Chapter 11. Discrimination and Classification.

Suppose we have a number of multivariate observations coming from two populations, just as in the standard two-sample problem. Very often we wish to, by taking advantage of the characteristics of the distribution of the observations from each population, derive a reasonable graphical or algebraic rules to separate the two population. These rules would be very useful when a new observation comes from one of the two populations and we wish to identify exactly which population. Examples 11.1 and 11.2 provide an illustration of the scenario.

Example 11.1 (OWNERS/NONOWNERS OF RIDING MOWERS.) (See Appendix D). The data set consists of lot/lawn size of income of 12 households that are riding mower owners and those of 12 households that are non-owners of riding mowers. Suppose now you are a salesman/saleswoman selling a new model of riding mower to the community. You figure that the potential of a household buying a riding mower solely depends on the income of the household and the size of their lot/lawn. It is then important to figure out what kind of households would be more likely to be a potential buyer. The given data set can be very helpful to draw characteristics of potential buyers/nonbuyers. Should a reasonable clear-cut rule to classify a given household, especially those newly move-in, into potential owners/nonowners, it shall be useful to develop an efficient business strategy of targeting appropriate clientele.

11.1 Population Classification.

Consider that the entire population consists of two sub-populations, denoted by π_1 and π_2. The percentage of π_1 (π_2) in the entire population, which is called *prior probability*, is p_1 (p_2). Obviously $p_1 + p_2 = 1$. Suppose a random variable coming from π_1 has density f_1 in p dimensional real space. In short, π_1 has density f_1. Likewise, let π_2 has density f_2.

Suppose X comes out of the entire population, namely one of the two sub-populations. (Note that the density of X is $p_1 f_1(\cdot) + p_2 f_2(\cdot)$.) A classification/seperation rule is defined by the classification region: R_1 and R_2 in the p-dimensional real space that complement each other. Then the rule is: If the value of X is in R_1, we classify it as from π_1; if the value of X is in R_2, we classify it as from π_2.

There are several criteria to measure whether a classification rule is good or not. The most straightforward one is to consider the *misclassification probabilities*:

$$P(1|2) = P(\text{classified as from } \pi_1 \ |\pi_2) = P(X \in R_1|\pi_2) = \int_{R_1} f_2(x)dx,$$

which is the chance that, given the observation/subject actually comes from π_2, it is misclassified as from π_1. Analogously,

$$P(2|1) = P(\text{classified as from } \pi_2 \ |\pi_1) = P(X \in R_2|\pi_1) = \int_{R_2} f_1(x)dx,$$

is the chance that, given the observation/subject actually comes from π_1, it is misclassified as from π_2. We can present the classification probabilities in the following table:

		Classified as	
		π_1	π_2
True population:	π_1	$P(1\|1)$	$P(2\|1)$
	π_2	$P(1\|2)$	$P(2\|2)$

Note that $P(1|1)$ and $P(2|2)$ are analogously defined, but they are not misclassification probabilities.

It is common that the a misclassification is considered a mistake that bears a penalty or cost. The two types of misclassifications: classify a subject as from π_2 but it actually comes from π_1 or classify a subject as from π_1 but it actually comes from π_2, often have different level of seriousness and deserves different penalty or cost. For example, in a deadly pandemic, misjudged a virus carrier as non-carrier is much severe a mistake than misjudged a non-carrier as carrier, since the former may pose grave hazard to public health. The misclassification costs is presented in the following table:

		Classified as	
		π_1	π_2
True population:	π_1	0	$C(2\|1)$
	π_2	$C(1\|2)$	0

where $C(1|2) > 0$ ($C(2|1) > 0$) are the costs or prices to pay if a subject from population π_2 (π_1) is misclassified as from π_1 (π_2).

