Analisis Diskriminan (Discriminant Analysisi)

Bahan Kuliah Secara Daring
Mahasiswa Departemen Statistika-FMIPA-IPB
Oleh: Dr. Ir. Budi Susetyo

Latar Belakang

- Jika dalam analisis gerombol kita melakukan pengelompokan individu/objek yang ada berdasarkan kemiripan kedalam beberapa gerombol, yang menjadi pertanyaan adalah bagaimana jika ada individu baru? Individu baru tersebut termasuk dalam gerombol yang mana?
- Untuk dapat memasukkan individu baru kedalam gerombol yang ada maka harus ada suatu fungsi yang dapat membedakan antar gerombol. Fungsi tersebut disebut dengan fungsi diskriminan.
- Jadi analisisi gerombol dan analisis diskriminan adalah dua metode yang erat hubungannya dalam mengelompokkan objek dan memasukkan objek baru dalam kelompok

Contoh Penerapan Fungsi Diskriminan

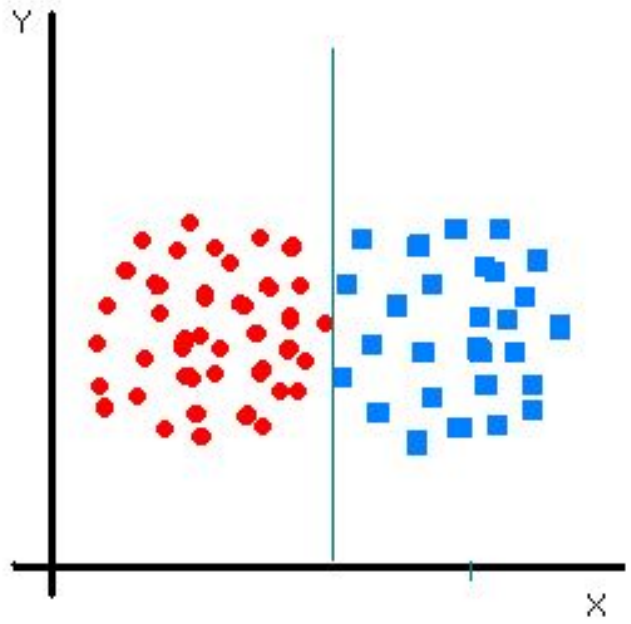
Fungsi diskriminan dipakai untuk memutuskan kasus berikut:

- Berdasarkan beberapa indicator nilai beberapa maple siswa SMA selama beberapa semester maka dapat diputuskan apakah mahasiswa ini bisa diterima atau tidak dalam salah satu PT tertentu?
- Berdasarkan data umur, pekerjaan, penghasilan, kepemilikan asset dan jumlah anggota keluarga, seseorang yang mengajukan kredit jika diterima apakah dapat dikategorikan punya potensi tidak bermasalah, sedikit bermasalah dan akan bermasalah dalam pengembalian pinjaman?
- Dengan melihat gejala-gejala yang nampak pada seseorang, bagaimana dokter bisa menduga penyakit apa yang diderita orang tersebut?

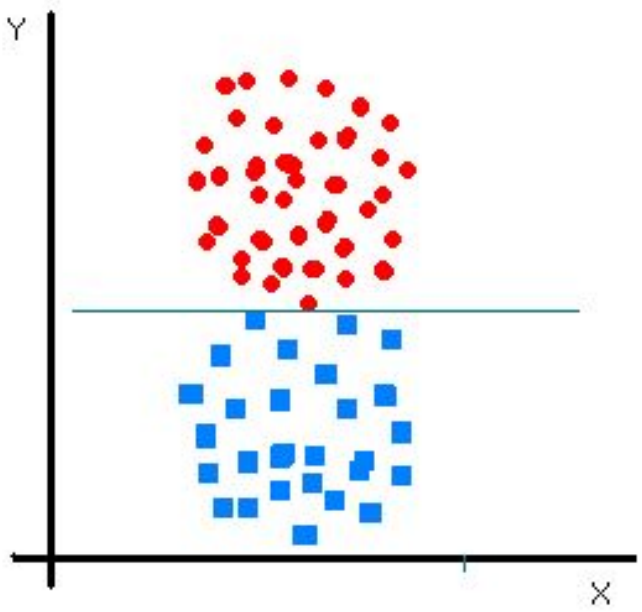
Apa Itu Fungsi Diskriminan?

