Bayesian Decision Theory

Slides compiled by Sanghamitra De

Introduction

- Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification.
- This approach is based on quantifying the tradeoffs between various classification decisions using probability and the costs that accompany such decisions.
- It makes the assumption that the decision problem is posed in probabilistic terms, and that all of the relevant probability values are known.

A **posterior probability** is the probability of assigning observations to groups given the data. A **prior probability** is the probability that an observation will fall into a group before you collect the data.

Introduction

- ➢ The sea bass/salmon example
  - – State of nature, prior/a priori probability P
    - ✓ State of nature is a random variable ω
    - ✓ The catch of salmon and sea bass is equiprobable
      - – $P(\omega_1) = P(\omega_2)$ (uniform priors)
      - – $P(\omega_1) + P(\omega_2) = 1$ (exclusivity and exhaustivity)

Introduction

- Decision rule with only the prior information
  - Decide ω_1 if $P(\omega_1) > P(\omega_2)$ otherwise decide ω_2

- Use of the class –conditional information

- $P(x \mid \omega_1)$ and $P(x \mid \omega_2)$ describe the difference in lightness between populations of sea bass and salmon

x is a continuous random variable whose distribution depends on the state of nature and is expressed as $p(x \mid \omega)$

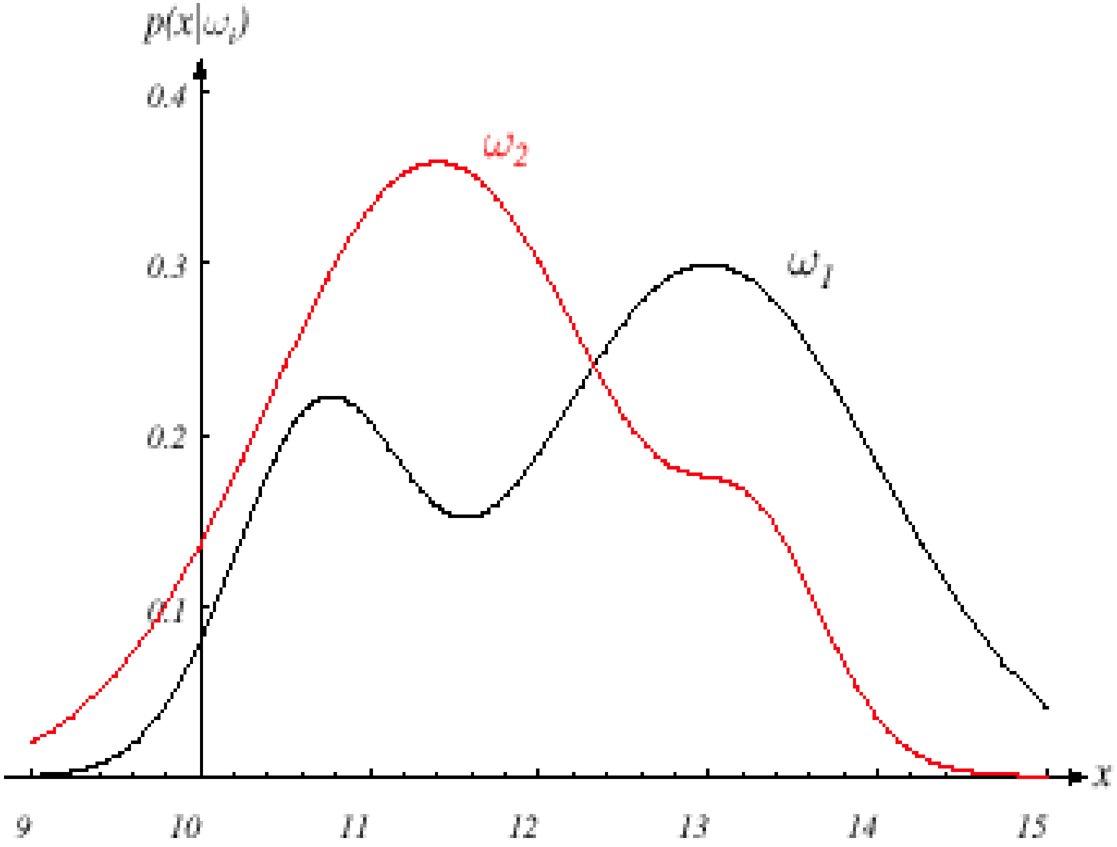


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two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

Introduction

- Posterior, likelihood, evidence
  - $P(\omega_j \mid x) = p(x \mid \omega_j)P(\omega_j) / p(x)$ (Bayes formula)
  - Where in case of two categories

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

  - Posterior = (Likelihood * Prior) / Evidence

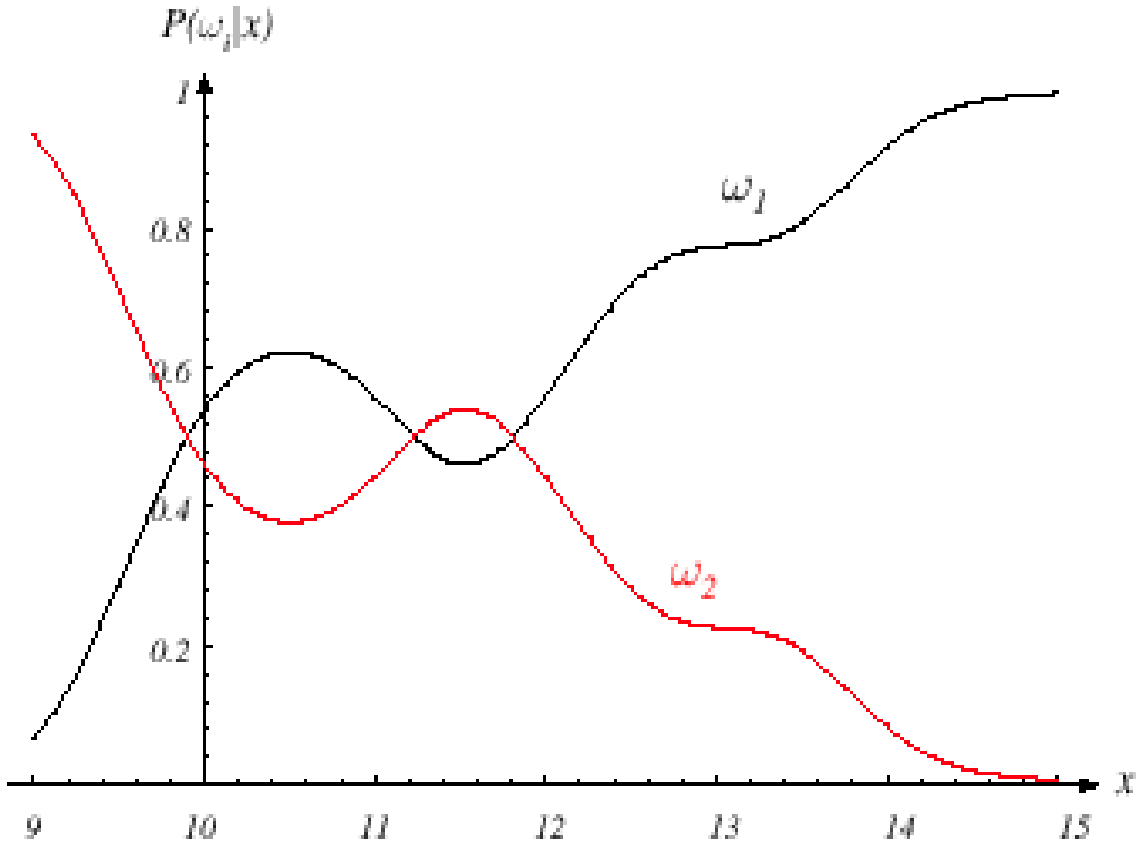


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly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

