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

Let's examine this discrimination function and resulting classification for three special cases.

Case 1: Same σ

If $\Sigma=\sigma^2 I=\begin{bmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \ldots & \\ & & & \sigma^2 \end{bmatrix}$

$\Sigma^{-1}=(1/\sigma^2)I$

$d_M^2(\underline{x},\underline{m})=(1/\sigma^2)(\underline{x}-\underline{m})^T(\underline{x}-\underline{m})= (1/\sigma^2)\, d_E^2(\underline{x},\underline{m})$

d_E: Euclidean distance

$$g_i(x) = -\frac{\|\mathbf{x}-\boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(S_i)$$

Since $\|\mathbf{x}-\boldsymbol{\mu}_i\|^2 = (\mathbf{x}-\boldsymbol{\mu}_i)^T(\mathbf{x}-\boldsymbol{\mu}_i)$

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}[\mathbf{x}^T\mathbf{x} - 2\boldsymbol{\mu}_i^T\mathbf{x} + \boldsymbol{\mu}_i^T\boldsymbol{\mu}_i] + \ln P(S_i) = \mathbf{w}_i^T\mathbf{x} + w_{n+1}$$

where $\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i$, $w_{n+1} = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i + \ln P(S_i)$

This equation shows that squared distance $\|x-\mu_i\|^2$ is normalized by the variance and offset by $\ln P(S_i)$. That is if x is equally near two different mean vectors, the optimal decision favors the a priori more likely category.

Equal Covariances

- Hyperspheres
- Radius scaled by σ

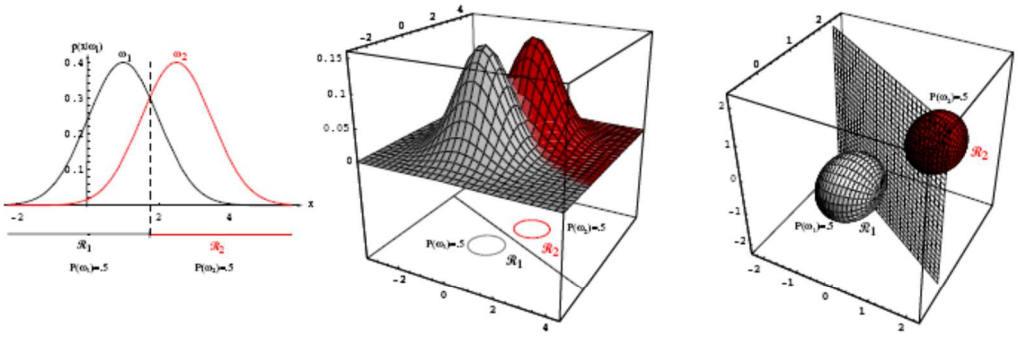


Figure 2.10: If the covariances of two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these 1-, 2-, and 3-dimensional examples, we indicate $p(\mathbf{x}|\omega_i)$ and the boundaries for the case $P(\omega_1) = P(\omega_2)$. In the 3-dimensional case, the grid plane separates $\mathcal{R}_1$ from $\mathcal{R}_2$.

