Analisis Diskriminan

Tujuan Utama

Memperoleh fungsi diskriminan, yaitu fungsi yang mampu digunakan membedakan suatu objek masuk ke dalam populasi tertentu berdasarkan pengamatan terhadap objek tersebut

Contoh Fungsi Diskriminan

- Dengan melihat gejala-gejala yang nampak pada seseorang, dokter bisa menduga penyakit apa yang diderita orang tersebut.
- Dengan melihat warna, merasakan, dan menghirup asap rokok, penilai bisa mengetahui kelas kualitas tembakau.
- Dengan mengetahui berbagai indikator yang berupa variabel derivatif keuangan sebuah bank, kita bisa menilai kesehatan bank tersebut.

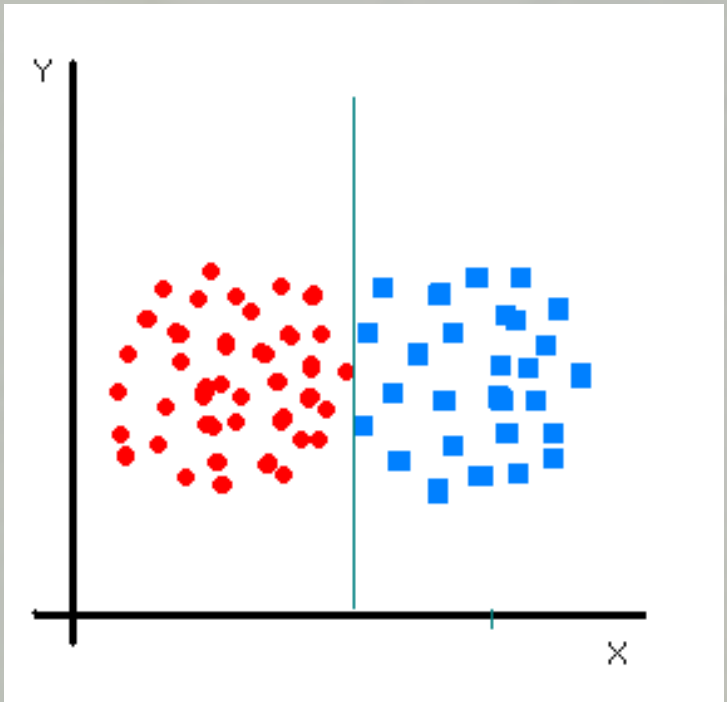
Fungsi Diskriminan

- Merupakan kombinasi dari beberapa peubah, satu peubah saja umumnya tidak mencukupi
- Dari banyak peubah, menggunakan fungsi diskriminan diperoleh sebuah indeks
- Berdasarkan kriteria tertentu, dengan indeks ini kita mengklasifikasikan objek