Introduction

- Decision given the posterior probabilities

  X is an observation for which:

  if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ ⇒ True state of nature = ω_1
  if $P(\omega_1 \mid x) < P(\omega_2 \mid x)$ ⇒ True state of nature = ω_2

  Therefore:

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

  $$P(error \mid x) = P(\omega_1 \mid x) \text{ if we decide } \omega_2$$
  $$P(error \mid x) = P(\omega_2 \mid x) \text{ if we decide } \omega_1$$

Introduction

- Minimizing the probability of error :

  Decide ω_1 *if* $P(\omega_1 \mid x) > P(\omega_2 \mid x);$
  otherwise decide ω_2

  Therefore:

$$P(error \mid x) = min\ [P(\omega_1 \mid x), P(\omega_2 \mid x)]$$

(Bayes decision)

Bayesian Decision Theory – Continuous Features

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

Bayesian Decision Theory – Continuous Features

- Allowing actions other than classification primarily allows the possibility of rejection
  - ✓ Rejection in the sense of abstention
  - ✓ Don’t make a decision if the alternatives are too close
  - ✓ This must be tempered by the cost of indecision

- The loss function states how costly each action taken is

Bayesian Decision Theory – Continuous Features

- Let $\{\omega_1, \omega_2, \ldots, \omega_c\}$ be the set of c states of nature (or "categories")

- Let $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ be the set of possible actions

- Let $\lambda(\alpha_i \mid \omega_j)$ be the loss incurred for taking

  action α_i when the state of nature is ω_j

Bayesian Decision Theory – Continuous Features

An expected loss is called a risk and $R(\alpha_i \mid \mathbf{x})$ is called *conditional risk*

Overall risk:

$R = \textit{Sum of all } \underbrace{R(\alpha_i \mid \mathbf{x}) \textit{ for}}_{\textbf{Conditional risk}}\, i = 1,\ldots,a$ ***and all x***

Minimizing R Minimizing $R(\alpha_i \mid x)$ *for* $i = 1,\ldots, a$

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

for each action α_i (i = 1,…,a)

Note: This is the risk specifically for observation *x*

Bayesian Decision Theory – Continuous Features

Select the action α_i for which $R(\alpha_i \mid \boldsymbol{x})$ is minimum

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

Bayesian Decision Theory – Continuous Features

- Two-category classification

$$\alpha_1 : \textit{deciding } \omega_1$$
$$\alpha_2 : \textit{deciding } \omega_2$$

$$\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$$

loss incurred for deciding ω_i when the true state of nature is ω_j

Conditional risk:

$$R(\alpha_1 \mid \mathbf{x}) = \lambda_{11} P(\omega_1 \mid \mathbf{x}) + \lambda_{12} P(\omega_2 \mid \mathbf{x})$$
$$R(\alpha_2 \mid \mathbf{x}) = \lambda_{21} P(\omega_1 \mid \mathbf{x}) + \lambda_{22} P(\omega_2 \mid \mathbf{x})$$

Bayesian Decision Theory – Continuous Features

Writing out the conditional risk we get:

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

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

The minimum-risk decision rule is: decide ω_1 if

$$R(\alpha_1 \mid \mathbf{x}) < R(\alpha_2 \mid \mathbf{x})$$

In terms of posterior probability decide ω_1 if:

$$\lambda_{11} P(\omega_1 \mid \mathbf{x}) + \lambda_{12} P(\omega_2 \mid \mathbf{x}) < \lambda_{21} P(\omega_1 \mid \mathbf{x}) + \lambda_{22} P(\omega_2 \mid \mathbf{x})$$

and decide ω_2 otherwise

Bayesian Decision Theory – Continuous Features

We can rewrite

$$\lambda_{11} P(\omega_1 \mid \mathbf{x}) + \lambda_{12} P(\omega_2 \mid \mathbf{x}) <$$
$$\lambda_{21} P(\omega_1 \mid \mathbf{x}) + \lambda_{22} P(\omega_2 \mid \mathbf{x})$$

As

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

Bayesian Decision Theory – Continuous Features

Using Bayes formula to replace posterior probabilities by prior probabilities and the conditional densities we can decide ω_1 if:

$$(\lambda_{21} - \lambda_{11})\, p(\mathbf{x} \mid \omega_1)\, P(\omega_1) > (\lambda_{12} - \lambda_{22})\, p(\mathbf{x} \mid \omega_2)\, P(\omega_2)$$

and decide ω_2 otherwise

Bayesian Decision Theory – Continuous Features

If $\lambda_{21} > \lambda_{11}$ then we can express our rule as a Likelihood ratio:

The preceding rule is equivalent to the following rule:

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

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

Classifiers, discriminant functions and decision surfaces

- One of the useful ways to represent pattern classifiers is in terms of discriminant functions $g_i(\mathbf{x})$, i=1, …,c.
- The classifier assigns a feature vector **x** to class ω_i if :

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

- The classifier can be viewed as a network or machine that computes c discriminant functions and selects the category corresponding to the largest discriminant. (see fig. next slide)
- For the Bayes classifier (the general case with risks), we can let $g_i(\mathbf{x}) \equiv - R(\alpha_i \mid \boldsymbol{x})$ because the maximum discriminant function will then correspond to the minimum conditional risk.
- For minimum error-rate we can take $g_i(\mathbf{x}) = P(\omega_i \mid \boldsymbol{x})$, so that maximum discriminant function corresponds to maximum posterior probability.
- In general, if we replace every $g_i(\mathbf{x})$ by $f(g_i(\mathbf{x}))$, where f(.) is a monotonically increasing function, then the resulting classification stays unchanged.
- The effect of any decision rule is to divide the feature space into *c decision regions*, R_1, …, R_c.
- If $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for all $j \neq i$, then x is in R_1, and the decision rule calls for us to assign **x** to ω_i
- The regions are separated by *decision boundaries,* surfaces in feature space where ties occur among the largest discriminant functions.

