Chapter 11 - Discrimination and Classification

Discrimination - Describe graphically or algebraically differentiate individual cases (observations) by population group. "Discriminants" are numerical values that seperate groups as much as possible. (Separation)

Classification - Separate cases into labelled groups, with goal being to create algorithm to classify new cases into their true class. (Allocation)

11.2 Separation/Classification of 2 Populations

Notations: Populations $\equiv \pi_1, \pi_2$

Meaured variables $\equiv \underset{\sim}{X}$

Case where some Groups are known/ others are not.

- Incomplete knowledge of final outcome (Financial fate of firms, medical school outcomes)
- Destruction of object needed for final result (Strength of lumber/fabric). Classify based on prelim measurements.
- Unavailable or perfect info - Federalist papers, Diagnostic Testing/Gold Standard

Example 1 Classification of NBA/WNBA players by ~~Height/Weight~~ Body Mass Index

$N_W = 139$, $\mu_W = 23.135$, $\sigma_W = 2.105$ $N_M = 505$, $\mu_M = 24.741$, $\sigma_M = 1.720$

Both Dists approx Normal

$$P(W) = \frac{139}{139+505} = .2158 \quad P(M) = \frac{505}{644} = .7842$$

See ppt presentation

Example 2 - Heights/Weights of $\overset{\text{"H"}}{\text{NHL}}$ / $\overset{\text{"S"}}{\text{EPL}}$ Players

$n_H = n_S = 100$ Heights add $U(-\frac{1}{2}, \frac{1}{2})$ to "continuize"

See R Program/Plot

- Prior probabilities of 2 population. If one population has higher prior probability, tend to classify individuals into that population unless strong evidence ow.
- Cost of misclassification. Is cost of misclassifying object in π_1 as being π_2 than misclassifying object in π_2 as being π_1?
- Probability density function of $\underset{\sim}{X}$ for each population

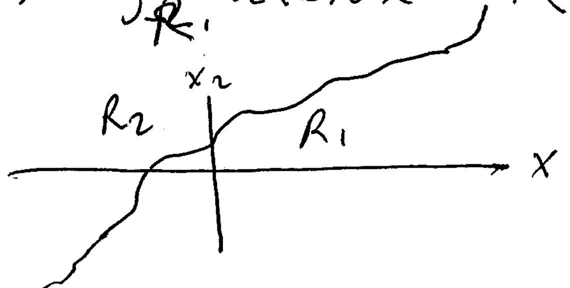
  Population 1: $f_1(\underset{\sim}{x})$ Population 2: $f_2(\underset{\sim}{x})$
- Observation with given $\underset{\sim}{x}$ must be assigned to either π_1 or π_2.
  $R_1 \equiv \{\underset{\sim}{x}\}$ s.t. $\underset{\sim}{x}$ is classified as π_1 $R_2 \equiv \Omega - R_1$
-

- **Conditional** Prob of classifying object to π_2 when really in π_1:

$$P(2|1) = \int_{R_2 = \Omega - R_1} f_1(\underset{\sim}{x})\, d\underset{\sim}{x} = P(X \in R_2 | \pi_1)$$

$$P(1|2) = \int_{R_1} f_2(\underset{\sim}{x})\, d\underset{\sim}{x} = P(X \in R_1 | \pi_2)$$

p=2



- **Overall** (unconditional) Probabilities : $p_1 \equiv$ Prior of π_1, $p_2 \equiv$ Prior of π_2
  $(p_1 + p_2 = 1)$

$Pr\{\text{obs comes from } \pi_1, \text{ classified as } \pi_1\} = P(\pi_1)\, P(x \in R_1 | \pi_1)$
$= p_1\, p(1|1)$

$Pr\{\text{obs misclassified as } \pi_1\} = P(\pi_2)\, P(x \in R_1 | \pi_2) = p_2\, p(1|2)$

$Pr\{\text{obs correctly classified as } \pi_2\} = P(\pi_2)\, P(x \in R_2 | \pi_2) = p_2\, p(2|2)$

