LAPORAN MULTIVARIAT

ANALISIS DISKRIMINAN

Disusun oleh :

Indra Syahriar	**055340**
Agung Dwi Suprapto	**055450**
Fathurochman	**055726**
Julianto	**055889**
Muhammad Nurviana	**057068**

FAKULTAS PENDIDIKAN MATEMATIKA

DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS PENDIDIKAN INDONESIA

2009

KATA PENGANTAR

Alhamdulillahirrabil'alamin. Puji dan syukur penulis panjatkan kepada Allah SWT, karena berkat limpahan rahmat dan kasih sayang-Nya penulis dapat menyelesaikan laporan ini. Penulis memahami bahwa penulisan laporan ini tidak akan berjalan dengan lancar tanpa adanya kerjasama dan bantuan dari berbagai pihak.

Penulis menyadari bahwa laporan ini jauh dari sempurna, oleh karena itu diharapkan adanya kritikan dan masukan yang membangun kepada penulis.

Akhir kata penulis berharap semoga laporan ini dapat bermanfaat bagi semua pihak yang memerlukan, khususnya bagi civitas akademika Fakultas Ilmu Matematika dan Pengetahuan Alam, Universitas Pendidikan Indonesia.

Bandung, Juni 2009

Penulis

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Analisis diskriminan dan klasifikasi merupakan teknik multivariat yang dipengaruhi dengan pemisahan himpunan-himpunan objek (observasi) yang berbeda dan dengan pengalokasian objek-objek (observasi) baru ke grup yang sebelumnya telah didefinisikan. Analisis diskriminan pada dasarnya bersifat exploratori. Sebagai prosedur pemisahan, analisis diskriminan seringkali dipakai pada one-time basis untuk menyelidiki perbedaan-perbedaan yang diamati saat hubungan kausal tidak dapat dimengerti dengan baik. Prosedur klasifikasi merupakan prosedur yang sedikit exploratori yang mengacu pada aturan well-defined yang bisa digunakan menentukan objek baru.

Klasifikasi biasanya membutuhkan struktur masalah lebih dari diskriminasi. Ketika melakukan klasifikasi, sangat mungkin akan terjadi kesalahan pengklasifikasian objek/observasi. Sebagai contoh, seorang nelayan akan mengelompokkan ikan salmon kedalam dua grup, misalnya ikan yang berasal dari perairan Canadia dan ikan yang berasal dari perairan Alaska, berdasarkan ukuran lingkaran pertumbuhan pada sisiknya. Ukuran lingkaran pertumbuhan pada sisik ikan salmon perairan Alaska lebih kecil dibandingkan ikan salmon perairan Canadia. Akan tetapi, jika nelayan tersebut tidak cukup jeli untuk membedakan ukuran lingkaran pertumbuhannya, maka akan mengakibatkan terjadinya kesalahan pengelompokkan. Ikan salmon perairan Alaska bisa dikelompokkan kedalam ikan salmon perairan Canadia, begitu juga sebaliknya. Kesalahan pengklasifikasian ini biasa disebut misklasifikasi.

Prosedur pengklasifikasian yang bagus harus memperkecil peluang terjadinya misklasifikasi. Dengan kata lain, peluang atau probabilitas misklasifikasi harus kecil. Oleh karena itu maka perlu dilakukan evaluasi terhadap prosedur klasifikasi.

Berdasarkan paparan di atas, perlu kiranya dibahas beberapa teori mengenai analisis diskriminan dan pengklasifikasian baik untuk yang dua

populasi maupun beberapa populasi serta metodenya dalam pendeskripsian prosedur klasifikasi.

1.2 Rumusan Masalah

Adapun rumuban masalah yang akan dibahas dalam makalah ini adalah sebagai berikut:

1. Bagaimana mendeskripsikan objek-objek (observasi-observasi) dari populasi yang diketahui?
2. Bagaimana menyusun objek-objek (observasi-observasi) menjadi dua atau lebih label kelas?
3. Bagaimana mengelompokan objek-objek dengan EOR, EAR, APPER, dan Metode Fisher?

1.3 Tujuan Penulisan

Tujuan dari penulisan ini adalah sebagai berikut:

1. Mampu mendeskripsikan baik secara aljabar atau secara aljabar differential features dari populasi yang diketahui.
2. Mampu menyusun objek-objek (observasi-observasi) menjadi dua atau lebih label kelas.
3. Mampu mengelompokan objek-objek dengan EOR, EAR, APPER, dan Metode Fisher.

1.4 Batasan Masalah

Pada makalah ini yang akan dibahas mengenai diskriminasi dan klasifikasi, akan tetapi dalam makalah ini penulis membatasinya hanya untuk yang berpopulasi normal.

1.5 Sistematika Penulisan

Makalah ini mengikuti sistematika penulisan sebagai berikut:

BAB I : Pendahuluan

Membahas latar belakang masalah, rumusan masalah, tujuan penulisan, batasan masalah, dan sistematika penulisan.

BAB II : Isi

Bab ini menyajikan tentang diskriminasi dan klasifikasi.

BAB III : Kesimpulan dan Saran

Bab ini menyajikan kesimpulan dan saran dari hasil makalah yang penulis tulis

BAB II
PEMBAHASAN
DISKRIMINAN DAN KLASIFIKASI

2.1 Pemisahan dan Klasifikasi Untuk 2 Populasi

Untuk memperbaiki ide, kita tulis secara berurut situasi di bawah ini dimana salah satunya bisa ternasuk dalam (1) Pemisahan dua kelas dari objek-objek, atau (2) Menentukan objek baru ke salah satu dari dua kelas (atau keduanya). Kita bisa melabelkan kelas π_1 dan π_2. Objek-objek ini biasanya dipisahKan atau diklasifikasikan pada basis pengukurannya, contohnya *p* dihubungkan variabel random $X' = [X_1, X_2, ..., X_P]$. Nilai hasil observasi untuk X berbeda besarnya dari satu kelas ke kelas lain. Kita dapat memikirkan keseluruhan nilai dari kelas pertama sebagai populasi dari nilai x untuk π_1 dan dari kelas kedua sebagai populasi dari nilai x untuk π_2. Dua populasi ini bisa digambarkan oleh peluang fungsi densitas $f_1(x)$ dan $f_2(x)$, dan, akibatnya kita bisa menyatakan bahwa penentuan observasi-observasi ke populasi-populasi atau objek-objek ke kelas-kelas dapat dipertukarkan.

Anda bisa mengingat kembali bahwa beberapa contoh berikut mengenai situasi pemisahan dan klasifikasi telah di perkenalkan di bagian 1.

No.	Populasi π_1 dan π_2	Variabel X yang di ukur
1.	Kesanggupan dan kesulitan membayar pertanggungjawaban properti perusahaan asuransi	Total aset, biaya saham dan obligasi, nilai pasar dari saham dan obligasi, biaya kerugian, kelebihan, jumlah hadiah yang tertulis.
2.	Nonulcer dyspeptics (orang yang bermasalah dengan penyakit perut) dan kontrol (normal).	Pengukuran dari kecemasan, ketergantungan, kesalahan, dan kesempurnaan.
3.	Krtas-kertas federalist yang ditulis oleh James Madison dan oleh	Frekuensi dari perbedaan kata dan panjangnya kalimat.

	Alexander Hamilton.	
4.	Dua species dari chickweed	Panjang kelopak dan daun bunga, ketebalan daun bunga, panjang ujung scarious, diameter tepung sari.
5.	Pembeli produk baru dan laggard (orang yang selalu datang terlambat)	Pendidikan, pendapatan, besar keluarga, banyaknya pergantian merek.
6.	Keberhasilan atau kegagalan mahasiswa	Skor ujian masuk, nilai rata-rata kenaikaan kelas SMA, banyaknya kegiatan di SMA.
7.	Laki-laki dan perempuan	Pengukuran secara antropologi seperti bentuk dan volume tengkorak purbakala.
8.	Resiko kredit yang bagus dan jelek	Pendapatan (income), usia, jumlah kartu kredit, besar keluarga.
9.	Alkohol dan nonalkohol	Aktivitas enzim monoamine oxidase, aktivitas enzim adenylate cyclase.

Kita lihat dari 5, misalnya objek-objek (para konsumen) dipisah menjadi dua label kelas (“purchaser” dan “laggards”) pada basis dari nilai observasi variabel yang dianggap relevan (pendidikan, pendapatan, dll). Dalam terminologi dari observasi dan populasi, kita ingin mengidentifikasi observasi dari x’ = [x_1(pendidikan), x_2(pendapatan), x_3(besar keluarga), x_4(banyaknya pergantian merek)] sebagai populasi π_1: purchasers, atau populasi π_2: laggards.

Pada poin ini, kita akan pada klasifikasi untuk dua populasi, kembali ke pemisahan di bagian 11.5

Aturan-aturan alokasi atau klasifikasi biasanya dikembangkan dari sampel "pembelajaran". Karakteristik-karakteristik yang diukur dari objek-objek yang dipilih secara acak diketahui berasal dari tiap dua populasi yang diuji dari perbedaan-perbedaannya. Sesungguhnya, himpunan dari semua hasil sampel yang mungkin dibagi dalam dua daerah, R_1 dan R_2, sehingga jika observasi baru jatuh di R_1, ini dialokasikan ke populasi π_1 dan jika jatuh di R_2, kita mengalokasikannya di populasi π_2. Dengan demikian, satu himpunan yang nilainya diobservasi menunjukkan π_1, himpunan nilai yang lain menunjukkan π_2.

Anda mungkin bertanya-tanya pada poin ini, bagaimana kita mengetahui beberapa observasi termasuk ke populasi tertentu tetapi kita tidak yakin dengan observasi yang lain (ini, tentu saja, akan membuat klasifikasi menjadi masalah). Ada beberapa syarat yang bisa memberi penjelasan terhadap penyimpangan ini.

1. Pengetahuan yang tidak lengkap dari kinerja yang akan datang.

   Contoh: Di masa lalu, nilai extrim dari variabel finansial tertentu diobservasi 2 tahun sebelum kemudian mengalami kebangkrutan. Pengklasifikasian firma lain sebagai firma yang *sehat* atau *sakit* pada basis nilai yang diobservasi dari indikator-indikator penting ini bisa membuat para pegawai dapat mengambil tindakan perbaikan, jika perlu, sebelum terlambat.

