



STA1342 – Teknik Peubah Ganda



ANALISIS DISKRIMINAN

Dhea Dewanti & Nur Khamidah

Pertemuan 10

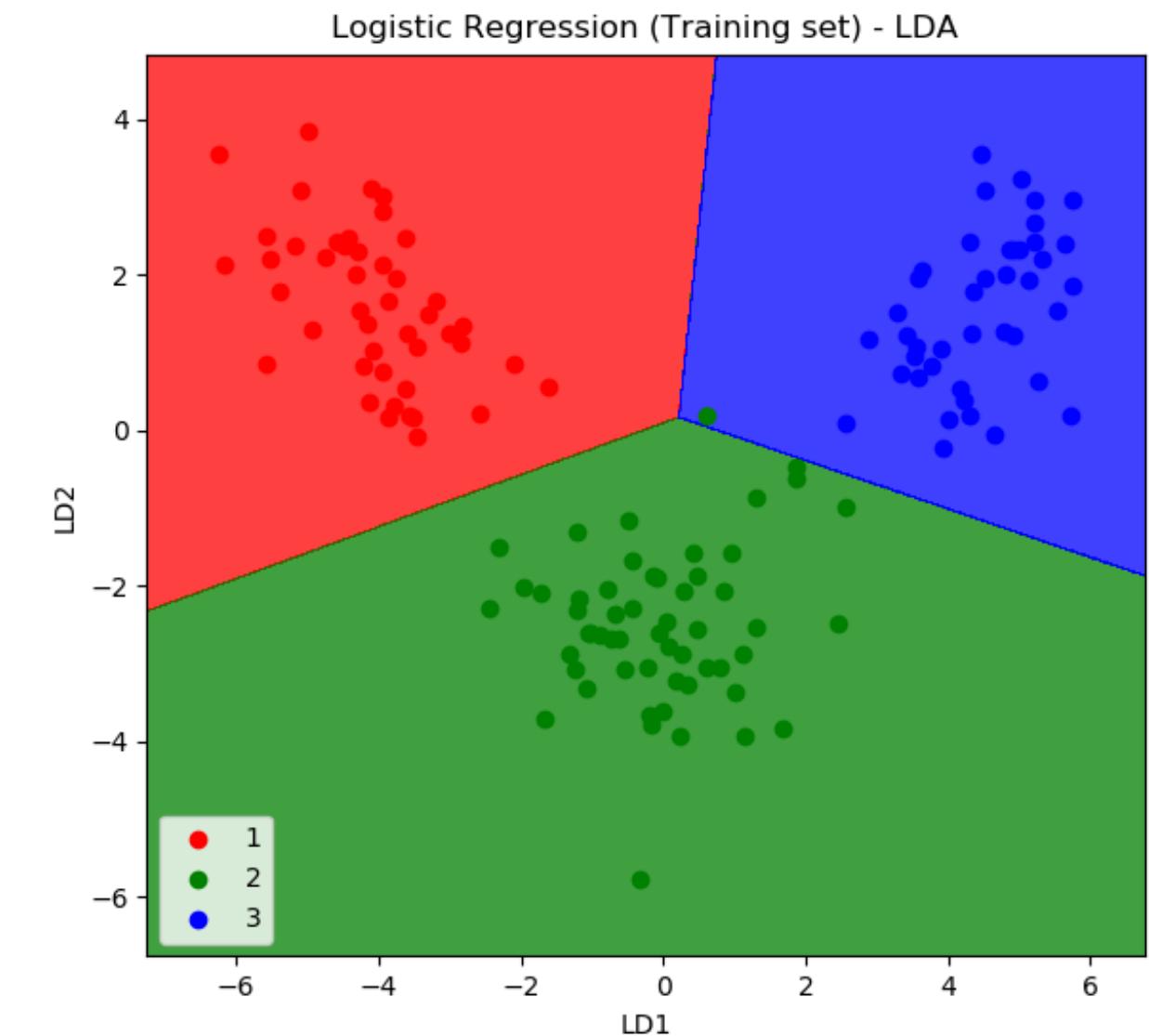
IDE DASAR

- Sudah ada pengelompokan objek
- Mencari fungsi yang bisa dijadikan dasar membedakan (mendiskriminakan) objek ke dalam kelompok-kelompok
- Menentukan ke kelompok mana suatu objek baru
- Peubah pembeda adalah Peubah yang ragamnya besar
- Pembedaan seringkali memerlukan kombinasi beberapa Peubah (satu peubahtidak cukup)

