Chapter 4 Bayesian Decision Theory

4.1 Introduction

Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification. It is considered the ideal case in which the probability structure underlying the categories is known perfectly. While this sort of stiuation rarely occurs in practice, it permits us to determine the optimal (Bayes) classifier against which we can compare all other classifiers. Moreover, in some problems it enables us to predict the error we will get when we generalize to novel patterns.

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.

Let us reconsider the hypothetical problem posed in Chapter 1 of designing a classifier to separate two kinds of fish: sea bass and salmon. Suppose that an observer watching fish arrive along the conveyor belt finds it hard to predict what type will emerge next and that the sequence of types of fish appears to be random. In decision-theoretic terminology we would say that as each fish emerges nature is in one or the other of the two possible states: Either the fish is a sea bass or the fish is a salmon. We let w denote the *state of nature,* with $w = w_1$ for sea bass and $w = w_2$ for salmon. Because the state of nature is so unpredictable, we consider w to be a variable that must be described probahilistically.

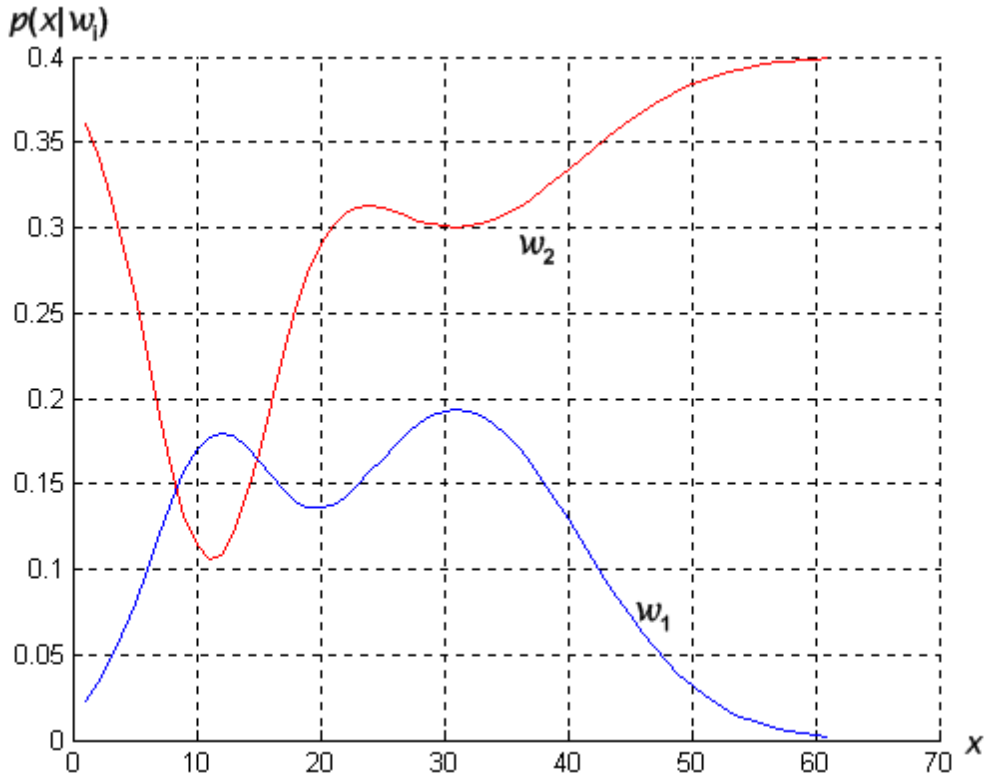


Figure 4.1: Class conditional density functions show the probabiltiy density of measuring a particular feature value x given the pattern is in category w_i.

If the catch produced as much sea bass as salmon, we would say that the next fish is equally likely to be sea bass or salmon. More generally, we assume that there is some **prior probability** $P(w_1)$ that the next fish is sea bass, and some prior probability $P(w_2)$ that it is salmon. If we assume there are no other types of fish relevant here, then $P(w_1)+ P(w_2)=1$. These prior probabilities reflect our prior knowledge of how likely we are to get a sea bass or salmon before the fish actually appears.

If we are forced to make a decision about the type of fish that will appear next just by using the value of the prior probahilities we will decide w_1 if $P(w_1)> P(w_2)$ otherwise decide w_2. This rule makes sense if we are to judge just one fish, but if we were to judge many fish, using this rule repeatedly, we would always make the same decision even though we know that *both* types of fish will appear. Thus, it does not work well depending upon the values of the prior probabilities.

In most circumstances, we are not asked to make decisions with so little information. We might for instance use a lightness measurement x to improve our classifier. Different fish will yield different lightness readings, and we express this variability: we consider x to be a continuous random variable whose distribution depends on the state of nature and is expressed as $p(x|w)$. This is the **class-conditional probability density** (state-conditional probability density) function, the probability density function for x given that the state of nature is in w. Then the difference between $p(x|w_1)$ and $p(x|w_2)$ describes the difference in lightness between populations of sea bass and salmon (Figure 4.1).

Suppose that we know both the prior probabilities $P(w_j)$ and the conditional densities $p(x|w_j)$ for $j = 1, 2$. Suppose further that we measure the lightness of a fish and discover that its value is x. How does this measurement influence our attitude concerning the true state of nature? We note first that the (joint) probability density of finding a pattern that is in category w_j and has feature value x can be written in two ways: $p(w_j,x) = P(w_j|x)\, p(x) = p(x|w_j)\, P(w_j)$. Rearranging these leads us to the answer to our question, which is called **Bayes formula**:

$$P(w_j \mid x) = \frac{p(x \mid w_j) P(w_j)}{p(x)} \tag{4.1}$$

where in this case of two categories

$$p(x) = \sum_{j=1}^{2} p(x \mid w_j) P(w_j) \tag{4.2}$$

Bayes formula can be expressed informally as

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \tag{4.3}$$

Bayes formula shows that by observing the value of x we can convert the prior probability $P(w_j)$ to the *posterior probability* $P(w_j|x)$ -the probability of the state of nature being w_j given that feature value x has been measured. $p(x|w_j)$ is called the **likelihood** of w_j with respect to x, a term chosen to indicate that, other things being equal, the category w_j, for which $p(x|w_j)$ is large is more "likely" to be the true category. Notice that it is the product of the likelihood and the prior probability that is most important in determining the posterior probability; the evidence factor p(x), can be viewed as a scale factor that guarantees that the posterior probabilities sum to one. The variation of *posterior probability* $P(w_j|x)$ with x is illustrated in Figure 4.2 for the case $P(w_1)=2/3$ and $P(w_2)=1/3$.

If we have an observation x for which $P(w_1|x)>P(w_2|x)$, we would naturally be inclined to decide that the true state of nature is w_1. The probability of error is calculated as

$$P(error|x) = \begin{cases} P(w_1|x) & \text{if we decide } w_2 \\ P(w_2|x) & \text{if we decide } w_1 \end{cases} \quad (4.4)$$

The *Bayes decision rule* is stated as

Decide w_1 if $P(w_1|x)>P(w_2|x)$; otherwise decide w_2 (4.5)

Under this rule eq.4.4 becomes

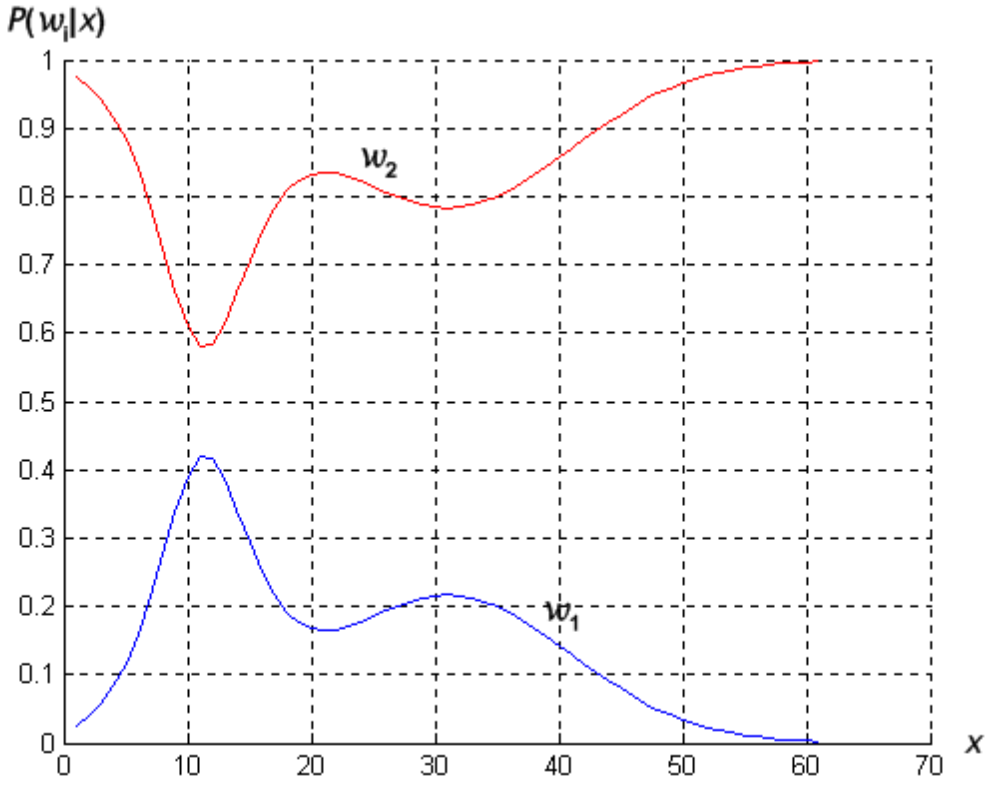


Figure 4.2: Posterior probabilities.

$P(\text{error}|x)=\min[P(w_1|x), P(w_2|x)]$

(4.6)

This form of decision rule emphasizes the role of the posterior probabilities. As being equivalent, the same rule can be expressed in terms of conditional and prior probabilities as:

Decide w_1 if $p(x|w_1)P(w_1) > p(x|w_2)P(w_2)$; otherwise decide w_2 (4.7)

4.2 Bayesian Decision Theory (continuous)

We shall now formalize the ideas just considered, and generalize them in four ways: by allowing the use of more than one feature, by allowing more than two states of nature, by allowing actions other than merely deciding the state of nature, and by introducing a loss function more general than the probability of error.

Allowing the use of more than one feature merely requires replacing the scalar x by the feature vector $\mathbf{x}$, where $\mathbf{x}$ is in a d-dimensional Euclidean space $\mathbf{R}^d$ called the *feature space.* Allowing more than two states of nature provides us with a useful generalization for a small notational expense as $\{w_1 \ldots w_c\}$. Allowing actions other than classification as $\{\alpha_1 \ldots \alpha_a\}$ allows the possibility of rejection-that is, of refusing to make a decision in close (costly) cases. The **loss function** states exactly how costly each action is, and is used to convert a probability determination into a decision. Cost functions let us treat situations in which some kinds of classification mistakes are more costly than others. Then the posterior probability can be computed by Bayes formula as:

$$P(w_j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid w_j) P(w_j)}{p(\mathbf{x})} \tag{4.8}$$

where the evidence is now

$$p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x} \mid w_j) P(w_j) \tag{4.9}$$

Suppose that we observe a particular x and that we contemplate taking action α_i. If the true state of nature is w_j by definition, we will incur the *loss* $\lambda(\alpha_i|w_j)$. Because $P(w_j|\mathbf{x})$ is the probability that the true state of nature is w_j, the expected loss associated with taking action α_i is

$$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \mathbf{x}) P(w_j \mid \mathbf{x}) \tag{4.10}$$

An expected loss is called a **risk**, and $R(\alpha_i|\mathbf{x})$ is called the **conditional risk**. Whenever we encounter a particular observation $\mathbf{x}$, we can minimize our expected loss by selecting the action that minimizes the conditional risk.