   Sebuah kantor sekolah aplikasi pengobatan ingin mengklasifikasi para pelamar sebagai *akan menjadi M.D.* dan *tidak akan menjadi M.D.* dengan basis skor hasil tes dan dokumen-dokumen sekolah tinggi lainnya. Untuk kasus ini, penentuannya hanya bisa di buat pada saat akhir pelatihan yang diadakan selama beberapa tahun.

2. Informasi yang "sempurna" memerlukan penghancuran objek.

   Contoh: Lama hidupnya suatu baterai kalkulator ditentukan oleh penggunaannya sampai baterai tersebut mati dan kekuatan dari sepotong kayu dapat dilihat dari lamanya kayu tersebut bertahan sampai rusak (lapuk).

3. Tidak tersedianya informasi atau mahalnya informasi.

   Contoh: Diasumsikan bahwa kertas-kertas Federalist tertentu ditulis oleh James Madison atau Alexander Hamilton karena pada kertas tersebut terdapat tanda tangan mereka. Tetapi ada kertas lain yang tidak ada tanda tangannya dan hal ini menarik untuk menentukan mana dari dua orang tersebut yang

telah menulis kertas yang tidak ada tangannya. Jelas, kita tidak bisa bertanya kepada mereka. Frekuensi kata dan panjang kalimat dapat membantu kita dalam mengklasifikasi kertas-kertas tersebut.

Banyak masalah pengobatan bisa diidentifikasi secara konklusif hanya dengan melakukan operasi yang mahal. Biasanya, seseorang akan mendiagnosa suatu penyakit dari kemudahannya diamati, namun berpotensi untuk terjadi kekeliruan, yaitu gejala luar. Pendekatan ini membantu untuk menghindari terjadinya operasi yang sebenarnya tidak perlu dilakukan dan juga mahal.

Dari contoh di atas, jelas bahwa aturan klasifikasi tidak selalu menyediakan metode yang bebas dari kesalahan penandaan. Hal ini bisa saja dikarenakan oleh tidak adanya penentuan yang jelas antara karakteristik-karakteristik dari populasi-populasi yang diukur; yaitu grup-grupnya hanya menutupi sebagian saja. Ini bisa saja terjadi, contohnya secara tidak benar mengklasifikasi objek π_2 termasuk dalam π_1 atau objek π_1 termasuk dalam π_2.

Misal $f_1(x)$ dan $f_2(x)$ merupakan peluang fungsi densitas yang dihubungkan dengan vektor $p \times 1$ variabel acak X untuk populasi π_1 dan π_2. Suatu objek yang dihubungkan pengukuran x, harus di masukkan ke π_1 atau π_2. Misal Ω adalah ruang sampel yang merupakan koleksi dari semua observasi x yang mungkin. Misal R_1 himpunan nilai x dimana kita mengklasifikasi objek-objek sebagai π_1 dan $R_2 = \Omega - R_1$ merupakan himpunan nilai x yang tersisa dimana kita mengklasifikasi objek-objek sebagai π_2. Karena setiap objek harus ditentukan ke salah satu dari dua populasi, himpunan R_1 dan R_2 bersifat saling menguntungkan dan saling melengkapi.

Peluang kondisional, $P(2|1)$. dari pengklasifikasian suatu objek sebagai π_2 dimana dalam kenyataannya dari π_1 adalah

$$P(2|1) = P(X \in R_2 | \pi_1) = \int_{R_2=\Omega-R_1} f_1(x)dx \qquad (11\text{-}1)$$

Dengan cara yang sama, peluang kondisional, $P(1|2)$, dari pengklasifikasian suatu objek sebagai π_1 dimana dalam kenyataannya dari π_2 adalah

$$P(1|2) = P(X \in R_1 | \pi_2) = \int_{R_1} f_2(x)dx \qquad (11\text{-}2)$$

Tanda integral dalam (11-1) menunjukkan volume yang dibentuk oleh fungsi densitas $f_1(x)$ di atas daerah R_2. Begitu juga, tanda integral dalam (11-2) menunjukkan volume yang dibentuk oleh $f_2(x)$ di atas daerah R_1.

Misal p_1 merupakan peluang prior dari π_1 dan p_2 merupakan peluang prior dari π_2 dimana $p_1 + p_2 = 1$.Peluang keseluruhan dari benar atau tidaknya pengklasifikasian objek dapat diperoleh sebagai hasil dari peluang klasifikasi prior dan kondisional:

P(correctly classified as π_1) = P(observation comes from π_1 and is correctly classified as π_1)

$$= P(X \in R_1 \mid \pi_1)P(\pi_1) = P(1 \mid 1)\,p_1$$

P(misclassified as π_1) = P(observation comes from π_2 and misclassified as π_1)

$$= P(X \in R_1 \mid \pi_2)P(\pi_2) = P(1 \mid 2)\,p_2$$

P(correctly classified as π_2) = P(observation comes from π_2 and is correctly classified as π_2)

$$= P(X \in R_2 \mid \pi_2)P(\pi_2) = P(2 \mid 2)\,p_2$$

P(misclassified as π_1) = P(observation comes from π_1 and misclassified as π_2)

$$= P(X \in R_2 \mid \pi_1)P(\pi_1) = P(2 \mid 1)\,p_1 \qquad (11\text{-}3)$$

Skema klasifikasi sering kali dinilai dari peluang kesalahan pengklasifikasiannya, tapi ini mengabaikan biaya misklasifikasi. Contohnya, untuk peluang yang terlihat kecil seperti 0,06 = $P(2|1)$ bisa menjadi terlalu besar jika biaya pembuatan assignment yang salah untuk π_2 terlalu tinggi. Sebuah aturan yang mengabaikan biaya dapat menimbulkan masalah.

Biaya misklasifikasi dapat didefinisikan oleh matriks cost (biaya).

		Klasifikasi sebagai:	
		π_1	π_2
Populasi sebenarnya	π_1	0	$C(2\|1)$
	π_2	$C(1\|2)$	0

(11-4)

Biayanya adalah: (1) 0 untuk klasifikasi yang benar, (2) $c(1|2)$ saat observasi π_2 salah diklasifikasikan sebagai π_1, dan (3) $c(2|1)$ saat observasi π_1 salah diklasifikasikan sebagai π_2.

Untuk setiap aturan manapun, rata-rata, atau biaya misklasifikasi yang diharapkan (*expexted cost of misclassification* (**ECM**)) di diberikan dengan mengalikan entri diagonaldalam (11-4) dengan peluang kejadiannya, yang didapatkan dari (11-3). Akibatnya,

$$\mathbf{ECM} = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \qquad (11\text{-}5)$$

Sebuah aturan klasifikasi yang berlasan akan mempunyai ECM yang kecil atau mendekati nilai kecil yang mungkin.

Akibat 11.1. Daerah R_1 dan R_2 yang meminimasi ECM didefinisikan oleh nilai x sedemekian sehingga pertidaksamaan di bawah ini dipenuhi.

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

$$\begin{pmatrix} densitas \\ ratio \end{pmatrix} \geq \begin{pmatrix} biaya \\ ratio \end{pmatrix} \begin{pmatrix} prior \\ probability \\ ratio \end{pmatrix}$$

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

$$\begin{pmatrix} densitas \\ ratio \end{pmatrix} < \begin{pmatrix} biaya \\ ratio \end{pmatrix} \begin{pmatrix} prior \\ probability \\ ratio \end{pmatrix}$$

(11-6)

Dari (11-6) jelas bahwa implementasi dari aturan ECM minimum memerlukan (1) rasio fungsi kepadatan dinilai pada observasi baru x_0, (2) rasio biaya, dan (3) rasio peluang prior. Kemunculan dari rasio-rasio dalam definisi dari daerah klasifikasi optimal adalah signifikan. Seringkali, lebih mudah untuk menspesifiksi rasio-rasio daripada komponen bagiannya.

Contohnya, mungkin sulit untuk menspesifikasi biaya (dalam unit-unit yang bersesuaian) dari pengklasifikasian seorang siswa sebagai material sekolah tinggi ketika, pada kenyataannya dia bukan dan pengklasifikasian siswa sebagai material nonsekolah tinggi ketika, pada kenyataannya dia iya. Biaya terhadap para pembayar pajak dari pendidikan sekolah tinggi dikeluarkan untuk 2 tahun, misalnya dapat ditetapkan secara kasar. Biaya untuk universitas dan masyarakat ysmh tidak melakukan pendidikan terhadap seorang siswa yang mampu lebih sulit untuk ditentukan. Akan tetapi bisa saja bahwa sebuah angka realistis untuk rasio dari biaya misklasifikasi ini dapat diperoleh. Apapun unit-unit dari pengukuran, tidak mengakui lulusan sekolah tinggi yang prospektif mungkin biayanya lebih besar 5 kali terhadap suatu horizon waktu yang sesuai, daripada mengakui dropout yang dapat terjadi. Pada kasus ini, rasio biaya adalah 5.

Ini menarik untuk memperhatikan daerah klasifikasi didefinisikan dalam (11-6) untuk beberapa kasus spesial.

Kasus Khusus Daerah-daerah Biaya Minimum yang Diharapkan

a) $p_2/p_1 = 1$ (peluang prior yang sama)

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

b) $c(1|2)/c(2|1) = 1$ (biaya-biaya misklasifikasi yang sama)

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1}; \qquad R_2 : \frac{f_1(x)}{f_2(x)} < \frac{p_2}{p_1} \tag{11-7}$$

c) $p_2/p_1 = c(1|2)/c(2|1) = 1$ atau $p_2/p_1 = 1/(c(1|2)/c(2|1))$ (peluang prior yang sama dan biaya-biaya misklasifikasi yang sama)

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

Saat peluang prior tidak diketahui, sering kali peluang prior tersebut disamakan dan aturan ECM minimum melibatkan pembandingan rasio dari kepadatan populasi terhadap rasio biaya-biaya misklasifikasi yang bersesuaian. Jika rasio biaya misklasifikasi tidak ditentukan, biasanya dijadikan satu dan rasio kepadatan populasi dibandingkan dengan rasio peluang prior. (Catat bahwa peluang prior adalah urutan terbalik dari kepadatan). Akhirnya, ketika rasio

peluang prior dan rasio biaya misklasifikasi adalah satu, atau rasio yang satu merupakan kebalikan dari rasio yang lainnya, daerah-daerah klasifikasi optimal ditentukan secara sederhana dengan membandingkan nilai-nilai dari fungsi-fungsi kepadatan. Pada kasus ini, jika x_0 adalah observasi baru dan $f_1(x_0)/f_2(x_0) \geq 1$, yaitu $f_1(x_0) \geq f_2(x_0)$, kita petakan x_0 ke π_1. Di sisi lain, jika $f_1(x_0)/f_2(x_0) < 1$, atau $f_1(x_0) < f_2(x_0)$, kita petakan x_0 ke π_2.