- Merupakan fungsi dari beberapa peubah penciri yang dapat membedakan karakteristik individu antar gerombol
- Dari banyak peubah tersebut, fungsi diskriminan akan menghasilkan sebuah indeks
- Berdasarkan kriteria tertentu, indeks ini digunakan untuk mengklasifikasikan suatu objek masuk dalam suatu gerombol
- Tidak selalu (bahkan jarang) diperoleh fungsi diskriminan dengan tingkat ketepatan yang sempurna
- Fungsi Diskriminan memiliki ukuran yang menggambarkan tingkat ketepatan/ketidaktepatan

Ilustrasi Fungsi Diskriminan

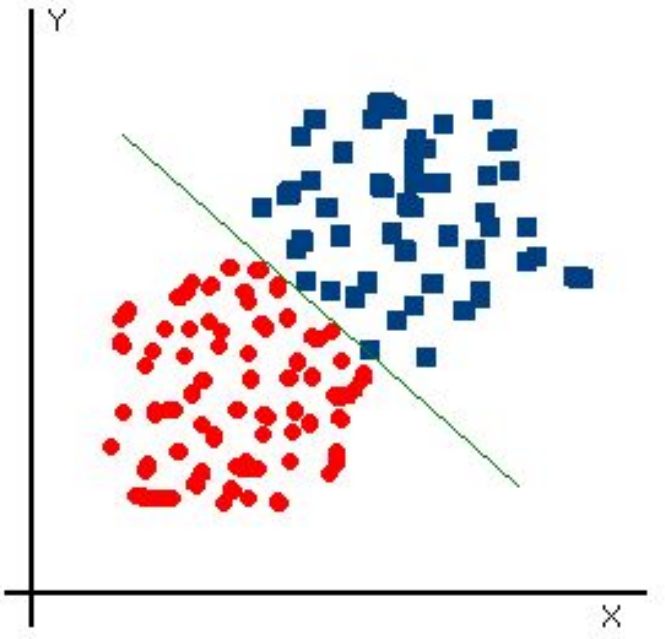


Peubah X mampu menjadi pembeda, tetapi Y tidak

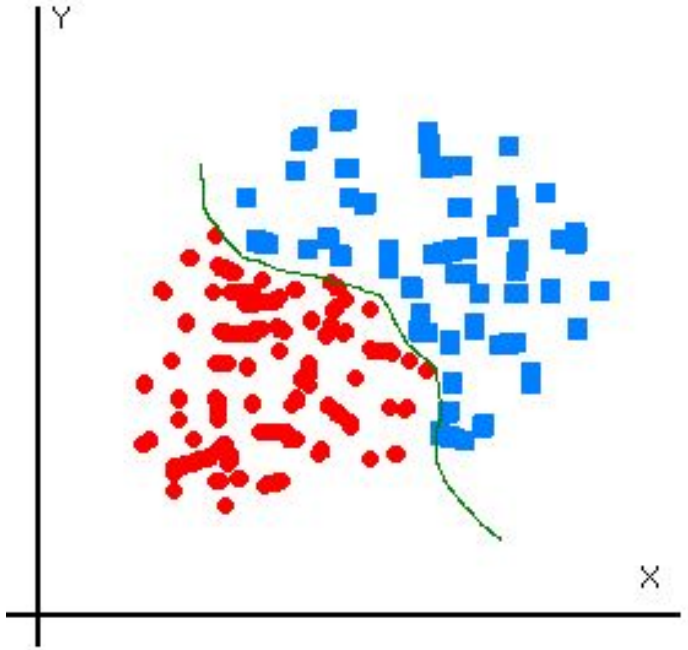


Peubah Y mampu menjadi pembeda, tetapi X tidak

Fungsi Diskriminan



Peubah X dan Y menjadi pembeda secara linear



Peubah X dan Y menjadi pembeda tetapi tidak linear

Bagaimana Cara Menduga Fungsi Diskriminan

Banyak metode untuk menduga fungsi diskriminan, yang tergantung kepada jumlah gerombol dan juga pola penggerombolan objek.

- **Pendekatan Fisher**

Pendekatn Fisher hanya digunakan jika jumlah gerombolnya hanya ada 2. Logika pendekatan Fisher bisa dituliskan sebagai berikut:

- Cari **a** sehingga jarak antara $E(\mathbf{a}'\mathbf{x}) = \mathbf{a}'\boldsymbol{\mu}_1$ di $\prod_1$ (gerombol 1) dengan $E(\mathbf{a}'\mathbf{x}) = \mathbf{a}'\boldsymbol{\mu}_2$ di $\prod_2$ (gerombol 2) maksimum, atau memaksimumkan $|\mathbf{a}'\boldsymbol{\mu}_1 - \mathbf{a}'\boldsymbol{\mu}_2|$ dengan kendala $\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} = 1$.

Pendekatan Fisher (lanjutan….)

- Diperoleh a = $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ dimana Σ adalah matriks ragam-peragam data