$Pr\{\text{obs misclassified as } \pi_2\} = P(\pi_1)\, P(x \in R_2 | \pi_1) = p_1\, p(2|1)$

- Costs of Misclassification Cost Matrix:

		Classified as:	
		π_1	π_2
True Population	π_1	0	$c(2\|1)$
	π_2	$c(1\|2)$	0

- Expected Cost of misclassification (ECM) :

$$ECM = Pr\{\pi_2 \text{ misclass as } \pi_1\} * c(1|2) + Pr\{\pi_1 \text{ misclass as } \pi_2\} \cdot c(2|1)$$
$$= p_2\, p(1|2)\, c(1|2) + p_1\, p(2|1)\, c(2|1)$$

Want ECM to be small.

Result 11.1 Regions R_1, R_2 that minimize ECM X s.t.

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

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

$\frac{f_1(\underline{x})}{f_2(\underline{x})}$ = density ratio

$\frac{c(1|2)}{c(2|1)} \equiv$ cost ratio

$\frac{p_2}{p_1} \equiv$ - Prior Prob ratio

cost ratio often "guessed" as ratio, no units.

Special Cases of Minimum Expected cost regions

a) $p_2/p_1 = 1$ (equal prior probs)

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)}{c(2|1)} \qquad R_2 : \frac{f_1(x)}{f_2(x)} < \frac{c(1|2)}{c(2|1)}$$

b) $\frac{c(1|2)}{c(2|1)} = 1$ (equal misclassify costs)

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1} \qquad R_2 : \frac{f_1(x)}{f_2(x)} < \frac{p_2}{p_1}$$

c) $\frac{p_2}{p_1} = \frac{c(1|2)}{c(2|1)} = 1$ or $\frac{p_2}{p_1} = \frac{1}{c(1|2)/c(2|1)}$

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq 1 \qquad \frac{f_1(x)}{f_2(x)} < 1$$

when no idea, prior prob and cost ratios typically set to 1.

- Ignoring costs, could choose R_1, R_2 to minimize Total Probability of Misclassification (TPM):

$$TPM = Pr\{\text{Misclassifying } \pi_1 \text{ or } \pi_2 \text{ observation}\}$$

$$= p_1 P(2|1) + p_2 P(1|2) = p_1 \int_{R_2} f_1(x)dx + p_2 \int_{R_1} f_2(x)dx$$

(Equivalent to Case b) above).

- Allocating new $\underset{\sim}{x}_0$ to population w/ highest posterior probability: $P(\pi_i|\underset{\sim}{x}_0)$. Using Bayes' Rule:

$$P\{\pi_1|\underset{\sim}{x}_0\} = \frac{Pr\{\pi_1 \text{ occurs } \underline{\text{and}} \text{ observe } x_0\}}{Pr\{\text{observe } x_0\}} =$$

$$\frac{Pr\{\text{observe } \underset{\sim}{x}_0|\pi_1\}\,P\{\pi_1\}}{Pr\{\text{observe } \underset{\sim}{x}_0|\pi_1\}\,P(\pi_1\} + Pr\{\text{observe } \underset{\sim}{x}_0|\pi_2\}\,P\{\pi_2\}} = \frac{p_1 f_1(x_0)}{p_1 f_1(x_0) + p_2 f_2(x_0)}$$

$$P\{\pi_2|\underset{\sim}{x}_0\} = 1 - P\{\pi_1|x_0\} = \frac{p_2 f_2(x_0)}{p_1 f_1(x_0) + p_2 f_2(x_0)}$$

Classify π_1 if $P(\pi_1|x_0) \geq P(\pi_2|x_0)$ (same as b) since equal denominators)

equivalently $p_1 f_1(x_0) \geq p_2 f_2(x_0)$

11.3 Classification w/ 2 MVN Populations

Case w/ $\Sigma_1 = \Sigma_2 = \Sigma$ $\quad \underset{\sim}{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$

$$f_i(\underset{\sim}{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_i)'\Sigma(\underset{\sim}{x}-\underset{\sim}{\mu}_i)\right\} \quad i=1,2$$