Ini adalah latihan umum terhadap kasus penggunaan sebarang c dalam (11-7) klasifikasi. Ini serupa dengan mengsumsikan peluang prior yang sama dan biaya misklasifikasi yang sama untuk aturan ECM minimum.

Example 11.2

Seorang ahli riset mempunyai data yang cukup tersedia untuk mengestimasi fungsi kepadatan $f_1(x)$ dan $f_2(x)$ yang masing-masing dihubungkan dengan populasi π_1 dan π_2. Misal $c(2|1)= 5$ unit dan $c(1|2) = 10$ unit. Sebagai tambahan, diketahui bahwa sekitar 20% dari semua objek-objek (sedemikian sehingga pengukuran-pengukuran x dapat dicatat) dimiliki oleh π_2. Dengan demikian, peluang prior adalah $p_1= 0,8$ dan $p_2= 0,2$.

Diberikan peluang-peluang prior dan biaya-biaya misklasifikasi, kita bisa menggunakan (11-6) untuk menurunkan daerah-daerah klasifikasi R_1 dan R_2. Secara spesifik,

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \left(\frac{10}{5}\right)\left(\frac{0,2}{0,8}\right) = 0,5$$

$$R_2 : \frac{f_1(x)}{f_2(x)} < \left(\frac{10}{5}\right)\left(\frac{0,2}{0,8}\right) = 0,5$$

Misalkan fungsi-fungsi kepadatan yang dinilai di suatu observasi baru x_0 menghasilkan $f_1(x_0) = 0,3$ dan $f_2(x_0) = 0,4$. Apakah kita mengklasifikasikan observasi baru sebagai π_1 atau π_2? Untuk menjawab pertanyaan ini, kita bentuk rasio

$$\frac{f_1(x_0)}{f_2(x_0)} = \frac{0,3}{0,4} = 0,75$$

Dan bandingkan dengan 0,5. Karena

$$\frac{f_1(x_0)}{f_2(x_0)} = 0,75 > \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) = 0,5$$

Kita temukan bahwa $x_0 \in R_1$dan klasifikasikan sebagai kepunyaan π_1.

Kriteria selain dari biaya misklasifikasi yang diharapkan dapat digunakan menurunkan prosedur-prosedur klasifikasi optimal. Contohnya, seseorang mungkin mengabaikan biaya-biaya misklasifikasi dan memilih R_1 dan R_2 untuk meminimumkan peluang total misklasifikasi (*total probability of misclassification* (TPM)),

TPM = *P*(memisklasifikasikan suatu observasi π_1 atau memisklasifikasikan suatu observasi π_2).

= *P*(observasi yang datang dari π_1 dan dimisklasifikasikan) + *P*(observasi yang datang dari π_2 dan dimisklasifikasikan)

$$= p_1 \int_{R_2} f_1(x)dx - p_2 \int_{R_1} f_2(x)dx \qquad (11\text{-}8)$$

Secara matematis, masalah ini ekivalen dengan meminimumkan biaya misklasifikasi yang diharapkan ketika biaya-biaya misklasifikasi sama. Akibatnya, daerah-daerah optimal dalam kasus ini, diberikan oleh (b) dalam (11-7).

Kita juga dapat mengalokasikan suatu observasi baru x_0 ke populasi dengan peluang posterior yang terbesar $P(\pi_1|x_0)$. Dengan aturan Bayes, peluang-peluang posterior adalah

$$P(\pi_1 \mid x_0) = \frac{P(\pi_1 \, terjadi \, dan \, mengobservasi \, x_0)}{P(mengobservasi \, x_0)}$$

$$P(\pi_1 \mid x_0) = \frac{P(mengobservasi \, x_0 \mid \pi_1)P(\pi_1)}{P(mengobservasi \, x_0 \mid \pi_1)P(\pi_1) + P(mengobservasi \, x_0 \mid \pi_2)P(\pi_2)}$$

$$= \frac{p_1 f_1(x_0)}{p_1 f_1(x_0) + p_2 f_2(x_0)}$$

$$P(\pi_2 \mid x_0) = 1 - P(\pi_1 \mid x_0) = \frac{p_2 f_2(x_0)}{p_1 f_1(x_0) + p_2 f_2(x_0)} \qquad (11\text{-}9)$$

Pengklasifikasian suatu observasi x_1 sebagai π_1 ketika $P(\pi_1 \mid x_0) > P(\pi_2 \mid x_0)$ adalah ekivalen dengan menggunakan aturan (b) untuk peluang total dari misklasifikasi dalam (11-7) karena penyebut-penyebut dalam (11-9) adalah sama. Akan tetapi penghitungan peluang dari populasi π_1 dan π_2 setelah mengobservasi x_0 (dengan demikian dinamakan peluang posterior) selalu berguna untuk tujuan dari pengidentifikasian assignment clear-cut yang sedikit.

2.2 Klasifikasi dengan dua multivariat berpopulasi normal

Prosedur klasifikasi yang didasarkan populasi normal menonjol dalam praktek statistika dikarenakan kesederhanaanya dan kelayakannya berefisiensi besar terhadap suatu model populasi yang luas. Kita sekarang mengasumsikan f_1(x) dan f_2(x) adalah multivariat berdensitas normal, vektor mean μ_1 dan kovarian matrik $\sum_1$ untuk yang pertama dan vektor mean μ_2 dan kovarian matrik $\sum_2$ untuk yang pertama kedua.

Kasus khusus dari persamaan matiks kovarian membawa ke sebuah klasifikasi statistik sederhana linear yang penting.

2.2.1 Klasifikasi dari populasi normal saat $\sum_1 = \sum_2 = \sum$

Asumsikan densitas gabungan dari $X^{'} = \left[X_1, X_2, ..., X_p \right]$ untuk populasi π_1 dan π_2 diberikan oleh

$$f_i(x) = \frac{1}{(2\pi)^{p/2} |\sum|^{1/2}} \exp\left[-\frac{1}{2} (x - \mu_i)^{'} \sum^{-1} (x - \mu_i) \right] \qquad \text{untuk } i = 1, 2$$

(11-10)

Andaikan parameter populasi μ_1, μ_2, dan $\sum$ diketahui.

Setelah penghapusan bentuk $(2\pi)^{p/2} |\sum|^{1/2}$, daerah minimum ECM dalam (11-6) menjadi

$$R_1 : \exp\left[-\frac{1}{2} (x - \mu_1)^{'} \sum^{-1} (x - \mu_1) + \frac{1}{2} (x - \mu_2)^{'} \sum^{-1} (x - \mu_2) \right] \geq \left(\frac{c(1|2)}{c(2|1)} \right) \left(\frac{p_2}{p_1} \right)$$

(11-11)

$$R_2 : \exp\left[-\frac{1}{2}(x-\mu_1)' \Sigma^{-1}(x-\mu_1)+\frac{1}{2}(x-\mu_2)' \Sigma^{-1}(x-\mu_2)\right] < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

Diberikan daerah R_1 dan R_2 di atas, kita dapat mengkontruksi kaidah klasifikasi berikut.

Hasil 11.2. Misalkan populasi π_1 dan π_2 dideskripsikan sebagai multivariat berpopulasi normal dari bentuk (11-10). Kaidah alokasi yang meminimumkan ECM diberikan oleh:

Alokasikan x_0 ke π_1 jika

$$(\mu_1-\mu_2)' \Sigma^{-1} x_0 - \frac{1}{2}(\mu_1-\mu_2)' \Sigma^{-1}(\mu_1+\mu_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (11\text{-}12)$$

Alokasikan x_0 ke π_2 untuk sebaliknya.

Bukti. Karena jumlah pada (11-11) nonnegatif untuk semua x, kita dapat mengambil logaritma natural mereka dan mempertahankan susunan dari pertidaksamaannya. Selain itu,

$$-\frac{1}{2}(x-\mu_1)' \Sigma^{-1}(x-\mu_1)+\frac{1}{2}(x-\mu_2)' \Sigma^{-1}(x-\mu_2)$$

$$= (\mu_1-\mu_2)' \Sigma^{-1} x - \frac{1}{2}(\mu_1-\mu_2)' \Sigma^{-1}(\mu_1+\mu_2) \quad (11\text{-}13)$$

dan, sebagai konsekuensi,

$$R_1 : (\mu_1-\mu_2)' \Sigma^{-1} x - \frac{1}{2}(\mu_1-\mu_2)' \Sigma^{-1}(\mu_1+\mu_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

$$R_2 : (\mu_1-\mu_2)' \Sigma^{-1} x - \frac{1}{2}(\mu_1-\mu_2)' \Sigma^{-1}(\mu_1+\mu_2) < \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

(11-14)

Kaidah klasifikasi ECM minimum mengikuti.

Dalam kebanyakan keadaan praktis, jumlah populasi μ_1, μ_2, dan $\sum$ tidak diketahui, jadi kaidah (11-12) harus dimodifikasi. Wald [25] dan Anderson [2] telah mengusulkan penggantian parameter populasi dengan sampel yang setara dengan mereka.

