Discriminant functions

Hiroshi Shimodaira*

4 March 2015

In the previous chapter we saw how we can combine a Gaussian probability density function with class prior probabilities using Bayes' theorem to estimate class-conditional posterior probabilities. For each point in the input space we can estimate the posterior probability of each class, assigning that point to the class with the maximum posterior probability. We can view this process as dividing the input space into decision regions, separated by decision boundaries. In the next section we investigate whether the maximum posterior probability rule is indeed the best decision rule (in terms of minimising the number of errors). In the following sections we introduce discriminant functions which define the decision boundaries, and investigate the form of decision functions induced by Gaussian pdfs with different constraints on the covariance matrix.

1 Decision boundaries

We may assign each point in the input space as a particular class. This divides the input space into *decision regions* $\mathcal{R}_c$, such that a point falling in $\mathcal{R}_c$ is assigned to class C. In the general case, a decision region $\mathcal{R}_c$ need not be contiguous, but may consist of several disjoint regions each associated with class C. The boundaries between these regions are called *decision boundaries*.

Figure 1 shows the decision regions that result from assigning each point to the class with the maximum posterior probability, using the Gaussians estimated for classes A, B and C from the example in the previous chapter.

1.1 Placement of decision boundaries

Estimating posterior probabilities for each class results in the input space being divided into decision regions, if each point is classified as the class with the highest posterior probability. But is this an optimal placement of decision boundaries?

Consider a 1-dimensional feature space (x) and two classes c_1 and c_2. A reasonable criterion for the placement of decision boundaries is one that *minimises the probability of misclassification*. To estimate the probability of misclassification we need to consider the two ways that a point can be classified wrongly:

1. assigning x to c_1 when it belongs to c_2 (x is in decision region $\mathcal{R}_1$ when it belongs to class c_2);
2. assigning x to c_2 when it belongs to c_1 (x is in $\mathcal{R}_2$ when it belongs to c_1).

*Heavily based on notes inherited from Steve Renals and Iain Murray.

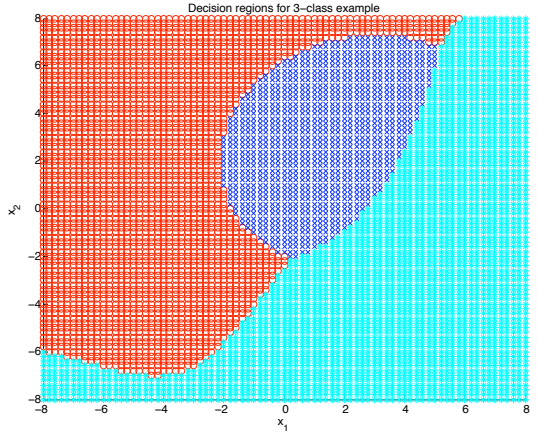

Figure 1: Decision regions for the three-class two-dimensional problem from the previous chapter. Class A (red), class B (blue), class C (cyan).

Thus the probability of the total error may be written as:

$$P(\text{error}) = P(x \in \mathcal{R}_2, c_1) + P(x \in \mathcal{R}_1, c_2)\,.$$

Expanding the terms on the right hand side as conditional probabilities, we may write:

$$P(\text{error}) = P(x \in \mathcal{R}_2 \mid c_1)\,P(c_1) + P(x \in \mathcal{R}_1 \mid c_2)\,P(c_2)\,. \tag{1}$$

1.2 Overlapping Gaussians

Figure 2 illustrates two overlapping Gaussian distributions (assuming equal priors). Two possible decision boundaries are illustrated and the two regions of error are coloured.

We can obtain $P(x \in \mathcal{R}_2 \mid c_1)$ by integrating $p(x|c_1)$ within $\mathcal{R}_2$, and similarly for $P(x \in \mathcal{R}_1 \mid c_2)$, and thus rewrite (1) as:

$$P(\text{error}) = \int_{\mathcal{R}_2} p(x|c_1)\,P(c_1)\,\mathrm{d}x + \int_{\mathcal{R}_1} p(x|c_2)\,P(c_2)\,\mathrm{d}x\,. \tag{2}$$

Minimising the probability of misclassification is equivalent to minimising $P(\text{error})$. From (2) we can see that this is achieved as follows, for a given x:

- if $p(x|c_1)\,P(c_1) > p(x|c_2)\,P(c_2)$, then point x should be in region $\mathcal{R}_1$;
- if $p(x|c_2)\,P(c_2) > p(x|c_1)\,P(c_1)$, then point x should be in region $\mathcal{R}_2$.

The probability of misclassification is thus minimised by assigning each point to the class with the maximum posterior probability.