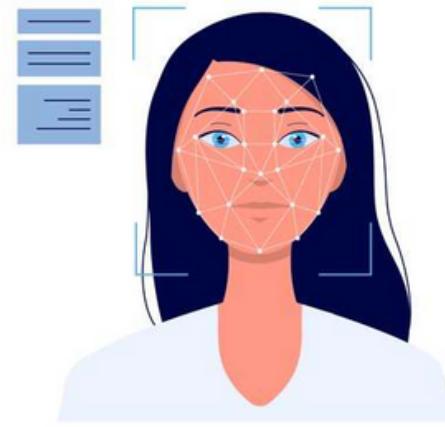


APA TUJUAN ANALISIS DISKRIMINAN?

Tujuan utama analisis ini adalah memperoleh fungsi diskriminan, yaitu fungsi yang mampu digunakan membedakan suatu objek masuk ke dalam populasi tertentu berdasarkan pengamatan terhadap objek tersebut



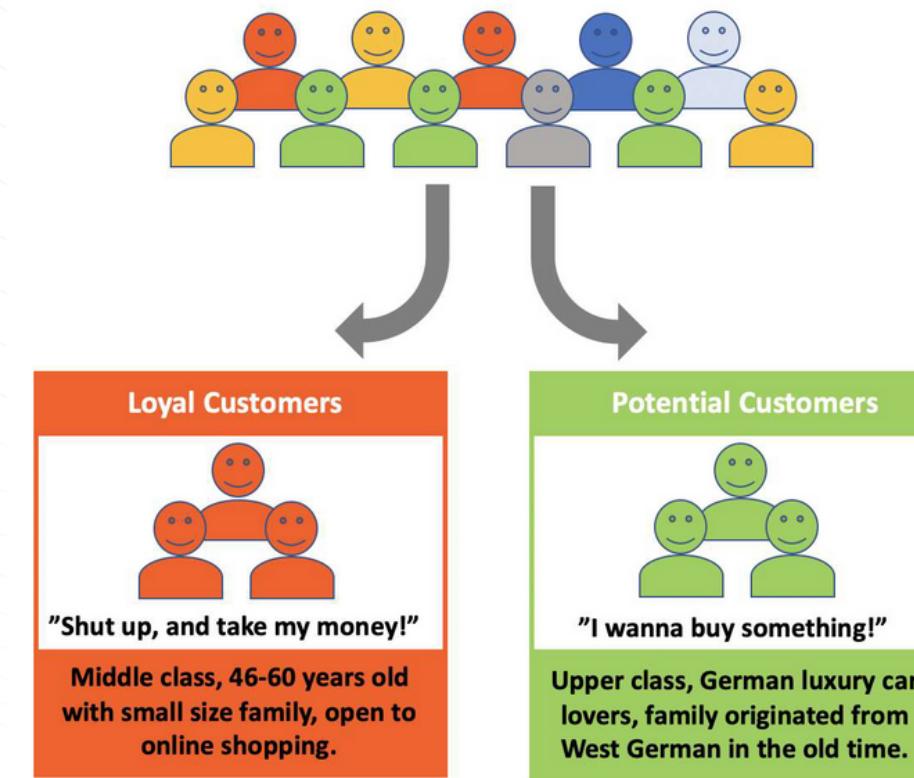
CONTOH PENERAPAN



FACE RECOGNITION



BANKRUPTCY PREDICTION

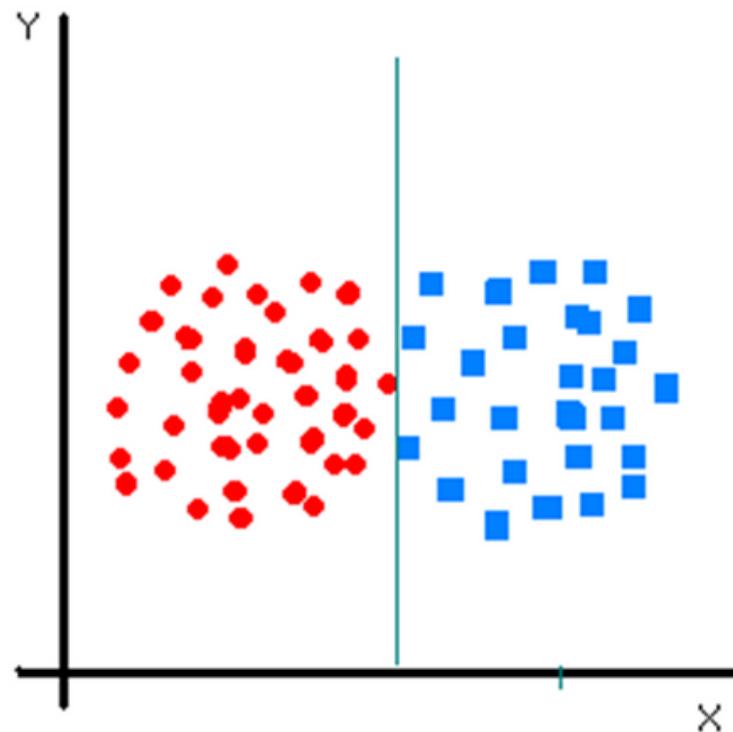


MARKETING

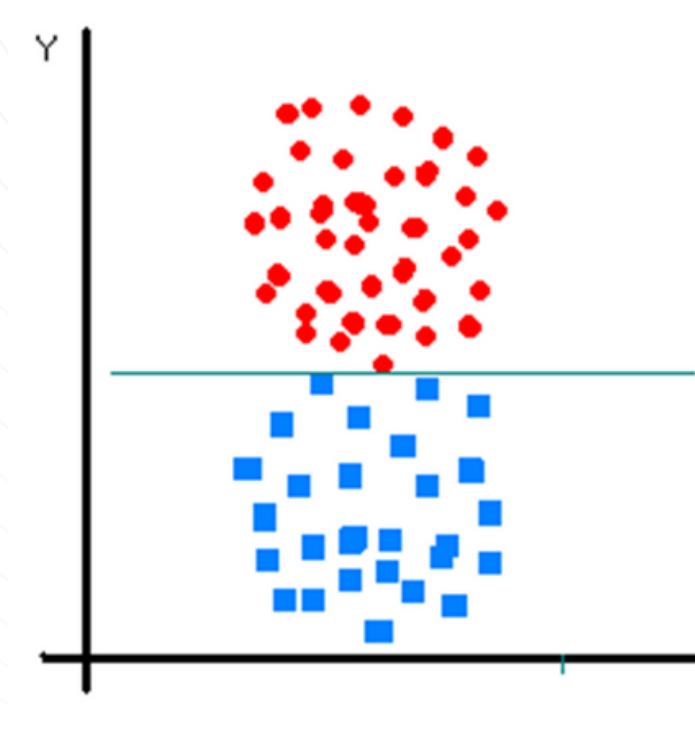


BIOMEDICAL STUDIES

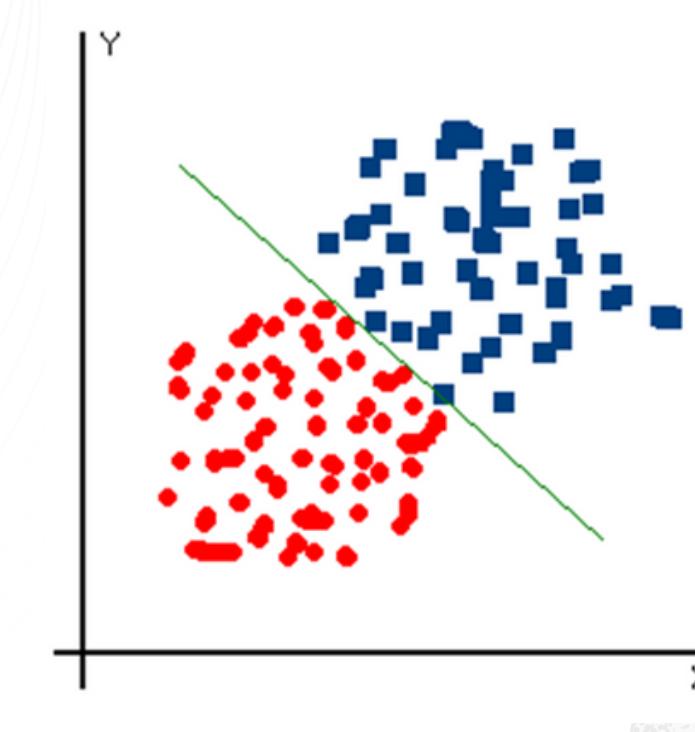
ILUSTRASI FUNGSI DISKRIMINAN



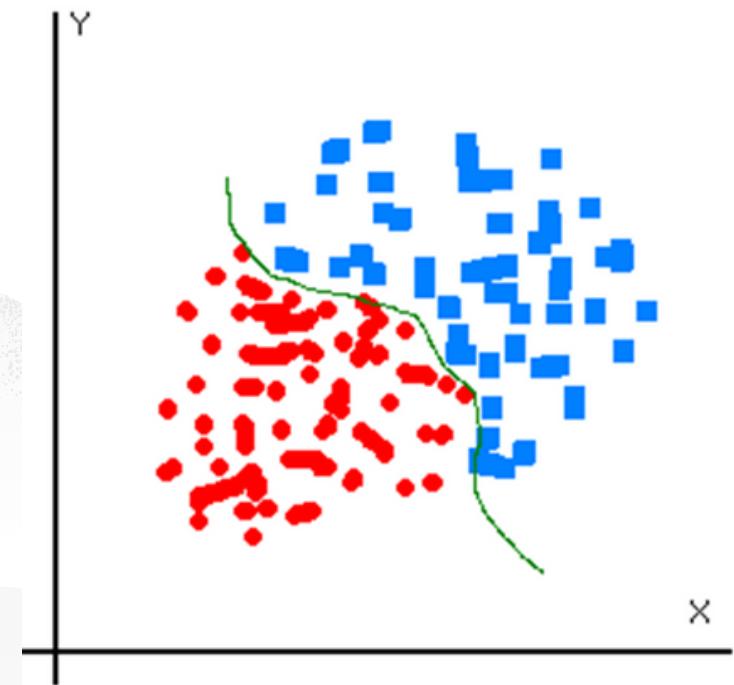
Peubah X mampu menjadi pembeda, tetapi Y tidak