Misalkan, waktu itu, kita memiliki n_1 observasi dari variabel acak multivariat $X' = \left[X_1, X_2, ..., X_p \right]$ dari π_1 dan pengukuran n_2 dari jumlah pada π_2 dengan $n_1 + n_2 - 2 \geq p$. Matriks data yang bersesuaian adalah

$$\underset{(p \ x \ n_1)}{X_1} = \left[x_{11}, x_{12}, ..., x_{1n_1} \right]$$

(11-15)

$$\underset{(p \ x \ n_2)}{X_2} = \left[x_{21}, x_{22}, ..., x_{2n_2} \right]$$

Dari matriks data ini, vektor mean sampel dan matriks kovarian dapat ditentukan oleh

$$\underset{(p \ x \ 1)}{\bar{x}_1} = \frac{1}{n_1} \sum_{j=1}^{n_1} x_{1j} \; ; \qquad \underset{(p \ x \ p)}{S_1} = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} \left(x_{1j} - \bar{x}_1 \right) \left(x_{1j} - \bar{x}_1 \right)'$$

(11-16)

$$\underset{(p \ x \ 1)}{\bar{x}_2} = \frac{1}{n_2} \sum_{j=1}^{n_2} x_{2j} \; ; \qquad \underset{(p \ x \ p)}{S_2} = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} \left(x_{2j} - \bar{x}_2 \right) \left(x_{2j} - \bar{x}_2 \right)'$$

Karena iru diasumsikan bahwa populasi induk menpunyai kovarian matriks $\sum$ yang sama, kovarian matriks sampel S_1 dan S_2 dikombinasikan (dikelompokkan) untuk memperoleh sebuah estimasi dari $\sum$ yang tunggal dan tidak bias seperti dalam (6-21). Secara khusus, bobot rata-rata

$$S_{pooled} = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right] S_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right] S_2$$

$$= \frac{(n_1 - 1) S_1 + (n_2 - 1) S_2}{(n_1 + n_2 - 2)}$$

(11-17)

adalah estimasti dari $\sum$ yang tidak bias jika data matriks X_1 dan X_2 mengandung sampel acak dari populasi π_1 dan π_2 , secara berturut-turut.

Mensubstitusikan $\bar{x}_1$ untuk π_1, $\bar{x}_2$ untuk π_2, dan S_{pooled} untuk $\sum$ dalam (11-12) memberikan kaidah klasifikasi sample.

Estimasi Kaidah ECM Minimum untuk Dua Pupolasi Normal

Alokasikan x_0 ke π_1 jika

$$(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x_0 - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} (\bar{x}_1 + \bar{x}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

(11-18)

Alokasikan x_0 ke π_2 untuk sebaliknya.

Jika, dalam (11-18), $\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)=1$ maka $\ln(1)=0$ dan estimasi kaidah ECM minimum untuk dua populasi normal dengan sejumlah perbandingan variabel skalar $y=(\bar{x}_1-\bar{x}_2)S_{pooled}x=\hat{l}'x$ (11-19)

yang dievaluasikan pada x_0, dengan bilangan $\hat{m}=\frac{1}{2}(\bar{x}_1-\bar{x}_2)'S_{pooled}^{-1}(\bar{x}_1+\bar{x}_2)$

$$= \frac{1}{2}(\bar{y}_1-\bar{y}_2) \qquad (11\text{-}20)$$

dimana $\bar{y}_1=(\bar{x}_1-\bar{x}_2)'S_{pooled}^{-1}\bar{x}_1=\hat{l}'\bar{x}_1$ dan $\bar{y}_2=(\bar{x}_1-\bar{x}_2)'S_{pooled}^{-1}\bar{x}_2=\hat{l}'\bar{x}_2$.

Itulah, estimasi kaidah ECM minimum untuk dua populasi normal yang sama dengan pembuatan dua univariat populasi untuk nilai y dengan mengambil kombinasi linear yang sesuai dari obserbasi dari populasi π_1 dan π_2 dan kemudian menentukan sebuah observasi baru, x_0 ke π_1 atau π_2 tergantung atas apakah $\bar{y}_0=\hat{l}'\bar{x}_0$ jatuh pada kanan atau kiri dari titik tengah, $\hat{m}$, diantara rata-rata dua univariat $\bar{y}_1$ dan $\bar{y}_2$.

Sewaktu estimasi paremater dimasukkan pada jumlah populasi yang tidak diketahui yang bersesuaian, tidak ada kepastian yang menghasilkan kaidah yang akan meminimukan biaya yang diharapkan dari misklasifikasi dalam aplikasi khusus. Ini karena kaidah optimal dalam (11-12) berasal pada asumsi multivariat densitas normal $f_1(x)$ dan $f_2(x)$ yang diketahui dengan sepenuhnya. Pernyataan (11-18) hanya sebuah estimasi dari kaidah optimal. Bagaimanapun, itu kelihatannya beralasan untuk mengharapkan bahwa itu akan berjalan baik jika ukuran sampel besar.

Untuk meringkas, jika data yang muncul adalah multivariat normal, statistik klasifikasi pada pertidaksamaan kiri dalam (11-18) dapat dihitung untuk

setiap observsi baru x_0. Observasi ini diklasifikasikan dengan membandingkan nilai mereka dengan nilai dari $\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$.

Contoh 11.3

Contoh ini diadaptasi dari sebuah pembelajaran [4] mengenai deteksi dari pembawa hemofilia A. Untuk membuat sebuah prosedur untuk mendeteksi potensi pen=mbawa hemofilia A, sampel darah diperiksa untuk dua grup wanita dan ukuran pada dua variabel, $X_1 = \log_{10}(AHF \quad activity)$ dan $X_2 = \log_{10}(AHF - like \quad antigen)$ dicatat. Grup pertama dari $n_1 = 30$ wanita dipilih dari sebuah populasi dari wanita yang tidak membawa gen hemofilia. Grup ini disebut grup normal. Grup kedua dari $n_2 = 22$ wanita dipilih dari pembawa hemofilia A yang diketahui (putri dari hemofilia, ibu dengan lebih dari satu putra hemofilia, dan ibu dengan satu putra hemofilia dan hubungan hemofilia lainnya). Grup ini disebut pembawa wajib. Pasangan observasi (x_1, x_2) untuk kedua grup digambarkan dalam Figur 11.4(lihat buku halaman 506). Juga ditunjukkan estimasi kontur mengandung 50% dan 95% dari probabilitas untuk bivariat distribusi normal yang berpusat pada $\bar{x}_1$ dan $\bar{x}_2$, secara berturut-turut. Matriks kovarian umum mereka diambil menurut matriks kovarian sampel yang dikelompokkan, S_{pooled}. Dalam contoh ini, distribusi normal bivariat kelihatannya sesuai dengan data cukup baik.

Pemeriksa (lihat [4]) menyediakan informasi

$$\bar{x}_1 = \begin{bmatrix} -0.0065 \\ -0.0390 \end{bmatrix}, \qquad \bar{x}_2 = \begin{bmatrix} -0.2483 \\ -0.0262 \end{bmatrix},$$

$$S_{pooled}^{-1} = \begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}$$

Oleh karena itu, biaya yang sama dan fungsi diskriminan prior yang sama [lihat(11-19)] adalah

$$y = \hat{l}' x = (\bar{x}_1 - \bar{x}_2) S_{pooled} x$$

$$= \begin{bmatrix} 0.2418 & -0.652 \end{bmatrix} \begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= 37.61x_1 - 28.92x_2$$

Selain itu, $\bar{y}_1 = \hat{l}' \bar{x}_1 = \begin{bmatrix} 37.61 & -28.92 \end{bmatrix} \begin{bmatrix} -0.0065 \\ -0.0390 \end{bmatrix} = 0.88$

$$\bar{y}_2 = \hat{l}' \bar{x}_2 = \begin{bmatrix} 37.61 & -28.92 \end{bmatrix} \begin{bmatrix} -0.2483 \\ 0.0262 \end{bmatrix} = -10.10$$

dan titik tengah antara rata2nya [lihat(11-20)] adalah $\hat{m} = \frac{1}{2}(\bar{y}_1 - \bar{y}_2) = \frac{1}{2}(0.88 \quad -10.10) = -4.61$.

Pengukuran dari *AHF activity* dan *AHF-like antigen* pada seorang wanita yang mungkin seorang pembawa hemofilia A memberikan $x_1 = -0.210$ dan $x_2 = -0.044$. Seharusnya wanita tersebut diklasifikasikan sebagai π_1 : normal atau π_2 : pembawa wajib?

Dengan menggunakan (11-18) dengan biaya yang sama dan prior yang sama maka $\ln(1) = 0$, kita mendapatkan: alokasikan x_0 ke π_1 jika $y_0 = \bar{l}' x_0 \geq \hat{m} = -4.61$

alokasikan x_0 ke π_2 jika $y_0 = \bar{l}' x_0 < \hat{m} = -4.61$

dimana $x_0' = \begin{bmatrix} -0.210 & -0.044 \end{bmatrix}$.

Karena $y_0 = \bar{l}' x_0 = \begin{bmatrix} 37.61 & -28.92 \end{bmatrix} \begin{bmatrix} -0.210 \\ -0.044 \end{bmatrix} = -6.62 < -4.61 = \hat{m}$ kita mengklasifikasikan wanita itu sebagai π_2 : pembawa wajib. Observasi baru diindikasikan oleh sebuah bintang pada Figur 11.4 (lihat buku hal 506). Kita melihat bahwa itu jatuh ke dalam estimasi 0.5 kontur probabiitas dari populasi π_2 dan sekitar pada estimasi 0.95 probabilitas kontur dari populasi π_1. Oleh karena itu, klasifikasi tersebut tidak *clearcut*.

Misalkan sekarang probabilitas prior dari jumlah anggota suatu grup diketahui. Sebagai contoh, misalkan pemerikasaan darah pengukuran x_1 dan x_2 di atas diambil dari saudara sepupu ibu dari seorang hemofilia. Kemungkinan genetik menjadi seorang pembawa hemofilia A dalam hal ni adalah 0.25. Sebagai konsekuensi, probabilitas prior dari jumlah anggota grup adalah $p_1 = 0.75$ dan $p_2 = 0.25$. Diasumsikan, setikit tidak secara realistis, bahwa biaya dari misklasifikasi adalah sebanding sedemikian sehingga $c(1|2) = c(2|1)$ dan dengan meggunakan statistik klasifikasi

$$w = \left(\bar{x}_1 - \bar{x}_2\right)' S_{pooled}^{-1} x_0 - \frac{1}{2}\left(\bar{x}_1 - \bar{x}_2\right)' S_{pooled}^{-1} \left(\bar{x}_1 + \bar{x}_2\right)$$

atau $w = \bar{l}' x_0 - \hat{m}$ dengan $x_0' = \begin{bmatrix} -0.210 & -0.044 \end{bmatrix}$, $\hat{m} = -4.61$, dan $\bar{l}' x_0 = -6.62$, kita mempunyai $w = -6.62 - \left(-4.61\right) = -2.01$

Aplikasikan (11-18), kita melihat bahwa $w = -2.01 < \ln\left[\frac{p_2}{p_1}\right] = \ln\left[\frac{0.25}{0.75}\right] = -1.10$ dan kita klasifikasikan wanita tersebut sebagai π_2 : pembawa wajib.

