Analisis Diskriminan

Program Studi Statistika dan Sains Data-IPB

Akbar Rizki S.Stat., M.Si.

Outline

1. Analisis Diskriminan
2. *Separation and Classification for Two Populations*
3. Classification with Two Multivariate Normal Populations
4. Evaluating Classification Functions

Analisis Diskriminan

Kadangkala kita harus menduga keanggotaan suatu kelompok sebagai suatu fungsi dari beberapa peubah kontinu. Misalnya :

- Sebuah perusahaan kartu kredit menginginkan penggunaan informasi keuangan untuk menentukan apakah seorang calon pelanggan memiliki resiko tinggi atau rendah, sebelum ditawarkan kepadanya kartu kredit
- Kantor Perpajakan menginginkan membuat keputusan apakah seorang wajib pajak termasuk orang yang jujur ataukah tidak, dengan tujuan menentukan apakah perlu mengaudit orang tersebut.
- Sebuah rumah sakit harus bisa menentukan apakah seorang korban yang parah kecelakaan mobil yang dibawa ke unit gawat darurat adalah seorang pecandu alkohol atau bukan. Ini karena perlakuan yang berbeda yang harus dilakukan bagi pecandu dan bukan. Meskipun diagnosis alkohol memerlukan pengamatan yang cukup ekstensif, yang tidak bisa dilakukan dalam keadaan terburu-buru, memungkinkan untuk mengambil darah dari pasien dan mencari fungsi dari test darah itu untuk membedakan pecandu dan bukan pecandu.
- Pertanyaannya adalah : Bagaimana kira bisa membuat fungsi linear peubah kontinu yang bisa menduga keanggotaan suatu individu pada kelompok tertentu ?

Analisis Diskriminan

- **Analisis diskriminan**
  Merupakan masalah klasifikasi, di mana dua atau lebih kelompok, kluster, atau populasi telah diketahui sebelumnya, dan satu atau lebih pengamatan baru diklasifikasikan ke dalam salah satu populasi yang diketahui berdasarkan karakteristik yang diukur.
- **Hubungan dan Perbedaan Analisis diskriminan dan Klasifikasi**
  Keduanya sama-sama merupakan teknik multivariat yang bertujuan memisahkan himpunan objek (observasi) yang berbeda dan mengalokasikan objek baru ke kelompok yang telah ditentukan. Analisis diskriminan bersifat lebih eksploratif, digunakan untuk memahami perbedaan antar kelompok ketika hubungan kausal belum jelas, sedangkan klasifikasi bersifat lebih terstruktur dengan tujuan membentuk aturan yang jelas untuk mengklasifikasikan objek baru.
- **Tujuan**
  seringkali tumpang tindih sehingga mengaburkan batasan
  ❑**Tujuan 1 (Discrimination):** Menggambarkan atau mengidentifikasi ciri pembeda antar kelompok objek agar perbedaan antar kelompok terlihat jelas, baik secara grafis maupun aljabar.
  ❑**Tujuan2 (Classification):** Menyusun aturan atau fungsi yang dapat digunakan untuk mengelompokkan objek baru ke dalam kelas yang telah diberi label.

Separation and Classification for Two Populations

- Misalkan terdapat dua populasi yaitu populasi 1 dan populasi 2 Dimana populasi 1 memiliki fungsi sebaran data $f_1(x)$ dan populasi 2 memiliki fungsi sebaran data $f_2(x)$. Selain itu R_1 dan R_2 merupakan wilayah Keputusan yaitu R_1 daerah di mana semua titik (observasi) akan dinyatakan milik Populasi 1 dan R_2 daerah di mana semua titik akan dinyatakan milik Populasi 2.
- Pembentukan R_1 dan R_2 ditentukan oleh aturan Bayes.

$$P(Populasi\ i|x_0) = \frac{f_i(x_0)p_i}{f(x_0)}; \quad f(x_0) = f_{1(x_0)}p_1 + f_2(x_0)p_2$$

Sehingga

Wilayah	Kriteria	Keputusan
R_1	$f_{1(x)}p_1 > f_2(x)p_2$	Klasifikasikan ke populasi 1
R_2	$f_{1(x)}p_1 < f_2(x)p_2$	Klasifikasikan ke populasi 2

- Dalam proses klasifikasi memungkinkan untuk terjadinya kesalahan klasifikasi, karena:

a. Pengetahuan tentang kinerja di masa depan tidak lengkap
b. Informasi sempurna hanya bisa didapat dengan merusak objek
c. Informasi tidak tersedia atau terlalu mahal untuk diperoleh