Peubah y mampu menjadi pembeda, tetapi X tidak



Peubah X dan Y menjadi pembeda secara linear



Peubah X dan Y menjadi pembeda tetapi tidak linear

CARA MENDUGA FUNGSI DISKRIMINAN

Pendekatan Fisher

Pendekatan Fisher hanya digunakan jika jumlah gerombolnya sebanyak 2

- Hitung $a = \Sigma^{-1}(\mu_1 - \mu_2)$ dimana Σ adalah matriks ragam-peragam data
- Kelompokan objek yang memiliki karakteristik vector x ke gerombol Π_1 jika $a'x \geq h$, dan masukkan dalam gerombol Π_2 jika $a'x < h$ dengan $h = a'(\mu_1 + \mu_2) / 2$. Dengan kata lain, x akan dimasukkan ke gerombol yang paling dekat kemiripannya.

Konsep Jarak

Individu baru akan dikelompokkan ke dalam gerombol yang jarak vektor rataan populasinya lebih dekat.

$$d_j(x) = \{[x - \bar{x}_j]'\mathbf{S}^{-1}[x - \bar{x}_j]\}^{1/2}$$

Konsep Peluang Posterior

Individu baru akan diklasifikasikan ke dalam gerombol yang peluang posteriornya lebih besar

$$P(j|x) = \frac{e^{-\frac{1}{2}d_j^2(x)}}{e^{-\frac{1}{2}d_1^2(x)} + e^{-\frac{1}{2}d_2^2(x)}}$$

PENDEKATAN ANALISIS DISKRIMINAN

Analisis Diskriminan Linear (LDA)

Asumsi:

1. Sebaran data dalam setiap kelompok adalah normal ganda
2. Memiliki matriks ragam-peragam yang sama, serta
3. Biaya salah klasifikasi sama besar untuk setiap populasi.