If a general *decision rule* $\alpha(\mathbf{x})$ tells us which action to take for every possible observation $\mathbf{x}$, the overall risk R is given by

$$R = \int R(\alpha(\mathbf{x}) \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \tag{4.11}$$

Thus, the *Bayes decision rule* states that to minimize the overall risk, compute the conditional risk given in Eq.4.10 for $i=1 \ldots a$ and then select the action α_i for which $R(\alpha_i|\mathbf{x})$ is minimum. The resulting minimum overall risk is called the **Bayes risk**, denoted R, and is the best performance that can be achieved.

4.2.1 Two-Category Classification

When these results are applied to the special case of two-category classification problems, action α_1 corresponds to deciding that the true state of nature is w_1, and action α_2 corresponds to deciding that it is w_2. For notational simplicity, let $\lambda_{ij}=\lambda(\alpha_i|w_j)$ be the loss incurred for deciding w_i, when the true state of nature is w_j. If we write out the conditional risk given by Eq.4.10, we obtain

$$R(\alpha_1 \mid \mathbf{x}) = \lambda_{11} P(w_1 \mid \mathbf{x}) + \lambda_{12} P(w_2 \mid \mathbf{x}) \tag{4.12}$$

$$R(\alpha_2 \mid \mathbf{x}) = \lambda_{21} P(w_1 \mid \mathbf{x}) + \lambda_{22} P(w_2 \mid \mathbf{x}) \tag{4.13}$$

There are a variety of ways of expressing the minimum-risk decision rule, each having its own minor advantages. The fundamental rule is to decide w_1 if $R(\alpha_1|\mathbf{x})<R(\alpha_2|\mathbf{x})$. In terms of the posterior probabilities,

we decide w_1 if

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

$$\lambda_{11}P(w_1|\mathbf{x}) + \lambda_{12}P(w_2|\mathbf{x}) < \lambda_{21}P(w_1|\mathbf{x}) + \lambda_{22}P(w_2|\mathbf{x})$$

$$(\lambda_{21} - \lambda_{11})P(w_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22})P(w_2|\mathbf{x}) \qquad (4.14)$$

or in terms of the prior probabilities

$$(\lambda_{21} - \lambda_{11})p(\mathbf{x}|w_1)P(w_1) > (\lambda_{12} - \lambda_{12})p(\mathbf{x}|w_2)P(w_2) \qquad (4.15)$$

or alternatively as **likelihood ratio**

$$\frac{p(\mathbf{x}|w_1)}{p(\mathbf{x}|w_2)} > \frac{(\lambda_{12} - \lambda_{12})}{(\lambda_{21} - \lambda_{11})}\frac{P(w_2)}{P(w_1)} \qquad (4.16)$$

This form of the decision rule focuses on the **x**-dependence of the probability densities. We can consider $p(\mathbf{x}|w_j)$ a function of w_j (i.e., the likelihood function) and then form the *likelihood ratio* $p(\mathbf{x}|w_1)/ p(\mathbf{x}|w_2)$. Thus the Bayes decision rule can be interpreted as calling for deciding w_1 if the likelihood ratio exceeds a threshold value that is independent of the observation **x**.

4.3 Minimum Error Rate Classification

In classification problems, each state of nature is associated with a different one of the classes, and the action α_i is usually interpreted as the decision that the true state of nature is w_i. If action α_i is taken and the true state of nature is w_j then the decision is correct if $i=j$ and in error if $i\neq j$. If errors are to be avoided it is natural to seek a decision rule, that minimizes the probability of error, that is the *error rate*.

This loss function is so called *symetrical* or *zero-one* loss function is given as

$$\lambda(\alpha_i|\mathbf{x}) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1,\ldots,c \qquad (4.17)$$

This loss function assigns no loss to a correct decision, and assigns a unit loss to any error: thus, all errors are equally costly. The risk corresponding to this loss function is precisely the average probability of error because the conditional risk for the two-category classification is

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

$$= \sum_{j\neq i}^{c} P(w_j|\mathbf{x})$$

$$= 1 - P(w_j|\mathbf{x}) \qquad (4.18)$$

and $P(w_j|\mathbf{x})$ is the conditional probability that action α_i is correct. The Bayes decision rule to minimize risk calls for selecting the action that minimizes the conditional risk. Thus, to minimize the average probability of

error, we should select the i that *maximizes* the posterior probability $P(w_j|\mathbf{x})$. In other words, for **minimum error rate:**

Decide w_i if $P(w_i|\mathbf{x})>P(w_j|\mathbf{x})$ for all $i\neq j$

(4.19)

This is the same rule as in Eq.4.5. The region in the input space where we decide w_1 is denoted $\mathcal{R}_1$.

We saw in Figure 4.1 some class-conditional probability densities and the posterior probabilities: Figure 4.3 shows the likelihood ratio for the same case. The threshold value θ_a marked is from the same prior probabilities but with a zero-one loss function. If we penalize mistakes in classifying w_1 patterns as w_2 more than the converse then Eq.4.14 leads to the threshold θ_b marked.

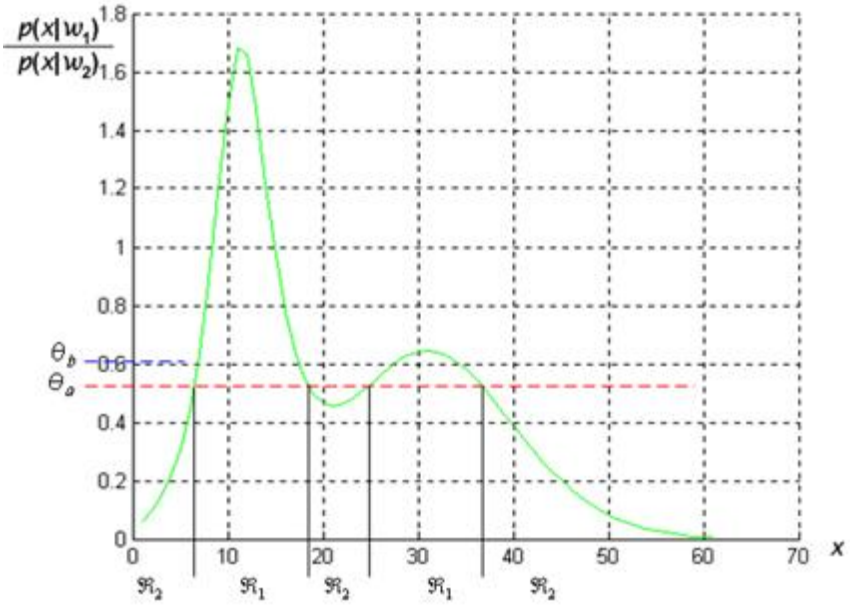


Figure 4.3: The likelihood ratio $p(\mathrm{x}|w_1)/p(\mathrm{x}|w_2)$ for the distributions shown in Figure 4.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold, if our loss function penalizes miscategorizing w_2 as w_1 patterns more than the converse, we get the larger threshold, and hence $\mathcal{R}_1$ becomes smaller.

4.3.1 Minimax Criterion

In-order to design a classifier to perform well over a range of prior probabilities the *worst* overall risk for any value of the priors is as small as possible that is, minimize the maximum possible overall risk. Let $\mathcal{R}_1$ denote that (as yet unknown) region in feature space where the classifier decides w_1 and likewise for $\mathcal{R}_2$ and w_2, and then we write our overall risk Eq.4.11 in terms of conditional risks:

$$R = \int R(\alpha(\mathbf{x}) \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

$$R = \int_{R_1} [\lambda_{11} P(w_1) p(\mathbf{x} \mid w_1) + \lambda_{12} P(w_2) p(\mathbf{x} \mid w_2)] d\mathbf{x} + \int_{R_2} [\lambda_{21} P(w_1) p(\mathbf{x} \mid w_1) + \lambda_{22} P(w_2) p(\mathbf{x} \mid w_2)] d\mathbf{x}$$

By using the fact that $P(w_2)$=1- $P(w_1)$ and that $\int_{R_1} p(\mathbf{x} \mid w_1) d\mathbf{x} = 1 - \int_{R_2} p(\mathbf{x} \mid w_1) d\mathbf{x}$ the function above can be rewritten the risk as:

$$R(P(w_1) = \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{R_1} p(\mathbf{x} \mid w_2) d\mathbf{x}$$

$$+ P(w_1) \left[(\lambda_{11} - \lambda_{22}) + (\lambda_{21} - \lambda_{11}) \int_{R_2} p(\mathbf{x} \mid w_1) d\mathbf{x} - (\lambda_{12} - \lambda_{22}) \int_{R_1} p(\mathbf{x} \mid w_2) d\mathbf{x} \right] \quad (4.20)$$

For minimax solution, the following condition should be satisfied.

$$\left[(\lambda_{11} - \lambda_{22}) + (\lambda_{21} - \lambda_{11}) \int_{R_2} p(\mathbf{x} \mid w_1) d\mathbf{x} - (\lambda_{12} - \lambda_{22}) \int_{R_1} p(\mathbf{x} \mid w_2) d\mathbf{x} \right] = 0$$

(4.21)

and

$$\lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{R_1} p(\mathbf{x} \mid w_2) d\mathbf{x}$$

(4.22)

in Eq.4.20 is known as **minimax risk**.

If we can find a boundary such that the constant of proportionality is 0, then the risk is independent of priors. This is the *minimax risk,* R_{mm}

$$R_{mm} = \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{R_1} p(\mathbf{x} \mid w_2) d\mathbf{x} = \lambda_{11} + (\lambda_{21} - \lambda_{11}) \int_{R_2} p(\mathbf{x} \mid w_1) d\mathbf{x}$$

(4.23)

4.4 The Gaussian (Normal) Density

The definition of the **expected value** of a scalar function $f(x)$ defined for some density $p(x)$ is given by

$$\varepsilon[f(x)] \equiv \int_{-\infty}^{\infty} f(x)p(x)dx$$

(4.24)

If the values of the feature x are restricted to points in a discrete set $\mathcal{D}$ we must sum over all samples as

$$\varepsilon[f(x)] = \sum_{x \in D} f(x)P(x)$$

(4.25)

where $P(x)$ is the **probability mass**.

The continuous **univariate normal density** is given by

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

(4.26)

where **mean** μ (expected value, average) is given by

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

(4.27)

and the special expectation that is **variance** (squared deviation) is given by

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

(4.28)

The univariate normal density is specified by its two parameters, its mean μ, and the variance σ. Samples from normal distributions tend to cluster about the mean, and the extend to which they spread out depends on the variance (Figure 4.4).