Separation and Classification for Two Populations

- Dalam proses klasifikasi terdapat dua jenis kesalahan yang mungkin terjadi yaitu:

a. P(2|1) : Data dari populasi 1 salah dikalsifikasikan ke populasi 2

b. P(1|2) Data dari populasi 2 salah diklasifikasikan ke populasi 1

Kedua kesalahan tersebut memiliki biaya. Matriks biaya kesalahan klasifikasi:

		Classify as:	
		π_1	π_2
True population:	π_1	0	$c(2\|1)$
	π_2	$c(1\|2)$	0

- Tujuan: meminimalkan biaya total kesalahan (Expected Cost of Misclassification / ECM)

$$ECM = c(2|1)p_1P(2|1) + c(1|2)p_2P(1|2)$$

Untuk meminimumkan ECM, maka jika terdapat objek baru (x_0), maka alokasikan ke populasi 1 jika:

$$\frac{f_1(x_0)}{f_2(x_0)} \geq \frac{c(1|2)p_2}{c(2|1)p_1}$$

Jika tidak ke populasi 2.

Separation and Classification for Two Populations

Example 11.2 (Classifying a new observation into one of the two populations) A researcher has enough data available to estimate the density functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ associated with populations π_1 and π_2, respectively. Suppose $c(2|1) = 5$ units and $c(1|2) = 10$ units. In addition, it is known that about 20% of *all* objects (for which the measurements $\mathbf{x}$ can be recorded) belong to π_2. Thus, the prior probabilities are $p_1 = .8$ and $p_2 = .2$.

Given the prior probabilities and costs of misclassification, we can use (11-6) to derive the classification regions R_1 and R_2. Specifically, we have

$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{10}{5}\right)\left(\frac{.2}{.8}\right) = .5$$

$$R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{10}{5}\right)\left(\frac{.2}{.8}\right) = .5$$

Suppose the density functions evaluated at a new observation $\mathbf{x}_0$ give $f_1(\mathbf{x}_0) = .3$ and $f_2(\mathbf{x}_0) = .4$. Do we classify the new observation as π_1 or π_2? To answer the question, we form the ratio

$$\frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} = \frac{.3}{.4} = .75$$

and compare it with .5 obtained before. Since

$$\frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} = .75 > \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) = .5$$

we find that $\mathbf{x}_0 \in R_1$ and classify it as belonging to π_1. ■

Classification with Two Multivariate Normal Populations

•Terdapat observasi baru, x_0 yang ingin diklasifikasikan ke dalam satu dari dua populasi Dimana setiap populasi menyebar normal multivariat.

$$\boldsymbol{X}|\pi_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i); i = 1,2$$

$$f_i(\boldsymbol{X}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp[-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu_i})'^{\boldsymbol{\Sigma_i^{-1}}}(\boldsymbol{X} - \boldsymbol{\mu_i}); i = 1,2$$

- Klasifikasi Ketika $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ (**Fungsi diskriminan Linier**) dan **parameter diketahui** maka aturan meminimumkan ECM yang memasukkan x_0 ke dalam populasi 1 adalah:

$$\frac{f_1(x_0)}{f_2(x_0)} \geq \frac{c(1|2)p_2}{c(2|1)p_1}$$

$$\frac{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)\right]}{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)\right]} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)\right] \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

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

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

Jika $\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) = 1$ maka $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\mathbf{x}_0 \geq \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$

Classification with Two Multivariate Normal Populations

•Skor diskriminan linier y adalah:

$$y = \mathbf{a}'\mathbf{x} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\mathbf{x}$$

- Titik Tengah/ midpoint (m) antara rata-rata y untuk kedua populasi adalah:

$$m = \frac{1}{2}(\mu_{1y} + \mu_{2y}) = \frac{1}{2}[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\boldsymbol{\mu}_1 + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\boldsymbol{\mu}_2] = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$$

Sehingga

a. Alokasikan x_0 ke populasi 1 (π_1) jika $y_0 = a'x_0 \geq m$
b. Alokasikan x_0 ke populasi 2 (π_2) jika $y_0 = a'x_0 < m$

Dimana $\boldsymbol{a} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ Adalah arah yang memisahkan kedua populasi sejauh mungkin