Suatu objek x diklasifikasikan kepada populasi yang terdekat, yang dihitung menggunakan formula di atas. Atau, x akan diklasifikasikan berasal dari populasi ke- t jika

$$d_t^2(x) = \min_{j=1,\dots,k} \{d_j^2(x)\}$$

di mana $d_t^2(x) = (x - \mu_t)' \Sigma^{-1} (x - \mu_t) - 2 \ln(\pi_t)$

Seperti halnya pada bagian terdahulu, **mengklasifikasikan objek pengamatan ke populasi yang terdekat setara dengan mengklasifikasikan objek ke populasi dengan peluang posterior yang paling besar**. Pada kasus k buah populasi, peluang tersebut besarnya diperoleh dari :

$$P(t | x) = \frac{e^{-\frac{1}{2}d_t^2(x)}}{\sum_{j=1}^k e^{-\frac{1}{2}d_j^2(x)}} \quad t = 1, 2, \dots, k$$

PENDEKATAN ANALISIS DISKRIMINAN

Analisis Diskriminan Kuadratik (QDA)

Asumsi:

1. Sebaran data dalam setiap kelompok adalah normal ganda
2. Memiliki matriks ragam-peragam tidak sama

Suatu objek x diklasifikasikan kepada populasi yang terdekat, yang dihitung menggunakan formula di atas. Atau, x akan diklasifikasikan berasal dari populasi ke- t jika

$$d_t^2(x) = \min_{j=1,\dots,k} \{d_j^2(x)\}$$

di mana

$$d_t^2(x) = (x - \mu_t)' \Sigma^{-1} (x - \mu_t) + \ln |\Sigma| - 2 \ln(\pi_t)$$

Rumus peluang posterior yang digunakan sama seperti LDA

TAHAPAN ANALISIS DISKRIMINAN

Algoritma:

1. **Membagi data menjadi dua bagian:** data latih (train data) dan data uji (test data) Cara untuk membagi dapat dilakukan dengan pengambilan acak. Proporsi data latih dapat sebesar 70% atau 80% dari data asal.
2. **Dengan menggunakan data latih,** lakukan:
 - Uji normal ganda
 - Uji asumsi kesamaan ragam. Jika uji ini menghasilkan kesimpulan matriks ragam-peragam sama, maka digunakan Linear Discriminant Analisis (LDA). Jika tidak, maka digunakan Quadratic Discriminant Analisis (QDA).
 - Estimasi koefisien analisis diskriminan
 - Evaluasi kemampuan klasifikasi analisis diskriminan
3. **Evaluasi kemampuan klasifikasi** menggunakan data uji

CONTOH SOAL

APPLIED MULTIVARIATE STATISTICAL ANALYSIS

BOOK BY DEAN W. WICHERN AND RICHARD A. JOHNSON

Exercises 11.19

- a) Using the original data sets \mathbf{X}_1 and \mathbf{X}_2 given in Example 11.6, calculate $\bar{\mathbf{x}}_i$, \mathbf{S}_i , $i = 1, 2$, and $\mathbf{S}_{\text{pooled}}$, verifying the results provided for these quantities in the example.

Example 11.6.