Koefisien vektor $\hat{l} = S_{pooled}^{-1}\left(\overline{x}_1 - \overline{x}_2\right)$ tidak unik. Itu unik hanya tergantung pada subuah konstanta multiplikatif, jadi untuk $c \neq 0$, sebarang vektor $c\hat{l}$ juga akan bertindak sebagai koefisien diskriminan.

Vektor $\hat{l}$ sering diskalakan atau dinormalkan untuk mempermudah interpretasi dari elemen-elemennya. Dua dari yang paling umum normalisasi yang digunakan adalah:

1. Himpunan $\hat{l}^* = \dfrac{\hat{l}}{\sqrt{\hat{l}'\hat{l}}}$

(11-21)

sehingga $\hat{l}^*$ mempunyai satuan panjang.

2. Himpunan $\hat{l}^* = \dfrac{\hat{l}}{\hat{l}_1}$

(11-22)

sehingga elemen pertama dari koefisien vektor $\hat{l}^*$ adalah 1.

Dalam kedua kasus, $\hat{l}^*$ merupakan bentuk dari $c\hat{l}$. Untuk normalisasi (1), $c = \left(\hat{l}'\hat{l}\right)^{-1/2}$ dan untuk (2), $c = \hat{l}_1^{-1}$.

Besarnya $\hat{l}_1^*, \hat{l}_2^*, ..., \hat{l}_p^*$ dalam (11-21) semuanya terletak dalam interval $[-1,1]$. Dalam (11-22), $\hat{l}_1^* = 1$ dan $\hat{l}_2^*, ..., \hat{l}_p^*$ diekspresikan sebagai kelipatan dari $\hat{l}_1^*$. Untuk mambatasi $\hat{l}_1^*$, pada interval $[-1,1]$ biasanya memudahkan sebuah perbandiangan visual dari koefisien. Dengan cara yang sama, Pengekspresian koefisien sebagai pengali dari $\hat{l}_1^*$ memenuhi satu untuk siap menaksir kepentingan relatif (sebagai lawan X_1) dari variabel $X_2, ..., X_p$ sebagai diskriminator.

Menormalkan direkomendasikan hanya jika varibel telah distandarisasikan. Jika bukan kasus ini, uraian yang besar dari ketelitian harud diperhatikan dalam menginterpretaskan hasilnya.

2.2.2 Klasifikasi dari populasi normal saat $\sum \neq \sum$

Seperti yang diharapkan, kaidah klasifikasi lebih rumit ketika matriks kovarian populasi tidak sama.

Pertimbangkan densitas normal multivariat dalam (11-10) dengan $\sum_i, i = 1, 2$, menggantikan $\sum$. Dengan begitu, matriks kovarian maupun vektor mean berbeda dati satu dengan yang lain untuk dua populasi. Seperti yang telah kita lihat, daerah ECM minimum dan probabilitas total dari misklasifikasi (TPM) minimum bergantung pada rasio dari densitas, $f_1(x) / f_2(x)$, atau sebanding dengan logaritma natural dari rasio densitas, $\ln\left[f_1(x) / f_2(x) \right] = \ln\left[f_1(x) \right] - \ln\left[f_2(x) \right]$. Ketika densitas normal multivariat mempunyai struktur kovarian yang berbeda, hubungan dalam rasio densitas yang menyangkut $\left|\sum_i\right|^{1/2}$ tidak dihilangkan seperti yang dilakukan ketika $\sum_1 = \sum_2$. Selain itu, bentuk kuadratik dalam eksponen dari $f_1(x)$ dan $f_2(x)$ tidak digabungkan untuk memberikan hasil yang lebih sederhana dalam (11-13).

Mensubstitusikan densitas normal multivariat dengan matriks kovarian yang berbeda ke dalam (11-6) memberikan, setelah mengambil logaritma natural dan menyederhanakannya, daerah klasifikasi

$$R_1 : -\frac{1}{2} x' \left(\sum_1^{-1} - \sum_2^{-1} \right) x + \left(\mu_1' \sum_1^{-1} - \mu_2' \sum_2^{-1} \right) x - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)} \right) \left(\frac{p_2}{p_1} \right) \right]$$

(11-23)

$$R_1 : -\frac{1}{2}x'\left(\Sigma_1^{-1}-\Sigma_2^{-1}\right)x+\left(\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1}\right)x-k<\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

dimana $$k=\frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right)+\frac{1}{2}\left(\mu_1'\Sigma_1^{-1}\mu_1-\mu_2'\Sigma_2^{-1}\mu_2\right) \qquad (11\text{-}24)$$

Daerah klasifikasi didefinisikan oleh fungsi kuadrat dari x. Ketika $\Sigma_1=\Sigma_2$, hubungan kuadratik, $-\frac{1}{2}x'\left(\Sigma_1^{-1}-\Sigma_2^{-1}\right)x$, menghilang dan daerahnya didefinisikan oleh (11-23) direduksi menjadi yang didefinisikan oleh (11-14).

Kaidah klasifikasi untuk populasi normal multivariat umum mengikuti secara langsung bentuk (11-23).

Hasil 11.3. Diberikan populasi π_1 dan π_2 dideskripsikan oleh densitas normal multivariat dengan vektor mean dan matriks kovarian μ_1, Σ_1, μ_2, dan Σ_2, secara berturut-turut. Kaidah alokasi yang meminimumkan biaya yang diharapkan dari misklasifikasi diberikan oleh:

Alokasikan x_0 ke π_1 jika

$$-\frac{1}{2}x_0'\left(\Sigma_1^{-1}-\Sigma_2^{-1}\right)x_0+\left(\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1}\right)x_0-k\geq\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

Alokasikan x_0 ke π_2 untuk sebaliknya.

Di sini, k ditentukan dalam (11-24).

Dalam latihan, kaidah klasifikasi dalam Hasil 11.3 diimplementasikan dengan mensubstitusikan besaran sampel $\bar{x}_1$, $\bar{x}_2$, S_1, dan S_2 (liaht (11-16)) untuk μ_1, μ_2, Σ_1, dan Σ_2, secara berturut-turut.

Kaidah Klasifikasi Kuadratik

(Populasi Normal dengan matriks kovarian yang tidak sama)

Alokasikan x_0 ke π_1 jika

$$-\frac{1}{2}x_0'\left(S_1^{-1}-S_2^{-1}\right)x_0+\left(\bar{x}_1'S_1^{-1}-\bar{x}_2'S_2^{-1}\right)x_0-k\geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (11\text{-}25)$$

Alokasikan x_0 ke π_2 untuk sebaliknya.

Klasifikasi dengan fungsi kuadratik lebih aneh dalam lebih dari dimensi dua dan dapan membuat ke hasil yang aneh. Ini secara khusus benar ketika data tidak (secara esensial) normal multivariat.

Figur 11.5(a) (lihat buku hal 511) menunjukkan biaya yang sama dan prior yang sama berdasarkan pada kasus ideal dari dua distribusi normal dnegan variansi yang berbeda. Kaidah kuadratik ini membawa pada sebuah daerah R_1 yang mengandung dua himpunan saling lepas dari titik.

Dalam banyak aplikasi, titik ujung untuk distribusi π_1 akan lebih kecil daripada yang berlaku pada distribusi normal. Lalu, seperti yang ditunjukkan dalam Figur 11.5(b) (lihat buku hal 511), bagian ujung dari daerah R_1, yang dihasilkan dengan prosedur kuadratik, tidak berjalan sebaik dengan distribusi populasi dan dapat membawa ke nilai error yang besar. Kelemahan yang penting dari kaidah kuadratik adalah itu sensitif pada awal dari normalitas.

Jika data tidak normal multivariat, ada dua pilihan yang tersedia. Pertama, data yang tidak normal dapat ditransformasikan menjadi data yang lebih mendekati normal dan sebuah tes untuk persamaan dari matriks kovarian dapat dilaksanakan untuk melihat apakah kaidah linear (11-18) atau kaidah kuadratik (11-25) yang tepat. Transformasi didiskusikan dalam Bab 4. (Tes yang umum untuk homogenitas kovarian sangat dipengaruhi oleh non normalitas. Konversi dari data non normal ke data normal harus dilakukan sebelum tes ini dilakukan.)

Kedua, kita dapat menggunakan kaidah linear (atau kuadratik) tanpa khawatir tentang bentuk dari populasi induk dan berharap bahwa itu akan bekerja secara layak dengan baik. Pembelajaran (lihat [20] dan [21]) telah menunjukkan, bagaimanapun jga, bahwa ada kasus non normal dimana fungsi klasifikasi linear berfingsi dengan buruk walaupun matriks kovarian dari populasinya sama. Etikanya adalah selalu mengecek hasil dari sebarang prosedur klasifikasi. Sekutang-kurangnya, ini harus dilakukan dengan himpunan data yang digunakan untuk membangun klasifikasinya. Idealnya, akan ada cukup data yang tersedia untuk memberikan terhadap sample latihan dan sampel validasi. Sampel latihan dapat digunakan untuk mengembangkan fungsi klasifikasi dan sampel validasi dapat digunakan untuk mengevaluasi hasilnya.

2.3 Mengevaluasi Fungsi Klasifikasi

Satu cara yang penting dari menilai hasil dari sebarang rosedur klasifikasi adalah dengan menghitung nilai errornya, atau probabilitas misklasifikasi. Ketika bentuk dari populasi induk dikenal secara menyeluruh, probabilitas misklasifikasi dapat dihitung dengan relatif mudah, seperti yang telah ditunjukkan dalam contoh 11.4. Karena populasi induk jarang sekali diketahui, kita seharusnya berkonsentrasi pada nilai error yang berhubungan dengan fungsi klasifikasi sampel. Segera sesudah fungsi klasifikasi ini dibentuk, pengukuran dari hasilnya dalam sampel yang akan datang merupakan perhatian kita.