Classification with Two Multivariate Normal Populations

- Klasifikasi Ketika $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ (**Fungsi diskriminan Linier**) dan **parameter tidak diketahui** maka menduga parameter dari data contoh sehingga,

$$\underset{(p\times 1)}{\bar{\mathbf{x}}_1} = \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbf{x}_{1j}, \quad \underset{(p\times p)}{\mathbf{S}_1} = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(\mathbf{x}_{1j}-\bar{\mathbf{x}}_1)(\mathbf{x}_{1j}-\bar{\mathbf{x}}_1)'$$

$$\underset{(p\times 1)}{\bar{\mathbf{x}}_2} = \frac{1}{n_2}\sum_{j=1}^{n_2}\mathbf{x}_{2j}, \quad \underset{(p\times p)}{\mathbf{S}_2} = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(\mathbf{x}_{2j}-\bar{\mathbf{x}}_2)(\mathbf{x}_{2j}-\bar{\mathbf{x}}_2)'$$

$$\mathbf{S}_{\text{pooled}} = \left[\frac{n_1-1}{(n_1-1)+(n_2-1)}\right]\mathbf{S}_1 + \left[\frac{n_2-1}{(n_1-1)+(n_2-1)}\right]\mathbf{S}_2$$

- Peminimuman ECM yang memasukkan x_0 ke dalam populasi 1 adalah

$$(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

Untuk kasus biaya dan prior yang sama, $\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) = 1$ maka:

a. Alokasikan x_0 ke populasi 1 (π_1) jika $y_0 = a'x_0 \geq m$
b. Alokasikan x_0 ke populasi 2 (π_2) jika $y_0 = a'x_0 < m$

Dimana

- $\boldsymbol{a}' = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'S_{pooled}^{-1}$ Dimana fungsi linier diskriminan adalah: $\hat{y} = (\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} = \hat{\mathbf{a}}'\mathbf{x}$
- $\hat{m} = \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\, S_{pooled}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) = \frac{1}{2}(\bar{y}_1 + \bar{y}_2)$
- $\bar{y}_1 = \hat{a}'\bar{\boldsymbol{x}}_1$; $\bar{y}_2 = \hat{a}'\bar{\boldsymbol{x}}_2$

Classification with Two Multivariate Normal Populations

Sebuah penelitian bertujuan untuk mengembangkan prosedur klasifikasi untuk mengidentifikasi **carrier hemofilia A** berdasarkan dua variabel pengukuran darah:

- $X_1 = log_{10}(\textit{Aktivitas AHF})$
- $X_2 = log_{10}(\textit{Antigen seperti AHF})$

Terdapat dua kelompok populasi:

π_1 :**Wanita normal** (tidak membawa gen hemofilia), $n_1 = 30$

π_2 :**Pembawa wajib (obligatory carriers)**, $n_2 = 22$

Bagaimana aturan klasifikasi untuk mengelompokkan wanita baru ke dalam salah satu kelompok berdasarkan nilai X_1 dan X_2.

Diketahui:

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} -.0065 \\ -.0390 \end{bmatrix}, \quad \bar{\mathbf{x}}_2 = \begin{bmatrix} -.2483 \\ .0262 \end{bmatrix} \quad \mathbf{S}_{\text{pooled}}^{-1} = \begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}$$

Langkah penyelesaian:

- Membentuk fungsi diskriminan linier
- Menghitung titik Tengah/ midpoint (m)
- Aturan klasifikasi
- Klasifikasi observasi baru
- Pertimbangan prior probability

Classification with Two Multivariate Normal Populations

Langkah penyelesaian:

- Membentuk fungsi diskriminan linier

$$\hat{y} = \hat{\mathbf{a}}'\mathbf{x} = [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2]'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}$$

$$= [.2418 \quad -.0652]\begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= 37.61x_1 - 28.92x_2$$

$$\bar{y}_1 = \hat{\mathbf{a}}'\bar{\mathbf{x}}_1 = [37.61 \quad -28.92]\begin{bmatrix} -.0065 \\ -.0390 \end{bmatrix} = .88$$

$$\bar{y}_2 = \hat{\mathbf{a}}'\bar{\mathbf{x}}_2 = [37.61 \quad -28.92]\begin{bmatrix} -.2483 \\ .0262 \end{bmatrix} = -10.10$$

- Menghitung titik Tengah/ midpoint (m)

and the midpoint between these means [see (11-20)] is

$$\hat{m} = \tfrac{1}{2}(\bar{y}_1 + \bar{y}_2) = \tfrac{1}{2}(.88 - 10.10) = -4.61$$