Minimum ECM Regions (common normalizing constants)

$$R_1: \exp\left\{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)'\Sigma(\underset{\sim}{x}-\underset{\sim}{\mu}_1) + \frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)\right\} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

R_2 if $<$ (otherwise)

Result 11.2 (Π_1, Π_2 described as previous MVN distn)

Minimize ECM: Allocate $\underset{\sim}{x}_0$ to π_1 if:

$$(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x}_0 - \frac{1}{2}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

Discrimination Based on Sample Data

Random Samples: n_1 from π_1, n_2 from π_2 (X_1, X_2)

$\bar{\underset{\sim}{x}}_1$, S_1, $\bar{\underset{\sim}{x}}_2$, S_2 $\quad S_{pooled} = \dfrac{(n_1-1)S_1 + (n_2-1)S_2}{(n_1-1)+(n_2-1)}$

Sample (Estimated) ECM Rule for 2 Normal Populations

Allocate $\underset{\sim}{x}_0$ to π_1 if:

$$(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)'S_{pooled}^{-1}\underset{\sim}{x}_0 - \frac{1}{2}(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)'S_{pooled}^{-1}(\bar{\underset{\sim}{x}}_1 + \bar{\underset{\sim}{x}}_2) \geq \ln\left[\frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right]$$

Allocate $\underset{\sim}{x}_0$ to π_2 otherwise.

Simplification if $\dfrac{c(1|2)}{c(2|1)}\dfrac{p_2}{p_1} = 1 \Rightarrow \ln(\cdot) = 0$

Let $\hat{y} = (\bar{x}_1 - \bar{x}_2)'S_{pooled}^{-1}\underset{\sim}{x} = \hat{\underset{\sim}{a}}'\underset{\sim}{x}$ $\qquad \hat{y}_0 = \hat{\underset{\sim}{a}}'\underset{\sim}{x}_0$

$\bar{y}_1 = \hat{\underset{\sim}{a}}'\bar{\underset{\sim}{x}}_1 \qquad \bar{y}_2 = \hat{\underset{\sim}{a}}'\bar{\underset{\sim}{x}}_2$

$\hat{m} = \frac{1}{2}(\bar{x}_1 - \bar{x}_2)'S_{pooled}^{-1}(\bar{x}_1 + \bar{x}_2) = \frac{1}{2}(\bar{y}_1 + \bar{y}_2)$

Decision based on whether $\hat{y}_0$ falls to the left or right of $\hat{m}$.

Fisher's Classification Method for 2 Populations

Linear transformations of $\underset{\sim}{x}$ to y (univariate) so that y's from π_1, π_2 are as different as possible.

$$\underset{\sim}{x}_{11}, \ldots, \underset{\sim}{x}_{1n_1} \rightarrow y_{11}, \ldots, y_{1n_1} \qquad \underset{\sim}{x}_{21}, \ldots, \underset{\sim}{x}_{2n_2} \rightarrow y_{21}, \ldots, y_{2n_2}$$

$$\text{Separation} = \frac{|\bar{y}_1 - \bar{y}_2|}{s_y} \qquad s_y^2 = \frac{\sum_{j=1}^{n_1}(y_{1j}-\bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j}-\bar{y}_2)^2}{n_1+n_2-2}$$

Result 11.3 $\hat{y} = \underset{\sim}{\hat{a}}'\underset{\sim}{x} = (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x$ maximizes:

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(\hat{a}'\bar{x}_1 - \hat{a}'\bar{x}_2)^2}{\hat{a}' S_{pooled} \hat{a}} = \frac{(\hat{a}'d)^2}{\hat{a}' S_{pooled} \hat{a}} \qquad (d = \bar{x}_1 - \bar{x}_2)$$

over all possible $\hat{a}$ w/ $\underset{\sim}{d} = \underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2$

max of ratio is $D^2 = (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} (\bar{x}_1 - \bar{x}_2)$

Proof: Maximization Lemma (eq. 2.50).