The *expected cost of misclassification (ECM)* is

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

A criterion of classification rule is then to minimize the ECM. Note that, if $C(2|1) = C(1|2)$, then minimizing ECM is the same as minimizing *total probability of misclassification (TPM)*, which is

$$\begin{aligned} TPM &= p_1 P(1|2) + p_2 P(1|2) = p_1 \int_{R_2} f_1(x)dx + p_2 \int_{R_1} f_2(x)dx \\ &= ECM \text{ with } C(2|1) = C(1|2) = 1. \end{aligned}$$

11.2 An optimal classification rule.

Given the misclassification costs $C(1|2)$ and $C(2|1)$, the prior probabilities p_1, p_2 and the densities $f_1(\cdot)$ of π_1 and $f_2(\cdot)$ of π_2, a likelihood ratio based classification rule is the optimal in the sense that it minimizes the ECM. This is the result of the following theorem.

Theorem 11.1 *Let*

$$\begin{aligned} R_1 &= \{x : \frac{f_1(x)}{f_2(x)} \geq \frac{C(1|2)}{C(2|1)} \frac{p_2}{p_1}\} \\ R_2 &= \{x : \frac{f_1(x)}{f_2(x)} < \frac{C(1|2)}{C(2|1)} \frac{p_2}{p_1}\} = R_1^c \quad \textit{(the complement of } R_1\textit{)} \end{aligned}$$

Then, the classification by (R_1, R_2) *minimizes ECM.*

Proof. For any classification rule defined by regions (R_1^*, R_2^*) with $R_2^* = R_1^{*c}$,

$$\begin{aligned} ECM &= C(2|1)P(2|1)p_1 + C(1|2)P(1|2)p_2 \\ &= \int_{R_2^*} C(2|1)p_1 f_1(x)dx + \int_{R_2^*} C(1|2)p_2 f_2(x)dx \\ &= \int \Big[C(2|1)p_1 f_1(x) 1_{\{x \in R_2^*\}} + C(1|2)p_2 f_2(x) 1_{\{x \in R_1^*\}}\Big] dx \\ &\geq \int \min\Big[C(2|1)p_1 f_1(x),\ C(1|2)p_2 f_2(x)\Big] dx \end{aligned}$$

where the last inequality becomes equality when $R_1^* = \{x : C(2|1)p_1 f_1(x) \geq C(1|2)p_2 f_2(x)\} = R_1$.
□

The above theorem implies that when the observation takes a value, which is more likely to be from π_1 than from π_2—measured by the likelihood ratio, then we should classify that as from π_1. The threshold $C(1|2)p_2/C(2|1)p_1$ is related with the costs of each type of misclassification as well as the prior probabilities. For example, the misclassify a subject from π_1 as from π_2 is a more severe mistake than the other type of misclassification, then the threshold should be lowered, allowing more to be classified as π_1. If p_1 is much larger than p_2, meaning that π_1 is much larger population than π_2, then the threshold should also be lowered.

Assume the population distributions of π_1 and π_2 are both normal:

$$\pi_1: \quad f_1 \sim MN(\mu_1, \Sigma_1) \qquad\qquad \pi_2: \quad f_2 \sim MN(\mu_2, \Sigma_2).$$

And suppose we have one sample $x_{11}, ..., x_{n_1 1}$ from π_1 and another sample $x_{12}, ..., x_{n_2 2}$ from π_2. The above likelihood ratio based optimal classification rule can be expressed more explicitly.

(1). Assume equal variances, i.e., $\Sigma_1 = \Sigma_2 = \Sigma$. By straightforward calculation of the likelihood ratio, the optimal classification rule is

$$\begin{cases} R_1: & (\mu_1 - \mu_2)'\Sigma^{-1}x - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq \log\left[\frac{C(1|2)p_2}{C(2|1)p_1}\right] \\ R_2: & R_1^c. \end{cases}$$

