DISCRIMINANT FUNCTIONS AND DECISION SURFACES

DISCRIMINANT FUNCTIONS: MULTI-CATEGORY CASE

There are many different ways to represent pattern classifiers. One of the most useful is in terms of a set of discriminant functions $g_i(x) : \mathcal{R}^d \rightarrow \mathcal{R}$, $i = 1, \dots, c$. These functions discriminate between the classes ω_i based upon the feature observations $x \in \mathcal{R}^d$.

The classifier is said to assign a feature observation x to class ω_i if $g_i(x) > g_j(x)$ for all j not equal to i. Ties can be broken arbitrarily. Thus, the classifier is viewed as a network or machine that computes c discriminant functions and selects the category corresponding to the largest discriminant.

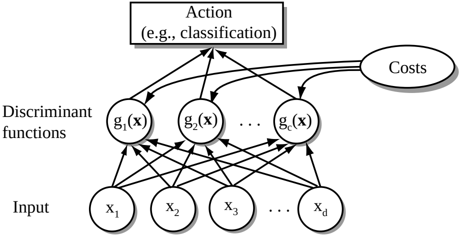

Fig. 1: The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(x)$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly.

DISCRIMINANT FUNCTIONS AND DECISION SURFACES

DISCRIMINANT FUNCTIONS: MULTI-CATEGORY CASE

A Bayes classifier is easily and naturally represented in this way. For the general case with risks, we can let $g_i(x) = -R(\alpha_i|x)$, where

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)p(\omega_j|x),$$

since the maximum discriminant function will then correspond to the minimum conditional risk. Here, $\alpha_1, \ldots, \alpha_n$ represent a series of actions that can be taken. In this expression, $\lambda(\alpha_i|\omega_j)$ represents a loss function that describes the cost associated with taking action α_i given ω_j.

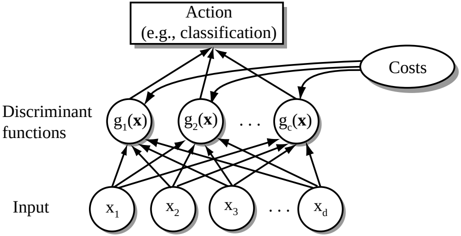

Fig. 1: The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(x)$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly.

DISCRIMINANT FUNCTIONS AND DECISION SURFACES

DISCRIMINANT FUNCTIONS: MULTI-CATEGORY CASE

For the minimum error-rate case, we can simplify things further by taking $g_i(x) = p(\omega_i|x)$, so that the maximum discriminant function corresponds to the maximum posterior probability. (I will explain why this is the minimum error-rate case at the end.)

The maximum posterior probability defines the most likely class ω_j given: (i) the prior assumptions for x (prior) and (ii) the probability of x belonging to ω_j (likelihood function).

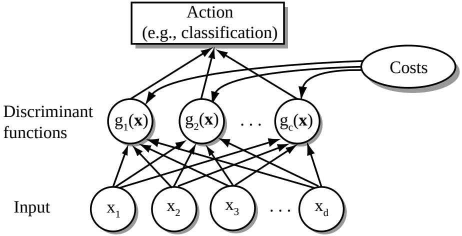

Fig. 1: The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(x)$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly.

DISCRIMINANT FUNCTIONS AND DECISION SURFACES

DISCRIMINANT FUNCTIONS: MULTI-CATEGORY CASE

The choice of discriminant functions is not unique. We can always multiply all the discriminant functions by the same positive constant or shift them by the same additive constant without influencing the decision.

More generally, if we replace every $g_i(x)$ by $f(g_i(x))$, where $f(\cdot)$ is a monotonically increasing function, e.g., the natural log $\ln(\cdot)$, the resulting classification is unchanged. This observation can lead to significant analytical and computational simplifications.