- Aturan klasifikasi

Allocate $\mathbf{x}_0$ to π_1 if $\hat{y}_0 = \hat{\mathbf{a}}'\mathbf{x}_0 \geq \hat{m} = -4.61$

Allocate $\mathbf{x}_0$ to π_2 if $\hat{y}_0 = \hat{\mathbf{a}}'\mathbf{x}_0 < \hat{m} = -4.61$

- Klasifikasi observasi baru

where $\mathbf{x}'_0 = [-.210, -.044]$. Since

$$\hat{y}_0 = \hat{\mathbf{a}}'\mathbf{x}_0 = [37.61 \quad -28.92]\begin{bmatrix} -.210 \\ -.044 \end{bmatrix} = -6.62 < -4.61$$

Masuk pada populasi 2/ obligatory carrier

Classification with Two Multivariate Normal Populations

- Pertimbangan prior probability

Suppose now that the prior probabilities of group membership are known. For example, suppose the blood yielding the foregoing x_1 and x_2 measurements is drawn from the maternal first cousin of a hemophiliac. Then the genetic chance of being a hemophilia A carrier in this case is .25. Consequently, the prior probabilities of group membership are $p_1 = .75$ and $p_2 = .25$. Assuming, somewhat unrealistically, that the costs of misclassification are equal, so that $c(1|2) = c(2|1)$, and using the classification statistic

$$\hat{w} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 - \tfrac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

or $\hat{w} = \hat{\mathbf{a}}'\mathbf{x}_0 - \hat{m}$ with $\mathbf{x}'_0 = [-.210, -.044]$, $\hat{m} = -4.61$, and $\hat{\mathbf{a}}'\mathbf{x}_0 = -6.62$, we have

$$\hat{w} = -6.62 - (-4.61) = -2.01$$

Applying (11-18), we see that

$$\hat{w} = -2.01 < \ln\left[\frac{p_2}{p_1}\right] = \ln\left[\frac{.25}{.75}\right] = -1.10$$

and we classify the woman as π_2, an obligatory carrier. ■

Fisher's Approach to Classification with Two Populations

•R.A. Fisher mengembangkan pendekatan yang berbeda dengan metode ECM (Expected Cost of Misclassification). Filosofi Fisher adalah:
"Transformasi variabel multivariat menjadi variabel univariat yang memisahkan kedua kelompok sebanyak mungkin"
•Berbeda dengan pendekatan normal theory yang memerlukan asumsi distribusi normal, pendekatan Fisher hanya memerlukan:
a.Asumsi matriks kovarian sama
b.Tidak memerlukan asumsi bentuk distribusi populasi

- Ukuran pemisah yang digunakan pada pendekatan Fisher untuk memaksimalkan pemisahan antar dua kelompok Adalah:

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

$$\max_{\hat{\mathbf{a}}} \frac{(\hat{\mathbf{a}}'\mathbf{d})^2}{\hat{\mathbf{a}}'\mathbf{S}_{\text{pooled}}\hat{\mathbf{a}}} = \mathbf{d}'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{d} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = D^2$$

$$\hat{y} = \hat{\mathbf{a}}'\mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} = 37.61x_1 - 28.92x_2$$

- Pada contoh sebelumnya, maka:

This linear discriminant function is Fisher's linear function, which maximally separates the two populations, and the maximum separation in the samples is

$$\begin{aligned} D^2 &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \\ &= [.2418, \quad -.0652]\begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}\begin{bmatrix} .2418 \\ -.0652 \end{bmatrix} \\ &= 10.98 \end{aligned}$$

■

Classification with Two Multivariate Normal Populations

- Klasifikasi Ketika $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ (**Fungsi diskriminan Kuadratik**)

Kembali lagi pada aturan meminimumkan ECM yang memasukkan x_0 ke dalam populasi 1 adalah:

$$\frac{f_1(x_0)}{f_2(x_0)} \geq \frac{c(1|2)p_2}{c(2|1)p_1}$$

Pada kasus kovarian tidak sama ini komponen $|\boldsymbol{\Sigma}_i|^{1/2}$ tidak saling menghilangkan, sehingga bentuk kuadratik dalam exponent tidak dapat disederhanakan.