- Kelompokan objek yang memiliki karakteristik vector x ke gerombol Π_1 jika $\mathbf{a}'\mathbf{x} \geq h$, dan masukkan dalam gerombol Π_2 jika <h dengan $h = \mathbf{a}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) / 2$. Dengan kata lain, **x** akan dimasukkan ke gerombol yang paling dekat kemiripannya.

Ilustrasi penerapan Pendekatan Fisher

Telah diketahui karakterisik ikan salmon yang berasal dari Alaska dan Kanada dengan melihat pertumbuhannya ketika hidup di air tawar (x1) dan ketika hidup di airl laut (x2). Dalam suatu penangkapan ikan salmon diinginkan bisa diidentifikasi apakah ikan yang tertangkap berasal dari Alaska atau Kanada. Lima puluh ikan diambil dari setiap tempat, kemudian diukur x1 dan x2 untuk mengetes model Fisher yang akan didapat(Minitab, Inc).

Ilustrasi penerapan Pendekatan Fisher

$$\mathbf{S} = \begin{bmatrix} 676.0 & -649.1 \\ -649.1 & 2138.1 \end{bmatrix},$$

serta vektor rataan untuk masing masing populasi

ikan dari Alaska $\quad \bar{\mathbf{x}}_1 = \begin{bmatrix} 98.38 \\ 429.66 \end{bmatrix}$

ikan dari Canada $\quad \bar{\mathbf{x}}_2 = \begin{bmatrix} 137.46 \\ 366.62 \end{bmatrix}$

Sehingga diperoleh vektor fungsi diskriminan

$$\mathbf{a} = \mathbf{S}^{-1}(\bar{x}_1 - \bar{x}_2) = \mathbf{a} = \mathbf{S}^{-1}\bar{\mathbf{x}}_1 - \mathbf{S}^{-1}\bar{\mathbf{x}}_2 = \begin{bmatrix} -0.0521 \\ 0.0137 \end{bmatrix},$$

dan batasan nilai bagi kedua populasi sebesar $\hat{k} = -0.5657$.

Ilustrasi penerapan Pendekatan Fisher

- Dengan demikian, jika kita memiliki suatu pengamatan baru **x** = (x_1, x_2) maka kita akan memasukkannya ke populasi 1 (ikan dari Alaska) jika

$$-0.0521\, x_1 + 0.0137\, x_2 \geq -0.5657$$

dan jika sebaliknya maka kita masukkan ke populasi ke-2 (Kanada).