DISCRIMINANT FUNCTIONS AND DECISION SURFACES

DISCRIMINANT FUNCTIONS: MULTI-CATEGORY CASE

For example, 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:

$$g_i(x) = p(\omega_i|x) = \frac{p(x|\omega_i)p(\omega_i)}{\sum_{j=1}^{c} p(x|\omega_j)p(\omega_j)}$$

$$g_i(x) = p(x|\omega_i)p(\omega_i)$$

$$g_i(x) = \ln(p(x|\omega_i)) + \ln(p(\omega_i))$$

Observe that all three describe the posterior probability. We can ignore the common normalizing constant. The last expression comes from taking the log of the second expression and writing out the multiplication as a sum of logarithms. We will use the last expression today, as it allows for us to write a simple discriminant expression for normal densities.

DISCRIMINANT FUNCTIONS AND DECISION SURFACES

DISCRIMINANT FUNCTIONS: MULTI-CATEGORY CASE

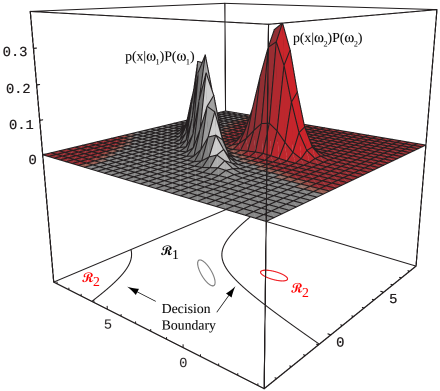

Fig. 2: In this two-dimensional two-category classifier, the probability densities are Gaussian (with $1/\epsilon$ ellipses shown), the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected.

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 regions $\mathcal{R}_1, \ldots, \mathcal{R}_c$. If $g_i(x) > g_j(x)$, for all j not equal to i, then x is in $\mathcal{R}_i$, 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.

DISCRIMINANT FUNCTIONS AND DECISION SURFACES

DISCRIMINANT FUNCTIONS: TWO-CATEGORY CASE

While the two-category case is just a special instance of the multicategory case, or a polychotomizer. A classifier that places a pattern in one of only two categories has a special name: a dichotomizer.

Instead of using two dichotomizer 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

$$g(x) = g_1(x) - g_2(x)$$

and to use the following decision rule: choose class ω_1 if $g(x) > 0$; otherwise choose class ω_2. Thus, 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 $g(x)$.

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

CAVEAT

I am going to just describe the form and properties of these discriminant functions today. I am not going to tell you how to obtain the parameters that define the discriminants. To do this, you can use a variety of methods, the most popular being gradient-ascent-based maximum likelihood estimators.

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

PRELIMINARIES

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

$$g_i(x) = \ln(p(x|\omega_i)) + \ln(p(\omega_i))$$

which are the multiplication of the likelihood and the prior (which forms the posterior). This expression can be readily evaluated if the likelihood densities $p(x|\omega_i)$ are multivariate normal (multivariate Gaussian). That is, if $p(x|\omega_i) \sim \textsc{Normal}(\mu_i, \Sigma_i)$. In this case, we have that

$$g_i(x) = \ln\left(\frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)\right) + \ln(p(\omega_i)).$$

We can expand the first term using properties of the natural log.

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

ISOTROPIC CASE: $\Sigma_i = \sigma^2 I$

The simplest case occurs when the features are statistically independent. The probability of observing one feature $x_i \in \mathcal{R}$ does not influence the probability of other features $x_j \in \mathcal{R}$, i.e., $p(x_1, x_2, \dots, x_d) = p(x_1)p(x_2)\dots p(x_d)$. We also assume that each feature has the same variance $\sigma^2 \in \mathcal{R}$.

In this case the covariance matrix $\Sigma \in \mathcal{R}^{d \times d}$ is diagonal, being merely σ^2 times the identity matrix I. Geometrically, this corresponds to the situation in which the samples fall in equal-size hyperspherical clusters, the cluster for the ith class being centered about the mean vector $\mu_i \in \mathcal{R}^d$.