It is possible to extend this justification for a decision rule based on the maximum posterior probability to d-dimensional feature vectors and K classes. In this case consider the probability of a pattern being

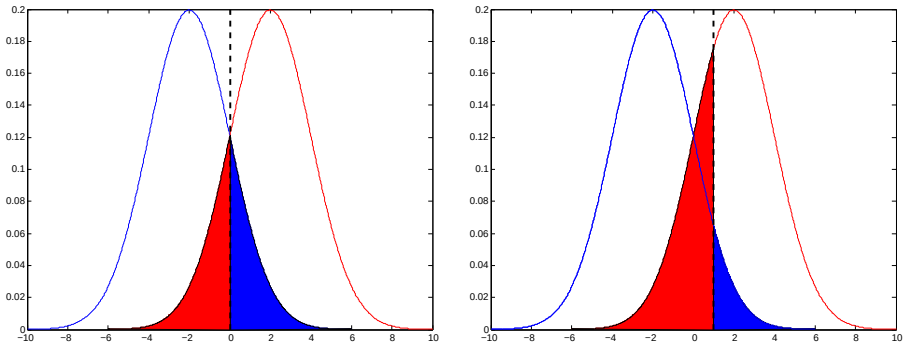

Figure 2: Overlapping Gaussian pdfs. Two possible decision boundaries are shown by the dashed line. The decision boundary on the left hand plot is optimal, assuming equal priors. The overall probability of error is given by the area of the shaded regions under the pdfs.

correctly classified:

$$\begin{aligned} P(\text{correct}) &= \sum_{k=1}^{K} P(\mathbf{x} \in \mathcal{R}_k, c_k) \\ &= \sum_{k=1}^{K} P(\mathbf{x} \in \mathcal{R}_k \mid c_k)\, P(c_k) \\ &= \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathbf{x} \mid c_k)\, P(c_k)\, \mathrm{d}x\,. \end{aligned}$$

This performance measure is maximised by choosing the $\mathcal{R}_k$ such that each $\mathbf{x}$ is assigned to the class k that maximises $p(\mathbf{x}|c_k)\,P(c_k)$. This procedure is equivalent to assigning each $\mathbf{x}$ to the class with the maximum posterior probability.

Thus the maximum posterior probability decision rule is equivalent to minimising the probability of misclassification. However, to obtain this result we assumed both that the class-conditional models are correct, and that the models are well-estimated from the data.

2 Discriminant functions

If we have a set of K classes then we may define a set of K *discriminant functions* $y_k(\mathbf{x})$, one for each class. Data point $\mathbf{x}$ is assigned to class c if

$$y_c(\mathbf{x}) > y_k(\mathbf{x}) \qquad \text{for all } k \neq c.$$

In other words: assign $\mathbf{x}$ to the class c whose discriminant function $y_c(\mathbf{x})$ is biggest.

This is precisely what we did in the previous chapter when classifying based on the values of the log posterior probability. Thus the log posterior probability of class c given a data point $\mathbf{x}$ is a possible discriminant function:

$$y_c(\mathbf{x}) = \ln P(c|\mathbf{x}) = \ln p(\mathbf{x}|c) + \ln P(c) + \text{const.}$$

The posterior probability could also be used as a discriminant function, with the same results: choosing the class with the largest posterior probability is an identical decision rule to choosing the class with the largest log posterior probability.

As discussed above, classifying a point as the class with the largest (log) posterior probability corresponds to the decision rule which minimises the probability of misclassification. In that sense, it forms an optimal discriminant function. A decision boundary occurs at points in the input space where discriminant functions are equal. If the region of input space classified as class c_k ($\mathcal{R}_k$) and the region classified as class c_ℓ ($\mathcal{R}_\ell$) are contiguous, then the decision boundary separating them is given by:

$$y_k(\mathbf{x}) = y_\ell(\mathbf{x})\,.$$

Decision boundaries are not changed by monotonic transformations (such as taking the log) of the discriminant functions.

Formulating a pattern classification problem in terms of discriminant functions is useful since it is possible to estimate the discriminant functions directly from data, without having to estimate probability density functions on the inputs. Direct estimation of the decision boundaries is sometimes referred to as *discriminative* modelling. In contrast, the models that we have considered so far are *generative* models: they could generate new 'fantasy' data by choosing a class label, and then sampling an input from its class-conditional model.

3 Discriminant functions for class-conditional Gaussians

What is the form of the discriminant function when using a Gaussian pdf? As before, we take the discriminant function as the log posterior probability:

$$\begin{aligned} y_c(\mathbf{x}) &= \ln P(c|\mathbf{x}) = \ln p(\mathbf{x}|c) + \ln P(c) + \text{const.} \\ &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_c)^T\boldsymbol{\Sigma}_c^{-1}(\mathbf{x}-\boldsymbol{\mu}_c) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_c| + \ln P(c)\,. \end{aligned} \tag{3}$$

We have dropped the term $-1/2\ln(2\pi)$, since it is a constant that occurs in the discriminant function for each class. The first term on the left hand side of (3) is quadratic in the elements of $\mathbf{x}$ (i.e., if you multiply out the elements, there will be some terms containing x_i^2 or $x_i x_j$).