- Sebagai teladan, jika diperoleh sebuah ikan dengan nilai pengamatan **x** = (103, 405), maka nilai

a'**x** = -0.0521 (103) + 0.0137 (405) = 10.918, dan kita masukkan ke dalam jenis ikan Alaska

Hasil Klasifikasi berdasarkan Pendekatan Fisher --

Bentuk tabel salah klasifikasinya adalah:

		Hasil Klasifikasi		% Salah Klasifikasi
		Alaska	Canada	
Seharusnya	Alaska	44	6	12%
	Canada	1	49	2%
Total				7%

Fungsi diskriminan yang dibentuk diatas memberikan tingkat kesalahan sebesar 7%

Metode Pendekatan Lain --

Cara lain untuk melakukan klasifikasi adalah menggunakan konsep jarak terhadap vektor rataan populasi yang paling dekat. Artinya jika ada suatu pengamatan baru $\mathbf{x} = (x_1, x_2)$, maka pengamatan atau objek baru ini akan kita masukkan ke dalam populasi ke-1 (Π_1) hanya jika jarak **x** terhadap vektor rataan populasi ke-1 lebih dekat daripada jarak **x** terhadap vektor rataan populasi ke-2. Jarak antara **x** terhadap vektor rataan diperoleh menggunakan formula mahalanobis, yaitu:

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

- Misalkan untuk pengamatan **x** = (103, 405) seperti pada ilustrasi sebelumnya. Setelah dihitung jarak dengan populasi 1 dan populasi 2 diperoleh:

  $d_1(\mathbf{x}) = 0.5421$

  $d_2(\mathbf{x}) = 1.3322$

- Karena $d_1(\mathbf{x}) < d_2(\mathbf{x})$ maka **x** diklasifikasikan berasal dari populasi 1 (ikan dari Alaska).

Metode Pendekatan Lain --

Pendekatan lain yang juga dapat digunakan adalah menggunakan peluang posterior. Suatu pengamatan **x** = (x_1, x_2) akan diklasifikasikan ke dalam populasi Π_1 hanya jika peluang posteiornya lebih besar dari pada peluang posterior masuk ke Π_2, dan sebaliknya. Peluang posterior masuk ke dalam Π_j adalah

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

- Misalkan untuk pengamatan **x** = (103, 405) seperti pada ilustrasi sebelumnya. Setelah dihitung peluang posteriornya diperoleh:

  $P(1|\mathbf{x}) = 0.677$

  $P(2|\mathbf{x}) = 0.323$.

- Karena $P(1|\mathbf{x}) > P(2|\mathbf{x})$ maka **x** sekali lagi diklasifikasikan berasal dari Alaska.

Analisis Diskriminan untuk k Populasi yang Menyebar Normal

- Ada konsep sebaran prior
- Seringkali juga perlu mempertimbangkan biaya salah klasifikasi sehingga perlu mencari fungsi yang meminimumkan expected cost of missclassification

$$\sum_{t=1}^{k} \pi_t \sum_{s=1}^{k} P(s \mid t) c(s \mid t)$$

Analisis Diskriminan Linear (1)

- Dengan asumsi bahwa setiap populasi menyebar normal ganda dan matriks ragam-peragam sama di setiap populasi serta biaya salah klasifikasi sama besar di setiap populasi, maka aturan yang paling sederhana pada klasifikasi bisa dinyatakan dalam fungsi kuadrat jarak yaitu

$$d_t(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_t)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_t) - 2\,ln(\pi_t)$$

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

$$d_t^2(\mathbf{x}) = \min_{j=1,\ldots,k} \left\{ d_j^2(\mathbf{x}) \right\}$$

Analisis Diskriminan Linear (2)