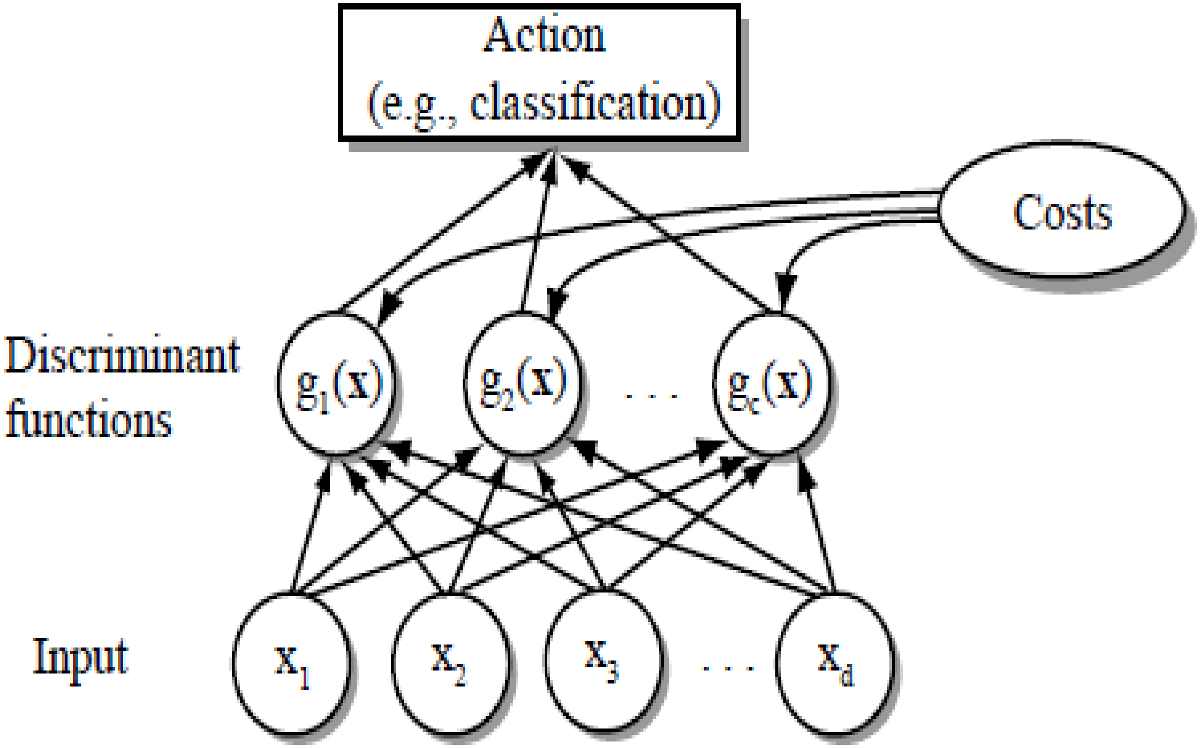
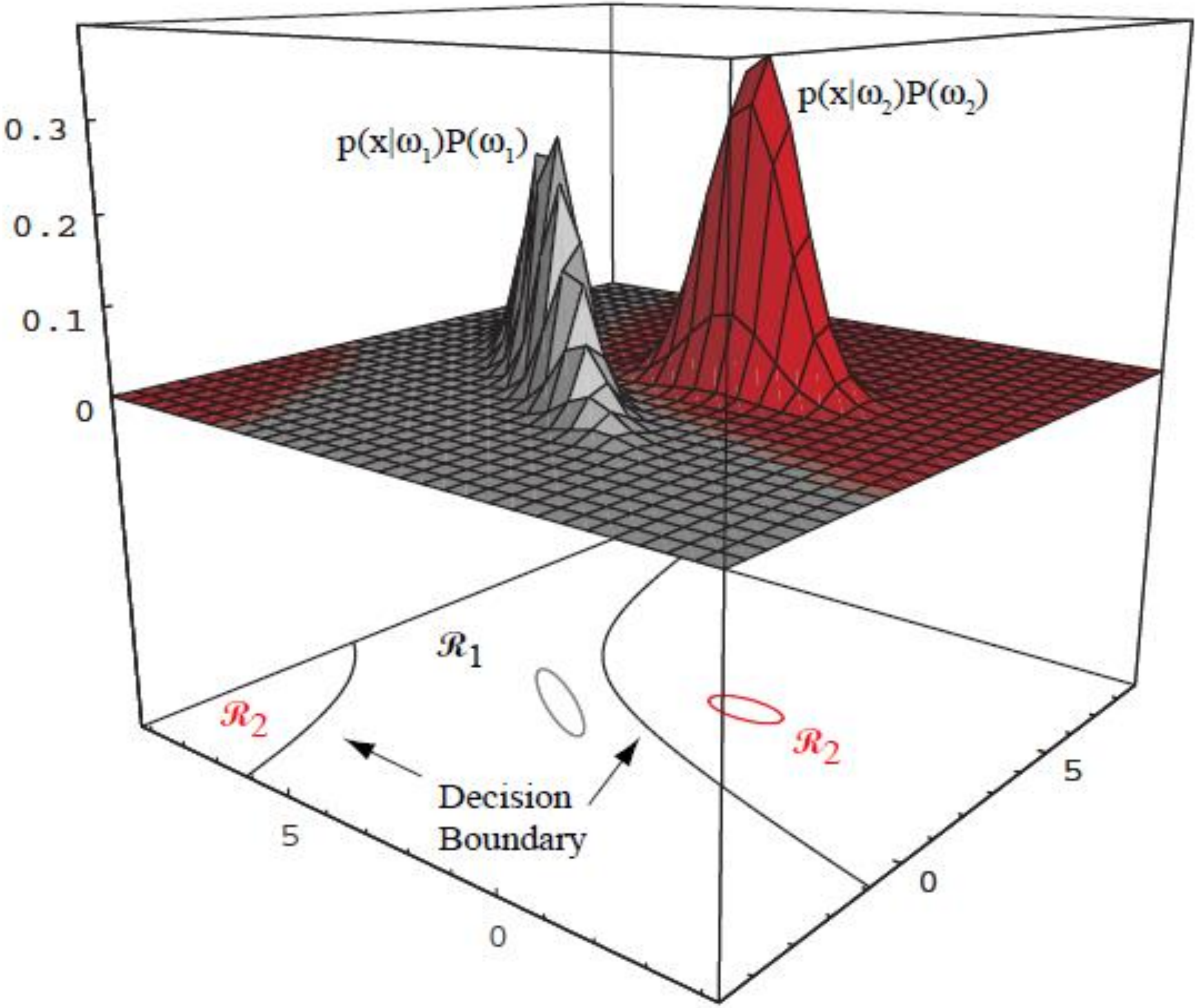


Figure 2.5: The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathrm{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident.



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

The Two-Category case

- A classifier that places a pattern in one of only two categories is called a *dichotomizer*.
- Instead of using two discriminant functions g_1 and g_2 and assigning x to ω_1 if $g_1>g_2$, it is more common to define a single discriminant function.

$$g(x) \equiv g_1(x) – g_2(x)$$

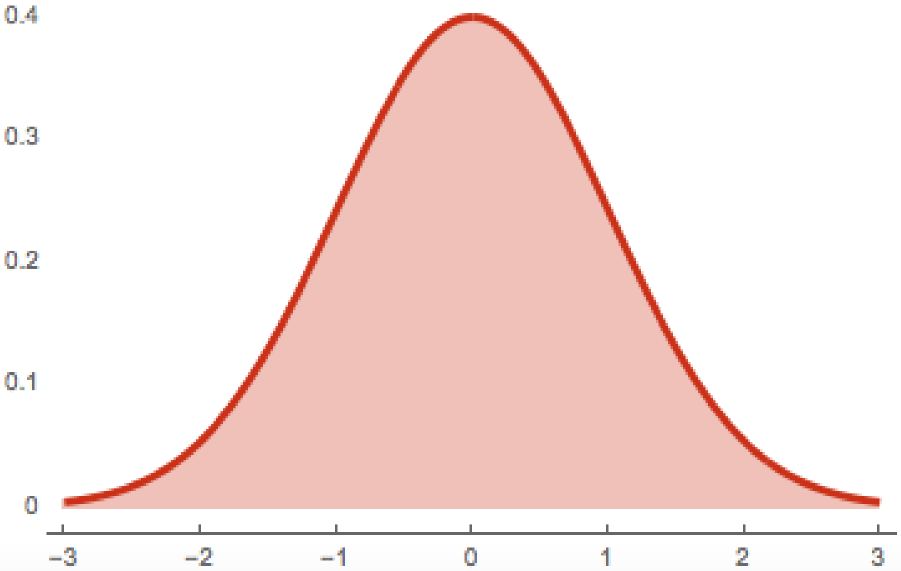
- Also we use the decision rule: Decide ω_1 if $g(x) > 0$; otherwise decide ω_2

The Two-Category case

- So, a dichotomizer can be viewed as a machine that computes a single discriminant function g(x) and classifies x according to the algebraic sign of the result.
- g(x) can also be represented as:

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} + \ln \frac{p(\omega_1)}{p(\omega_2)}$$