Allocation rule for Fisher's Discriminant Function

Allocate $\underset{\sim}{x_0}$ to π_1 if:

$$\hat{y}_0 = (\underset{\sim}{\bar{x}}_1 - \bar{x}_2)' S_{pooled}^{-1} \underset{\sim}{x_0} \geq \hat{m} = \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} (\underset{\sim}{\bar{x}}_1 + \underset{\sim}{\bar{x}}_2)$$

or $\hat{y}_0 - \hat{m} \geq 0$

Allocate $\underset{\sim}{x_0}$ to π_2 if:

$\hat{y}_0 < \hat{m}$ or $\hat{y}_0 - \hat{m} < 0$

Classification of MVN Populations when $\Sigma_1 \neq \Sigma_2$

$$R_1: \; -\frac{1}{2} x'\left(\Sigma_1^{-1}-\Sigma_2^{-1}\right)\underset{\sim}{x} + \left(\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1}\right)\underset{\sim}{x} - k \geq$$

$$\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

R_2 if $<$

where: $k = \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \frac{1}{2}\left(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2\right)$

When $\Sigma_1 = \Sigma_2$ drops back to equal Σ_i case.

Rule for Allocating $\underset{\sim}{x}_0$ based on sample data

Allocate $\underset{\sim}{x}_0$ to π_1 if:

$$-\frac{1}{2}\underset{\sim}{x}_0'\left(S_1^{-1}-S_2^{-1}\right)\underset{\sim}{x}_0 + \left(\underset{\sim}{\bar{x}}_1' S_1^{-1} - \underset{\sim}{\bar{x}}_2 S_2^{-1}\right)\underset{\sim}{x}_0 - k \geq$$

$$\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

11.4 Evaluating Classification Functions

Total Prob of Misclass: $TPM = p_1\int_{R_1} f_1(x)dx + p_2\int_{R_2} f_2(x)dx.$

Optimum error rate $\equiv$ TPM evaluated w/ R_1, R_2 determined

by: $\quad R_1: \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1} \qquad R_2: \frac{f_1(x)}{f_2(x)} < \frac{p_2}{p_1}$

In practice population density functions not known.

Estimates replacing parameters:

<u>Actual Error Rate</u>: $AER = p_1 \int_{\hat{R}_2} f_1(x)\,dx + p_2 \int_{\hat{R}_1} f_2(x)\,dx$

$\hat{R}_1 : (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} (\bar{x}_1 + \bar{x}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$

$\hat{R}_2$ if $<$.

(Still depends on unknown $f_1(x)$, $f_2(x)$).

- Measure that can be applied for any procedure (training data).

Apparent Error Rate (APER) based on Confusion Matrix

		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

$n_{ic} \equiv$ # of π_i observations correctly classified $i=1,2$

$n_{im} \equiv$ " " " " incorrectly " $i=1,2$

Apparent Error Rate: $APER = \dfrac{n_{1m} + n_{2m}}{n_1 + n_2}$

Tends to underestimate AER unless n_1, n_2 LARGE.

(Overfitting of classification rule to training data).

Can split total sample into
- Training
- Validation

Doesn't work well in classification unless huge samples.

Lauchenbruch's holdout method (jacknife/crossvalidation)

1) Begin w/ π_1 Group. Delete one case (observation) and develop classification function on remaining $n_1 - 1$, n_2 cases.

2) Classify the "holdout" observation based on classification function in 1).

3) Repeat steps 1) and 2) until all of the π_1 cases have been heldout and classified

   $n_{1m}^{(H)} \equiv$ # of Holdout cases misclassified

4) Repeat steps 1) - 3) for all cases from π_2.

   $n_{2m}^{(H)} \equiv$ # of Holdout cases misclassified.

- Conditional misclassification probabilities:

$$\hat{P}(2|1) = \frac{n_{1m}^{(H)}}{n_1} \qquad \hat{P}(1|2) = \frac{n_{2m}^{(H)}}{n_2}$$

Total proportion misclassified gives nearly unbiased estimate of expected actual error rate $E\{AER\}$: (for moderate-large samples)

$$\hat{E}\{AER\} = \frac{n_{1m}^{(H)} + n_{2m}^{(H)}}{n_1 + n_2}$$