Role of the Priors

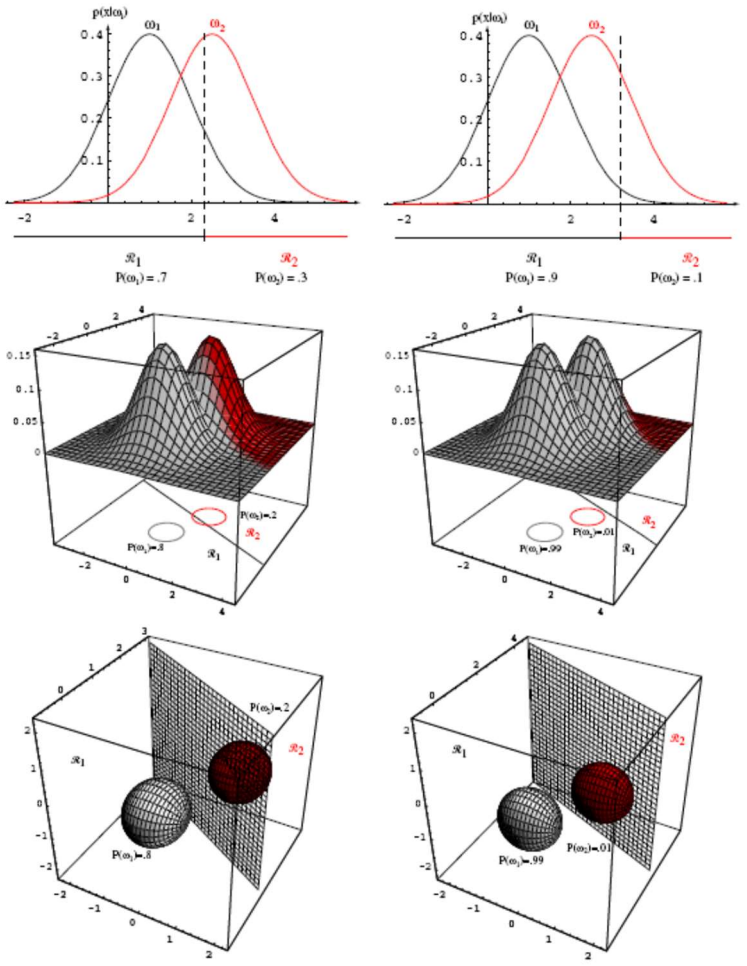


Figure 2.11: As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these 1-, 2- and 3-dimensional spherical Gaussian distributions.

Case 2: Different σ

If $\Sigma = \begin{bmatrix} \sigma_{11}^2 & & & \\ & \sigma_{22}^2 & & \\ & & \cdots & \\ & & & \sigma_{NN}^2 \end{bmatrix}$

$\Sigma^{-1} = \begin{bmatrix} 1/\sigma_{11}^2 & & & \\ & 1/\sigma_{22}^2 & & \\ & & \cdots & \\ & & & 1/\sigma_{NN}^2 \end{bmatrix}$

$d_M^2(\underline{x},\underline{m}) = (\underline{x}-\underline{m})^T\Sigma^{-1}(\underline{x}-\underline{m}) = \sum_{i=1}^{N} (1/\sigma_{ii}^2)(\underline{x}_i-\underline{m}_i)^2$

(i^{th} term of d_E is scaled by σ_{ii})

For the 2-D Case

$d_M^2=(x_1-m_1)^2/\sigma_{11}^2 + (x_2-m_2)^2/\sigma_{22}^2$

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

- To classify a feature vector $\underline{x}$, measure the squared Mahalanobis distance from $\underline{x}$ to each of the mean vectors, and assign $\underline{x}$ to the category of the nearest mean.
- Classifier becomes linear and decision boundaries become hyperplanes.

Probability Densities and Decision Regions for Equal, but Asymmetric Normal Distribtions

- Hyperellipsoids
- Axes parallel to coordinate axes.
- Note the effect of priors.