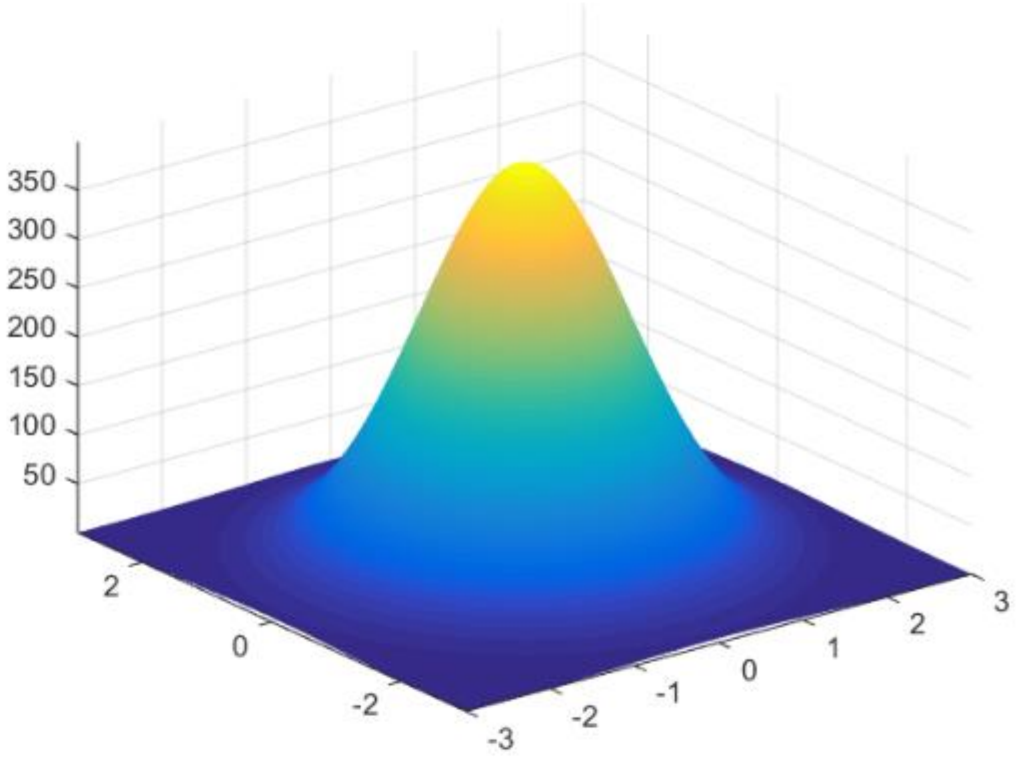
Univariate Normal Density



The graph of a normal distribution with mean of 0 *and standard deviation of* 1

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} .$$

Multivariate Normal Distribution



A normal (or Gaussian) distribution in 2 variables

$$p(\mathbf{x};\mu,\Sigma)=\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

Discriminant Functions for the Normal Density

- Minimum error-rate classification can also be achieved by the discriminant function:

  $$g_i(x) = \ln p(x \mid \omega_i) + \ln P(\omega_i)$$

- If the densities $p(x \mid \omega_i)$ are multivariate normal:

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

Multivariate normal distribution, **multivariate Gaussian distribution**, or **joint normal distribution** is a generalization of the one-dimensional (univariate) normal distribution to higher dimensions. One definition is that a random vector is said to be k-variate normally distributed if every linear combination of its k components has a univariate normal distribution. The multivariate normal distribution is often used to describe, at least approximately, any set of (possibly) correlated real-valued random variables each of which clusters around a mean value.

Case 1: $\Sigma_i = \sigma^2 I$ (I stands for the identity matrix)

- $\Sigma_i = \sigma^2 I$ simplest case when the features are statistically independent & when each feature has the same variance, σ^2
- In this case, the covariance matrix is diagonal, being σ^2 times the identity matrix I
- This corresponds to the situation where the samples fall in equal-size hyperspherical clusters, the cluster for the ith class being centered about the mean vector μ_i

Note : both $|\Sigma_i|$ and (d/2) ln π are independent of i in

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

Thus we can simplify to :

$$g_i(x) = -\frac{\|x-\mu_i\|^2}{2\sigma^2} + \ln P(\omega_i)$$

where $\|\cdot\|$ denotes the Euclidean norm

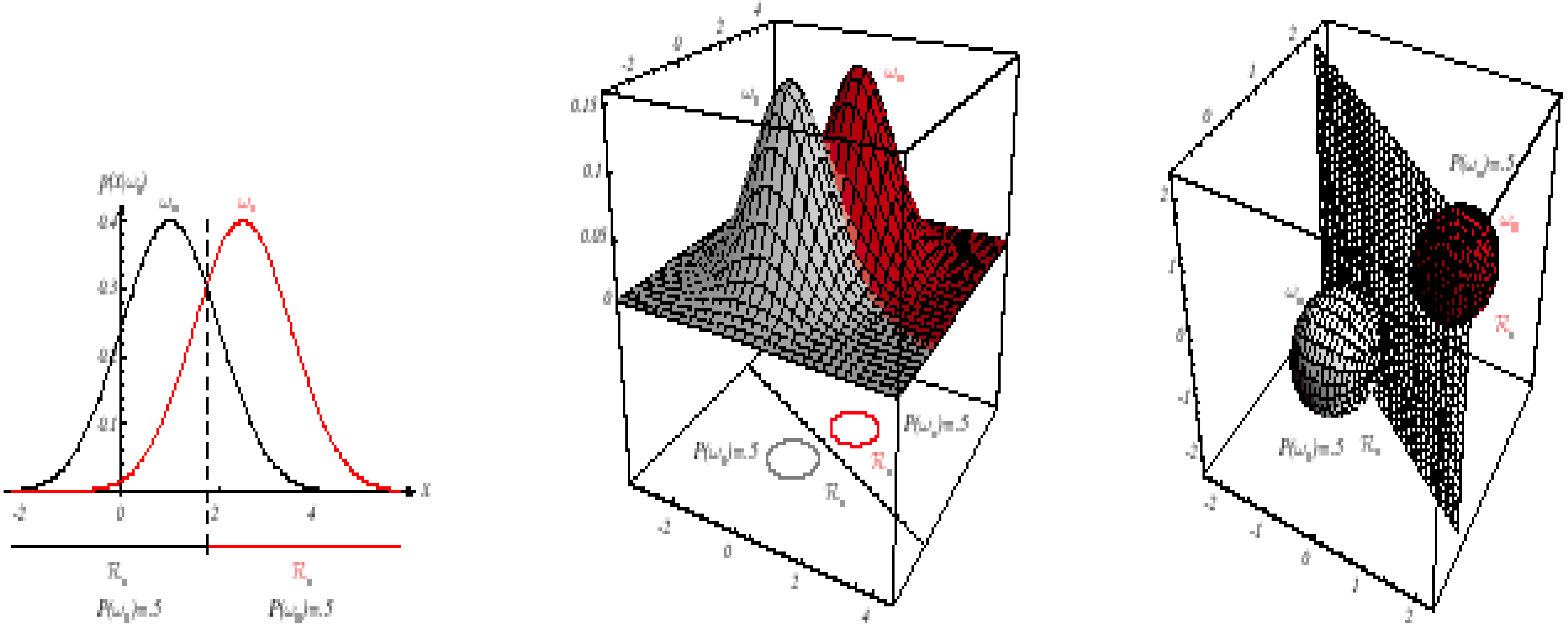


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-,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

- We can further simplify by recognizing that the quadratic term ***x*^t*x*** implicit in the Euclidean norm is the same for all i, making it an ignorable additive constant.
- So we obtain the equivalent linear discriminant function:

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

where:

$$\mathbf{w}_i = \frac{\boldsymbol{\mu}_i}{\sigma^2}; \; w_{i0} = -\frac{1}{2\sigma^2} \boldsymbol{\mu}_i^t \boldsymbol{\mu}_i + \ln P(\omega_i)$$

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

- A classifier that uses linear discriminant functions is called "a linear machine"

- The decision surfaces for a linear machine are pieces of hyperplanes defined by:

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

The equation can be written as:

$$\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0)=0 \quad \text{where } w = \mu_i - \mu_j$$

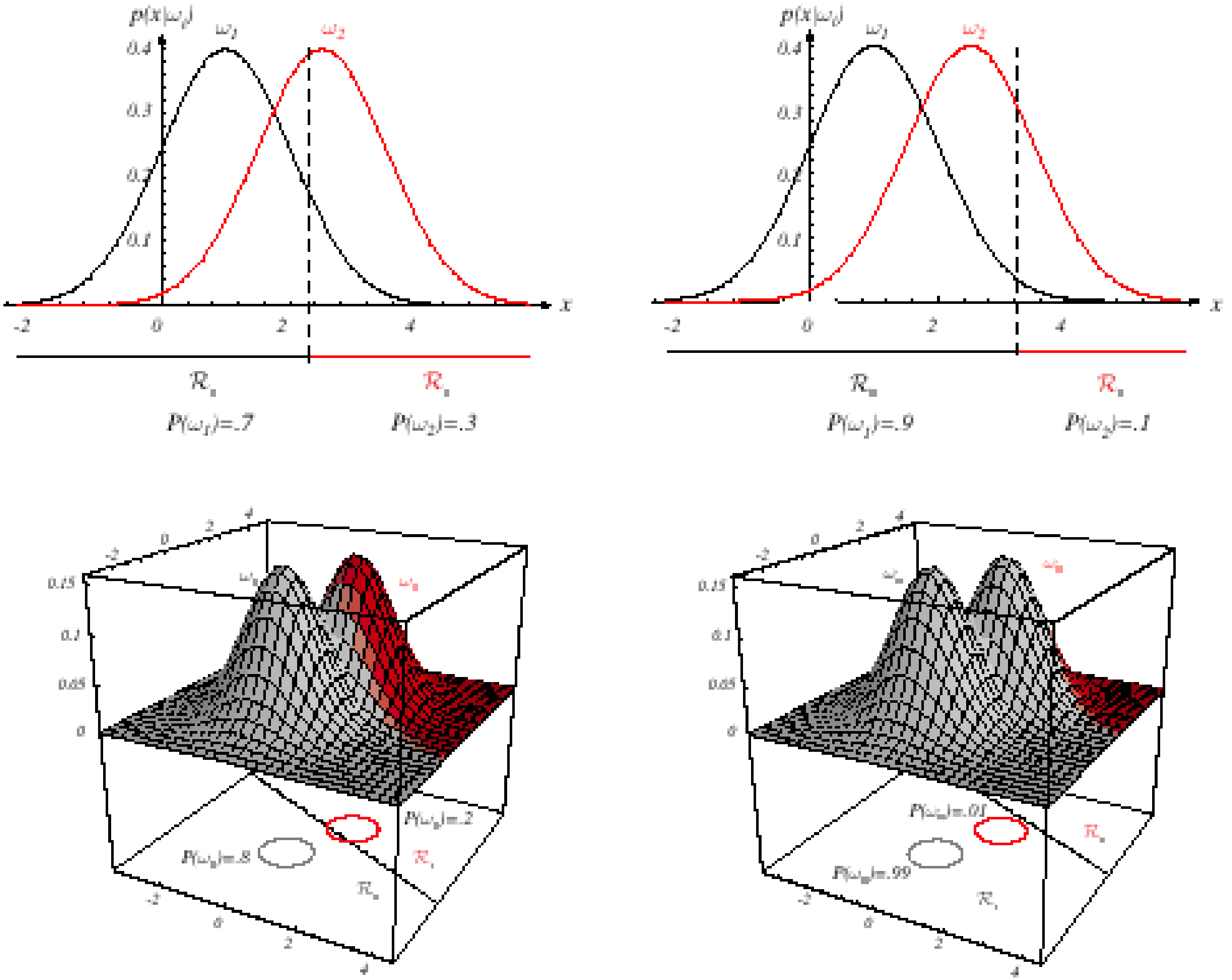
– The hyperplane separating R_i and R_j

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

The hyperplane is through the point $\mathbf{x}_0$ and always orthogonal to the line linking the means!

$$if \quad P(\omega_i) = P(\omega_j) \ \text{ then } \ \mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)$$

That is, the second term on the r.h.s of the above eqn. vanishes.


$p(x|\omega_i)$
ω_1
ω_2
0.4
0.3
0.2
0.1
-2
0
2
4
x
$\mathcal{R}_1$
$\mathcal{R}_2$
$P(\omega_1)=.7$
$P(\omega_2)=.3$
$p(x|\omega_i)$
ω_1
ω_2
0.4
0.3
0.2
0.1
-2
0
2
4
x
$\mathcal{R}_1$
$\mathcal{R}_2$
$P(\omega_1)=.9$
$P(\omega_2)=.1$
0.15
0.1
0.05
0
-2
0
2
4
ω_1
ω_2
$P(\omega_2)=.2$
$P(\omega_1)=.8$
$\mathcal{R}_1$
$\mathcal{R}_2$
0.15
0.1
0.05
0
-2
0
2
4
ω_1
ω_2
$P(\omega_2)=.01$
$P(\omega_1)=.99$
$\mathcal{R}_1$
$\mathcal{R}_2$