Sehingga:

$$R_1: \; -\frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} + (\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})\mathbf{x} - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

$$R_2: \; -\frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} + (\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})\mathbf{x} - k < \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (11\text{-}27)$$

where

$$k = \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) + \frac{1}{2}(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2) \quad (11\text{-}28)$$

- Untuk nilai parameter yang tidak diketahui maka x_0 akan dimasukkan ke dalam populasi 1 jika:

$$-\frac{1}{2}\mathbf{x}_0'(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})\mathbf{x}_0 + (\bar{\mathbf{x}}_1'\mathbf{S}_1^{-1} - \bar{\mathbf{x}}_2'\mathbf{S}_2^{-1})\mathbf{x}_0 - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

Evaluating Classification Functions

- Evaluasi dilakukan dengan menghitung **tingkat kesalahan klasifikasi** (*error rates*) atau **probabilitas misklasifikasi**.

a. Tingkat Kesalahan Optimal (Optimum Error Rate - OER) – fungsi kepadatan populasi diketahui

Total Probability of Misclassification (TPM): Probabilitas salah mengklasifikasikan sebuah observasi.

$$\text{TPM} = p_1 \int_{R_2} f_1(\mathbf{x})\, d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x}$$

Optimum Error Rate (OER): Nilai TPM terkecil yang dapat dicapai dengan memilih daerah klasifikasi R_1 dan R_2 secara optimal (biasanya menggunakan aturan biaya salah klasifikasi minimum).

$$\text{Optimum error rate (OER)} = p_1 \int_{R_2} f_1(\mathbf{x})\, d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x}$$

b. Tingkat Kesalahan Aktual (Actual Error Rate - AER)

$$\text{AER} = p_1 \int_{\hat{R}_2} f_1(\mathbf{x})\, d\mathbf{x} + p_2 \int_{\hat{R}_1} f_2(\mathbf{x})\, d\mathbf{x}$$

AER menunjukkan bagaimana performa fungsi klasifikasi sampel pada sampel-sampel di masa depan. Seperti OER, secara umum, AER tidak dapat dihitung, karena tergantung pada fungsi kepadatan yang tidak diketahui.

Evaluating Classification Functions

c. Tingkat Kesalahan Nyata (Apparent Error Rate - APER)

		Predicted membership π_1	Predicted membership 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

(11-33)

where

n_{1C} = number of π_1 items correctly classified as π_1 items
n_{1M} = number of π_1 items misclassified as π_2 items
n_{2C} = number of π_2 items correctly classified
n_{2M} = number of π_2 items misclassified

The apparent error rate is then

$$\text{APER} = \frac{n_{1M} + n_{2M}}{n_1 + n_2} \qquad (11\text{-}34)$$

which is recognized as the *proportion* of items in the training set that are misclassified.

		Predicted membership π_1: riding-mower owners	Predicted membership π_2: nonowners	
Actual membership	π_1: riding-mower owners	$n_{1C} = 10$	$n_{1M} = 2$	$n_1 = 12$
	π_2: nonowners	$n_{2M} = 2$	$n_{2C} = 10$	$n_2 = 12$

The apparent error rate, expressed as a percentage, is

$$\text{APER} = \left(\frac{2+2}{12+12}\right)100\% = \left(\frac{4}{24}\right)100\% = 16.7\%$$

■

Kekurangan: overfitting,

(i) It requires large samples.

(ii) The function evaluated is not the function of interest. Ultimately, almost *all* of the data must be used to construct the classification function. If not, valuable information may be lost.

Evaluating Classification Functions

d. Estimasi yang Tidak Bias dengan Metode Lachenbruch (Holdout/Jackknife)

Prosedur:

1. Start with the π_1 group of observations. Omit one observation from this group, and develop a classification function based on the remaining $n_1 - 1, n_2$ observations.
2. Classify the "holdout" observation, using the function constructed in Step 1.
3. Repeat Steps 1 and 2 until all of the π_1 observations are classified. Let $n_{1M}^{(H)}$ be the number of holdout (H) observations misclassified in this group.
4. Repeat Steps 1 through 3 for the π_2 observations. Let $n_{2M}^{(H)}$ be the number of holdout observations misclassified in this group.

Estimates $\hat{P}(2|1)$ and $\hat{P}(1|2)$ of the conditional misclassification probabilities in (11-1) and (11-2) are then given by

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

$$\hat{P}(1|2) = \frac{n_{2M}^{(H)}}{n_2} \qquad (11\text{-}35)$$

and the total proportion misclassified, $(n_{1M}^{(H)} + n_{2M}^{(H)})/(n_1 + n_2)$, is, for moderate samples, a nearly unbiased estimate of the *expected* actual error rate, $E(\text{AER})$.

$$\hat{E}(\text{AER}) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2} \qquad (11\text{-}36)$$

Lihat contoh 11.7 di buku Johnson & Wichern