The **multivariate normal density** in d dimensions is written as

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

(4.29)

where $\mathbf{x}$ is a d-component column vector, $\boldsymbol{\mu}$ is the d-component *mean vector*, $\mathbf{\Sigma}$ is $d \times d$ *covariance matrix*, $|\mathbf{\Sigma}|$ is its determinant and $\mathbf{\Sigma}^{-1}$ is its inverse and mean vector $\boldsymbol{\mu}$ becomes

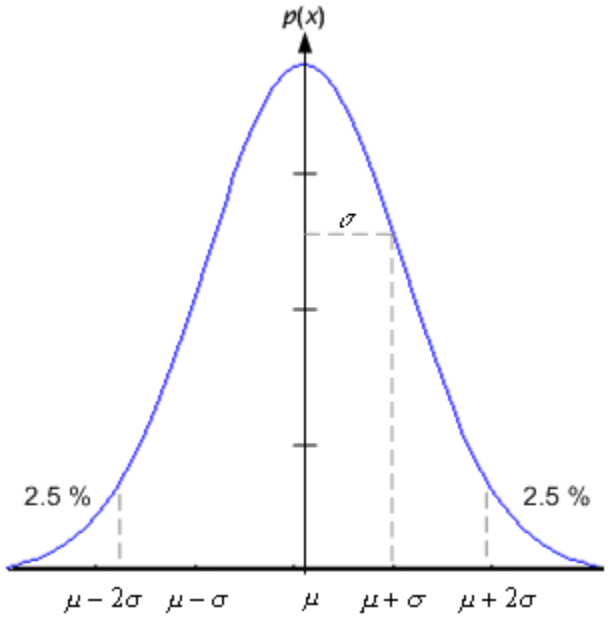


Figure 4.4: A univariate Gaussian distribution has roughly 95% of its area in the range $|x-\mu| \leq 2\sigma$.

$$\boldsymbol{\mu} = \varepsilon[\mathbf{x}] = \begin{bmatrix} \varepsilon[x_1] \\ \varepsilon[x_2] \\ \cdots \\ \varepsilon[x_d] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_d \end{bmatrix} = \int_{-\infty}^{\infty} \mathbf{x} p(\mathbf{x}) d\mathbf{x} = \sum_{x} \mathbf{x} P(\mathbf{x})$$

(4.30)

The covariance matrix $\boldsymbol{\Sigma}$ is defined as the square matrix whose ij^{th} element σ_{ij} is the covariance of x_i and x_j: The covariance of two features measures their tendency to vary together, i.e., to co-vary.

$$\sigma_{ij} = \varepsilon[(x_i - \mu_i)(x_j - \mu_j)] = \int_{-\infty}^{\infty} (x_i - \mu_i)(x_j - \mu_j) p(x) dx = \sum_{x \in D} (x_i - \mu_i)(x_j - \mu_j) P(x) \quad (4.31)$$

where $i, j=1,\ldots,d$. Therefore, in expanded form we have

$$\boldsymbol{\Sigma} = \begin{bmatrix} \varepsilon[(x_1 - \mu_1)(x_1 - \mu_1)] & \varepsilon[(x_1 - \mu_1)(x_2 - \mu_2)] & \cdots & \varepsilon[(x_1 - \mu_1)(x_d - \mu_d)] \\ \varepsilon[(x_2 - \mu_2)(x_1 - \mu_1)] & \varepsilon[(x_2 - \mu_2)(x_2 - \mu_2)] & \cdots & \varepsilon[(x_2 - \mu_2)(x_d - \mu_d)] \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon[(x_d - \mu_d)(x_1 - \mu_1)] & \varepsilon[(x_d - \mu_d)(x_2 - \mu_2)] & & \varepsilon[(x_d - \mu_d)(x_d - \mu_d)] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix} \tag{4.32}$$

We can use vector product $(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T$ to write the covariance matrix as

$$\boldsymbol{\Sigma} = \varepsilon[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T] \tag{4.33}$$

Thus $\boldsymbol{\Sigma}$ is symmetric, and its diagonal elements are the *variances* of the respective individual elements x_i of $\mathbf{x}$ (e.g., σ_i^2), which can never be negative; the off-diagonal elements are the *covariances* of x_i and x_j, which can be positive or negative. If the variables x_i and x_j are statistically independent, the covariances σ_{ij} are zero, and the covariance matrix is diagonal. If all the off-diagonal elements are zero, $p(\mathbf{x})$ reduces to the product of the univariate normal densities for the components of $\mathbf{x}$. The analog to the Cauchy-Schwarz inequality comes from recognizing that if $\mathbf{w}$ is any d-dimensional vector, then the variance of $\mathbf{w}^T\mathbf{x}$ can never be negative. This leads to the requirement that the quadratic form $\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}$ never be negative. Matrices for which this is true are said to be *positive semidefinite*; thus, the covariance matrix is positive semidefinite.

Linear combinations of jointly normally distributed random variables, independent or not, are normally distributed.

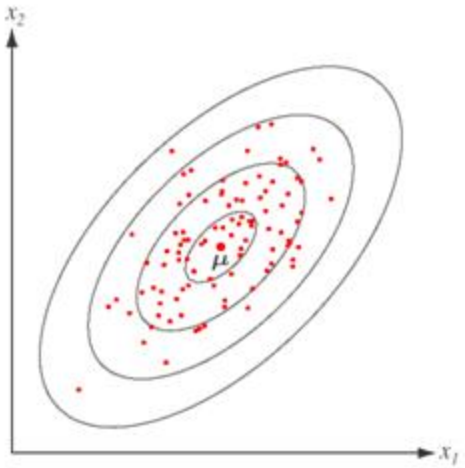


Figure 4.5: Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean. The ellipses show lines of equal probability density of the Gaussian.

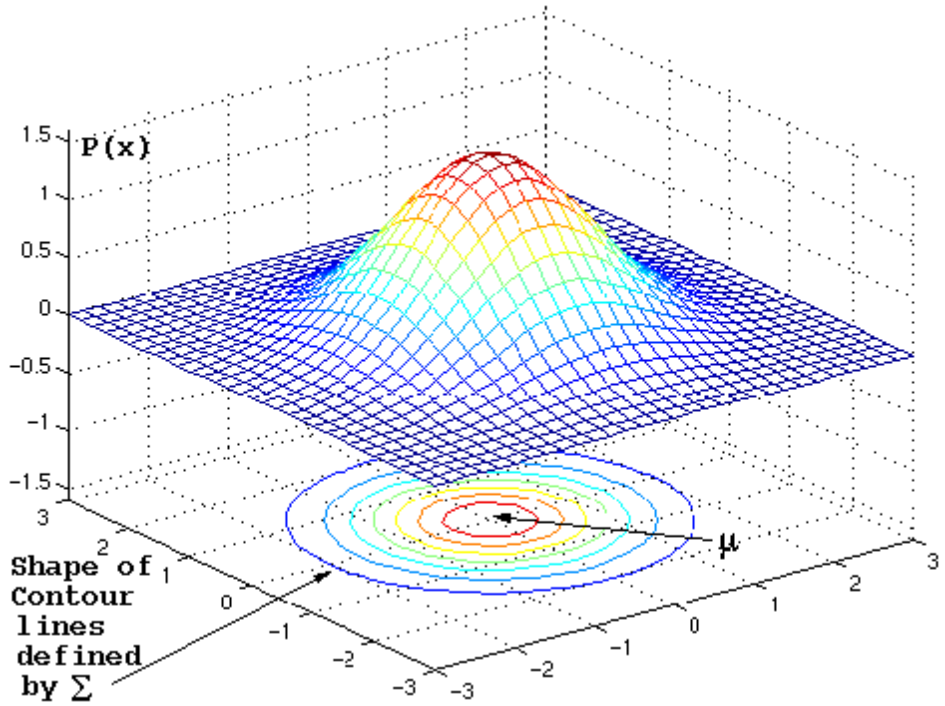


Figure 4.6: The contour lines show the regions for which the function has constant density. From the equation for the normal density, it is apparent that points, which have the same density, must have the same constant term $(x - \mu)^{-1}\Sigma(x - \mu)$.

As with the univariate density, samples from a normal population tend to fall in a single cloud or cluster centered about the mean vector, and the shape of the cluster depends on the covariance matrix (see Figure 4.5 and Figure 4.6).

From the multivariate normal density formula in Eq.4.27 notice that the density is constant on surfaces where the squared distance (Mahalanobis distance)$(\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ is constant. These paths are called **contours** (hyperellipsoids). The principle axes of these contours are given by the eigenvectors of $\boldsymbol{\Sigma}$, where the eigenvalues determine the lengths of these axes.

The covariance matrix for 2 features x and y is diagonal (which implies that the 2 features don't co-vary), but feature x varies more than feature y. The contour lines are stretched out in the x direction to reflect the fact that the distance spreads out at a lower rate in the x direction than it does in the y direction. The reason that the distance decreases slower in the x direction is because the variance for x is greater and thus a point that is far away in the x direction is not quite as *distant* from the mean as a point that is far away in the y direction (see Figure 4.9).

4.4.1 Interpretation of eigenvalues and eigenvectors

It is sometimes convenient to perform a coordinate transformation that converts an arbitrary multivariate normal distribution into a spherical one –that is, one having a covariance matrix proportional to the identity matrix **I**. If we define $\boldsymbol{\Phi}$ to be the matrix whose columns are the orthonormal eigenvectors of $\boldsymbol{\Sigma}$, and $\boldsymbol{\Lambda}$ the diagonal matrix of the corresponding eigenvalues, then the transformation $A=\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2}$ applied to the coordinates ensures that the transformed distribution has covariance matrix equal to the identity matrix **I**. If we view matrix A as a linear transformation, an eigenvector represents an invariant direction in the vector space. When transformed by A, any point lying on the direction defined by **v** will remain on that direction, and its magnitude will be multipled by the corresponding eigenvalue (see Figure 4.7).

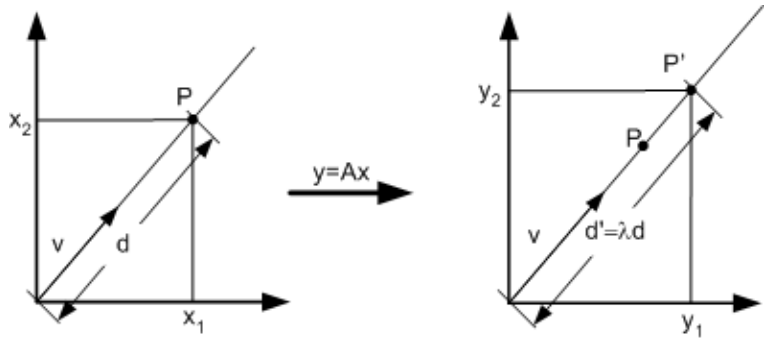


Figure 4.7: The linear transformation of a matrix.