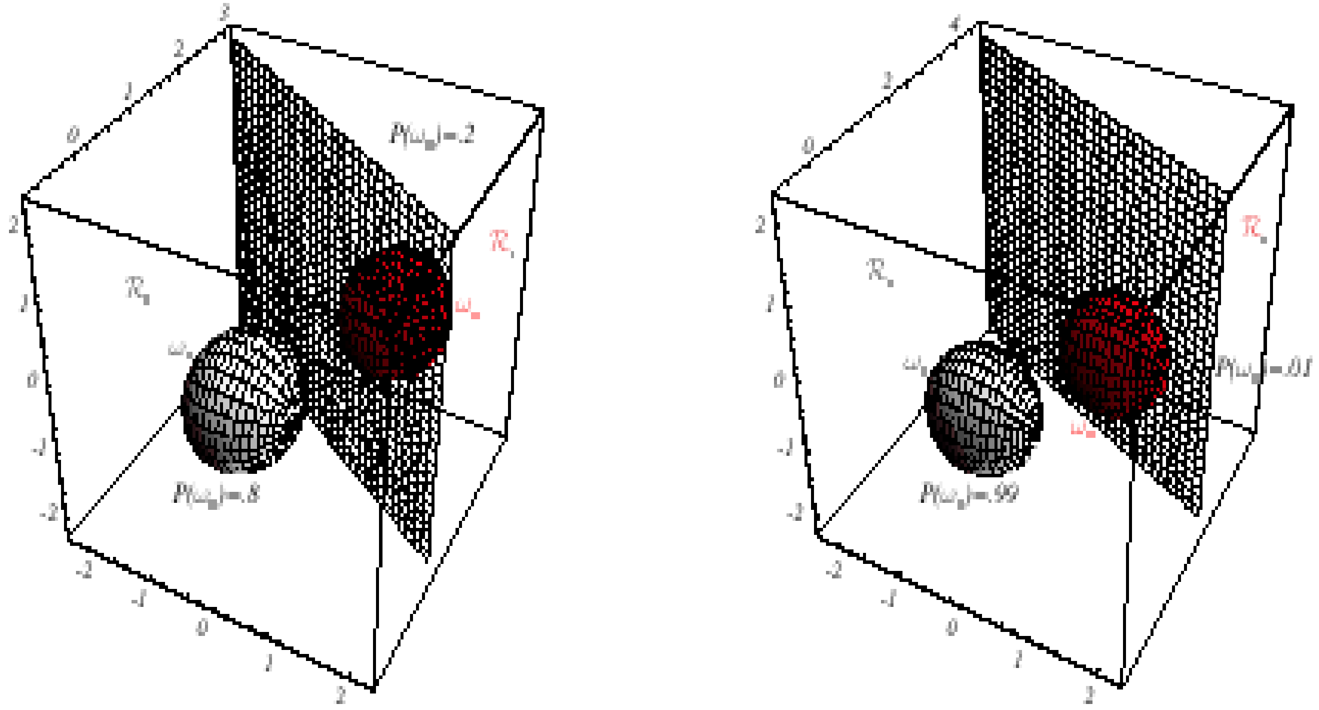


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

Case 2: $\Sigma_i = \Sigma$ (covariance of all classes are identical but arbitrary!)

- Hyperplane separating R_i and R_j
- Corresponds to the situation where all samples fall in hyperellipsoidal clusters of equal size and shape and the cluster for the ith class is centered about the mean vector μ_i.
- Decision boundaries are hyperplanes (since the discriminant function is linear). If regions are contiguous, the boundary separating them has the following equation:

$$\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0)=0$$

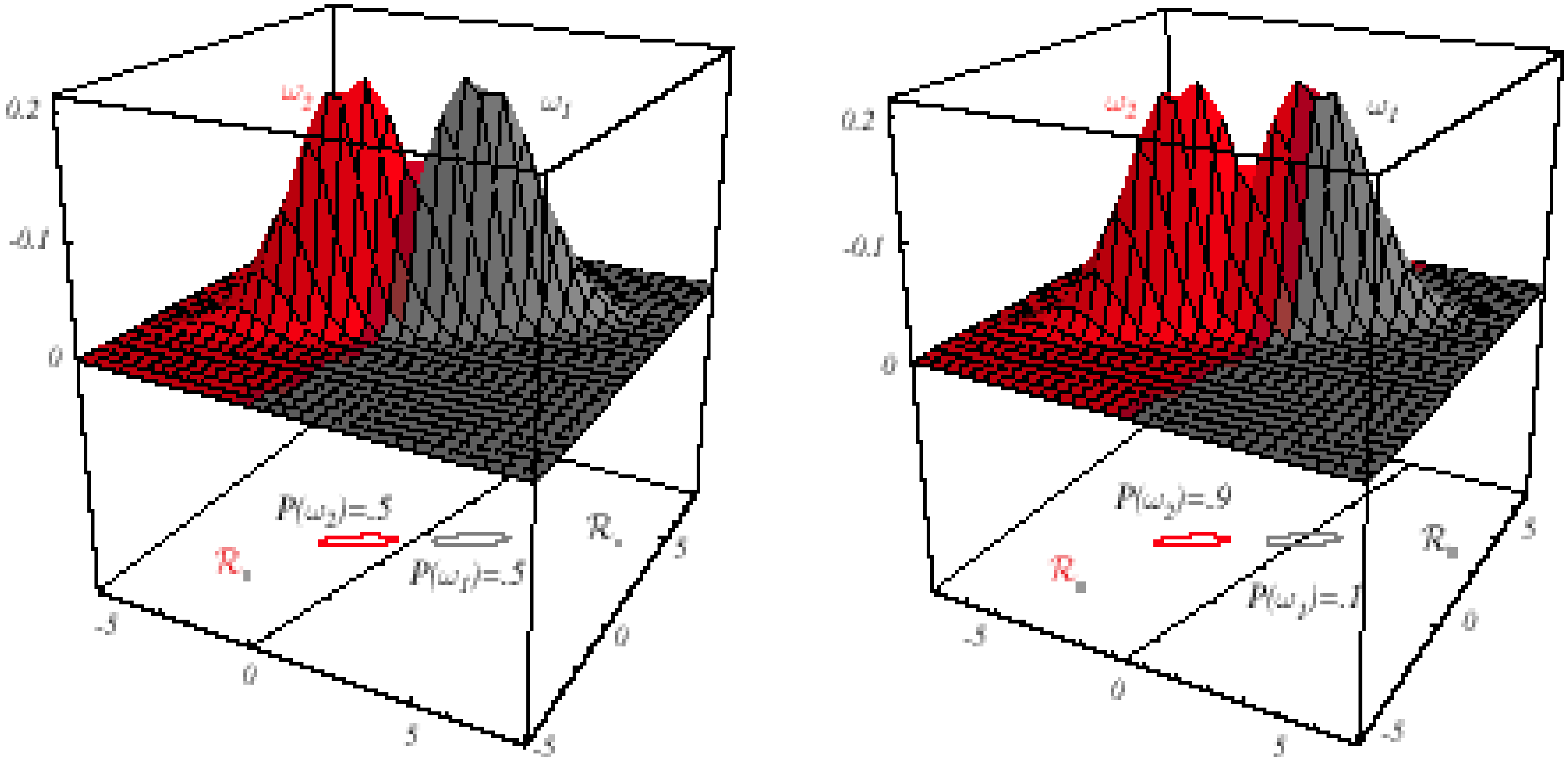
Where

$$\mathbf{w}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$$

and

$$\mathbf{x}_0=\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}.(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$$

(the hyperplane separating R_i and R_j is generally not orthogonal to the line between the means!)


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

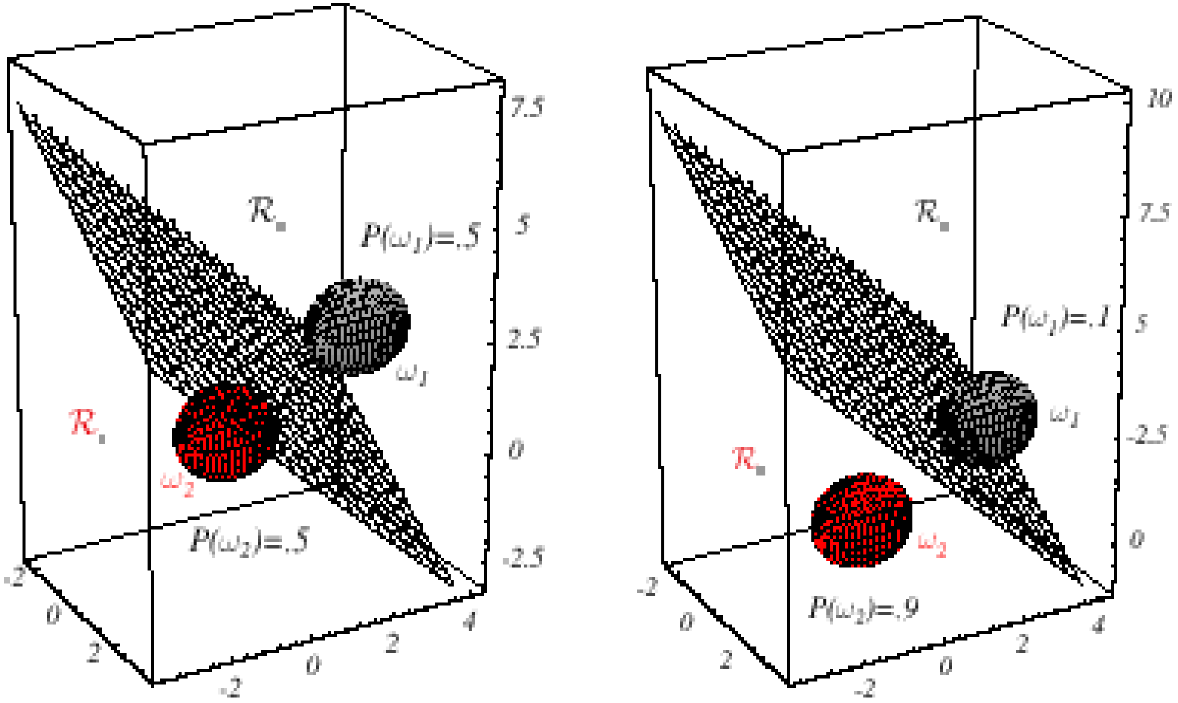


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda,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