4 Linear discriminants

Let's consider the case in which the Gaussian pdfs for each class all share the same covariance matrix. That is, for all classes c, $\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}$. In this case $\boldsymbol{\Sigma}$ is class-independent (since it is equal for all classes), therefore the term $-1/2\ln|\boldsymbol{\Sigma}|$ may also be dropped from the discriminant function and we have:

$$y_c(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_c)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_c) + \ln P(c)\,.$$

If we explicitly expand the quadratic matrix-vector expression we obtain the following:

$$y_c(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c - \boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c) + \ln P(c)\,. \tag{4}$$

The mean $\boldsymbol{\mu}_c$ depends on class c, but (as stated before) the covariance matrix is class-independent. Therefore, terms that do not include the mean or the prior probabilities are class independent, and may be dropped. Thus we may drop $\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}$ from the discriminant.

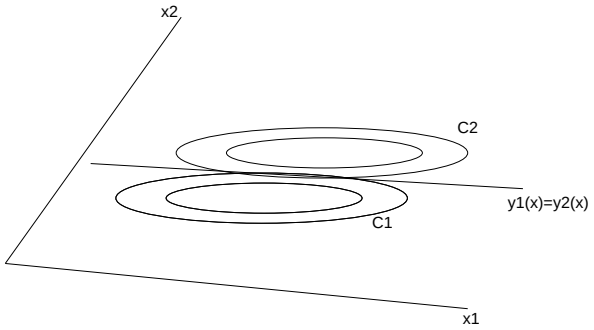

Figure 3: Discriminant function for equal covariance Gaussians

We can simplify this discriminant function further. It is a fact that for a symmetric matrix $\mathbf{M}$ and vectors $\mathbf{a}$ and $\mathbf{b}$:

$$\mathbf{a}^T\mathbf{M}\mathbf{b} = \mathbf{b}^T\mathbf{M}\mathbf{a}\,.$$

Now since the covariance matrix $\mathbf{\Sigma}$ is symmetric, it follows that $\mathbf{\Sigma}^{-1}$ is also symmetric[1]. Therefore:

$$\mathbf{x}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_c = \boldsymbol{\mu}_c^T\mathbf{\Sigma}^{-1}\mathbf{x}\,.$$

We can thus simplify (4) as:

$$y_c(\mathbf{x}) = \boldsymbol{\mu}_c^T\mathbf{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_c^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_c + \ln P(c)\,. \tag{5}$$

This equation has three terms on the right hand side, but only the first depends on $\mathbf{x}$. We can define two new variables $\mathbf{w_c}$ (d-dimension vector) and w_{c0}, which are derived from $\boldsymbol{\mu}_c$, $P(c)$, and $\mathbf{\Sigma}$:

$$\mathbf{w}_c^T = \boldsymbol{\mu}_c^T\mathbf{\Sigma}^{-1} \tag{6}$$

$$w_{c0} = -\frac{1}{2}\boldsymbol{\mu}_c^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_c + \ln P(c) = -\frac{1}{2}\mathbf{w}_c^T\boldsymbol{\mu}_c + \ln P(c)\,. \tag{7}$$

Substituting (6) and (7) into (5) we obtain:

$$y_c(\mathbf{x}) = \mathbf{w}_c^T\mathbf{x} + w_{c0}\,. \tag{8}$$

This is a linear equation in d dimensions. We refer to $\mathbf{w}_c$ as the *weight vector* and w_{c0} as the *bias* for class c.

We have thus shown that the discriminant function for a Gaussian which shares the same covariance matrix with the Gaussians pdfs of all the other classes may be written as (8). We call such discriminant functions *linear discriminants*: they are linear functions of $\mathbf{x}$. If $\mathbf{x}$ is two-dimensional, the decision boundaries will be straight lines, illustrated in Figure 3. In three dimensions the decision boundaries will be planes. In d-dimensions the decision boundaries are called *hyperplanes*.

5 Spherical Gaussians with equal covariance

Let's look at an even more constrained case, where not only do all the classes share a covariance matrix, but that covariance matrix is *spherical*: the off-diagonal terms (covariances) are all zero, and

[1] It also follows that $\mathbf{x}^T\Sigma^{-1}\mathbf{x} \geq 0$ for any $\mathbf{x}$.

the diagonal terms (variances) are equal for all components. In this case the matrix may be defined by a single number, σ^2, the value of the variances:

$$\mathbf{\Sigma} = \sigma^2\mathbf{I}$$

$$\mathbf{\Sigma}^{-1} = \frac{1}{\sigma^2}\mathbf{I}$$

where $\mathbf{I}$ is the identity matrix.