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

ISOTROPIC CASE: $\Sigma_i = \sigma^2 I$

The computation of the determinant and the inverse of Σ_i is straightforward, since we are working with an identity matrix: $|\Sigma_i| = \sigma^{2d}$ and $\Sigma_i^{-1} = (1/\sigma^2)I$. We can therefore re-write

$$g_i(x) = \ln\left(\frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)^\top\right)\right) + \ln(p(\omega_i))$$

as

$$g_i(x) = -\frac{1}{2\sigma^2}(x-\mu)^\top(x-\mu) + \ln(p(\omega_i)).$$

Observe that we've lost two terms: $-(d/2)\ln(2\pi)$ and $-(1/2)\ln(\det(\Sigma_i))$; we don't even need to compute the determinant. These terms are independent of the index i and will therefore not change the discriminant result (recall that adding or multiplying all discriminant functions by a constant leaves it unchanged).

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

ISOTROPIC CASE: $\Sigma_i = \sigma^2 I$

Surprisingly, it is not necessary to compute the scaled Euclidean distance $(x-\mu)^\top(x-\mu)/\sigma^2$. If we expand this term, we find that the discriminant is

$$g_i(x) = -\frac{1}{2\sigma^2}\left(x^\top x - 2\mu_i^\top x + \mu_i^\top \mu_i\right) + \ln(p(\omega_i))$$

which appears to be a quadratic function of x. However, the quadratic term $x^\top x$ (inner product) is the same for all i, making it an ignorable additive constant.

We thus obtain the equivalent linear discriminant function

$$g_i(x) = w_i^\top x + w_{i,0}$$

where $w_i = \mu_i/\sigma^2$ and with the hyperplane offset in the ith direction given by $w_{i,0} = -\mu_i^\top \mu_i/2\sigma^2 + \ln(p(\omega_i))$.

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

ISOTROPIC CASE: $\Sigma_i = \sigma^2 I$

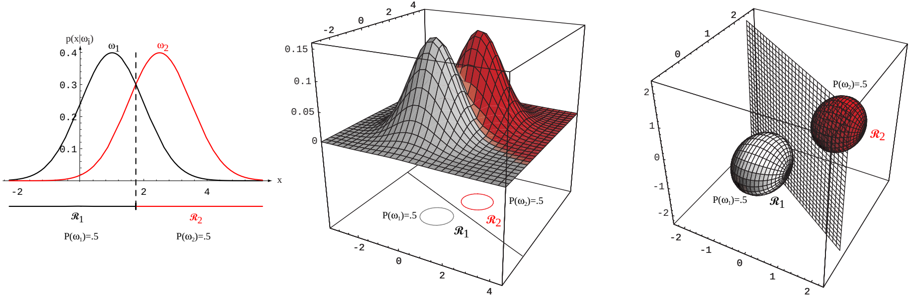

Fig. 3: If the covariances of two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these 1-, 2-, and 3-dimensional examples, we indicate $p(x|\omega_i)$ 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 $\mathcal{R}_2$.

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

ISOTROPIC CASE: $\Sigma_i = \sigma^2 I$

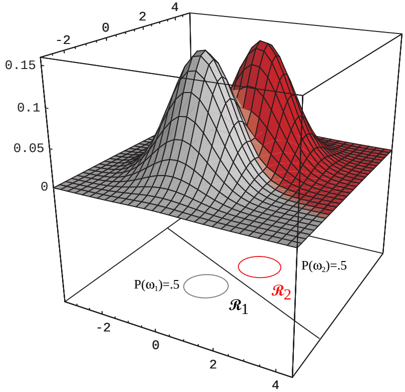

Fig. 3 (continued)

A classifier that uses linear discriminant functions is called a linear classifier. The decision surfaces for a linear classifier are pieces of hyperplanes defined by the linear equations $g_i(x) = g_j(x)$ for the two categories with the highest posterior probabilities. For our particular case, this equation can be written as $w^\top(x - x_0) = 0$, where $w = \mu_i - \mu_j$ and

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

This equation defines a hyperplane through the point x_0 that is orthogonal to the vector w.