- Case Σ_i = arbitrary
  - The covariance matrices are different for each category

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

$$where:$$

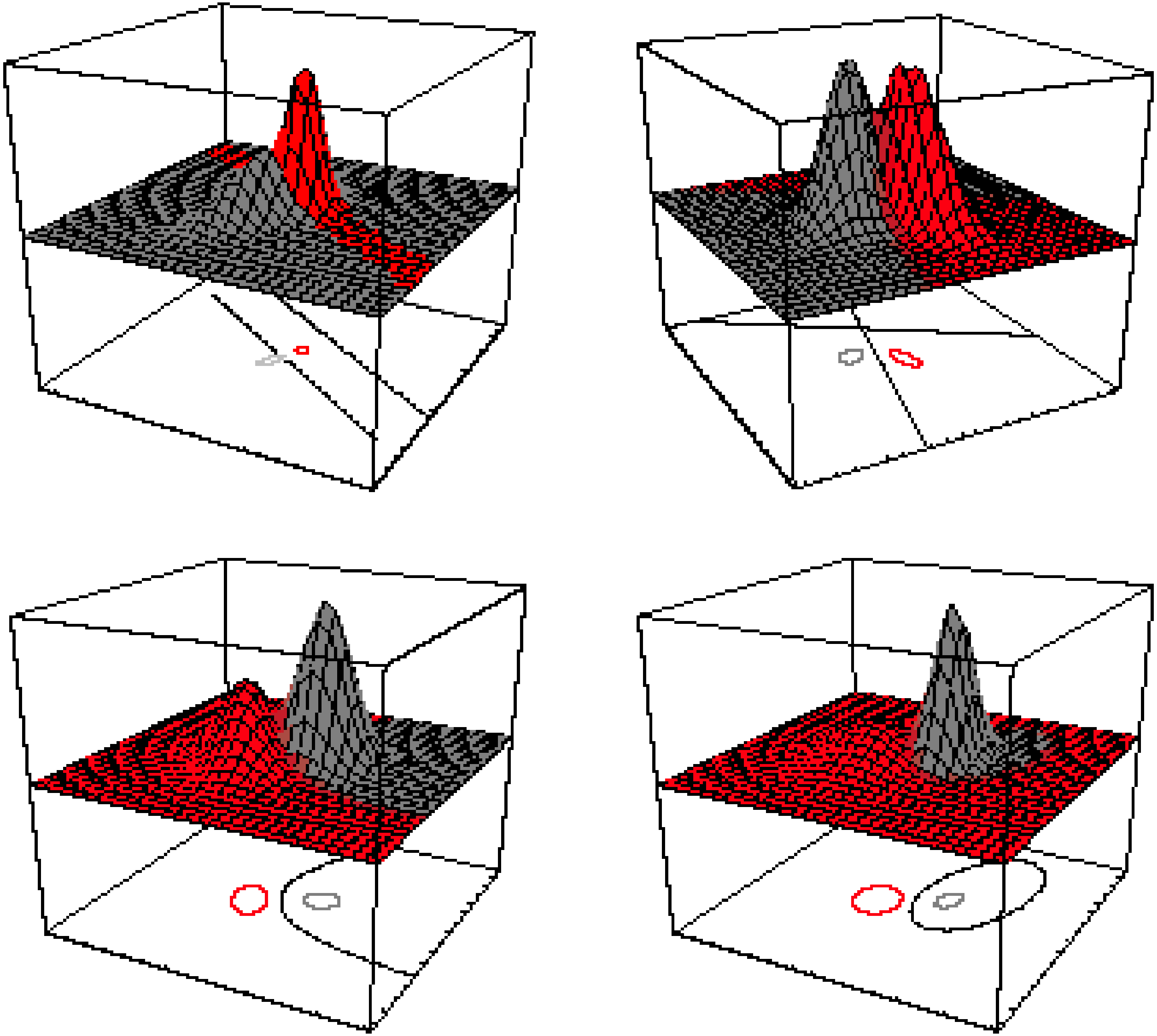
$$W_i = -\frac{1}{2}\Sigma_i^{-1}$$

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

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

The decision surfaces are hyperquadratics

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



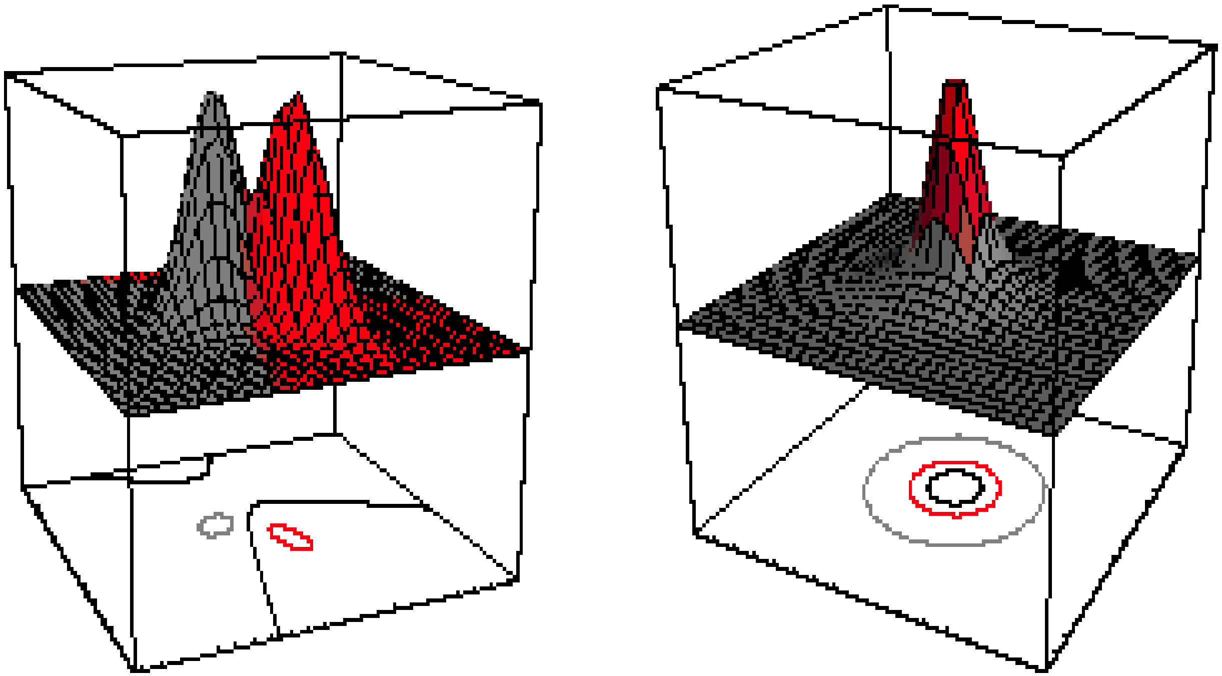


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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