Given the covariance matrix $\boldsymbol{\Sigma}$ of a Gaussian distribution, the eigenvectors of $\boldsymbol{\Sigma}$ are the principal directions of the distribution, and the eigenvalues are the variances of the corresponding principal directions. The linear transformation defined by the eigenvectors of $\boldsymbol{\Sigma}$ leads to vectors that are uncorrelated regardless of the form of the distribution. If the distribution happens to be Gaussian, then the transformed vectors will be statistically independent.

$$\boldsymbol{\Sigma}\mathrm{M} = \mathrm{M}\boldsymbol{\Lambda} \quad \text{with} \quad \mathrm{M} = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ v_1 & v_2 & \cdots & v_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$$

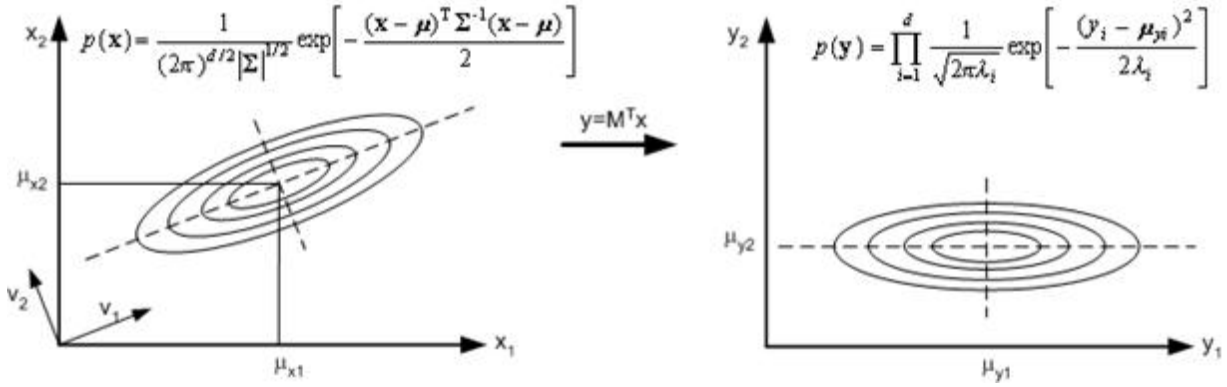


Figure 4.8: The linear transformation.

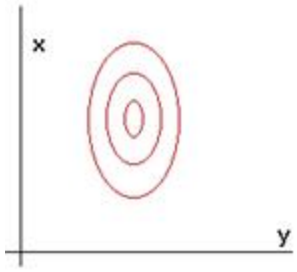


Figure 4.9: The covariance matrix for two features x and y do not co-vary, but feature x varies more than feature y.

The covariance matrix for two features x and y is diagonal, and x and y have the exact same variance. This results in euclidean distance contour lines (see Figure 4.10).

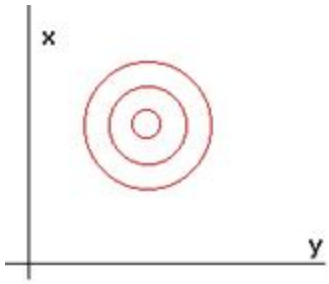


Figure 4.10: The covariance matrix for two features that have exact same variances.

The covariance matrix is *not* diagonal. Instead, x and y have the same variance, but x varies with y in the sense that x and y tend to increase together. So the covariance matrix would have identical diagonal elements, but the off-diagonal element would be a strictly positive number representing the covariance of x and y (see Figure 4.11).

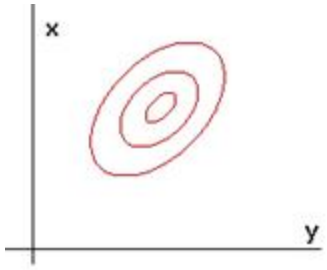


Figure 4.11: The covariance matrix for two features that has exact same variances, but x varies with y in the sense that x and y tend to increase together.

4.5 Discriminant Functions, and Decision Surfaces

There are many different ways to represent pattern classifiers. One of the most useful is in terms of a set of *discriminant functions* $g_i(\mathbf{x})$, i=1,…,c. The classifier is said to assign a feature vector $\mathbf{x}$ to class w_i if

$$g_i(\mathbf{x}) > g_j(\mathbf{x}),\ i \neq j \qquad (4.34)$$

A Bayes classifier is easily and naturally represented in this way. For the general case with risks, we can let $g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$, because the maximum discriminant function will then correspond to the minimum conditional risk. For the minimum error-rate case, we can simplify things further by taking $g_i(\mathbf{x}) = P(w_i|\mathbf{x})$, so that the maximum discriminant function corresponds to the maximum posterior probability. Clearly, the choice of discriminant functions is not unique. In particular, for minimum-error rate classification, any of the following choices gives identical classification results, but some can be much simpler to understand or to compute than others:

$$g_i(\mathbf{x}) = P(w_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid w_i)P(w_i)}{\sum_{j=1}^{c} p(\mathbf{x} \mid w_j)P(w_j)} \quad (4.35)$$

$$g_i(\mathbf{x}) = p(\mathbf{x} \mid w_i)P(w_i) \quad (4.36)$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{x} \mid w_i) + \ln P(w_i) \quad (4.37)$$

Even though the discriminant functions can be written in a variety of forms, the decision rules are equivalent. The effect of any decision rule is to divide the feature space into *c decision boundaries*, $\mathcal{R}_1,\ldots, \mathcal{R}_c$. If $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for all $i \neq j$, then $\mathbf{x}$ is in $\mathcal{R}_i$, and the decision rule calls for us to assign $\mathbf{x}$ to w_i. The regions are separated by decision boundaries, surfaces in feature space where ties occur among the largest discriminant functions.

While the two-category case is just a special instance of the multicategory case, instead of using two discriminant functions g_1 and g_2 and assigning $\mathbf{x}$ to w_1 if $g_1 > g_2$, it can be treated by using a single discriminant function

$$g(\mathbf{x}) \equiv g_1(\mathbf{x}) - g_2(\mathbf{x}) \quad (4.38)$$

and to use the following decision rule: Decide w_1 if $g(\mathbf{x})>0$; otherwise decide w_2. One of the various forms in which the minimum-error rate discriminant function can be written, the following two are particularly convenient:

$$g(\mathbf{x}) = P(w_1 \mid \mathbf{x}) - P(w_2 \mid \mathbf{x}) \quad (4.39)$$

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x} \mid w_1)}{p(\mathbf{x} \mid w_2)} + \ln \frac{P(w_1)}{P(w_2)} \quad (4.40)$$

4.6 Discriminant Functions For The Normal Density

Eq.4.37 can be easily evaluated if the densities are multivariate normal. In this case, from eq.4.29 we have

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln |\boldsymbol{\Sigma}_i| + \ln P(w_i) \quad (4.41)$$

Case 1: $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}$

The simplest case occurs when the features that are measured is independent of each other, that is, statistically independent, and when each feature has the same variance, σ^2. For example, if we were trying to recognize an apple from an orange, and we measured the colour and the weight as our feature vector, then chances are that there is no relationship between these two properties. The non-diagonal elements of the covariance matrix are the covariances of the two features x_1=colour and x_2=weight. But because these features are independent, their covariances would be 0. Therefore, the covariance matrix for both classes would be diagonal, being merely σ^2 times the identity matrix $\mathbf{I}$.

As a second simplification, assume that the variance of colours is the same is the variance of weights. This means that there is the same degree of spreading out from the mean of colours as there is from the mean of weights. If this is true for some class *i* then the covariance matrix for that class will have identical diagonal elements. Finally, suppose that the variance for the colour and weight features is the same in both classes. This means that the degree of spreading for these two features is independent of the class from which you

draw your samples. If this is true, then the covariance matrices will be identical. When normal distributions are plotted that have a diagonal covariance matrix that is just a constant multplied by the identity matrix, their cluster points about the mean are shperical in shape.

Geometrically, this corresponds to the situation in which the samples fall in equal-size hyperspherical clusters, the cluster for the i^{th} class being centered about the mean vector $\boldsymbol{\mu}_i$ (see Figure 4.12). The computation of the determinant and the inverse of $\boldsymbol{\Sigma}_i$ is particularly easy:

$$|\boldsymbol{\Sigma}_i| = \sigma^{2d} \quad \text{and} \quad \boldsymbol{\Sigma}_i^{-1} = \frac{1}{\sigma^2}\mathbf{I} \tag{4.42}$$

Because both $|\boldsymbol{\Sigma}_i|$ and the $(d/2)\ln 2\pi$ term in Eq.4.41 are independent of i, they are unimportant additive constants that can be ignored. Thus, we obtain the simple discriminant functions

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(w_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\frac{1}{\sigma^2}\mathbf{I}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^{2d} + \ln P(w_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\mathbf{I}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln P(w_i)$$

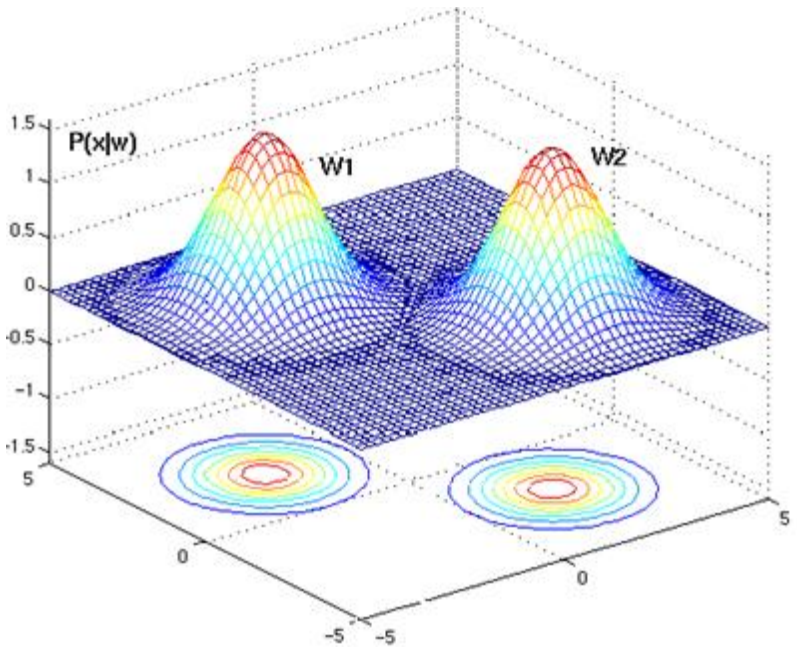


Figure 4.12: Since the bivariate normal densities have diagonal covariance matrices, their contours are spherical in shape. Each class has the exact same covariance matrix, the circular lines forming the contours are the same size for both classes. This is because identical covariance matrices imply that the two classes have identically shaped clusters about their mean vectors.

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

(4.43)

where

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

(4.44)

If the prior probabilities are not equal, then Eq.4.43 shows that the squared distance $\|\mathbf{x}-\boldsymbol{\mu}_i\|^2$ must be normalized by the variance σ^2 and offset by adding ln $P(w_i)$; thus, if **x** is equally near two different mean vectors, the optimal decision will favor the a priori more likely category.

Regardless of whether the prior probabilities are equal or not, it is not actually necessary to compute distances. Expansion of the quadratic form $(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu}_i)$ yields

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

(4.45)

which appears to be a quadratic function of **x**. However, the quadratic term $\mathbf{x}^{\mathrm{T}}\mathbf{x}$ is the same for all *i*, making it an ignorable additive constant. Thus, we obtain the equivalent *linear discriminant functions*

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

(4.46)

where

$$\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i$$

(4.47)

and

$$w_{i0} = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\mu}_i + \ln P(w_i)$$

(4.48)

We call w_{i0} the *threshold* or *bias* for the i^{th} category.