Since this is a special case of Gaussians with equal covariance, the discriminant functions are linear, and may be written as (8). However, we can get another view of the discriminant functions if we write them as:

$$y_c(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_c\|^2}{2\sigma^2} + \ln P(c)\,. \tag{9}$$

If the prior probabilities are equal for all classes, the decision rule simply assigns an unseen vector to the nearest class mean (using the Euclidean distance). In this case the class means may be regarded as class *templates* or *prototypes*.

Exercise: Show that (9) is indeed reduced to a linear discriminant.

6 Two-class linear discriminants

To get some more insights into linear discriminants, we can look at another special case: two-class problems. Two class problems occur quite often in practice, and they are more straightforward to think about because we are considering a single decision boundary between the two classes.

In the two-class case it is possible to use a single discriminant function: for example one which takes value zero at the decision boundary, negative values for one class and positive values for the other. A suitable discriminant function in this case is the log odds (log ratio of posterior probabilities):

$$\begin{aligned} y(\mathbf{x}) &= \ln\frac{P(c_1|\mathbf{x})}{P(c_2|\mathbf{x})} = \ln\frac{p(\mathbf{x}|c_1)}{p(\mathbf{x}|c_2)} + \ln\frac{P(c_1)}{P(c_2)} \\ &= \ln p(\mathbf{x}|c_1) - \ln p(\mathbf{x}|c_2) + \ln P(c_1) - \ln P(c_2)\,. \end{aligned} \tag{10}$$

Feature vector $\mathbf{x}$ is assigned to class c_1 when $y(\mathbf{x}) > 0$; $\mathbf{x}$ is assigned to class c_2 when $y(\mathbf{x}) < 0$. The decision boundary is defined by $y(\mathbf{x}) = 0$.

If the pdf for each class is a Gaussian, and the covariance matrix is shared, then the discriminant function is linear:

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0\,,$$

where $\mathbf{w}$ is a function of the class-dependent means and the class-independent covariance matrix, and the w_0 is a function of the means, the covariance matrix and the prior probabilities.

The decision boundary for the two-class linear discriminant corresponds to a $(d-1)$-dimensional hyperplane in the input space. Let $\mathbf{x}_n a$ and $\mathbf{x}_n b$ be two points on the decision boundary. Then:

$$y(\mathbf{x}_n a) = 0 = y(\mathbf{x}_n b)\,.$$

And since $y(\mathbf{x})$ is a linear discriminant:

$$\mathbf{w}^T\mathbf{x}_n a + w_0 = 0 = \mathbf{w}^T\mathbf{x}_n b + w_0.$$

And a little rearranging gives us:

$$\mathbf{w}^T(\mathbf{x}_n a - \mathbf{x}_n b) = 0\,. \tag{11}$$

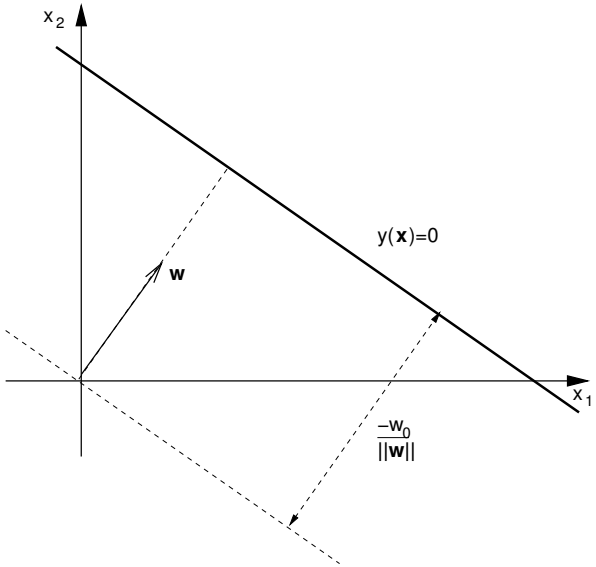

Figure 4: Geometry of a two-class linear discriminant

In three dimensions (11) is the equation of a plane, with $\mathbf{w}$ being the vector normal to the plane. In higher dimensions, this equation describes a hyperplane, and $\mathbf{w}$ is normal to any vector lying on the hyperplane. The hyperplane is the decision boundary in this two-class problem.

If $\mathbf{x}$ is a point on the hyperplane, then the normal distance from the hyperplane to the origin is given by:

$$\ell = \frac{\mathbf{w}^T\mathbf{x}}{\|\mathbf{w}\|} = -\frac{w_0}{\|\mathbf{w}\|} \qquad (\text{using } y(\mathbf{x}) = 0)\,,$$

which is illustrated in Figure 4.