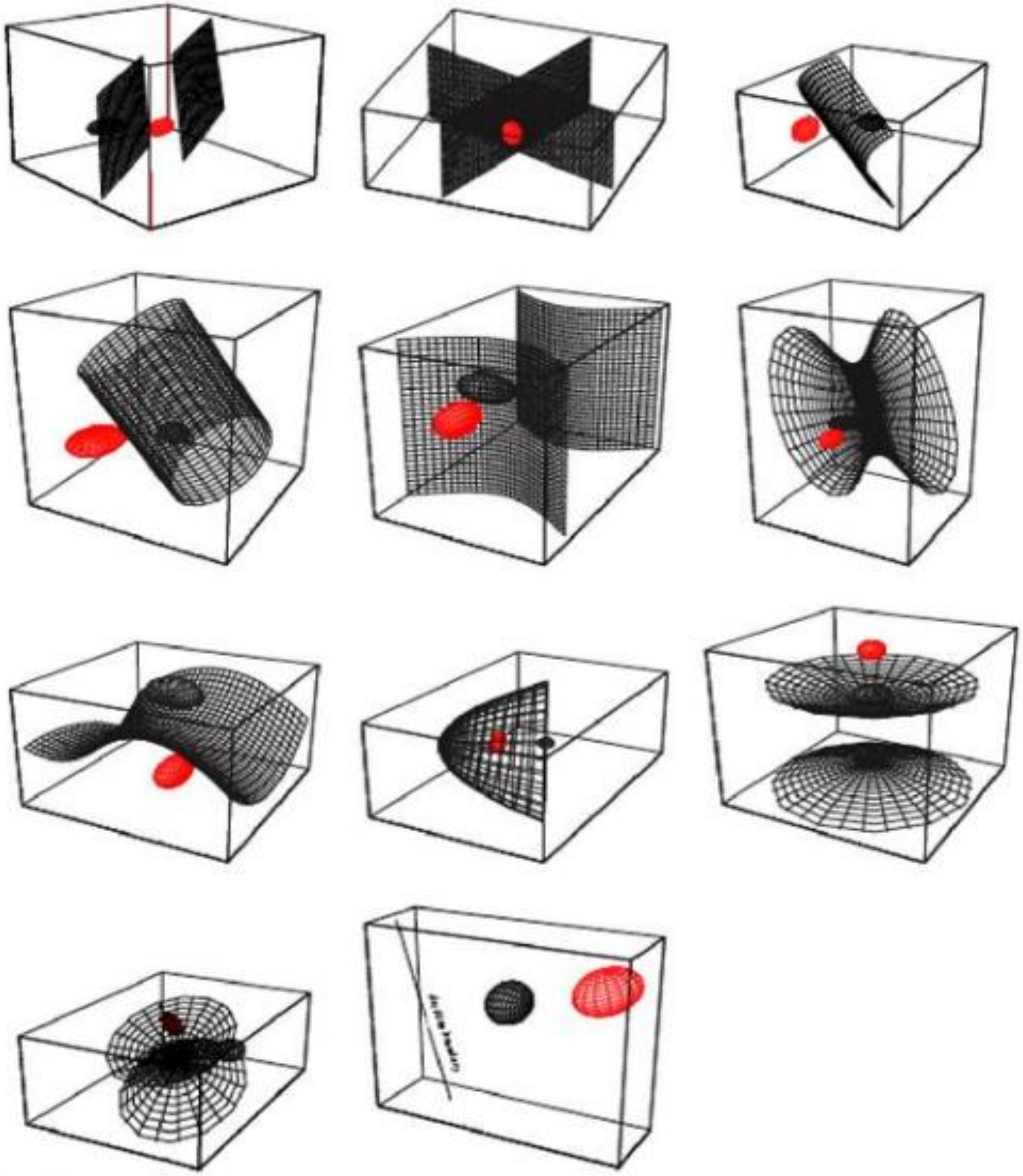


FIGURE 2.15. Arbitrary three-dimensional Gaussian distributions yield Bayes decision boundaries that are two-dimensional hyperquadrics. There are even degenerate cases in which the decision boundary is a line. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright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

Bayes Decision Theory – Discrete Features

- Components of feature vector **x** are binary, ternary or higher integer valued, so **x** can take only one of m discrete values

$$v_1,\ v_2,\ \ldots,\ v_m$$

Bayes formula then involves probabilities rather than probability densities:

$$P(\omega_j \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \omega_j)P(\omega_j)}{P(\mathbf{x})}$$

where

$$P(\mathbf{x}) = \sum_{j=1}^{c} P(\mathbf{x} \mid \omega_j)P(\omega_j)$$

Bayes Decision Theory – Discrete Features

- Conditional risk is defined as before: R(α|**x**)

- Approach is still to minimize risk:

$$\alpha^* = \arg\min_i R(\alpha_i \mid \mathbf{x})$$

Bayes Decision Theory – Discrete Features

- Case of independent binary features in 2 category problem (i.e. where the components of the feature vector are binary-valued and conditionally independent.)

  Let $\mathbf{x} = [x_1, x_2, \ldots, x_d]^t$ where each x_i is either 0 or 1, with probabilities:

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

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

Bayes Decision Theory – Discrete Features

- Assuming conditional independence, P(**x**|ω$_i$) can be written as a product of component probabilities:

$$P(\mathbf{x}\,|\,\omega_1) = \prod_{i=1}^{d} p_i^{x_i}(1-p_i)^{1-x_i}$$

and

$$P(\mathbf{x}\,|\,\omega_2) = \prod_{i=1}^{d} q_i^{x_i}(1-q_i)^{1-x_i}$$

yielding a likelihood ratio given by :

$$\frac{P(\mathbf{x}\,|\,\omega_1)}{P(\mathbf{x}\,|\,\omega_2)} = \prod_{i=1}^{d}\left(\frac{p_i}{q_i}\right)^{x_i}\left(\frac{1-p_i}{1-q_i}\right)^{1-x_i}$$

Bayes Decision Theory – Discrete Features

- Taking our likelihood ratio

$$\frac{P(\mathbf{x}\,|\,\omega_1)}{P(\mathbf{x}\,|\,\omega_2)} = \prod_{i=1}^{d}\left(\frac{p_i}{q_i}\right)^{x_i}\left(\frac{1-p_i}{1-q_i}\right)^{1-x_i}$$

and plugging it into Eq. 31

$$g(\mathbf{x}) = \ln\frac{p(\mathbf{x}\,|\,\omega_1)}{p(\mathbf{x}\,|\,\omega_2)} + \ln\frac{p(\omega_1)}{p(\omega_2)}$$

yields :

$$g(\mathbf{x}) = \sum_{i=1}^{d}\left[x_i \ln\frac{p_i}{q_i} + (1-x_i)\ln\frac{1-p_i}{1-q_i}\right] + \ln\frac{P(\omega_1)}{P(\omega_2)}$$

- The discriminant function in this case is:

$$g(\mathbf{x}) = \sum_{i=1}^{d} \left[w_i x_i \right] + w_0$$

$$where :$$

$$w_{\text{i}} = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad i = 1, \ldots, d$$

$$and :$$

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

$$decide\ \omega_1 \text{ if } \text{g(x)} > 0 \text{ and } \omega_2 \text{ if } \text{g(x)} \leq 0$$

References

- https://brilliant.org/wiki/multivariate-normal-distribution/
- https://brilliant.org/wiki/normal-distribution/#:~:text=The%20normal%20distribution%2C%20also%20called,and%20test%20scores.

End.