Class Probabilities

For classification, the response variable, G, takes values in a finite set, $\{1, \ldots, K\}$.

Our aim is to estimate the class probabilities for given inputs, X, written as

$$P(G = k \,|\, X = x), \quad \text{for } k = 1, \ldots, K$$

The most obvious thing to do with these probabilities (if we knew them) is to make a guess at G, via

$$\hat{G}(x) \;\; = \;\; \operatorname*{argmax}_{k=1,\ldots,K} P(G = k \,|\, X = x)$$

But often some errors are worse than others, in which case we should guess by

$$\hat{G}(x) \;\; = \;\; \operatorname*{argmin}_{k=1,\ldots,K} \sum_{k'=1}^{K} L(k, k')\, P(G = k' \,|\, X = x)$$

where $L(k, k')$ is the "loss" from guessing $G = k$ when really $G = k'$.

Another option is to allow a guess of "don't know": eg,

$$\hat{G}(x) \;\; = \;\; \left\{ \begin{array}{cl} k & \text{if } P(G = k \,|\, X = x) > 0.9 \\ \text{“don't know”} & \text{if } P(G = k \,|\, X = x) \leq 0.9 \text{ for all } k \end{array} \right\}$$

Discriminant Functions

Some methods are unable to produce realistic class probabilities. This is bad, since you can't do lots of things you'd usually like to do.

Nevertheless, if you just want to make a guess, all you really need is a set of *discriminant functions*, $\delta_k(x)$. You can then guess according to

$$\hat{G}(x) \quad = \quad \underset{k=1,\ldots,G}{\operatorname{argmax}} \; \delta_k(x)$$

For any classification method that produces class probabilities, we can produce a set of discriminant functions given by $\delta_k(x) = P(G = k \,|\, X = x)$. In fact, any monotonically increasing function of the probabilities will also work.

If there are only two classes, all that matters is $\delta_1(x) - \delta_2(x)$, so we really need only one function.

Similarly, with K classes, we could get away with $K-1$ functions (eg, we might fix $\delta_K(x) = 0$). The resulting asymmetry may distort our thinking, however.

Methods with Linear Discriminant Functions

Chapter 4 in the text looks at classification methods for which the discriminant functions are linear functions of the inputs (plus intercept).

If there are only two classes, the boundary between x where $\hat{G}(x) = 1$ and where $\hat{G}(x) = 2$ is a hyperplane.

With K classes, the regions in the input space where $\hat{G}(x) = k$, for $k = 1, \ldots, K$, are bounded by hyperplanes, but can be quite complex.

Figure 4.11 shows an example with $K = 11$ and two inputs (after reduction).

Three Types of Classification Methods

1) Methods that produce discriminant functions, $\delta_k(x)$, without modeling any probabilities.

2) Methods that model the conditional probability of G given X, and from that obtain discriminant functions of the form

$$\delta_k(x) \quad = \quad g(P(G=k\,|\,X=x))$$

where g is some monotonically increasing function.

3) Methods that model the joint probability of G and X, from that obtain the conditional probability of G given X, and from that obtain discriminant functions. The conditional probability is usually found by Bayes' Rule:

$$P(G=k\,|\,X=x) \quad = \quad \frac{P(G=k)\,P(X=x\,|\,G=k)}{P(X=x)}$$

(modified to use densities when X is continuous).

Estimating $P(G=k)$ is pretty easy — just count how many training cases are in each class. Modeling $P(G=k\,|\,X=x)$ or $P(X=x\,|\,G=g)$ is the hard part.