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

ISOTROPIC CASE: $\Sigma_i = \sigma^2 I$

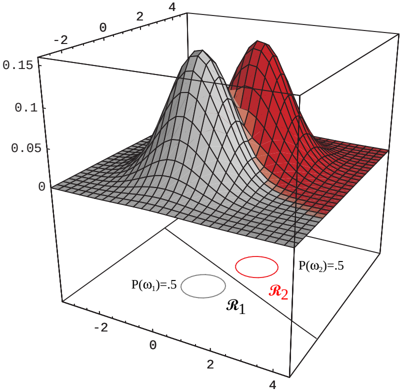

Fig. 3 (continued)

Since $w = \mu_i - \mu_j$, the hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$ is orthogonal to the line linking the means. If $p(\omega_i) = p(\omega_j)$, the second term for x_0 vanishes. Thus, the point x_0 is halfway between the means and the hyperplane is the perpendicular bisector of the line between the means μ_i and μ_j.

Note that if the variance σ^2 is small relative to the distance between μ_i and μ_j, then the position of the decision boundary is relatively insensitive to the exact values of the prior probabilities (the decision boundary should be about halfway between the means).

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

ISOTROPIC CASE: $\Sigma_i = \sigma^2 I$

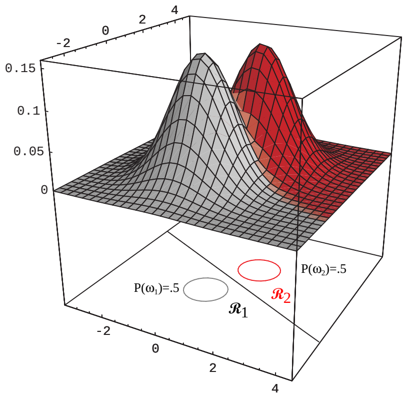

Fig. 3 (continued)

If the prior probabilities $p(\omega_i)$ are the same for all c classes, then the $\ln(p(\omega_i))$ term becomes another unimportant additive constant that can be ignored. The optimum decision rule can be stated very simply: to classify a feature vector x, measure the Euclidean distance $(x - \mu_i)^\top(x - \mu_i)$ from each x to each of the c mean vectors, and assign x to the category of the nearest mean. This is a minimum-distance classifier.

The minimum distance classifier is one of the simplest parametric classifiers (we assume a parametric distribution model for the classes, e.g., Gaussian with means and variances). The nearest neighbor classifier is one of the simplest non-parametric classifiers (we assume either no distribution model or a model with infinite parameters for the classes).

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

ISOTROPIC CASE: $\Sigma_i = \sigma^2 I$

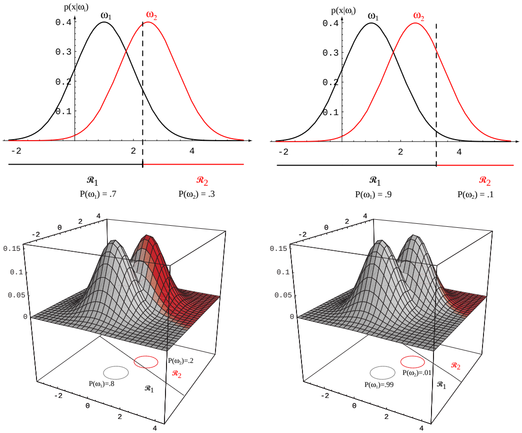

Fig. 4: As the priors are changed, the decision boundary shifts.

If $p(\omega_i)$ does not equal $p(\omega_j)$, then the point x_0 shifts away from the more likely mean. That is, if $p(\omega_i) > p(\omega_j)$, then the point x_0 is further from μ_i and closer to μ_j.

In this case, we want to assign a larger volume to $\mathcal{R}_i$ compared to $\mathcal{R}_j$ to reduce the chance of misclassification for the class that we're most likely to see.

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

EQUAL COVARIANCE CASE: $\Sigma_i = \Sigma$

Another simple case arises when the covariance matrices for all of the classes are identical but otherwise arbitrary. In this case, we have statistical dependence. The probability of observing one feature $x_i \in \mathcal{R}$ does influence the probability of observing other features $x_j \in \mathcal{R}$, i.e., $p(x_1, x_2, \dots, x_d)$ is not equal to $p(x_1)p(x_2) \dots p(x_d)$.