The decision boundaries for these discriminant functions are found by intersecting the functions $g_i(x)$ and $g_j(x)$ where i and j represent the 2 classes with the highest a posteriori probabilites. As in the univariate case, this is equivalent to determining the region for which $g_i(x)$ is the maximum of all the discriminant functions. By setting $g_i(x) = g_j(x)$ we have that:

$$(\mathbf{w}_i^T - \mathbf{w}_j^T)\mathbf{x} + (w_{i0} - w_{j0}) = 0 \tag{4.49}$$

Consider the term w_{i0} - w_{j0}:

$$= -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i + \ln P(w_i) + \frac{1}{2\sigma^2}\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j - \ln P(w_j)$$

$$= -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i + \ln\frac{P(w_i)}{P(w_j)} + \frac{1}{2\sigma^2}\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j$$

$$= -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i + \frac{1}{2\sigma^2}\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j + \frac{(\boldsymbol{\mu}_i^T - \boldsymbol{\mu}_j^T)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}\ln\frac{P(w_i)}{P(w_j)}$$

$$= -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i + \frac{1}{2\sigma^2}\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j + \frac{\boldsymbol{\mu}_i^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}\ln\frac{P(w_i)}{P(w_j)} - \frac{\boldsymbol{\mu}_j^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}\ln\frac{P(w_i)}{P(w_j)} \tag{4.50}$$

Now, by adding and subtracting the same term, we get:

$$= -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i + \frac{1}{2\sigma^2}\boldsymbol{\mu}_j^T\boldsymbol{\mu}_i + \frac{\boldsymbol{\mu}_i^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}\ln\frac{P(w_i)}{P(w_j)} + \frac{1}{2\sigma^2}\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j$$
$$- \frac{1}{2\sigma^2}\boldsymbol{\mu}_j^T\boldsymbol{\mu}_i - \frac{\boldsymbol{\mu}_j^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}\ln\frac{P(w_i)}{P(w_j)}$$

$$= \frac{\boldsymbol{\mu}_j^T}{\sigma^2}\left[\frac{1}{2}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_i) - \sigma^2\frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}\ln\frac{P(w_i)}{P(w_j)}\right] - \frac{\boldsymbol{\mu}_i^T}{\sigma^2}\left[\frac{1}{2}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_i) - \sigma^2\frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}\ln\frac{P(w_i)}{P(w_j)}\right] \tag{4.51}$$

By letting:

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

(4.52)

the result is:

$$= \left(\frac{\boldsymbol{\mu}_j^{\mathrm{T}}}{\sigma^2} - \frac{\boldsymbol{\mu}_i^{\mathrm{T}}}{\sigma^2} \right) \mathbf{x}_0$$

(4.53)

But because of the way we define w_i and w_j, this is just:

$$(\mathbf{w}_j - \mathbf{w}_i)^{\mathrm{T}} \mathbf{x}_0$$

(4.54)

So from the originial equation we have :

$$(\mathbf{w}_i - \mathbf{w}_j)^{\mathrm{T}} \mathbf{x} - (\mathbf{w}_i - \mathbf{w}_j)^{\mathrm{T}} \mathbf{x}_0$$

(4.55)

and after multiplying through by variance the final decision boundary is given by:

$$(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\mathbf{x} - (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\mathbf{x}_0$$

(4.56)

Now let $\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$. Then this boundary can be written as:

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

(4.57)

where

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

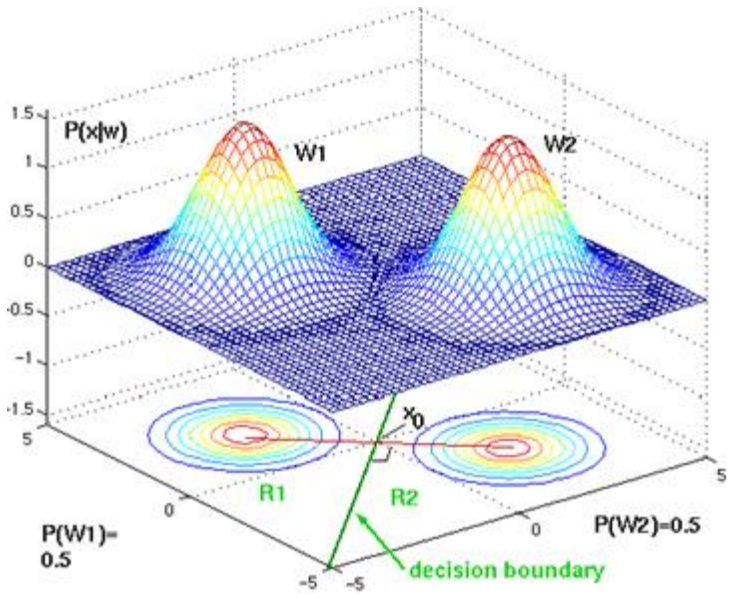
(4.58)

and

$$\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$$

(4.59)

This is called the *normal form* of the boundary equation. Geometrically, equations 4.57, 4.58, and 4.59 define a hyperplane throught the point $\mathbf{x}_0$ that is orthogonal to the vector $\mathbf{w}$. But since $\mathbf{w}=\boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ then the hyperplane which seperates $\mathcal{R}_i$ and $\mathcal{R}_j$ is orthogonal to the line that links their means. If $P(w_i)=P(w_j)$, the second term on the right of Eq.4.58 vanishes, and thus the point $\mathbf{x}_0$ is halfway between the means (equally divide the distance between the 2 means, with a decision region on either side), and the hyperplane is the perpendicular bisector of the line between the means (see Figure 4.13).



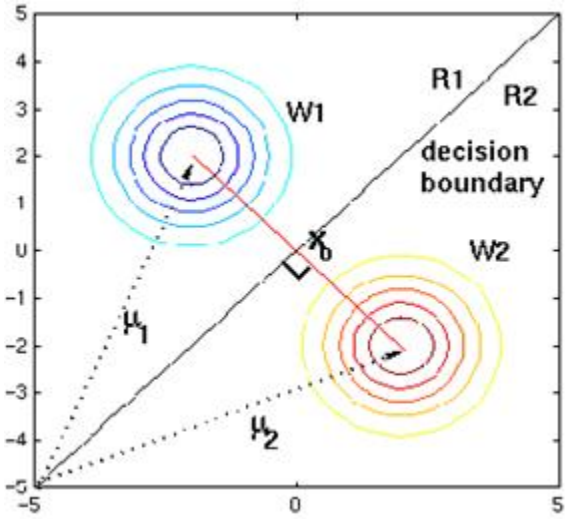


Figure 4.13: Two bivariate normal distributions, whose priors are exactly the same. Therefore, the decision bondary is exactly at the midpoint between the two means. The decision boundary is a line orthogonal to the line joining the two means.

If $P(w_i) \neq P(w_j)$ the point $\mathbf{x}_0$ shifts away from the more likely mean. Note, however, that if the variance is small relative to the squared distance $\|\mu_i - \mu_j\|^2$, then the position of the decision boundary is relatively insensitive to the exact values of the prior probabilities. In other words, there are 80% apples entering the store. If you observe some feature vector of color and weight that is just a little closer to the mean for oranges than the mean for apples, should the observer classify the fruit as an orange? The answer depends on how far from the apple mean the feature vector lies. In fact, if $P(w_i) > P(w_j)$ then the second term in the equation for $\mathbf{x}_0$ will subtract a positive amount from the first term. This will move point $\mathbf{x}_0$ away from the mean for $\mathcal{R}_i$. If $P(w_i) < P(w_j)$ then $\mathbf{x}_0$ would tend to move away from the mean for $\mathcal{R}_j$. So for the above example and using the above decision rule, the observer will classify the fruit as an apple, simply because it's not very close to the mean for oranges, and because we know there are 80% apples in total (see Figure 4.14 and Figure 4.15).

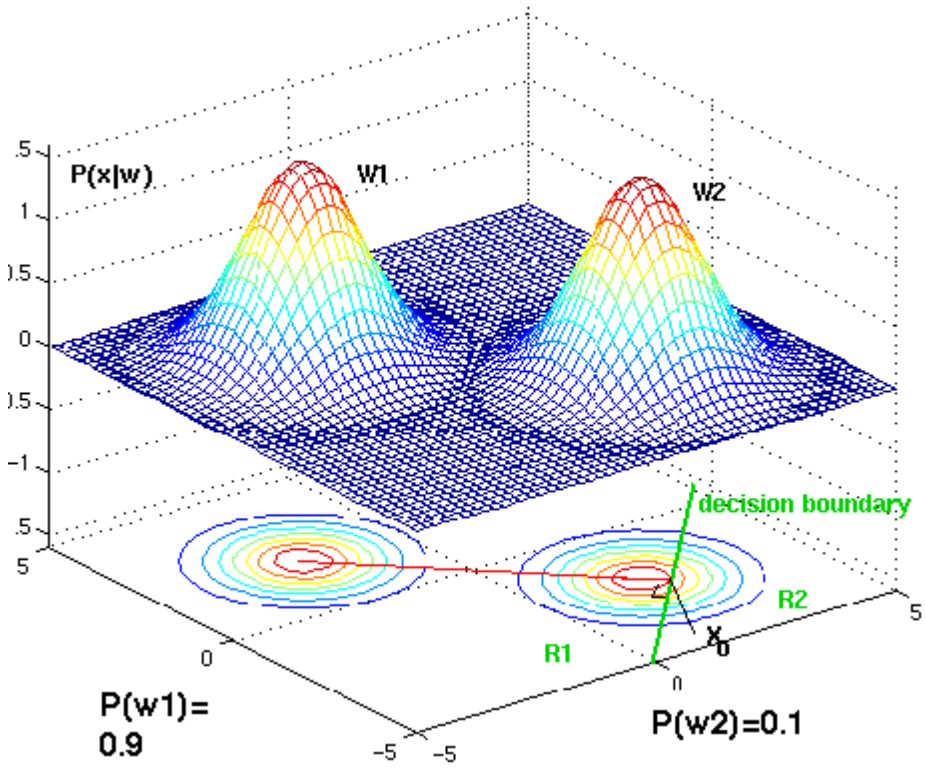


Figure 4.14: As the priors change, the decision boundary throught point x_0 shifts away from the more common class mean (two dimensional Gaussian distributions).