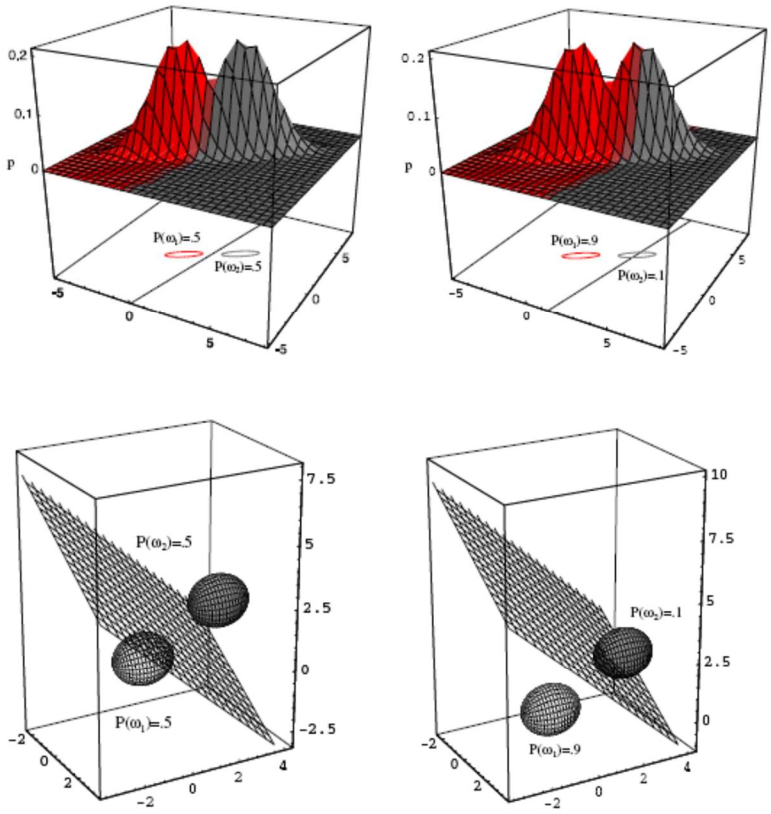


Figure 2.12: Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means.

But, no simple decision regions for Gaussians with unequal variance

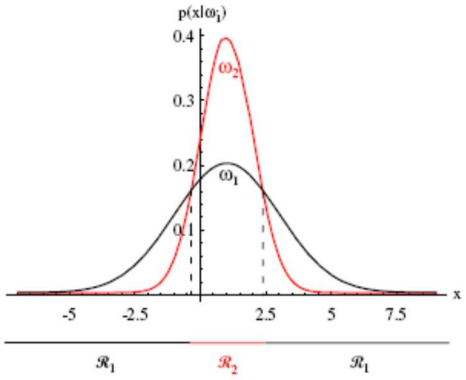


Figure 2.13: Non-simply connected decision regions can arise in one dimensions for Gaussians having unequal variance.

Case 3: If ∑i=general or arbitrary

If the covariance matrices are different for each category, the resulting discriminant functions are inherently quadratic

$$g_i(\underline{x}) = -\underline{x}^T \frac{1}{2}\Sigma_i^{-1}\underline{x} + \Sigma_i^{-1}\underline{m}_i\underline{x} + w$$

$$w = -\frac{1}{2}\underline{m}_i^T\Sigma_i^{-1}\underline{m}_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(S_i)$$

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & & \\ \sigma_{21}^2 & \sigma_{22}^2 & & \\ & & \cdots & \\ & & & \sigma_{NN}^2 \end{bmatrix}$$

$d_M^2(\underline{x},\underline{m}) = (\underline{x}-\underline{m})^T\Sigma^{-1}(\underline{x}-\underline{m})$

2-D Case

$$d_M^2 = \frac{(x_1 - m_1)^2}{\sigma_{11}'^2} + \frac{(x_2 - m_2)^2}{\sigma_{22}'^2}$$