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 is centered about the mean vector $\mu_i \in \mathcal{R}^d$ and has a shape determined by $\Sigma \in \mathcal{R}^{d \times d}$.

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

EQUAL COVARIANCE CASE: $\Sigma_i = \Sigma$

For this case, we can re-write the discriminant

$$g_i(x) = \ln\left(\frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)^\top\right)\right) + \ln(p(\omega_i))$$

as

$$g_i(x) = -\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu) + \ln(p(\omega_i))$$

Like before, we've lost two terms: $-(d/2)\ln(2\pi)$ and $-(1/2)\ln(\det(\Sigma_i))$; we don't even need to compute the determinant, since it is the same for all Σ_i. These terms are independent of the index i and will therefore not change the discriminant result (recall that adding or multiplying all discriminant functions by a constant leaves it unchanged).

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

EQUAL COVARIANCE CASE: $\Sigma_i = \Sigma$

If the prior probabilities $p(\omega_i)$ are the same for all c classes, then the $\ln(p(\omega_i))$ term in

$$g_i(x) = -\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu) + \ln(p(\omega_i))$$

can be ignored. In this case, the optimal decision rule can once again be stated very simply: to classify a feature vector x, we should measure the squared Mahalanobis distance $(x-\mu_i)^\top \Sigma^{-1}(x-\mu_i)$ from x to each of the c mean vectors, and assign x to the category of the nearest mean. This is again a minimum-distance classifier.

As before, unequal prior probabilities bias the decision in favor of the a priori more likely category.

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

EQUAL COVARIANCE CASE: $\Sigma_i = \Sigma$

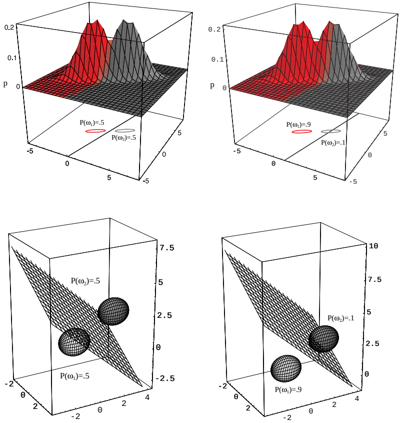

Fig. 5: Probability densities and decision regions for equal but asymmetric Gaussian distributions.

Expansion of the quadratic form $(x - \mu_i)^\top \Sigma^{-1}(x - \mu_i)$ results in a sum involving a quadratic term $x^\top \Sigma^{-1} x$ that is constant for all i; it can be dropped. The resulting discriminant functions are again linear: $g_i(x) = w_i^\top x + w_{i,0}$, where $w_i = \Sigma^{-1}\mu_i$ and the decision boundary offset is $-\mu_i^\top \Sigma^{-1}\mu_i/2 + \ln(p(\omega_i))$. (Without the offset, the decision boundary would just rotate about the origin.)

Since the discriminants are linear, the resulting decision boundaries are again hyperplanes. If the prior probabilities are not equal, the optimal boundary hyperplane is shifted away from the more likely mean.

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

EQUAL COVARIANCE CASE: $\Sigma_i = \Sigma$

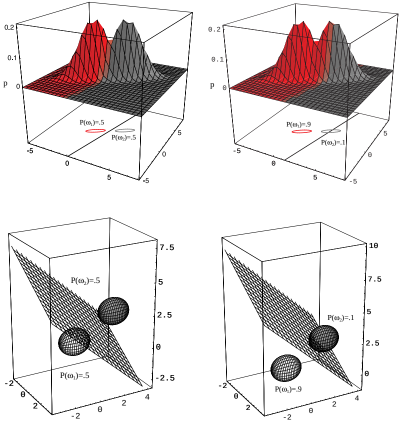

Fig. 5: Probability densities and decision regions for equal but asymmetric Gaussian distributions.