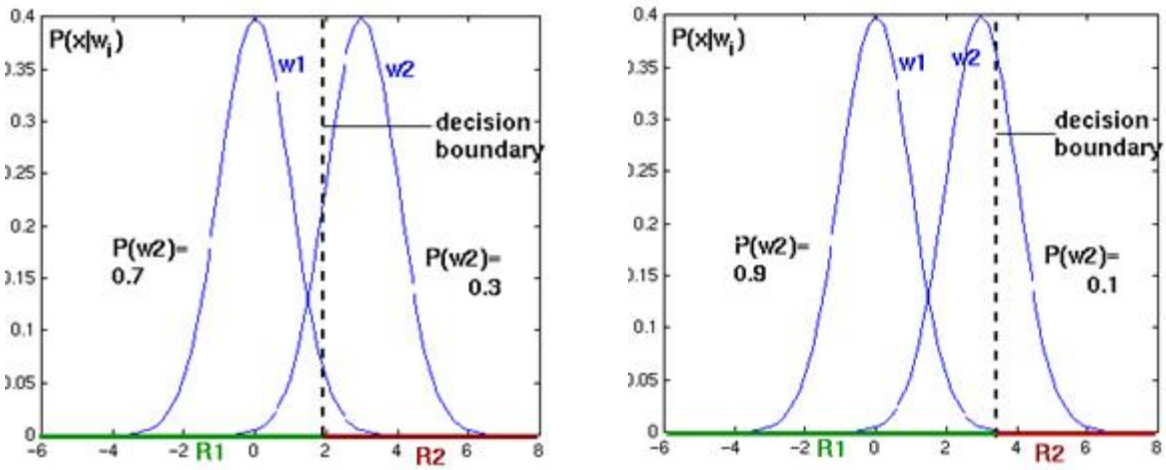


Figure 4.15: As the priors change, the decision boundary throught point x_0 shifts away from the more common class mean (one dimensional Gaussian distributions).

If the prior probabilities $P(w_i)$ are the same for all c classes, then the ln $P(w_i)$ term becomes another unimportant additive constant that can be ignored. When this happens, the optimum decision rule can be stated very simply: the decision rule is based entirely on the distance from the feature vector **x** to the different mean vectors. The object will be classified to $\mathcal{R}_i$ if it is closest to the mean vector for that class. To classify a feature vector **x**, measure the Euclidean distance $\|\mathbf{x} - \boldsymbol{\mu}_i\|$ from each **x** to each of the c mean vectors, and assign **x** to the category of the nearest mean. Such a classifier is called a *minimum-distance classifier.* If each mean vector is thought of as being an ideal prototype or template for patterns in its class, then this is essentially a *template-matching* procedure.

Case 2: $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$

Another simple case arises when the covariance matrices for all of the classes are identical but otherwise arbitrary. Since it is quite likely that we may not be able to measure features that are independent, this section allows for any arbitrary covariance matrix for the density of each class. In order to keep things simple, assume also that this arbitrary covariance matrix is the same for each class w_i. This means that we allow for the situation where the color of fruit may covary with the weight, but the way in which it does is exactly the same for apples as it is for oranges. Instead of having shperically shaped clusters about our means, the shapes may be any type of hyperellipsoid, depending on how the features we measure relate to each other. However, the clusters of each class are of equal size and shape and are still centered about the mean for that class.

Geometrically, this corresponds to the situation in which the samples fall in hyperellipsoidal clusters of equal size and shape, the cluster for the ith class being centered about the mean vector $\boldsymbol{\mu}_i$. Because both $\boldsymbol{\Sigma}_i$ and the $(d/2)\ln 2\pi$ terms in eq. 4.41 are independent of i, they can be ignored as superfluous additive constants. Using the general discriminant function for the normal density, the constant terms are removed. This simplification leaves the discriminant functions of the form:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(w_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln P(w_i) \qquad (4.60)$$

Note that, the covariance matrix no longer has a subscript i, since it is the same matrix for all classes.

If the prior probabilities $P(w_i)$ are the same for all c classes, then the $\ln P(w_i)$ term can be ignored. In this case, the optimal decision rule can once again be stated very simply: To classify a feature vector x, measure the squared Mahalanobis distance $(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$ from $\mathbf{x}$ to each of the c mean vectors, and assign $\mathbf{x}$ to the category of the nearest mean. As before, unequal prior probabilities bias the decision in favor of the a priori more likely category.

Expansion of the quadratic form $(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$ results in a sum involving a quadratic term $\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}$ which here is independent of i. After this term is dropped from eq.4.41, the resulting discriminant functions are again linear.

After expanding out the first term in eq.4.60,

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} - \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i - \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \;\; \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}) + \ln P(w_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i)^{\mathrm{T}}\mathbf{x} + \ln P(w_i) \qquad (4.61)$$

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

(4.62)

where

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

(4.63)

and

$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(w_i) \tag{4.64}$$

The boundary between two decision regions is given by

$$g_i(\mathbf{x}) - g_j(\mathbf{x}) = \mathbf{w}_i^{\mathrm{T}}\mathbf{x} + w_{i0} - \mathbf{w}_j^{\mathrm{T}}\mathbf{x} - w_{j0} = (\mathbf{w}_i - \mathbf{w}_j)^{\mathrm{T}}\mathbf{x} + (w_{i0} - w_{j0}) \tag{4.65}$$

Now examine the second term $(w_{i0} - w_{j0})$ from eq.4.64. Substituting the values for w_{i0} and w_{j0} yields:

$$(w_{i0} - w_{j0}) = -\frac{1}{2}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(w_i) + \frac{1}{2}\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j - \ln P(w_j)$$

$$= -\frac{1}{2}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln\frac{P(w_i)}{P(w_j)} + \frac{1}{2}\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j$$

$$= -\frac{1}{2}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \frac{1}{2}\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \frac{(\boldsymbol{\mu}_i^{\mathrm{T}} - \boldsymbol{\mu}_j^{\mathrm{T}})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{(\boldsymbol{\mu}_i^{\mathrm{T}} - \boldsymbol{\mu}_j^{\mathrm{T}})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}\ln\frac{P(w_i)}{P(w_j)}$$

$$= -\frac{1}{2}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \frac{1}{2}\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \frac{\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{(\boldsymbol{\mu}_i^{\mathrm{T}} - \boldsymbol{\mu}_j^{\mathrm{T}})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}\ln\frac{P(w_i)}{P(w_j)} - \frac{\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{(\boldsymbol{\mu}_i^{\mathrm{T}} - \boldsymbol{\mu}_j^{\mathrm{T}})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}\ln\frac{P(w_i)}{P(w_j)}$$

(4.66)

Then, by adding and subtracting the same term:

$$= \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \left[\frac{1}{2}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_i) - \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)} \ln \frac{P(w_i)}{P(w_j)} \right]$$
$$- \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \left[\frac{1}{2}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_i) - \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)} \ln \frac{P(w_i)}{P(w_j)} \right]$$

(4.67)

Now if we let

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_i) - \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)} \ln \frac{P(w_i)}{P(w_j)}$$

(4.68)

Then the above line reduces to:

$$= (\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} - \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1}) \mathbf{x}_0$$

which is actually just:

$$(\mathbf{w}_j - \mathbf{w}_i)^T \mathbf{x}_0 = 0$$

(4.69)

Now, starting with the original equation and substituting this line back in, the result is:

$$(\mathbf{w}_i - \mathbf{w}_j)^T \mathbf{x} - (\mathbf{w}_j - \mathbf{w}_i)^T \mathbf{x}_0 = 0$$

(4.70)

So let

$\mathbf{w} = \mathbf{w}_i - \mathbf{w}_j$

which means that the equation for the decision boundary is given by:

$$\mathbf{w}^T \mathbf{x} - \mathbf{w}^T \mathbf{x}_0 = 0$$

(4.71)

Because the discriminants are linear, the resulting decision boundaries are again hyperplanes. If $\mathcal{R}_i$ and $\mathcal{R}_j$ are contiguous, the boundary between them has the equation eq.4.71 where

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

(4.72)

and

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_i) - \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)} \ln \frac{P(w_i)}{P(w_j)} \qquad (4.73)$$

Again, this formula is called the *normal form* of the decision boundary.

As in case 1, a line through the point $\mathbf{x}_0$ defines this decision boundary between $\mathcal{R}_i$ and $\mathcal{R}_j$. If the prior probabilities are equal then $\mathbf{x}_0$ is halfway between the means. If the prior probabilities are not equal, the optimal boundary hyperplane is shifted away from the more likely mean The decision boundary is in the direction orthogonal to the vector $\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$; The difference lies in the fact that the term **w** is no longer exactly in the direction of $\boldsymbol{\mu}_i - \boldsymbol{\mu}_j$. Instead, the vector between $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ is now also multipled by the inverse of the covariance matrix. This means that the decision boundary is no longer orthogonal to the line joining the two mean vectors. Intstead, the boundary line will be tilted depending on how the 2 features covary and their respective variances (see Figure 4.19). As before, with sufficient bias the decision plane need not lie between the two mean vectors.

To understand how this tilting works, suppose that the distributions for class *i* and class *j* are bivariate normal and that the variance of feature 1 is σ_1^2 and that of feature 2 is σ_2^2. Suppose also that the covariance of the 2 features is 0. Finally, let the mean of class *i* be at (a,b) and the mean of class *j* be at (c,d) where a>c and b>d for simplicity. Then the vector **w** will have the form:

$$[\frac{1}{\sigma_1^2}(a-c), \frac{1}{\sigma_2^2}(b-d)]$$

This equation can provide some insight as to how the decision boundary will be tilted in relation to the covariance matrix. Note though, that the direction of the decision boundary is orthogonal to this vector, and so the direction of the decision boundary is given by:

$$[-\frac{1}{\sigma_2^2}(b-d), \frac{1}{\sigma_1^2}(a-c)]$$

Now consider what happens to the tilt of the decision boundary when the values of σ_1^2 or σ_2^2 are changed (Figure 4.16). Although the vector form of **w** provided shows exactly which way the decision boudary will tilt, it does not illustrate how the contour lines for the 2 classes are changing as the variances altered.

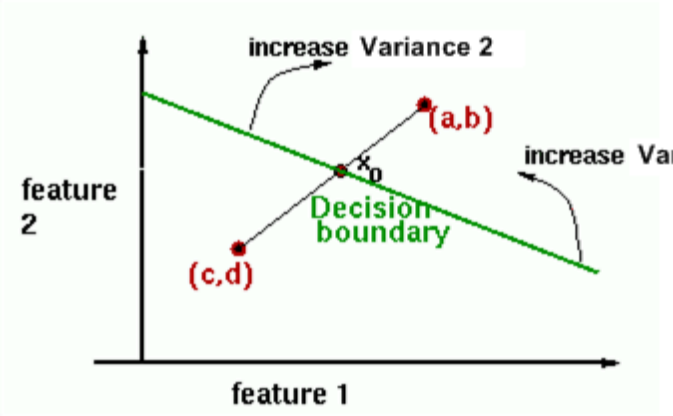


Figure 4.16: As the variance of feature 2 is increased, the x term in the vector will become less negative. This means that the decision boundary will tilt vertically. Similarly, as the variance of feature 1 is increased, the y term in the vector will decrease, causing the decision boundary to become more horizontal.

Does the tilting of the decision boundary from the orthogonal direction make intuitive sense? With a little thought, it is easy to see that it does. For example, suppose that you are again classifying fruits by measuring their color and weight. Suppose that the color varies much more than the weight does. Then consider making a measurement at point P in Figure 4.17:

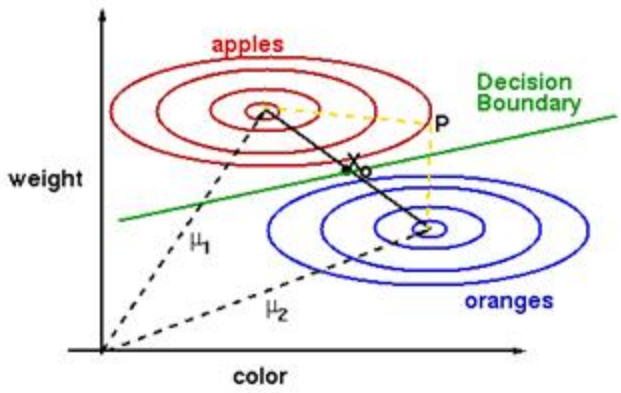


Figure 4.17: The discriminant function evaluated at *P* is smaller for class apple than it is for class orange.