Consider the following data matrices. We shall assume that the $n_1 = n_2 = 3$ bivariate observations were selected randomly from two populations π_1 and π_2 with a common covariate matrix.

$$\mathbf{X}_1 = \begin{bmatrix} 2 & 12 \\ 4 & 10 \\ 3 & 8 \end{bmatrix} \quad \text{dan} \quad \mathbf{X}_2 = \begin{bmatrix} 5 & 7 \\ 3 & 9 \\ 4 & 5 \end{bmatrix}$$

CONTOH SOAL

$$\mathbf{x}_1 = \begin{bmatrix} 2 & 12 \\ 4 & 10 \\ 3 & 8 \end{bmatrix}; \quad \bar{\mathbf{x}}_1 = \begin{bmatrix} 3 \\ 10 \end{bmatrix}, \quad \mathbf{s}_1 = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

$$\mathbf{x}_2 = \begin{bmatrix} 5 & 7 \\ 3 & 9 \\ 4 & 5 \end{bmatrix}; \quad \bar{\mathbf{x}}_2 = \begin{bmatrix} 4 \\ 7 \end{bmatrix}, \quad \mathbf{s}_2 = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

Menghitung $\mathbf{S}_{\text{pooled}}$ dapat menggunakan persamaan berikut:

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

sehingga matriks kovarian gabungannya adalah

$$\begin{aligned} \mathbf{S}_{\text{pooled}} &= \left[\frac{3 - 1}{(3 - 1) + (3 - 1)} \right] \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix} \\ &\quad + \left[\frac{3 - 1}{(3 - 1) + (3 - 1)} \right] \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix} \\ \mathbf{S}_{\text{pooled}} &= \frac{1}{2} \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix} \end{aligned}$$

CONTOH SOAL

b) Using the calculations in Part a, compute **Fisher's linear discriminant function**, and use it to classify the sample observations according to Rule (11-35).

Diketahui nilai \bar{x}_1 , \bar{x}_2 dan S_{pooled} dibagian (a). Kemudian hitung S_{pooled}^{-1} dimana

$$S_{\text{pooled}}^{-1} = \begin{bmatrix} 1.333 & 0.333 \\ 0.333 & 0.333 \end{bmatrix}$$

Fungsi diskriminan prior yang sama adalah

$$\begin{aligned}\hat{y} &= \hat{a}'x = [\bar{x}_1 - \bar{x}_2]' S_{\text{pooled}}^{-1} x \\ &= [-1 \quad 3] \begin{bmatrix} 1.333 & 0.333 \\ 0.333 & 0.333 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= -0.333x_1 + 0.667x_2\end{aligned}$$

dengan

$$\begin{aligned}\bar{y}_1 &= \hat{a}'\bar{x}_1 = [-0.333 \quad 0.667] \begin{bmatrix} 3 \\ 10 \end{bmatrix} = 5.667 \\ \bar{y}_2 &= \hat{a}'\bar{x}_2 = [-0.333 \quad 0.667] \begin{bmatrix} 4 \\ 7 \end{bmatrix} = 3.333\end{aligned}$$

dan titik tengah rata-ratanya

$$\bar{m} = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \frac{1}{2}(5.667 + 3.333) = 4.5$$

CONTOH SOAL

- Alokasikan x_0 ke π_1 jika $-0.333x_1 + 0.667x_2 - 4.5 \geq 0$ dan
- alokasikan x_0 ke π_2 jika $-0.333x_1 + 0.667x_2 - 4.5 < 0$ sehingga diperoleh

π_1		π_2		
$\hat{a}'x - \bar{m}$	klasifikasi	$\hat{a}'x - \bar{m}$	klasifikasi	
$=(-0.333*2)+(0.667*12)-4.5$	2.838	π_1	-1.496	π_2
$=(-0.333*4)+(0.667*10)-4.5$	0.838	π_1	0.505	π_1
$=(-0.333*3)+(0.667*8)-4.5$	-0.163	π_2	-2.497	π_2
				$=(-0.333*5)+(0.667*7)-4.5$
				$=(-0.333*3)+(0.667*9)-4.5$
				$=(-0.333*4)+(0.667*5)-4.5$

$$\mathbf{X}_1 = \begin{bmatrix} 2 & 12 \\ 4 & 10 \\ 3 & 8 \end{bmatrix} \quad \text{dan} \quad \mathbf{X}_2 = \begin{bmatrix} 5 & 7 \\ 3 & 9 \\ 4 & 5 \end{bmatrix}$$

CONTOH SOAL

c) Classify the sample observations on the basis of **smallest squared distance** of the observations from the group means x_1 and x_2 • [See (11-54) .] Compare the results with those in Part b. Comment.