Equating two discriminants together, $g_i(x) = g_j(x)$, we find that the boundary between contiguous regions $\mathcal{R}_i$ and $\mathcal{R}_j$ is given by $w^\top(x - x_0) = 0$, where $w = \Sigma^{-1}(\mu_i - \mu_j)$ and

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln(p(\omega_i)/p(\omega_j))}{(\mu_i - \mu_j)^\top \Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j).$$

Since $w = \Sigma^{-1}(\mu_i - \mu_j)$ is generally not in the direction of $\mu_i - \mu_j$, the hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$ is generally not orthogonal to the line between the means. However, it does intersect that line at the point x_0 which is halfway between the means if the prior probabilities are equal.

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

GENERAL CASE: Σ_i IS ARBITRARY

In the general multivariate normal case, the covariance matrices are different for each of the c classes. The only term that can be dropped from the minimum-error-rate discriminant

$$g_i(x) = \ln\left(\frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)^\top\right)\right) + \ln(p(\omega_i))$$

is the $-(d/2)\ln(2\pi)$ term. The resulting discriminant functions are no longer linear: they are quadratic, which can be seen if we re-write this discriminant as

$$g_i(x) = x^\top W_i x + w_t^\top x + w_{i,0}$$

where $w_i = \Sigma_i^{-1}/2$, $w_i = \Sigma_i^{-1}\mu_i$ and $w_{i,0} = -\mu_i^\top\Sigma_i^{-1}\mu_i/2 - \ln(|\Sigma_i|)/2 + \ln(p(\omega_i))$. The variable $w_{i,0}$ again acts a bias away from the origin.

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

GENERAL CASE: Σ_i IS ARBITRARY

The decision surfaces are hyperquadrics, and can assume any of the general forms: hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, and quadric hyperboloids of various types. Even in one dimension, for arbitrary covariance the decision regions $\mathcal{R}_1, \dots, \mathcal{R}^n$ need not be simply connected: they can be disjoint. The two-dimensional examples to the right indicate how these different forms can arise.

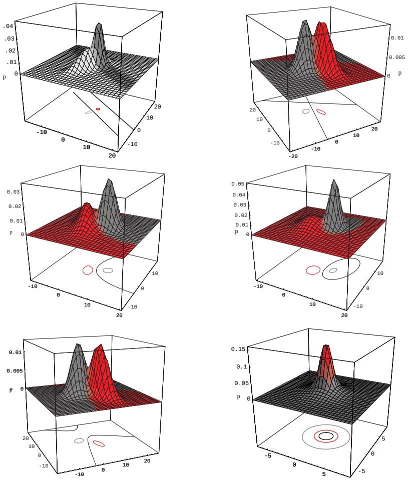

Fig. 6: General Bayes decision boundaries.

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

GENERAL CASE: Σ_i IS ARBITRARY

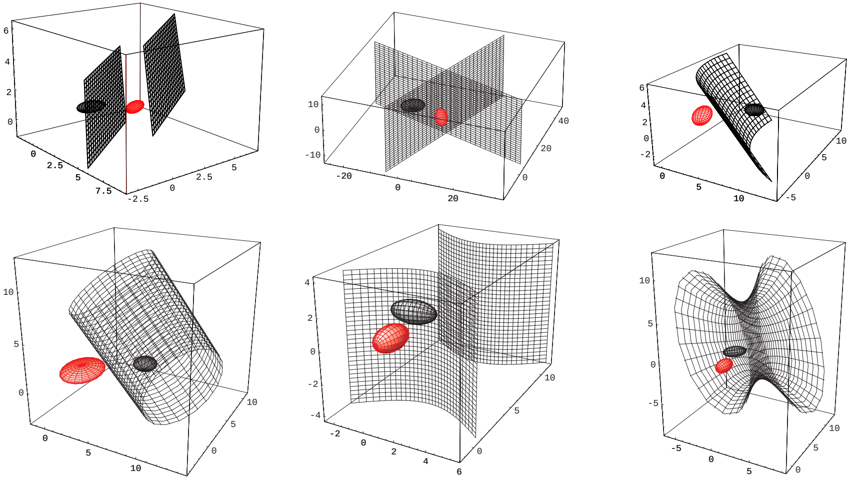

Fig. 7: Arbitrary three-dimensional Gaussian distributions yield Bayes decision boundaries that are two-dimensional hyperquadrics.