In Figure 4.17, the point P is at actually closer euclideanly to the mean for the orange class. But as can be seen by the ellipsoidal contours extending from each mean, the discriminant function evaluated at P is smaller for class 'apple' than it is for class 'orange'. This is because it is much worse to be farther away in the *weight* direction, then it is to be far away in the *color direction*. Thus, the total 'distance' from P to the means must consider this. For this reason, the decision bondary is tilted.

The fact that the decision boundary is not orthogonal to the line joining the 2 means, is the only thing that seperates this situation from case 1. In both cases, the decision boundaries are straight lines that pass through the point $\mathbf{x}_0$. The position of $\mathbf{x}_0$ is effected in the exact same way by the a priori probabilities.

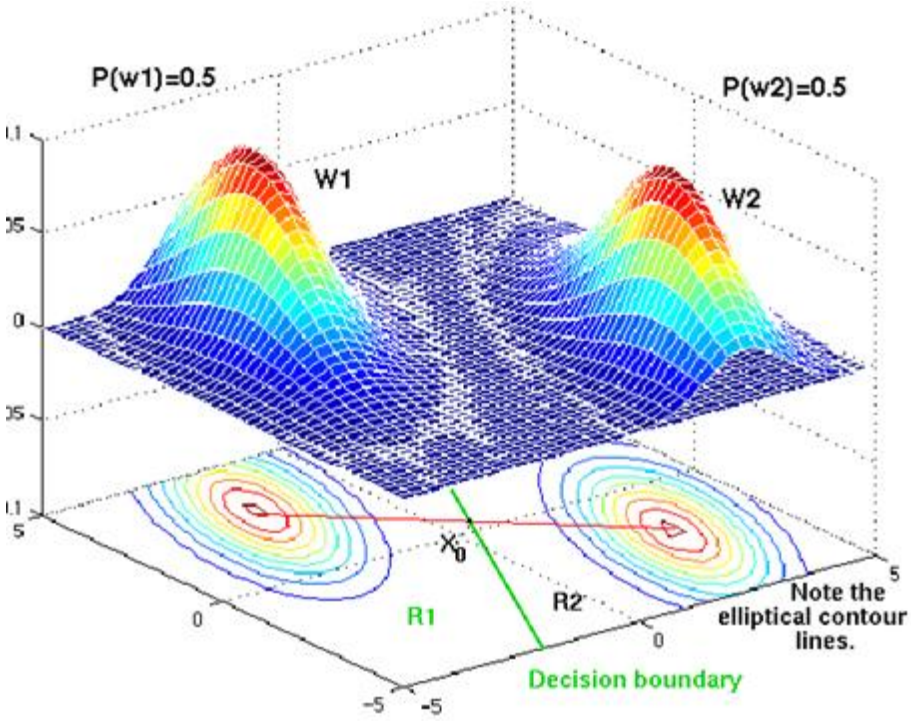
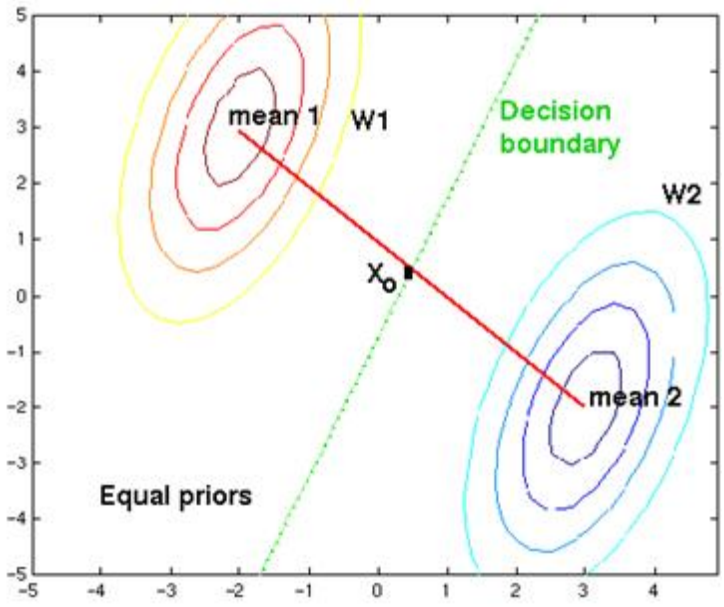


Figure 4.18: The contour lines are elliptical in shape because the covariance matrix is not diagonal. However, both densities show the same elliptical shape. The prior probabilities are the same, and so the point x_0 lies halfway between the 2 means. The decision boundary is not orthogonal to the red line. Instead, it is is tilted so that its points are of equal distance to the contour lines in w_1 and those in w_2.



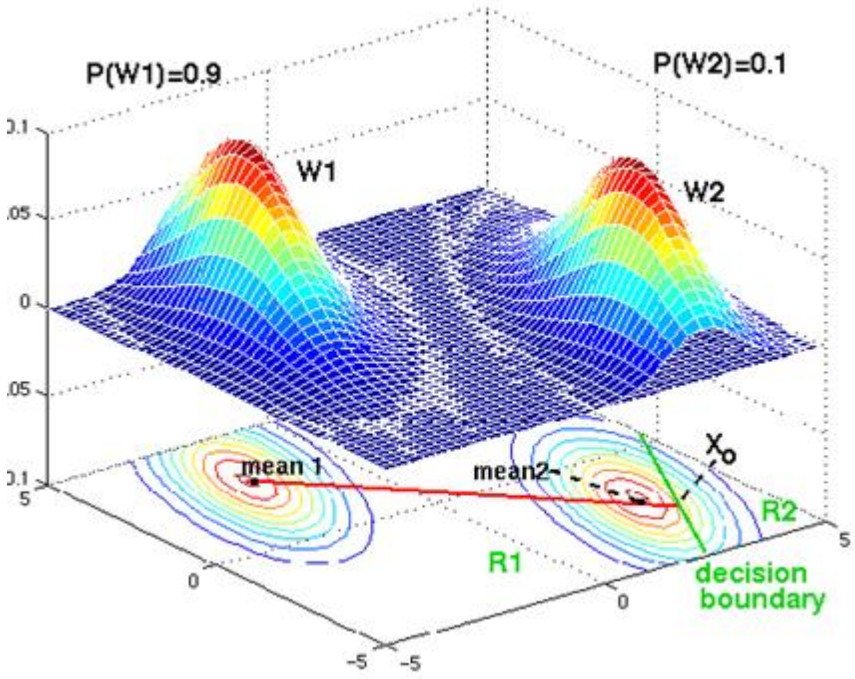


Figure 4.19: The contour lines are elliptical, but the prior probabilities are different. Although the decision boundary is a parallel line, it has been shifted away from the more likely class. With sufficient bias, the decision boundary can be shifted so that it no longer lies between the 2 means:

Case 3: $\boldsymbol{\Sigma}_i$ = arbitrary

In the general multivariate normal case, the covariance matrices are different for each category. This case assumes that the covariance matrix for each class is arbitrary. The discriminant functions cannot be simplified and the only term that can be dropped from eq.4.41 is the (d/2) ln 2π term, and the resulting discriminant functions are inherently quadratic.

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(w_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(w_i) \qquad (4.74)$$

which can equivalently be written as:

$$\boldsymbol{g}_i(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^{\mathrm{T}}\mathbf{x} + w_{i0}$$

where

$$\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1} \quad \text{and} \quad \mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i \qquad (4.75)$$

and

$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(w_i) \tag{4.76}$$

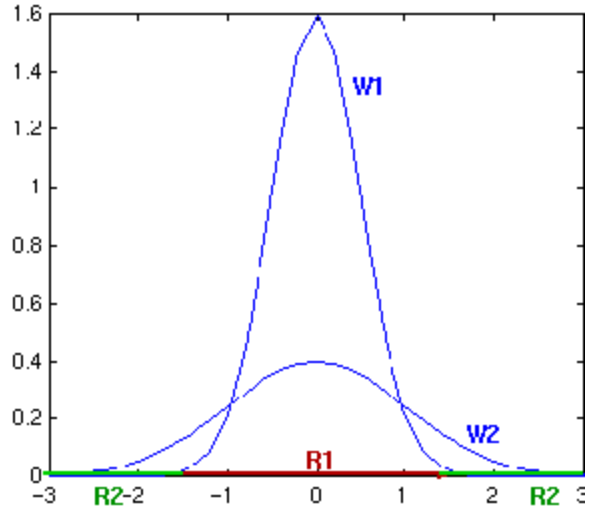


Figure 4.20: Typical single-variable normal distributions showing a disconnected decision region $\mathcal{R}_2$

Because the expression for the $g_i(\mathbf{x})$ has a quadratic term in it, the decision surfaces are no longer linear. Instead, they are *hyperquadratics*, and they can assume any of the general forms: hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, and hyperhyperboloids of various types. The decision regions vary in their shapes and do not need to be connected. Even in one dimension, for arbitrary variance the decision regions need not be simply connected (Figure 4.20). The two-dimensional examples with different decision boundaries are shown in Figure 4.23, Figure 4.24, and in Figure 4.25.

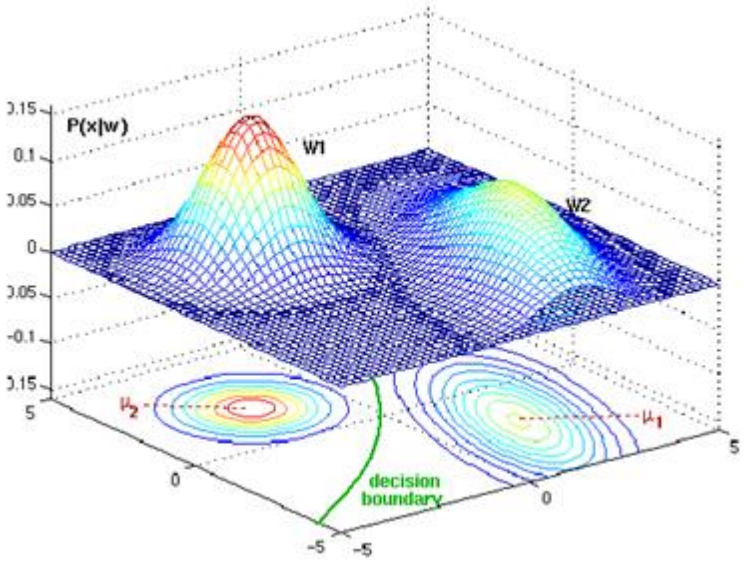


Figure 4.21: Two bivariate normals, with completely different covariance matrix, are showing a hyperquatratic decision boundary.