Dari (11-8) TPMnya adalah $p_1\int_{R_2} f_1(x)dx - p_2\int_{R_1} f_2(x)dx$ (11-28)

Nilai terkecil dari jumlah ini, diperoleh dengan pilihan bijaksana dari R_1 dan R_2, disebut nilai error oprimum (OER).

Nilai error optimum $OER = p_1\int_{R_2} f_1(x)dx - p_2\int_{R_1} f_2(x)dx$

(11-26)

dimana R_1 dan R_2 ditentukan oleh kasus (b) dalam (11.7)

Oleh karena itu, OER adalah nilai error untuk kaidah klasifikasi TPM minimum.

Contoh 1.4

Diberikan sebuah pernyataan untuk nilai error optimum ketika $p_1 = p_2 = \frac{1}{2}$ dan $f_1(x)$ dan $f_2(x)$ merupakan densitas normal multivariat dalam (11-10).

Sekarang, kaidah klasifikasi ECM minimum dan TPM minimum serupa ketika $c(1|2) = c(2|1)$. Karena probabilitas prior juga sama, daerah klasifikasi TPM minimum didefinisikan untuk populasi normal oleh (11-12), dengan $\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] = 0$. Kita mendapatkan

$$R_1 : (\mu_1 - \mu_2)' \Sigma^{-1} x - \frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 + \mu_2) \geq 0$$

$$R_2 : (\mu_1 - \mu_2)' \Sigma^{-1} x - \frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 + \mu_2) < 0$$

Himpuan ini dapat dinyatakan dalam hubungan dari $y = (\mu_1 - \mu_2)' \Sigma^{-1} x = l' x$ sebagai

$$R_1(y) : y \geq \frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 + \mu_2)$$

$$R_2(y) : y < \frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 + \mu_2)$$

Tetapi Y merupakan sebuah kombinasi linear dari variabel acak normal, jadi probabilitas densitas dari Y, $f_1(y)$, dan $f_2(y)$, adalah normal univariat (lihat hasil 4.2) dengan rata-rata dan variansi diberikan oleh

$$\mu_{1Y} = l' \mu_1 = (\mu_1 - \mu_2)' \Sigma^{-1} \mu_1$$

$$\mu_{2Y} = l' \mu_2 = (\mu_1 - \mu_2)' \Sigma^{-1} \mu_2$$

$$\sigma_Y^2 = l' \Sigma l = (\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2) = \Delta^2$$

Sekarang, TPM = $\frac{1}{2}P$ [kesalahan mengklasifikasi observasi π_1 sebagai π_2]

+ $\frac{1}{2}P$ [kesalahan mengklasifikasi observasi π_2 sebagai π_1]

Tetapi, P [kesalahan mengklasifikasi observasi π_1 sebagai π_2] = $P(2|1)$

$$= P\left[Y < \frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 + \mu_2) \right]$$

$$= P\left(\frac{Y - \mu_{1Y}}{\sigma_Y} < \frac{\frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 + \mu_2) - (\mu_1 - \mu_2)' \Sigma^{-1} \mu_1}{\Delta} \right)$$

$$= P\left(Z < \frac{\frac{1}{2}\Delta^2}{\Delta} \right) = \Phi\left(\frac{-\Delta}{2} \right)$$

dimana $\Phi(.)$ merupakan fungsi distribusi kumulatif dari variabel acak normal standar. Dengan cara yang sama, P [kesalahan mengklasifikasi observasi π_2 sebagai π_1] = $P(1|2)$

$$= P\left[Y \geq \frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 + \mu_2) \right]$$

$$= P\left(Z \geq \frac{\Delta}{2} \right) = 1 - \Phi\left(\frac{\Delta}{2} \right) = \Phi\left(\frac{-\Delta}{2} \right)$$

Probabilitas misklasifikasi ini ditunjukkan dalam Figur 11.6 (liaht buku hal 513). Oleh karena itu, nilai error optimumnya adalah OER = TPM minimum =

$$\frac{1}{2}\Phi\left(\frac{-\Delta}{2} \right) + \frac{1}{2}\Phi\left(\frac{-\Delta}{2} \right) = \Phi\left(\frac{-\Delta}{2} \right) \qquad (11\text{-}27)$$

Jika, sebagai contoh, $\Delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) = 2.56$, lalu $\Delta = \sqrt{2.56} = 1.6$ dan mengunakan tabel appendiks 1, TPM minimum = $\Phi\left(\frac{-1.6}{2}\right) = \Phi(-0.8) = 0.2119$

Kaidah klasifikasi optimal di sini akan dengan tidak tepat dialokasikan, ke satu populasi atau yang lainnya, sekitar 21% dari materi.

Contoh 11.4 mengilustrasikan bagaimana nailai error optimum dapad dihitung ketika fungsi densitas populasinya diketahui. Jika, pada umumnya seperti halnya dalam kasus, paremeter populasi khusus yang muncul dalam kaidah alokasi harus diestimasi dari sampel, maka evaluasi dari nilai error tidak langsung.

Hasil dari fungsi klasifikasi sampel dapat, dalam prinsipnya, dievaluasi dengan mengkalkulasikan nilai error aktual (AER),

$$AER = p_1 \int_{\hat{R}_2} f_1(x)dx - p_2 \int_{\hat{R}_1} f_2(x)dx \qquad (11\text{-}28)$$

dimana $\hat{R}_1$ dan $\hat{R}_2$ mewaliki daerah klasifikasi yang ditentukan oleh ukuran sampel n_1 dan n_2, secara berturut-turut. Sebagai contoh, jika fungsi klasifikasi dalam (11-18) dipakai, daerah $\hat{R}_1$ dan $\hat{R}_2$ didefinisikan oleh himpunan dari x dimana pertidaksamaan berikut ini memenuhi.

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

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

AER mengndikasikan bagaimana fungsi klasifikasi sampel akan menghasilkan dalam sampel yang akan datang. Seperti nilai error optimum, itu tidak dapat, secara umum, dihirung karena itu tergantung pada fungsi densitas

yang tidak diketahui $f_1(x)$ dan $f_2(x)$. Bagaimanapun juga, sebuah estimasi dari jumlah yang berhubungan dengan nilai error aktual dapat dihitung, dan estimasi ini akan dibahas secara singkat.

Ada sebuah pengukuran dari hasil yang tidak bergantung pada bentuk dari populasi induk dan dapat dihitung untuk sebarang prosedur klasifikasi. Pengukuran ini, disebut nilai error nyata (APER), didefinisikan sebagai fraksi dari observasi dalam sampel latihan yang merupakan misklasifikasi oleh fungsi klasifikasi sampel.

Nilai error nyata dapat dengan mudah dihitung dari matriks *confusion*, yang mana menunjukkan grup anggota aktual melawan prediksi. Untuk n_1 observasi dari π_1 dan n_2 observasi dari π_2, matriks *confusion* mempunyai bentuk

Anggota prediksi

		π_1	π_2
Anggota Nyata	π_1	n_{1C}	$n_{1M} = n_1 - n_{1C}$
	π_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}

(11-29)

dimana n_{1C} = bilangan dari materi π_1 secara benar diklasifikasikan sebagai materi π_1

n_{1M} = bilangan dari materi π_1 salah diklasifikasikan sebagai materi π_2

n_{2C} = bilangan dari materi π_2 secara benar diklasifikasikan

n_{2M} = bilangan dari materi π_2 salah diklasifikasikan

Nilai error nyatanya adalah $APER = \frac{n_{1M} + n_{2M}}{n_1 - n_2}$ (11-30)

yang mana dikenali sebagai proporsi dari materi dalam himpunan latihan yang salah diklasifikasi.

Walau APER mudah dihitung, tapi APER terlalu rendah menaksir AER dan masalah ini tidak akan hilang kecuali jika ukuran sampel n_1 dan n_2 sangat besar. Hal ini terjadi karena data yang digunakan untuk membuat fungsi klasifikasi juga digunakan untuk mengevaluasinya.

Taksiran *error rate* dapat dibuat lebih baik dari AER, relative tetap mudah dihitung dan tidak memerlukan asumsi distribusional.

Satu prosedur untuk memisahkan sampel total kedalam sampel training dan sampel validasi. Sampel training/*training sampel* digunakan untuk mengkontruksi fungsi klasiffikasi dan sampel validasi digunakan untuk mengevaluasinya.

Error rate ditentukan oleh proporsi kesalahan klasifikasi pada sampel validasi. Walaupun metode ini mengatasi masalah bias atau penyimpangan dengan tidak menggunakan data yang sama baik untuk membentuk fungsi klasifikasi dan menilainya, namun metode ini mempunyai dua cacat utama, yaitu :

1. Membutuhkan sampel yang berukuran besar,
2. Fungsi yang dievaluasi bukan fungsi yang dihasilkan. Pada akhirnya hampir semua data harus digunakan untuk membentuk fungsi klasifikasi. Jika tidak, informasi berharga dapat saja hilang.

Pendekatan lainnya, selain memisahkan sampel total kedalam sampel training dan sampel validasi, yang nampaknya bekerja dengan baik disebut prosedur “holdout” Lachenbruch. Algoritma prosedur ini adalah sebagai berikut :

1. Mulai dengan pengamatan pada grup π_1. Abaikan satu observasi dari grup ini dan hasilkan fungsi klasifikasi berdasarkan pada sisa $n_1 - 1, n_2$ observasi.
2. Klasifikasi observasi yang ditahan (*the ‘holdout’ observation*) dengan menggunakan fungsi yang dihasilkan pada langkah 1.

3. Ulangi langkah 1 dan 2 sampai semua observasi π_1 diklasifikasikan. Misal $n_{1M}^{(H)}$ adalah banyaknya observasi holdout dalam grup ini (H) yang salah diklasifikasikan.
4. Ulangi langkah 1 sampai 3 untuk observasi π_2. Misal $n_{2M}^{(H)}$ adalah banyaknya observasi holdout dalam grup ini yang salah diklasifikasikan.

