Classification

Introduction

- Classification is concerned with separating distinct sets of objects (or observations) and with allocating new objects (observations) to previously defined groups.
- The goal of classification is to sort objects (observations) into two or more labeled classes.

Classification for Two Populations

- Label two classes as π_1 and π_2.
- The objects are classified on the basis of measurements on p associated random variables $\mathbf{X}^T=[\mathbf{X}_1, \ldots, \mathbf{X}_p]$. The observed values of $\mathbf{X}$ differ to some extent from one class to the other.
- The probability density functions of $\mathbf{X}$ for π_1 and π_2 are $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$, respectively.
- Classification rules are developed from "learning" samples *known* to come from each of the two populations (similar to phase I control chart). The set of all possible sample outcomes is divided into two regions, R_1 and R_2.
- If a *new* observation falls in R_1, it is allocated to population π_1; and if it falls in R_2, we allocate it to population π_2.

Example 11.1

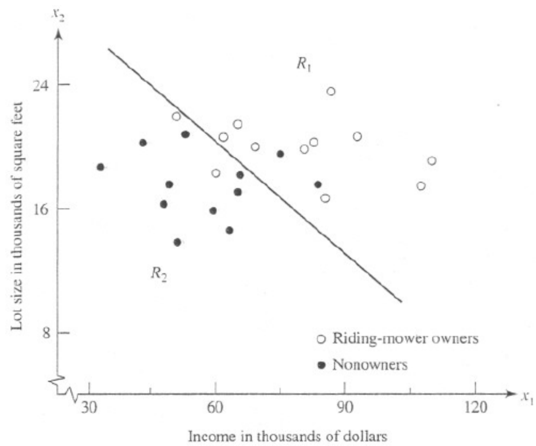
Consider two groups in a city: π_1, riding-mower owners, and π_2, those without riding mowers—that is, nonowners. In order to identify the best sales prospects for an intensive sales campaign, a riding-mower manufacturer is interested in classifying families as prospective owners or nonowners on the basis of x_1 = income and x_2 = lot size. Random samples of $n_1 = 12$ current owners and $n_2 = 12$ current nonowners yield the values in Table 11.1.

TABLE 11.1

π_1: Riding-mower owners		π_2: Nonowners	
x_1 (Income in \$1000s)	x_2 (Lot size in 1000 ft^2)	x_1 (Income in \$1000s)	x_2 (Lot size in 1000 ft^2)
60.0	18.4	75.0	19.6
85.5	16.8	52.8	20.8
64.8	21.6	64.8	17.2
61.5	20.8	43.2	20.4
87.0	23.6	84.0	17.6
110.1	19.2	49.2	17.6
108.0	17.6	59.4	16.0
82.8	22.4	66.0	18.4
69.0	20.0	47.4	16.4
93.0	20.8	33.0	18.8
51.0	22.0	51.0	14.0
81.0	20.0	63.0	14.8

Example 11.1 (Cont.)

- Some overlap between the two groups, resulting in misclassification errors..
- A good classification procedure should result in few misclassifications.



Expected Cost of Misclassification (ECM)

Expected cost of misclassification (ECM) is

$ECM=c(2|1)P(2|1)p_1+c(1|2)P(1|2)p_2$

where p_1 is the *prior* probability of π_1 and p_2 be the *prior* probability of π_2, with $p_1+p_2=1$.

Regions Minimizing ECM

- The regions R_1 and R_2 that minimize the ECM are defined by the values **x** for the following inequalities hold:

$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

- Special cases:
  - $p_2=p_1$
  - $c(1|2)=c(2|1)$
- When the prior probabilities are unknown, they are often taken to be equal. If the misclassification cost ratio is indeterminate, it is usually taken to be unity.

Example 11.2

Suppose c(2|1)=5 and c(1|2)=10. Also, 20% of all objects belong to group 2. Derive the classification regions to minimize ECM.