This rule is useful only when μ_1, μ_2 and Σ are known. In practice, they are unknown and are estimated by $\bar{X}_1, \bar{X}_2$ and $\mathbf{S}_{\text{pooled}}$, which are sample means and the pooled estimator of the population variance. Then the sample analogue of the above theoretical optimal classification rule is, by replacing μ_1, μ_2 and Σ by their estimators,

$$\begin{cases} R_1: & (\bar{X}_1 - \bar{X}_2)'\mathbf{S}_{\text{pooled}}^{-1}x - \frac{1}{2}(\bar{X}_1 - \bar{X}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{X}_1 + \bar{X}_2) \geq \log\left[\frac{C(1|2)p_2}{C(2|1)p_1}\right] \\ R_2: & R_1^c. \end{cases}$$

Note that R_1 and R_2 are separated by a hyperplane, and therefore may be regarded as half spaces.

(2). Unequal variances.($\Sigma_1 \neq \Sigma_2$)

The population optimal classification rule is

$$\begin{cases} R_1: & -\frac{1}{2}x'(\Sigma_1^{-1} - \Sigma_2^{-1})x + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})x - k \geq \log\left[\frac{C(1|2)p_2}{C(2|1)p_1}\right] \\ R_2: & R_1^c. \end{cases}$$

where

$$k = \frac{1}{2}\log\Big(\frac{|\Sigma_1|}{|\Sigma_2|}\Big) + \frac{1}{2}(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2).$$

Then the sample analogue is

$$\begin{cases} R_1: & -\frac{1}{2}x'(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})x + (\bar{X}_1'\mathbf{S}_1^{-1} - \bar{X}_2'\mathbf{S}_2^{-1})x - \hat{k} \geq \log\left[\frac{C(1|2)p_2}{C(2|1)p_1}\right] \\ R_2: & R_1^c. \end{cases}$$

where

$$\hat{k} = \frac{1}{2}\log\Big(\frac{|\mathbf{S}_1|}{|\mathbf{S}_2|}\Big) + \frac{1}{2}(\bar{X}_1'\mathbf{S}_1^{-1}\bar{X}_1 - \bar{X}_2'\mathbf{S}_2^{-1}\bar{X}_2).$$

Associated with the theoretical or sample classification rules, there are several quantities that can be considered as criteria to measure the rules.

The *optimal error rate* (OER) is the minimum total probability of misclassification over all classification rules.

$$\begin{aligned} OER &\equiv \text{minimum } TPM = \text{ minimum total probability of misclassification} \\ &= \min_{(R_1, R_2)} \Big[p_1 \int_{R_2} f_2(x)dx + p_2 \int_{R_1} f_1(x)dx\Big] \end{aligned}$$

The *actual error rate* (AER) is the total probability of misclassification for a given classification rule, say $(\hat{R}_1, \hat{R}_2)$, which is usually constructed based on given data.

$$AER \equiv TPM \text{ for } (\hat{R}_1, \hat{R}_2) = p_1 \int_{\hat{R}_2} f_2(x)dx + p_2 \int_{\hat{R}_1} f_1(x)dx$$

The *confusion matrix* listed below presents, for a given classification rule, say $(\hat{R}_1, \hat{R}_2)$, the number of correct and mistaken classified observations in the data when this rule is applied to the subjects in the dataset.

		Predicted membership		
		π_1	π_2	
Actual membership	π_1	n_{1C}	$n_{1M} = n_1 - n_{1C}$	n_1
	π_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2

where n_1 (n_2) are number of observations from π_1 (π_2) in the dataset, n_{1C} (n_{2C}) is the number of subjects from π_1 (π_2) and correctly classified as from π_1 (π_2) by this rule, and n_{1M} (n_{2M}) is the number of subjects from π_1 (π_2) but mistakenly classified as from π_2 (π_1) by this rule,

The *apparent error rate* (APER) is an estimator of the a given rule $(\hat{R}_1, \hat{R}_2)$:

$$APER \equiv \frac{n_{1M} + n_{2M}}{n_1 + n_2}.$$

The confusion matrix and APER is easily available from the dataset once a classification rule is given, and they are commonly used to justify whether a rule is good or not.