Taksiran $\hat{P}(2|1)$ dan $\hat{P}(1|2)$ dari probabilitas misklasifikasi bersyarat pada (11-1) dan (11-2) diberikan oleh :

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

$$\hat{P}(2|1) = \frac{n_{2M}^{(H)}}{n_2} \qquad (11\text{-}31)$$

dan total proporsi misklasifikasi,$\left(n_{1M}^{(H)} + n_{2M}^{(H)} \right)/(n_1 + n_2)$ adalah, hampir tak bias mengestimasi AER yang diharapkan, *E*(AER).

$$\hat{E}(AER) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2} \qquad (11-32)$$

2.4 Fungsi Diskriminan Fisher – Pemisahan Populasi

Fisher baru menyampaikan statistik pengklasifikasian yang linier (11-19) dengan menggunakan suatu yang sama sekali yang berbeda. Gagasan Fisher mengubah bentuk pengamatan-pengamatan multivariate *x* ke pengamatan-pengamatan univariate *y* bahwa *y* berasal dari populasi-populasi $\pi_1$dan π_1 yang telah dipisahkan. Fisher mengusulkan pengambilan kombinasi linier dari *x* untuk membentuk *y* karena fungsi-fungsi *x* cukup sederhana untuk ditangani dengan mudah. Pendekatan Fisher tidak berasumsi bahwa populasi-populasi itu bersifat normal. bagaimanapun, secara implisit mengasumsikan matriks kovarians populasi bersifat sama karena suatu perkiraan dari matriks kovarians yang disatukan yang umum digunakan.

Suatu kombinasi liniar yang ditetapkan diperbaiki *x* diambil dari nilai-nilai $y_{11}, y_{12}, ..., y_{1n_1}$ untuk pengamatan-pengamatan dari populasi yang pertama dan

nilai-nilai $y_{21}, y_{22},..., y_{2n_2}$ untuk pengamatan-pengamatan dari populasi yang kedua. Pemisahan dua himpunan ini dari univariate *y* ditaksir dari selisih antara $\bar{y}_1$ *dan* $\bar{y}_2$ diyatakan dalam simpangan baku. Yaitu;

$$pemisahan = \frac{|\bar{y}_1 - \bar{y}_2|}{s_y}; \quad \text{dimana } s_y^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n2}(y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2}$$

adalah estimasi yang disatukan dari varians. Tujuannya untuk memilih kombinasi linier *x* untuk mencapai pemisahan yang maksimum dari sampel $\bar{y}_1$ *dan* $\bar{y}_2$.

Result 11.4. Kombinasi linier $y = \hat{\ell}' x = (\bar{x}_1 - \bar{x})' S_{H.pooled}^{-1} x$ memksimalkan rasio

$$\frac{\left(\begin{matrix}\text{Squared distance} \\ \text{between sample means of } y\end{matrix}\right)}{(\text{Sample variance of } y)} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(\hat{\ell}'\bar{\mathbf{x}}_1 - \hat{\ell}'\bar{\mathbf{x}}_2)^2}{\hat{\ell}'\mathbf{S}_{\text{pooled}}\hat{\ell}} = \frac{(\hat{\ell}'\mathbf{d})^2}{\hat{\ell}'\mathbf{S}_{\text{pooled}}\hat{\ell}} \qquad (11\text{-}33)$$

atas semua vektor-vektor koefisien $\hat{\ell}$ yang mungkin dimana $d = (\bar{x}_1 - \bar{x}_2)$. Maksimum dari perbandingan (11-33) adalah $D^2 = (\bar{x}_1 - \bar{x}_2)' S_{H.pooled}^{-1} (\bar{x}_1 - \bar{x}_2)$

Bukti.

Maksimum dari perbandingan di dalam (11-33) diberi dengan menerapkan (2-50) secara langsung. Jadi; Dengan demikian, menentukan $d = (\bar{x}_1 - \bar{x}_2)$

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

di mana D^2 adalah jarak kuadrat sampel antara kedua rerata.

Catat bahwa s_y^2 di dalam (11-33) bisa dihitung sebagai

$$s_y^2 = \frac{\sum_{j=1}^{n_1} (y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2} \qquad (11\text{-}34)$$

Dengan $y_{1j} = \hat{\ell}' x_{1j}$ *dan* $y_{2j} = \hat{\ell}' x_{2j}$

Contoh 11.8

Suatu studi terkait dengan pendeteksian hemofili A diperkenalkan di Example 11.3. Ingat bahwa *cost* dan fungsi diskriminan linear prior yang sama adalah

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

Fungsi diskriminan linear di atas adalah fungsi linear Fisher, yang secara maksimal memisahkan kedua populasi-populasi, dan pemisahan yang maksimum di dalam sampel-sampel itu adalah

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

Solusi Fisher dalam masalah pemisahan dapat juga digunakan untuk menggolongkan pengamatan-pengamatan baru.

Satu aturan alokasi yang didasarkan pada Fungsi Diskriminan Fisher

alokasikan x_o ke π_1 jika

$$\begin{aligned} y_0 &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 \\ &\geq \hat{m} = \tfrac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \end{aligned}$$

atau

$$y_0 - \hat{m} \geq 0$$

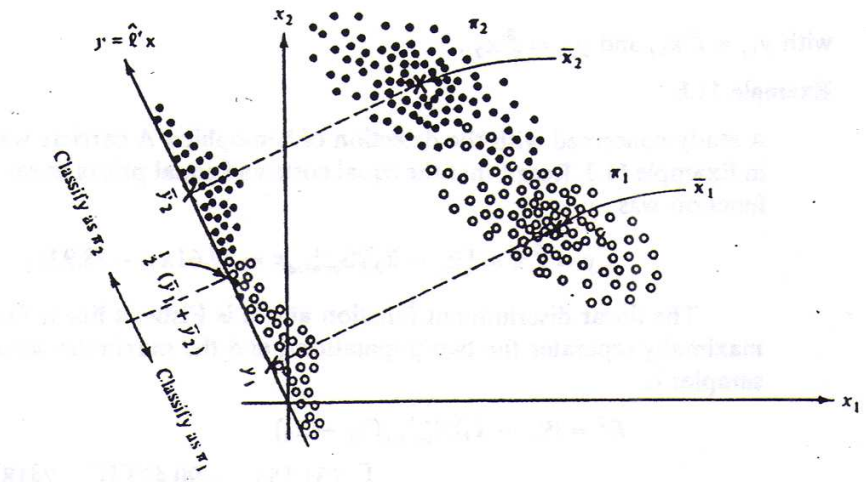
alokasikan x_o ke π_2 jika

$$y_0 < \hat{m} \qquad (11\text{-}35)$$

atau

$$y_0 - \hat{m} < 0$$

Prosedur (11-33) digambarkan, secara sistematiks untuk p =2 di dalam Gambar 118. Semua titik-titik di dalam scatterplots itu diproyeksikan ke satu baris di dalam arah $\hat{\ell}$ dan arah ini bervariasi sampai sampel berpisah secara maksimal.



Gambar 11.8 penyajian yang bergambar prosedur Fisher untuk dua populasi

Fungsi diskriminan linear Fisher di dalam (11-35) dikembangkan di bawah asumsi dua populasi, yang mempunyai suatu matriks kovarians yang umum. Sebagai konsekwensi, mungkin tidak mengejutkan bahwa metoda Fisher berkorespodensi dengan kasus tertentu dari aturan ekspektasi *cost misclassification* minimum. Istilah yang pertama, $\hat{y} = (\bar{x}_1 - \bar{x}_2)' S_{H.pooled}^{-1} x$ di dalam aturan pengklasifikasian (11-18) adalah fungsi linear yang diperoleh oleh Fisher bahwa memaksimalkan univariate "antara" sampel variabilitas relative dengan "di dalam" sampel variabilitas [lihat (11-33)]. Yaitu;

$$\begin{aligned} w &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} - \tfrac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \\ &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}[\mathbf{x} - \tfrac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)] \end{aligned} \qquad (11\text{-}36)$$

sering disebut fungsi pengklasifikasian Anderson (statistik). Sekali lagi, jika $\left[\left(c(1|2)/c(2|1)\right)(p_1/p_2)\right] = 1$ sehingga $\ln\left[\left(c(1|2)/c(2|1)\right)(p_1/p_2)\right] = 0$. Aturan (11-18) dapat diperbandingkan dengan aturan (11-35) berdasar pada fungsi diskriminan linear Fisher. Jadi; Dengan demikian, dengan syarat kedua populasi-

populasi normalyang mempunyai matriks kovarians yang sama, aturan pengklasifikasian Fisher's adalah setara dengan ECM yang minimum dengan peluang prior dan *cost* misclassification sama.

Secara ringkas, selama dua populasi, pemisahan relatif maksimum dapat diperoleh dengan mempertimbangkan kombinasi linier pengamatan-pengamatan multivariate dengan jarak D^2 yang sama. Ini tepat karena D^2 dapat digunakan, di dalam situasi-situasi yang tertentu, untuk menguji apakah rerata populasi μ_1 dan μ_2 berbeda secara signifikan. Sebagai akibatnya, suatu test untuk selisih rerata di dalam vektor-vektor dapat dipandang sebagai suatu test "signifikan" dari pemisahan yang dapat dicapai.

Misalkan populasi-populasi π_1 dan π_2 adalah normal multivariate dengan suatu matriks kovarians yang umum $\sum$ .Lalu, seperti di Section 6.3, suatu test dari Ho: $\mu_1=\mu_2$ melawan H_1: $\mu_1 \neq \mu_2$ dapat ditunjukan oleh;

$$\left(\frac{n_1+n_2-p-1}{(n_1+n_2-2)\,p}\right)\left(\frac{n_1 n_2}{n_1+n_2}\right)D^2$$

Suatu F-distribution dengan df $v_1 = p \; dan \; v_2 = n_1 + n_2 - p - 1$. Jika H_o ditolak, kita dapat menyimpulkan pemisahan antara kedua populasi-populasi $\pi_1$dan π_2 adalah signifikan

Pemisahan signifikan tidak perlu menyiratkan pengklasifikasian baik. Seperti kita sudah melihat di Section 11.4, kemanjuran dari suatu prosedur pengklasifikasian dapat dievaluasi bebas dari setiap test perpisahan. Sebaliknya, jika pemisahan itu tidak signifikan, pencarian suatu aturan pengklasifikasian yang bermanfaat mungkin akan membuktikan tidak berarti.