ATURAN *smallest squared distance* $D_i^2(x)$

- Hitung $D_i^2(x) = (x - \bar{x}_i)' S_{\text{pooled}}^{-1} (x - \bar{x}_i)$, $i = 1, 2$ untuk semua kelompok
- Nilai D_i yang terkecil merupakan lokasi dari kelompoknya

CONTOH SOAL

c) Classify the sample observations on the basis of **smallest squared distance** of the observations from the group means x_1 and x_2 • [See (11-54) .] Compare the results with those in Part b. Comment.

Dari persamaan diatas untuk \bar{x}_1 diperoleh hasil

$$D_1^2(\mathbf{x}) = \begin{pmatrix} 2 - 3 \\ 12 - 10 \end{pmatrix}' \begin{bmatrix} 1.333 & 0.333 \\ 0.333 & 0.333 \end{bmatrix} \begin{pmatrix} 2 - 3 \\ 12 - 10 \end{pmatrix} = 1.333$$

$$D_1^2(\mathbf{x}) = \begin{pmatrix} 4 - 3 \\ 10 - 10 \end{pmatrix}' \begin{bmatrix} 1.333 & 0.333 \\ 0.333 & 0.333 \end{bmatrix} \begin{pmatrix} 4 - 3 \\ 10 - 10 \end{pmatrix} = 1.333$$

$$D_1^2(\mathbf{x}) = \begin{pmatrix} 3 - 3 \\ 8 - 10 \end{pmatrix}' \begin{bmatrix} 1.333 & 0.333 \\ 0.333 & 0.333 \end{bmatrix} \begin{pmatrix} 3 - 3 \\ 8 - 10 \end{pmatrix} = 1.332$$

$$D_1^2(\mathbf{x}) = \begin{pmatrix} 5 - 3 \\ 7 - 10 \end{pmatrix}' \begin{bmatrix} 1.333 & 0.333 \\ 0.333 & 0.333 \end{bmatrix} \begin{pmatrix} 5 - 3 \\ 7 - 10 \end{pmatrix} = 4.333$$

$$D_2^2(\mathbf{x}) = \begin{pmatrix} 3 - 3 \\ 9 - 10 \end{pmatrix}' \begin{bmatrix} 1.333 & 0.333 \\ 0.333 & 0.333 \end{bmatrix} \begin{pmatrix} 3 - 3 \\ 9 - 10 \end{pmatrix} = 0.333$$

$$D_2^2(\mathbf{x}) = \begin{pmatrix} 4 - 3 \\ 5 - 10 \end{pmatrix}' \begin{bmatrix} 1.333 & 0.333 \\ 0.333 & 0.333 \end{bmatrix} \begin{pmatrix} 4 - 3 \\ 5 - 10 \end{pmatrix} = 6.328$$

untuk \bar{x}_2 diperoleh hasil

$$\rightarrow D_1^2(\mathbf{x}) = \begin{pmatrix} 2 - 4 \\ 12 - 7 \end{pmatrix}' \begin{bmatrix} 1.333 & 0.333 \\ 0.333 & 0.333 \end{bmatrix} \begin{pmatrix} 2 - 4 \\ 12 - 7 \end{pmatrix} = 6.997$$

$$\rightarrow D_1^2(\mathbf{x}) = \begin{pmatrix} 4 - 4 \\ 10 - 7 \end{pmatrix}' \begin{bmatrix} 1.333 & 0.333 \\ 0.333 & 0.333 \end{bmatrix} \begin{pmatrix} 4 - 4 \\ 10 - 7 \end{pmatrix} = 2.997$$

$$\rightarrow D_1^2(\mathbf{x}) = \begin{pmatrix} 3 - 4 \\ 8 - 7 \end{pmatrix}' \begin{bmatrix} 1.333 & 0.333 \\ 0.333 & 0.333 \end{bmatrix} \begin{pmatrix} 3 - 4 \\ 8 - 7 \end{pmatrix} = 1$$