Arbitrary Gaussian distributions with decision boundaries

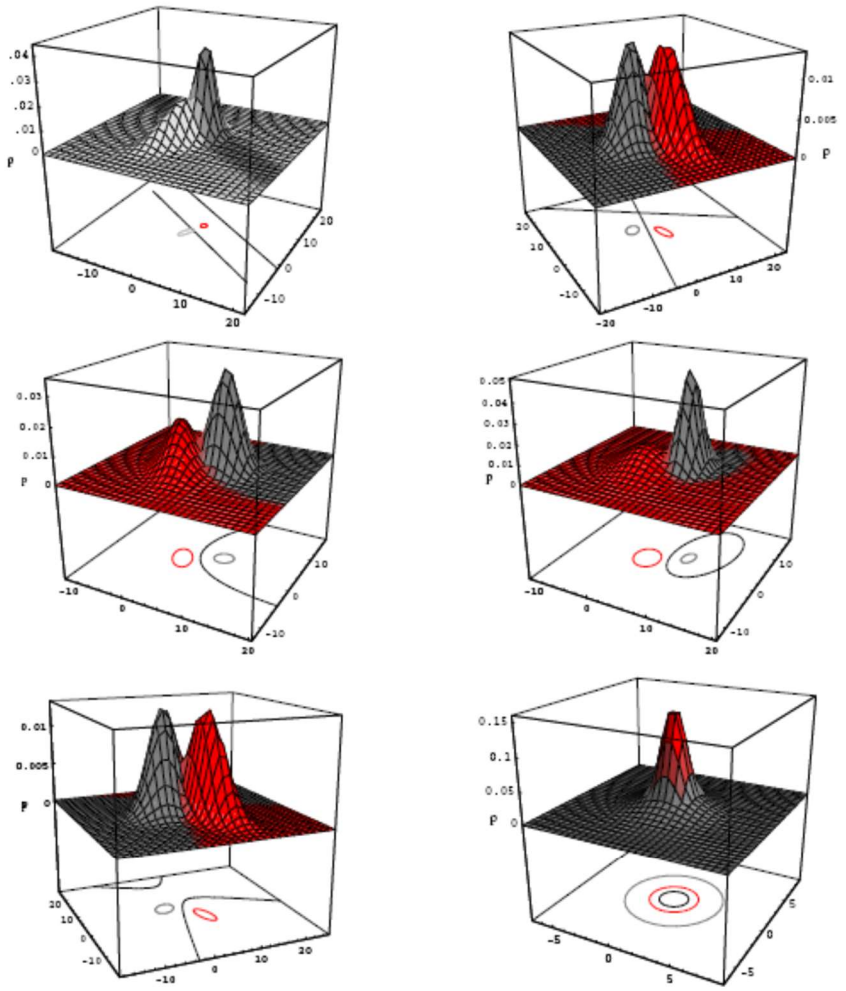


Figure 2.14: Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadratic, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric.