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

GENERAL CASE: Σ_i IS ARBITRARY

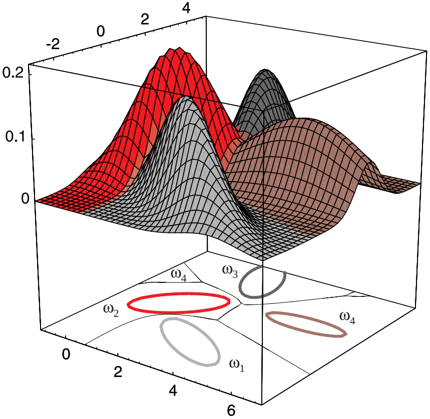

Fig. 8: The decision regions for four normal distributions. The shapes of the boundary regions can be rather complex.

Thus far, we have focused on two classes. The extension of these results to more than two classes is straightforward, though we need to keep clear which two of the total c categories are responsible for any boundary segment. The figure on the right shows the decision surfaces (where $g_i(x) = g_j(x)$) for a four-category case made up of Gaussian distributions. Of course, if the distributions are more complicated, then the decision regions can be even more complex.

ERROR PROBABILITIES

TWO-CLASS CASE

We can obtain insight into the operation of a classifier if we consider the sources of its error. Consider first the two-category case, and suppose the dichotomizer has divided the space into two regions $\mathcal{R}_1$ and $\mathcal{R}_2$ in a possibly non-optimal way.

There are two ways in which a classification error can occur: (i) either an observation x falls in $\mathcal{R}_2$ and the true state of nature is ω_1 or (ii) x falls in $\mathcal{R}_1$ and the true state of nature is ω_2. Since these events are mutually exclusive (one or the other can happen, not both), the probability of error is given by the marginal distribution. Knowing the marginals allows us to infer the posterior, so

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

ERROR PROBABILITIES

TWO-CLASS CASE

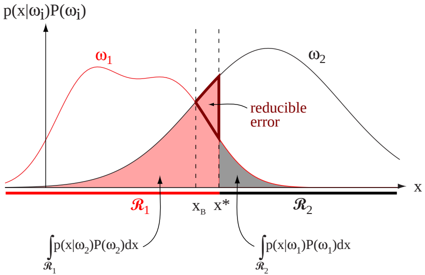

Fig. 9: Components of the probability of error for equal priors and (non-optimal) decision point x^*. The pink area corresponds to the probability of errors for deciding ω_1 when the state of nature is in fact ω_2; the gray area represents the converse.

This result is illustrated in the one-dimensional case. The two integrals represent the pink and the gray areas in the tails of the functions $p(x|\omega_i)p(\omega_i)$. Because the decision point x^* (and hence the regions $\mathcal{R}_1$ and $\mathcal{R}_2$) were chosen arbitrarily for that figure, the probability of error is not as small as it might be. The triangular area marked "reducible error" can be eliminated if the decision boundary is moved from x^* to x_B. This is the Bayes optimal decision boundary and gives the lowest probability of error. In general, if $p(x|\omega_1)p(\omega_1) > p(x|\omega_2)p(\omega_2)$, it is advantageous to classify x as in $\mathcal{R}_1$ so that the smaller quantity will contribute to the error integral; this is exactly what the Bayes decision rule achieves.

ERROR PROBABILITIES
MULTI-CLASS CASE

In the multicategory case, there are more ways to be wrong than to be right, and it is simpler to compute the probability of being correct. Since knowledge of the marginal allows us to obtain the posterior, we have that

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

Bayesian classifiers (choosing class ω_i when $p(\omega_i|x) > p(\omega_j|x)$ for all j) maximizes $p(\text{correct})$. Naive Bayes classifiers (where the features are conditionally independent from each other for the class) also maximizes $p(\text{correct})$ for its assumptions. The latter may perform more poorly than the former, though, since the features may not be independent.

COPYRIGHT NOTICE