11.3 Examples.

We shall only apply the sample optimal classification rule with the two population distribution assumed following normal distributions with equal varaince. We summarize the results as:

Population:	π_1	π_2
Data:	x_{11}	x_{12}
	$\vdots$	$\vdots$
	$x_{n_1 1}$	$x_{n_2 2}$
Sample means:	$\bar{X}_1$	$\bar{X}_2$
Sample variances	$\mathbf{S}_1$	$\mathbf{S}_2$

Under $\Sigma_1 = \Sigma_2$, the pooled estimator of $\Sigma_1 = \Sigma_2 = \Sigma$ is

$$\mathbf{S}_{\text{pooled}} = \frac{(n-1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}.$$

The classification rule is

$$R_1 = \Big\{x : \hat{\mathbf{a}}' x \geq \frac{1}{2}(\bar{X}_1 - \bar{X}_2)' \mathbf{S}_{\text{pooled}}^{-1} (\bar{X}_1 + \bar{X}_2) + \log\Big[\frac{C(1|2)p_2}{C(2|1)p_1}\Big]$$

where $\hat{\mathbf{a}} = \mathbf{S}_{\text{poooled}}^{-1}(\bar{X}_1 - \bar{X}_2)$.

Example 11.1 (Owners/Nonowners of riding mower) See Appendix D for the data set.

$n_1 = n_2 = 12$. π_1: owners; and π_2 non-owners. And

$$\bar{X}_1 = \begin{pmatrix} 79.5 \\ 20.27 \end{pmatrix} \ \mathbf{S}_1 = \begin{pmatrix} 32.06 & -1.08 \\ -1.08 & 0.38 \end{pmatrix} \ \bar{X}_2 = \begin{pmatrix} 57.4 \\ 17.6 \end{pmatrix} \ \mathbf{S}_2 = \begin{pmatrix} 18.25 & -.24 \\ -.24 & 0.41 \end{pmatrix}.$$

Then,

$$\mathbf{S}_{\text{pooled}} = \begin{pmatrix} 25.15 & -.65 \\ -.65 & .39 \end{pmatrix} \ \mathbf{S}_{\text{pooled}}^{-1} = \begin{pmatrix} .04 & .07 \\ .07 & 2.69 \end{pmatrix}$$

and

$$\hat{\mathbf{a}} = \mathbf{S}_{\text{pooled}}^{-1}(\bar{X}_1 - \bar{X}_2) = \begin{pmatrix} 1.10 \\ 8.64 \end{pmatrix}, \qquad \frac{1}{2}(\bar{X}_1 - \bar{X}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{X}_1 + \bar{X}_2) = 239.12$$

Suppose $x = (x_1, x_2)$ is a new observation with x_1 as income and x_2 as size of lot. We classify it as from π_1, the owners sub-population, if

$$1.10x_1 + 8.64x_2 \geq 239.12 + \log\Big[\frac{C(1|2)p_2}{C(2|1)p_1}\Big]$$

and as from π_2, the nonowners sub-population, otherwise.

Case 1: Suppose $p_1 = p_2$ and $C(1|2) = C(2|1)$ (equal costs of two types of mistakes). Then,

$$R_1 = \{(x_1, x_2) : 1.10x_1 + 8.64x_2 \geq 239.12\}$$

And

$$APER = \frac{1+2}{24} = 1/8 = 0.125$$

Case 2: Suppose $p_1 = p_2$ and $C(1|2) = 50C(2|1)$ (Classifying a non-owner as owner is a 50 times more severe a mistake than the other type mistake). Then,

$$R_1 = \{(x_1, x_2) : 1.10x_1 + 8.64x_2 \geq 243.03\}$$

And

$$APER = \frac{2+2}{24} = 1/6 = 0.167$$

Case 3: Suppose $p_1 = p_2$ and $C(2|1) = 50C(1|2)$ (Classifying an owner as non-owner is 50 times more severe a mistake than the other type mistake). Then,

$$R_1 = \{(x_1, x_2) : 1.10x_1 + 8.64x_2 \geq 235.21\}$$

And

$$APER = \frac{3+1}{24} = 1/6 = 0.167.$$

See Appendix D for plots.

