Classifiers, Discriminant Functions, and Decision Surfaces

The Multicategory Case

Suppose we have a set of *discriminant functions* $g_i(\mathbf{x}), \quad i = 1, \cdots, M$.

The classifier is said to assign a feature vector $\mathbf{x}$ to class ω_i if

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

Thus, the classifier is viewed as a network or machine that computes M discriminant functions and selects the category corresponding to the largest discriminant. A network representation of a classifier is illustrated in figure 1 below.

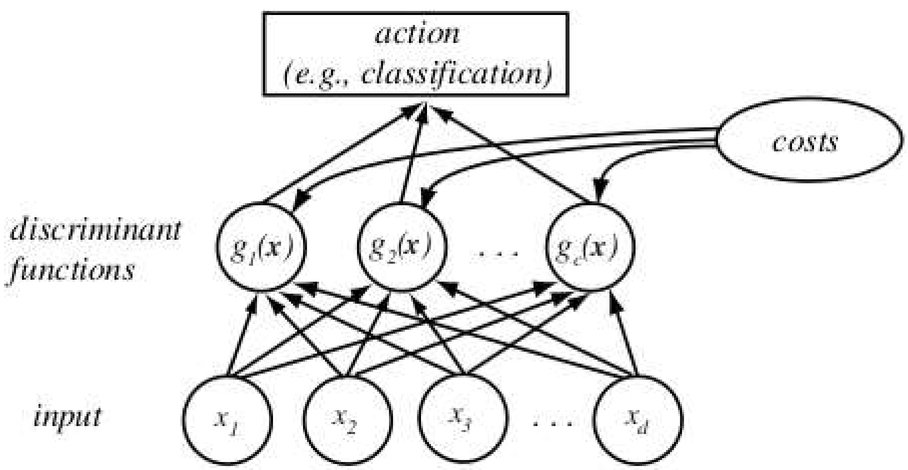

Figure 1. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A

subsequent step determines which of the discriminant values is the maximum, and categorises the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident.

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 $g_i(\mathbf{x}) = P(\omega_i|\mathbf{x})$, so that the maximum discriminant function corresponds to the maximum posterior probability.

Clearly, 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 withought influencing the decision. More generally, if we replace every $g_i(\mathbf{x})$ by $f(g_i(\mathbf{x}))$, where $f(\cdot)$ is monotonically increasing function, the resulting classification is unchanged.

Such observation can lead to significant analytical and computational simplications. 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:

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{i=1}^{M} p(\mathbf{x}|\omega_j)P(\omega_j)}$$
$$g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i)$$
$$g_i(\mathbf{x}) = \log p(\mathbf{x}|\omega_i) + \log P(\omega_i)$$

Note

- Even though the discriminant functions can be written in a variety of forms, their corresponding decision rules are equivalent.
- A classifier that places a pattern in one of only two categories is called a *dichotomiser*. On the other hand, a classifier for more than two categories is called a *polychotomiser*.

The Two-Category Case

Consider the two-category case, a special instance of the multicategory case.

Instead of using two discriminant functions g_1 and g_2 and assigning $\mathbf{x}$ to ω_1 if $g_1 > g_2$, it is more common to define a single discriminant function

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

and to use the following decision rule: Decide ω_1 if $g(\mathbf{x}) > 0$; otherwise decide ω_2.

Thus, a dichotomiser can be viewed as a machine that computes a single discriminant function $g(\mathbf{x})$, and classifies $\mathbf{x}$ according to the algebraic sign of the result.

Of the various forms in which the minimum-error-rate discriminant function can be written, the following two are particularly convenient:

$$g(\mathbf{x}) = P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x})$$
$$g(\mathbf{x}) = \log\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \log\frac{P(\omega_1)}{P(\omega_2)}$$

Def A *decision region* R_i is a set of feature vectors $\mathbf{x}$ such that $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for all $j \neq i$.

Note

- A decision region $\mathcal{R}_i$ need not be simply connected.
- The effect of any decision rule: assign $\mathbf{x}$ to ω_i if $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for all $j \neq i$, is to divide the feature space into M decision regions, $\mathcal{R}_1, \cdots, \mathcal{R}_M$.

Def The *decision boundaries* are surfaces in feature space where ties occur among the largest discriminant functions.

e.g.

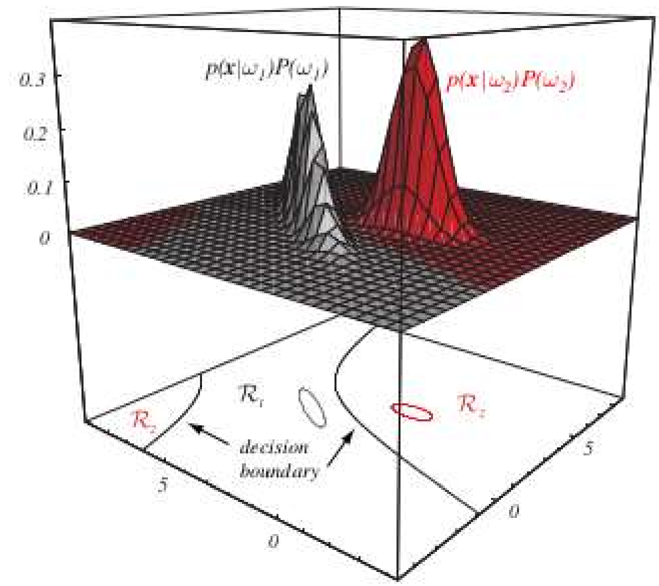

Figure 2. In this two-dimensional tow-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $\frac{1}{e}$ times that at the peak of the distribution.

Note