- 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 :

$$P(t \mid x) = \frac{e^{-\frac{1}{2}d_t^2(x)}}{\sum_{j=1}^{k} e^{-\frac{1}{2}d_j^2(x)}} \qquad t = 1, 2, \ldots, k$$

Menduga Tingkat Salah Klasifikasi

- *Error Rate*, dugaan tingkat kesalahan di populasi ke-s adalah:

$$\hat{E}R(s) = \sum_{t=1, t \neq s}^{k} P(t \mid s)$$

Menduga Tingkat Salah Klasifikasi (1)

Pendugaan Tingkat Kesalahan dengan Validasi Silang

- jika ada *n* objek pengamatan, maka hanya (*n* – 1) pengamatan yang digunakan sebagai gugus data pembentukan fungsi diskriminan
- satu pengamatan sisanya digunakan untuk evaluasi
- proses di atas diulang sebanyak n kali, satu kali untuk setiap data yang disisihkan
- proporsi kesalahan adalah dugaan tingkat kesalahan

Menduga Tingkat Salah Klasifikasi (2)

posterior probability error rate

$$\mathrm{PPER}_{1t} = 1 - \frac{1}{\pi_t \sum_{j=1}^{k} n_j} \sum_{D_t} P(t \mid \mathbf{x})$$

Simple PPER

$$\mathrm{PPER}_{2t} = 1 - \frac{1}{\pi_t} \sum_{j=1}^{k} \frac{\pi_j}{n_j} \sum_{D_{jt}} P(t \mid x),$$

Stratified PPER

Analisis Diskriminan Kuadratik

Asumsi: populasi menyebar normal ganda, namun matriks ragam-peragamnya tidak sama

Suatu objek pengamatan tertentu **x** akan dimasukkan ke populasi ke-t jika

$$d_t^2(\mathbf{x}) = \min_{j=1,\ldots,k} \left\{ d_j^2(\mathbf{x}) \right\}$$

dengan $d_j^2(\mathbf{x})$ adalah kuadrat jarak yang didefinisikan (sedikit berbeda dengan kasus fungsi diskriminan linear) sebagai:

$$d_j^2(\mathbf{x}) = [\mathbf{x} - \bar{\mathbf{x}}_j]' \mathbf{S}^{-1} [\mathbf{x} - \bar{\mathbf{x}}_j] + ln\, |\mathbf{S}_j| - 2\, ln(\pi_t);\ j = 1, 2, \ldots, k.$$

Formula peluang posterior sama persis dengan formula peluang posterior untuk kasus diskriminan linear kecuali pada formula $d_j^2(\mathbf{x})$.

Pemilihan Peubah Yang digunakan pada Analisis Diskriminan (1)

- Agar supaya peubah yang digunakan dalam fungsi diskriminan tidak terlalu banyak, maka dapat dipilih beberapa peubah yang paling penting dan memiliki daya pembeda yang kuat
- Beberapa metode yang dapat digunakan seperti pada penerapan regresi liniear berganda:

forward selection: dimulai dengan memilih satu peubah yang paling penting, dan dilanjutkan dengan pemilihan peubah penting lain satu demi satu menggunakan suatu kriteria tertentu.

Pemilihan Peubah Yang digunakan pada Analisis Diskriminan (2)

backward selection: dimulai dengan model penuh, yaitu memuat semua peubah. Di setiap tahap dilakukan pembuangan peubah yang paling tidak penting satu demi satu dengan kriteria yang sama dengan prosedur forward. Proses diteruskan hingga tidak ada lagi peubah yang dikeluarkan. Prosedur ini dikenal sebagai prosedur.

stepwise selection: Kombinasi antara kedua prosedur di atas. Di setiap tahap dimungkinkan ada peubah yang masuk sekaligus ada peubah yang dikeluarkan, berdasarkan kriteria tertentu yang ditetapkan pada awal proses.

Terimakasih