Arbitrary 3-D Gaussian distributions with decision boundaries

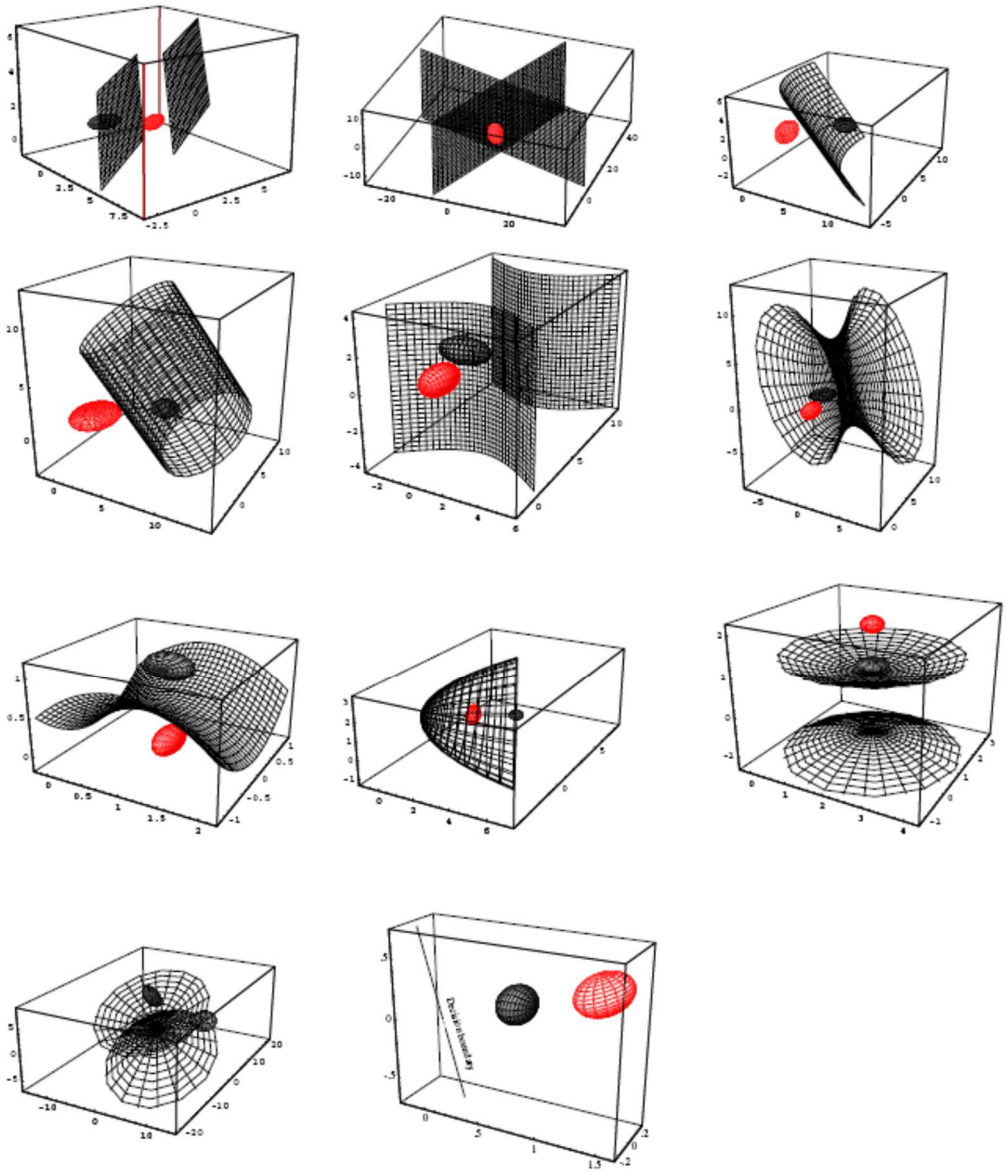


Figure 2.15: Arbitrary three-dimensional Gaussian distributions yield Bayes decision boundaries that are two-dimensional hyperquadrics. There are even degenerate cases in which the decision boundary is a line.

Decision regions for four normal distributions

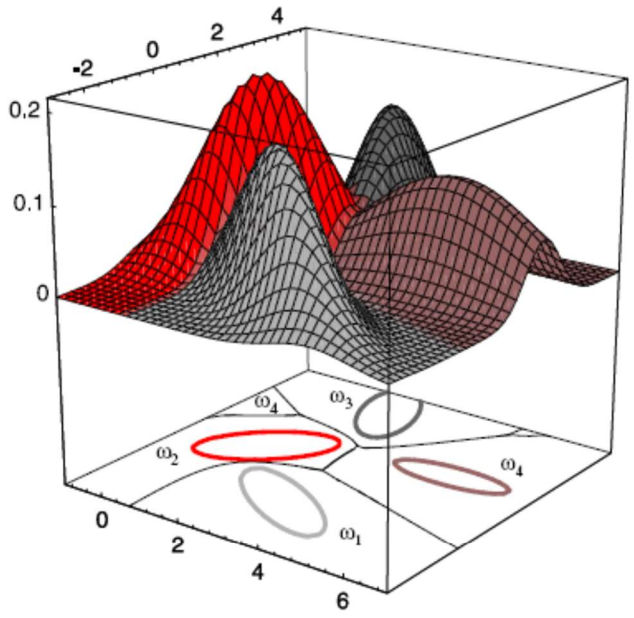


Figure 2.16: The decision regions for four normal distributions. Even with such a low number of categories, the shapes of the boundary regions can be rather complex.