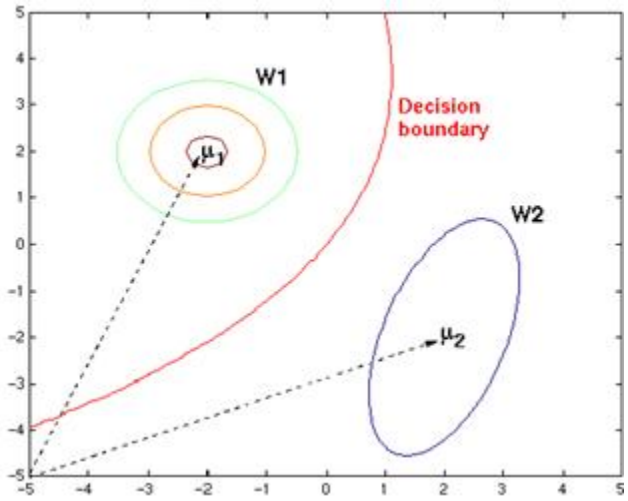


Figure 4.22: The contour lines and decision boundary from Figure 4.21

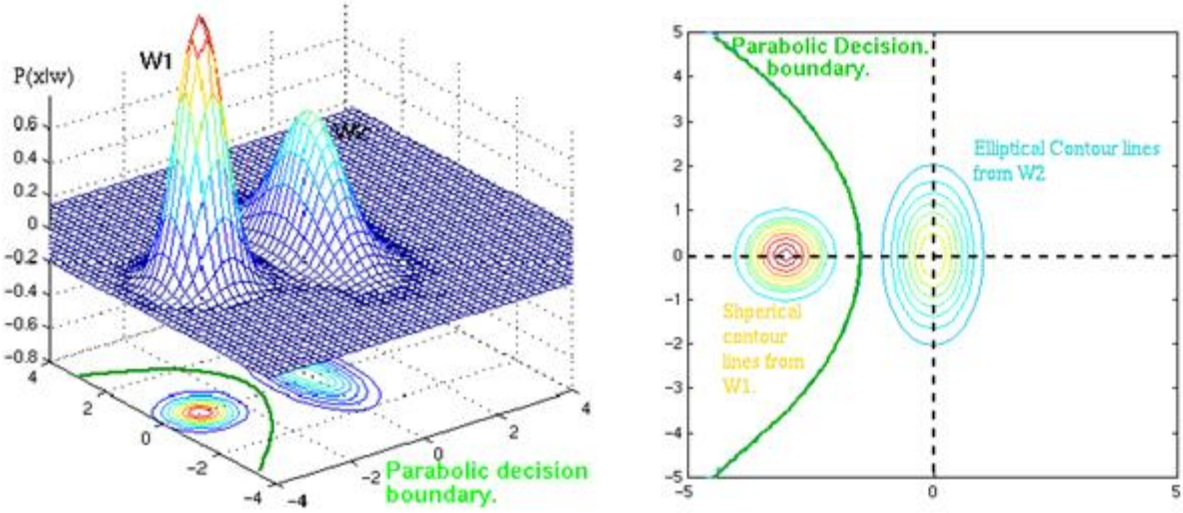


Figure 4.23: Example of parabolic decision surface.

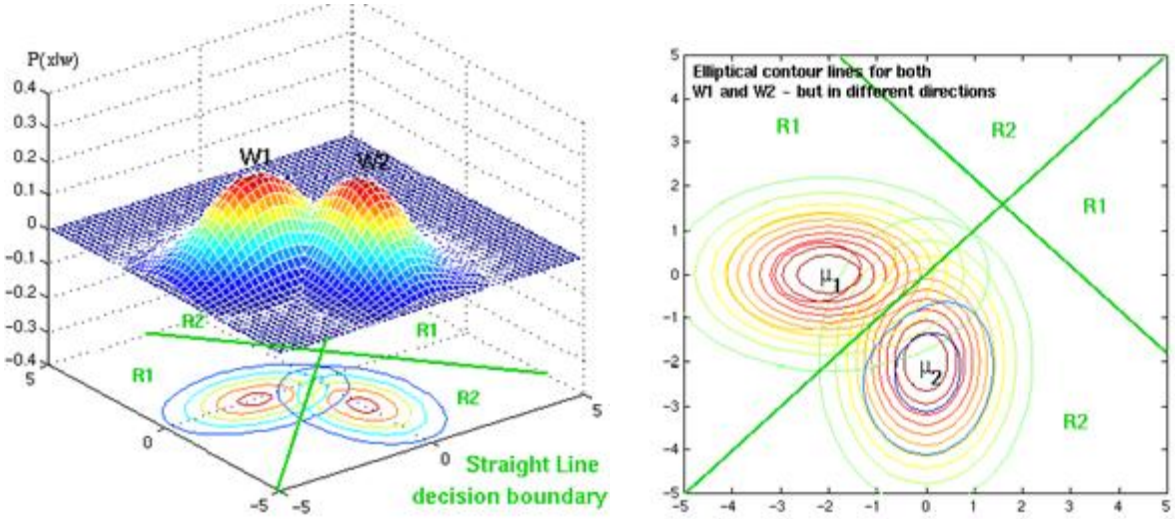


Figure 4.24: Example of straight decision surface.

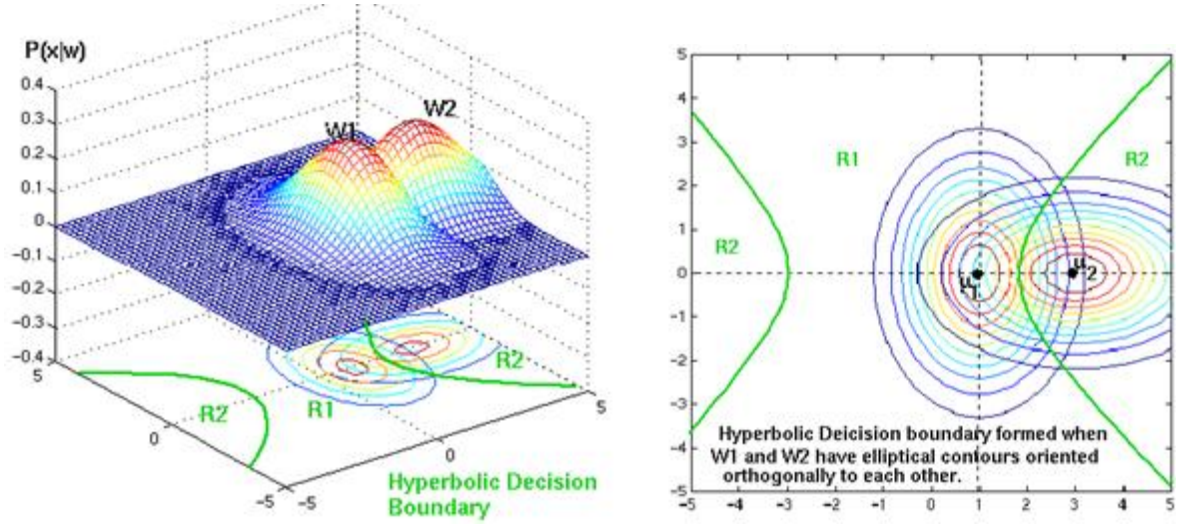


Figure 4.25: Example of hyperbolic decision surface.

4.7 Bayesian Decision Theory (discrete)

In many practical applications, instead of assuming vector **x** as any point in a *d*-dimensional Euclidean space, the components of **x** are binary valued integers, so that **x** can assume only one of *m* discrete values $\mathbf{v}_1,\ldots,\mathbf{v}_m$.

In such cases, the probability density function $p(\mathbf{x}|w_j)$ becomes singular; integrals of the from given by

$$\int p(\mathbf{x}|w_j)\, d\mathbf{x} \tag{4.77}$$

must then be replaced by corresponding sums, such as

$$\sum_{\mathbf{x}} P(\mathbf{x}|w_j) \tag{4.78}$$

where we understand that the summation is over all values of **x** in the discrete distribution. Bayes formula then involves probabilities, rather than probability densities:

$$P(w_j|\mathbf{x}) = \frac{P(\mathbf{x}|w_j)P(w_j)}{P(\mathbf{x})} \tag{4.79}$$

where

$$P(\mathbf{x}) = \sum_{j=1}^{c} P(\mathbf{x}|w_j)P(w_j) \tag{4.80}$$

The definition of conditional risk is unchanged, and the fundamental Bayes decision rule remains the same: To minimize the overall risk, select the action for which is minimum. The basic rule to minimize the error rate by mazimizing the posterior probability is also unchanged as are the discriminant functions.

As an example of a classification involving discrete features, consider two categry case with $\mathbf{x}=(x_1 \ldots x_d)$, where the components x_i are either 0 or 1, and with probabilities

$p_i=\Pr[x_i=1|\, w_1]$ and $q_i=\Pr[x_i=1|\, w_2]$

(4.81)

This is a model of a classification problem in which, each feature gives a yes/no answer about the pattern. If $p_i>q_i$, we expect the i^{th} feature to give a yes answer when the state of nature is w_1. By assuming conditional independence we can write $P(\mathbf{x}|\, w_i)$ as the product of the probabilities for the components of $\mathbf{x}$ as:

$$P(\mathbf{x}|\, w_1)=\prod_{i=1}^{d} p_i^{x_i}(1-p_i)^{1-x_i} \tag{4.82}$$

and

$$P(\mathbf{x}|\, w_2)=\prod_{i=1}^{d} q_i^{x_i}(1-q_i)^{1-x_i} \tag{4.83}$$

Then the likelihood ratio is given by

$$\frac{P(\mathbf{x}|\, w_1)}{P(\mathbf{x}|\, w_2)}=\prod_{i=1}^{d}\left(\frac{p_i}{q_i}\right)^{x_i}\left(\frac{1-p_i}{1-q_i}\right)^{1-x_i} \tag{4.84}$$

and the discriminant function yields

$$g(\mathbf{x})=\sum_{i=1}^{d}\left[x_i \ln\frac{p_i}{q_i}+(1-x_i)\ln\frac{1-p_i}{1-q_i}\right]+\ln\frac{P(w_1)}{P(w_2)} \tag{4.85}$$

The discriminant function is linear and thus can be writeen as

$$g(\mathbf{x})=\sum_{i=1}^{d} w_i x_i + w_0 \tag{4.86}$$

where

$$w_i=\ln\frac{p_i(1-q_i)}{q_i(1-p_i)} \tag{4.87}$$

and

$$w_0=\sum_{i=1}^{d}\ln\frac{1-p_i}{1-q_i}+\ln\frac{P(w_1)}{P(w_2)} \tag{4.88}$$

REFERENCES

[1] *Statistics Toolbox For Use With MATLAB,* User's Guide, (2003) Mathworks Inc., ver.4.0

[2] Prees W. H., Teukolsky S. A., Vetterling W. T., and Flannery B. P. *Numerical Recipies in C*: *The Art Scientific Computing,* User's Guide, (2nd Ed.) Cambridge: Cambridge University Press

[3] Duda, R.O., Hart, P.E., and Stork D.G., (2001). *Pattern Classification*. (2nd ed.). New York: Wiley-Interscience Publication.

[4] Duda, R.O. *Pattern Recognition for Human Computer Interface,* Lecture Notes, web site, http://www-engr.sjsu.edu/~knapp/HCIRODPR/PR-home.htm