Example 11.2 (Discrimination analysis of hemophilia data) See Appendix D.

Hemophilia is an abnormal condition of males inherited from the mother, characterized by a tendency to bleed excessively. Whether is person is a hemophilia carrier of non-carrier is reflected in the two indices of anti-hemophilia factor (AHF) antigen and AHF activity.

$n_1 = 30$, $n_2 = 45$. π_1: Noncarriers; π_2: Obligatory carriers. By simple calculation, we have And

$$\bar{X}_1 = \begin{pmatrix} -.1349 \\ -.0778 \end{pmatrix} \ \mathbf{S}_1 = \begin{pmatrix} .0209 & .0155 \\ .0155 & .0179 \end{pmatrix} \ \bar{X}_2 = \begin{pmatrix} -.3079 \\ -.0060 \end{pmatrix} \ \mathbf{S}_2 = \begin{pmatrix} .0238 & .0153 \\ .0153 & .0240 \end{pmatrix}.$$

Then,

$$\mathbf{S}_{\text{pooled}} = \begin{pmatrix} .0226 & .0154 \\ .0154 & .0216 \end{pmatrix} \ \mathbf{S}_{\text{pooled}}^{-1} = \begin{pmatrix} 86.09 & -61.49 \\ -61.49 & 90.20 \end{pmatrix}$$

and

$$\hat{\mathbf{a}} = \mathbf{S}_{\text{pooled}}^{-1}(\bar{X}_1 - \bar{X}_2) = \begin{pmatrix} 19.319 \\ -17.124 \end{pmatrix}, \qquad \frac{1}{2}(\bar{X}_1 - \bar{X}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{X}_1 + \bar{X}_2) = -3.559$$

Suppose $x = (x_1, x_2)'$ is a new observation with x_1 as $\log_{10}$(AHFactivity) and x_2 as $\log_{10}$(AHFantigen). We classify it as from π_1, the noncarriers sub-population, if

$$19.319x_1 - 17.124x_2 \geq -3.559 + \log\Big[\frac{C(1|2)p_2}{C(2|1)p_1}\Big]$$

and as from π_2, the obligatory carriers sub-population, otherwise.

Case 1: Suppose $p_1 = p_2$ and $C(1|2) = C(2|1)$ (equal costs of two types of mistakes). Then,

$$R_1 = \{(x_1, x_2) : 19.319x_1 - 17.124x_2 \geq -3.559\}.$$

The confusion matrix is

		Predicted membership π_1	π_2	
Actual	π_1	27	3	30
membership	π_2	8	37	45

And

$$APER = \frac{3+8}{75} = 11/75 = 14.67\%$$

Case 2: Suppose $p_1 = p_2$ and $C(1|2) = 10C(2|1)$ (Classifying an obligatory carrier as noncarrier is a 10 times more severe a mistake than the other type mistake, which indeed makes sense). Then,

$$R_1 = \{(x_1, x_2) : 19.319x_1 - 17.124x_2 \geq -1.296\}.$$

The confusion matrix is

		Predicted membership π_1	π_2	
Actual	π_1	17	13	30
membership	π_2	0	45	45

And

$$APER = \frac{0+13}{75} = 17.33\%$$

Case 3: Suppose $p_1 = 100p_2$ and $C(1|2) = 10C(2|1)$ (Same penalties as in Case 2, but assuming, perhaps more realistically, only 1/101 of the entire population are obligatory carries). Then,

$$R_1 = \{(x_1, x_2) : 19.319x_1 - 17.124x_2 \geq -5.862\}.$$

The confusion matrix is

		Predicted membership π_1	π_2	
Actual	π_1	30	0	30
membership	π_2	23	22	45

And

$$APER = \frac{23+0}{75} = 30.7\%.$$

See Appendix D for plots.