$$\rightarrow D_2^2(\mathbf{x}) = \begin{pmatrix} 5 - 4 \\ 7 - 7 \end{pmatrix}' \begin{bmatrix} 1.333 & 0.333 \\ 0.333 & 0.333 \end{bmatrix} \begin{pmatrix} 5 - 4 \\ 7 - 7 \end{pmatrix} = 1.333$$

$$\rightarrow D_2^2(\mathbf{x}) = \begin{pmatrix} 3 - 4 \\ 9 - 7 \end{pmatrix}' \begin{bmatrix} 1.333 & 0.333 \\ 0.333 & 0.333 \end{bmatrix} \begin{pmatrix} 3 - 4 \\ 9 - 7 \end{pmatrix} = 1.333$$

$$\rightarrow D_2^2(\mathbf{x}) = \begin{pmatrix} 4 - 4 \\ 5 - 7 \end{pmatrix}' \begin{bmatrix} 1.333 & 0.333 \\ 0.333 & 0.333 \end{bmatrix} \begin{pmatrix} 4 - 4 \\ 5 - 7 \end{pmatrix} = 1.333$$

CONTOH SOAL

c) Classify the sample observations on the basis of **smallest squared distance** of the observations from the group means x_1 and x_2 • [See (11-54) .] Compare the results with those in Part b. Comment.

π_1			π_2		
$D_1^2(\mathbf{x})$	$D_2^2(\mathbf{x})$	klasifikasi	$D_1^2(\mathbf{x})$	$D_2^2(\mathbf{x})$	klasifikasi
1.333	6.997	π_1	4.333	1.333	π_2
1.333	2.997	π_1	0.333	1.333	π_1
1.332	1	π_2	6.328	1.333	π_2

Kedua rule memberikan hasil klasifikasi yang sama

STUDI KASUS 1

Nama peubah	Keterangan
Type	Tipe anggur, terdiri dari: 1 (59 amatan), 2 (71 amatan), dan 3 (48 amatan)
Alcohol	Kadar alkohol
Malic	Kadar asam malat
Ash	Abu
Alcalinity	Alkalinitas abu
Magnesium	Kadar Magnesium
Phenols	Total fenol
Flavonoids	Kadar fenol flavonoid
Nonflavonoids	Kadar fenol nonflavonoid
Proanthocyanins	Proantosianidin
Color	Intensitas warna
Hue	Spektrum warna
Dilution	Dilusi anggur D280/OD315
Proline	Kadar prolin

Dataset Wine (Package R **rattle**) ini berisi hasil analisis kimia anggur yang tumbuh di daerah tertentu di Italia. Tiga jenis anggur direpresentasikan dalam 178 sampel, dengan hasil dari 13 analisis kimia yang dicatat untuk setiap sampel. Peubah Type telah diubah menjadi peubah kategorik.

STUDI KASUS 2

IPK	Skor tes	Status
2.96	596	Lanjut
3.14	473	Lanjut
3.22	482	Lanjut
3.29	527	Lanjut
3.69	505	Lanjut
3.46	693	Lanjut
3.03	626	Lanjut
3.19	663	Lanjut
...
3.01	453	Pindah
3.03	414	Pindah
3.04	446	Pindah

- Sebuah Program Pascasarjana di sebuah perguruan tinggi melakukan evaluasi terhadap keberhasilan studi mahasiswa S2 di semester pertama.
- Data menunjukkan bahwa ada **tiga** pengelompokan status mahasiswa yang berkaitan dengan kelanjutan studi di semester 2, yaitu **Lanjut** ke semester 2, **Pindah** ke program studi lain, dan **Drop Out**.
- Banyak mahasiswa yang tidak dapat melanjutkan ke semester dua dengan lancar, diduga karena kurang seleksinya kualitas input mahasiswa.
- Oleh karena itu, Ketua Program Studi tersebut berencana membuat kriteria seleksi untuk menyaring mahasiswa yang berkualitas.
- Variabel yang digunakan sebagai dasar penentuan kriteria seleksi adalah IPK S1 dan skor hasil tes masuk S2. Lakukan analisis yang sesuai untuk tujuan tersebut.



THANKYOU

Any Question, Just Ask don't be shy