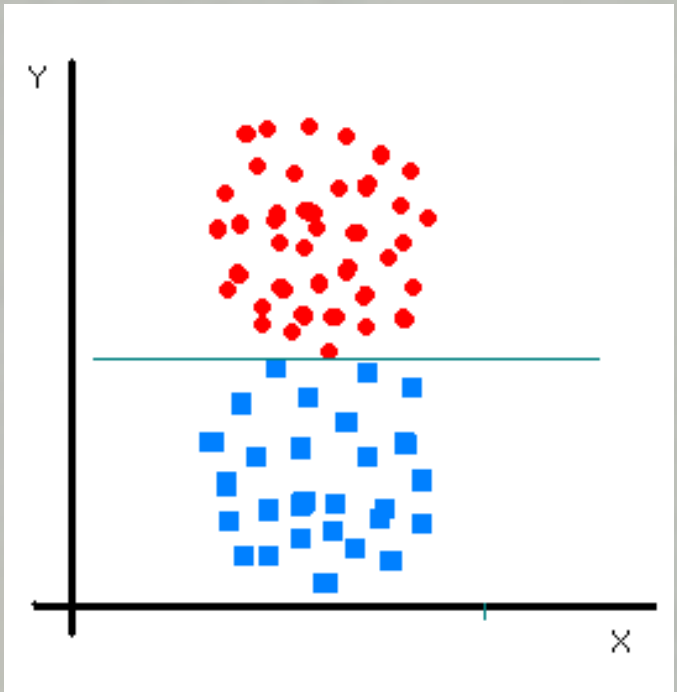
Fungsi Diskriminan

- Tidak selalu (bahkan jarang) diperoleh fungsi diskriminan dengan tingkat ketepatan yang sempurna
- Fungsi Diskriminan memiliki ukuran yang menggambarkan tingkat ketepatan

Fungsi Diskriminan

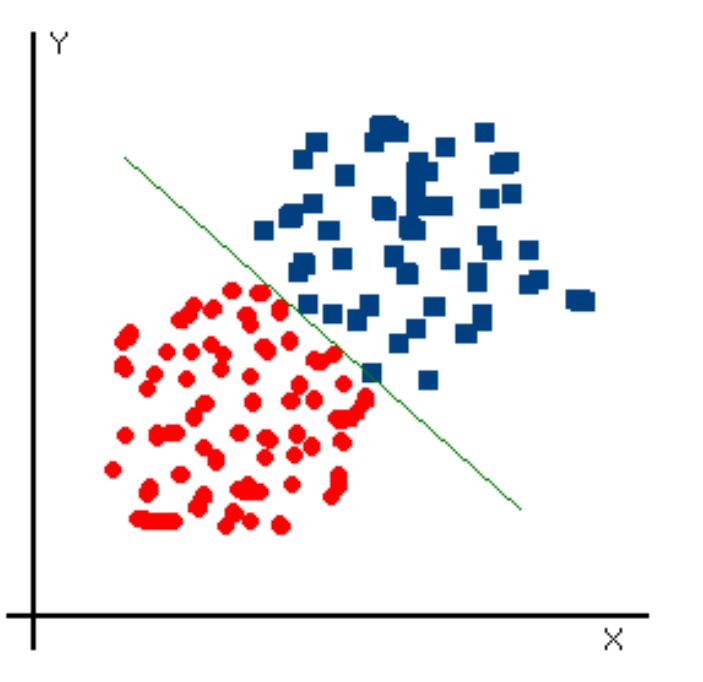


X mampu menjadi pembeda, tetapi Y tidak

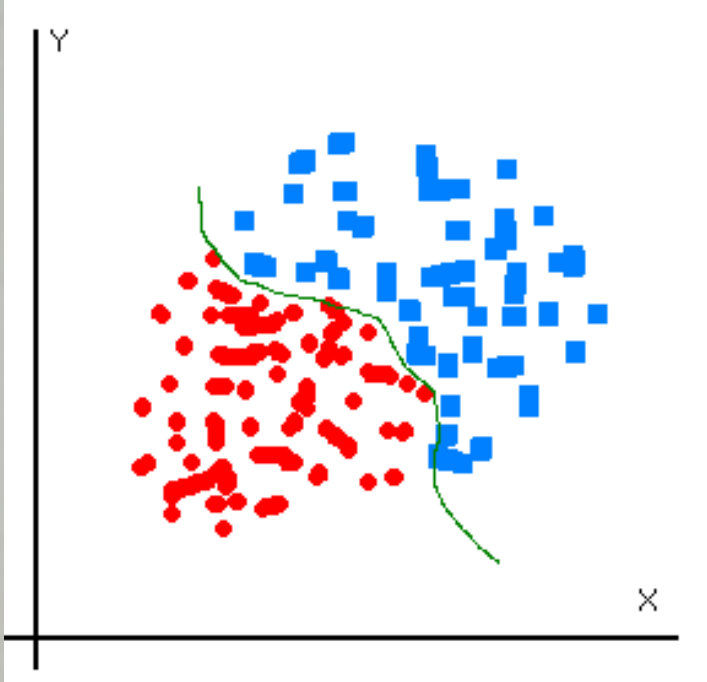


Y mampu menjadi pembeda, tetapi X tidak

Fungsi Diskriminan



X dan Y saja tidak mampu menjadi pembeda, tetapi kombinasi linearnya bisa



Membutuhkan fungsi non-linear dari X dan Y untuk bisa membedakan

Pendekatan Fisher

- Hanya untuk 2 populasi
- pendekatan Fisher bisa dituliskan sebagai berikut:

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

Pendekatan Fisher

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

dan kita akan mengelompokkan $\mathbf{x}$ ke Π_1 jika $\mathbf{a}'\mathbf{x} \geq h$, dan kebalikannya kita masukkan $\mathbf{x}$ ke dalam Π_2, dengan $h = \mathbf{a}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) / 2$. Dengan kata lain, $\mathbf{x}$ akan dimasukkan ke populasi yang paling dekat dengannya.

Pendekatan Fisher -- ILUSTRASI

Dalam rangka mengatur penangkapan ikan salmon, sangat diinginkan bisa mengidentifikasi apakah ikan yang tertangkap berasal dari Alaska atau Kanada. Lima puluh ikan diambil dari setiap tempat, dan pertumbuhan diameternya diukur ketika ikan-ikan itu hidup di air tawar dan ketika hidup di air laut. Tujuannya adalah untuk mengetahui apakah ikan yang tertangkap di kemudian hari berasal dari Alaska atau dari Kanada (Minitab, Inc).

Pendekatan Fisher -- ILUSTRASI

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

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 $\hat{k} = -0.5657$.

Pendekatan Fisher -- ILUSTRASI

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. 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 populasi 1

Pendekatan Fisher -- ILUSTRASI

Bentuk tabel salah klasifikasinya adalah:

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

Pendekatan Fisher -- ILUSTRASI

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 $\mathbf{x}$ terhadap vektor rataan populasi ke-1 lebih dekat daripada jarak $\mathbf{x}$ terhadap vektor rataan populasi ke-2. Jarak antara $\mathbf{x}$ terhadap vektor rataan diperoleh menggunakan formula mahalanobis, yaitu:

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

Pendekatan Fisher -- ILUSTRASI

Misalkan untuk pengamatan $\mathbf{x}$ = (103, 405) seperti pada ilustrasi sebelumnya

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

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

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

Pendekatan Fisher -- ILUSTRASI

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

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

Pendekatan Fisher -- ILUSTRASI

Kembali pada **x** ilustrasi di atas dihasilkan $P(1|\mathbf{x}) = 0.677$ dan $P(2|\mathbf{x}) = 0.323$. Sehingga 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
- 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

- Asumsi : multivariate normal dengan matriks ragam-peragam sama di setiap populasi
- Asumsi : Biaya salah klasifikasi sama besar di setiap populasi

Analisis Diskriminan Linear

- 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

- 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

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

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

Multivariate normal 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 $d_j^2(\mathbf{x})$.

Penyeleksian Peubah pada Analisis Diskriminan

Dimulai dengan memilih satu peubah yang paling penting, dan dilanjutkan dengan pemilihan peubah penting lain satu demi satu menggunakan suatu kriteria tertentu. Salah satu kriterianya adalah dengan menentukan taraf nyata tertentu sebagai batas. Kriteria lain adalah dengan menganggap peubah yang sudah terpilih bersifat tetap, dan menghitung korelasi parsial peubah yang akan dipilih, serta sebelumnya sudah ditentukan batasan besaran korelasi parsial yang bisa diterima. Proses ini akan berhenti jika tidak ada lagi peubah yang memenuhi kriteria yang telah ditentukan. Prosedur yang seperti ini dikenal sebagai prosedur ***forward selection***.

Penyeleksian Peubah pada Analisis Diskriminan

Dimulai dengan model penuh, yaitu memuat semua peubah. Di setiap tahap dilakukan pembauangan 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 ***backward selection***.

Penyeleksian Peubah pada Analisis Diskriminan

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

Terima Kasih

atas perhatiannya