2.5 Klasifikasi Untuk Beberapa Populasi

Teorinya, secara umum prosedur klasifikasi dari 2 sampai $g \geq 2$ grup akan dijelaskan disini. Meskipun, tidak banyak yang mengetahui tentang syarat dari korespondensi fungsi klasifikasi sampel dan, faktanya, nilai kekeliruan dari klasisfikasi di atas tidak dapat dihitung.

Teori "Robust" dari statistik klasifikasi linier dua grup, misalnya, kovarian tidak sama atau tidak berdistribusi normal dapat kita pelajari dengan menggunakan program komputer yang mendukung sampling eksperimen. Untuk populasi lebih dari dua, pendekatan ini tidak dapat menghasilkan kesimpulan secara umum karena syaratnya bergantung pada di mana populasi itu akan ditempatkan dan masih banyak bentuk yang dipelajari lebih dalam.

Sebelumnya pendekatan kita pada bahasan sekarang akan dibentuk aturan optimal secara teori selanjutnya modifikasi yang diperoleh pada aplikasi kehidupan nyata.

Nilai Harapan Minimum dari Misclassification Method

Misal $f_i(x)$ fungsi kepadatan dari populasi $\pi_i, i = 1, 2, ..., g$.[sebagian besar kita akan menggunakan $f_i(x)$ berupa fungsi kepadatan multivariat normal, tetapi tidak diperlukan pada pembentukan teori secara umum.]. Misal

$p_i =$ probabilitas prior dari populasi $\pi, i = 1, 2, ..., g$

$c(k \mid i) =$ nilai alokasi dari item π_k, kenyataannya berkaitan dengan π_i

untuk $k, i = 1, 2, ..., g$

untuk $k = 1, c(i \mid i) = 0$.

Nilai harapan kondisional dari klasifikasi x dari

π_1 sampai π_2, atau $\pi_3, ...,$ π_g adalah

$$\begin{aligned} \text{ECM}(1) &= P(2 \mid 1)c(2 \mid 1) + P(3 \mid 1)c(3 \mid 1) + ... + P(g \mid 1)c(g \mid 1) \\ &= \sum_{k=2}^{g} P(k \mid 1)c(k \mid 1) \end{aligned}$$

Perkalian dari ECM kondisional masing-masing oleh probabilitas priornya dan dijumlahkan akan menghasilkan

$$\begin{aligned}\text{ECM} &= p_1\text{ECM}(1)+p_2\text{ECM}(2)+\ldots+p_g\text{ECM}(g)\\ &= p_1\left(\sum_{k=2}^{g}P(k\mid 1)c(k\mid 1)\right)+p_2\left(\sum_{k=2}^{g}P(k\mid 2)c(k\mid 2)\right)+\ldots+p_g\left(\sum_{k=2}^{g-1}P(k\mid g)c(k\mid g)\right)\\ &= \sum_{i=1}^{g}p_i\left(\sum_{\substack{k=2\\k\neq i}}^{g}P(k\mid i)c(k\mid i)\right) \qquad (11\text{-}37)\end{aligned}$$

Perhitungan jumlah klasifikasi optimal pada pemilihan kualitas daerah klasifikasi $R_1, R_2, \ldots, R_g$ khusus dan mendalam seperti pada (11-37) adalah minimum.

Hasil 11.5 daerah klasifikasi untuk nilai ECM minimum (11-37) didefinisikan oleh pengalokasian x pada populasi $\pi_k, k = 1, 2, \ldots, g$ di mana

$$\sum_{\substack{i=1\\i\neq k}}^{g} p_i f_i(x) c(k\mid i) \qquad (11\text{-}38)$$

adalah yang terkecil. Jika berkaitan, x dapat dimasukkan ke populasi yang berkaitan itu.

Bukti. Lihat Anderson [2].

Misalkan semua nilai misklasifikasi sama, pada kasus nilai harapan minimum dari aturan misklasifikasi adalah probabilitas total minimum dari aturan misklasifiasi. (Tanpa menghilangkan sifat umum kita dapat menentukan semua nilai misklasifikai sama dengan 1.) Menggunakan alasan sebelumnya pada (11-38), kita akan mengalokasikan x pada $\pi_k, k = 1, 2, \ldots, g$, di mana

$$\sum_{\substack{i=1\\i\neq k}}^{g} p_i f_i(x) \qquad (11\text{-}39)$$

adalah yang terkecil. (11-39) akan menjadi yang terkecil ketika mengabaikan bentuk $p_k f_k(x)$ yang terbesar. Akibatnya, ketika semua nilai misklsifikasi sama, nilai harapan minimum dari aturan misklasifikasi ditunjukkan bentuk sederhana.

Syarat klasifikasi ECM minimum dengan nilai misklasifikasi sama

alokasi x pada π_k jika

$$p_k f_k(x) > p_i f_i(x) \quad \text{untuk semua i} \neq k \qquad (11\text{-}40)$$

atau, ekivalen dengan

alokasi x pada π_k jika

$$\ln p_k f_k(x) > \ln p_i f_i(x) \quad \text{untuk semua i} \neq k \qquad (11\text{-}41)$$

Hal ini menarik untuk diingat bahwa aturan klasifikasi pada (11-40) adal identik dengan memaksimumkan probabilitas posterior, $\mathrm{P}(\pi_k \mid x) =$ P(x dari π_k diganti dengan x dari observasi), dimana

$$\mathrm{P}(\pi_k \mid x) = \frac{p_k f_k(x)}{\sum_{i=1}^{g} p_i f_i(x)} = \frac{(\text{prior})\,\mathrm{x}\,(\text{likelihood})}{\sum\left[(\text{prior})\,\mathrm{x}\,(\text{likelihood})\right]} \quad \text{untuk } k = 1, 2, ..., g \tag{11-42}$$

Persamaan (11-42) adalah bentuk umum dari persamaan (11-9) untuk $g \geq 2$ grup.

Anda seharusnya mengingat hal tersebut, umumnya, aturan ECM minimum mempunyai tiga komponen: probabilitas prior, nilai misklasifikasi, dan fungsi kepadatan. Komponen ini sudah diketahui (atau diestimasi) sebelum aturan ini digunakan.

Contoh 11.9

Misalkan didapat observasi x_0 ke suatu $g = 3$ populasi π_1, π_2, atau π_3, diberikan hipotesis probabilitas prior, nilai misklasifikasi, dan nilai fungsi kepadatan. Kita gunakan prosedur ECM minimum.

		Populasi		
		π_1	π_2	π_3
Klasifikasi :	π_1	$c(1\mid 1) = 0$	$c(1\mid 2) = 500$	$c(1\mid 3) = 100$
	π_2	$c(2\mid 1) = 10$	$c(2\mid 2) = 0$	$c(2\mid 3) = 50$
	π_3	$c(3\mid 1) = 50$	$c(3\mid 2) = 200$	$c(3\mid 3) = 0$
Probabilitas prior:		$p_1 = 0,05$	$p_2 = 0,60$	$p_3 = 0,35$
Kepadatan di x_0:		$f_1(x_0) = 0,01$	$f_2(x_0) = 0,85$	$f_2(x_0) = 2$

Nilai dari $\sum_{\substack{i=1 \\ i \neq k}}^{3} p_i f_i(x_0) c(k \mid i)$ $\left[\text{lihat } (11\text{-}38)\right]$ adalah

$$\begin{aligned}
k = 1&: p_2 f_2(x_0) c(1\mid 2) + p_3 f_3(x_0) c(1\mid 3) \\
&= (0,60)(0,85)(500) + (0,35)(2)(100) = 325 \\
k = 2&: p_1 f_1(x_0) c(2\mid 1) + p_3 f_3(x_0) c(2\mid 3) \\
&= (0,05)(0,01)(10) + (0,35)(2)(50) = 35,055 \\
k = 3&: p_1 f_1(x_0) c(3\mid 1) + p_2 f_2(x_0) c(3\mid 2) \\
&= (0,05)(0,01)(50) + (0,60)(0,85)(200) = 102,025
\end{aligned}$$

Karena $\sum_{\substack{i=1 \\ i\neq k}}^{3} p_i f_i(x_0) c(k \mid i)$ yang terkecil untuk $k=2$, kita akan alokasikan x_0 ke π_2.

Jika semua nilai misklasifikasinya sama, kita akan memasukkan x_0 berdasarkan (11-40), sehingga diperoleh

$$p_1 f_1(x_0) = (0,05)(0,01) = 0,0005$$
$$p_2 f_2(x_0) = (0,60)(0,85) = 0,510$$
$$p_3 f_3(x_0) = (0,35)(2) = 0,700$$

Karena

$$p_3 f_3(x_0) = 0,700 \geq p_i f_i(x_0), \qquad i = 1,2$$

Kita harus mengalokasikan x_0 ke π_3. Ekivalen dengan, perhitungan probabilitasnya, [lihat (11-42)]

$$\mathrm{P}(\pi_1 \mid x_0) = \frac{p_1 f_1(x_0)}{\sum_{i=1}^{3} p_1 f_1(x_0)} = \frac{(0,05)(0,01)}{(0,05)(0,01)+(0,60)(0,85)+(0,35)(2)} = 0,0004$$

$$\mathrm{P}(\pi_2 \mid x_0) = \frac{p_2 f_2(x_0)}{\sum_{i=1}^{3} p_1 f_1(x_0)} = \frac{(0,60)(0,85)}{(0,05)(0,01)+(0,60)(0,85)+(0,35)(2)} = 0,421$$

$$\mathrm{P}(\pi_3 \mid x_0) = \frac{p_3 f_3(x_0)}{\sum_{i=1}^{3} p_1 f_1(x_0)} = \frac{(0,35)(2)}{(0,05)(0,01)+(0,60)(0,85)+(0,35)(2)} = 0,578$$

Kita lihat bahwa x_0 dialokasikan ke π_3, populasi dengan probabilitas posterior terbesar.

Klasifikasi dengan Populasi Normal

Kasus khusus terjadi ketika

$$f_i(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\mu_i)' \Sigma_i^{-1} (x-\mu_i) \right], \quad i = 1,2,...,g \qquad (11\text{-}43)$$

Adalah kepadatan multivariat normal dengan vektor mean μ_i dan matriks Σ_i. Jika didapat $c(i|i)=0, c(k|i)=1, k \neq i$ (atau, ekivalen, nilai misklasifikasinya sama semua), (11-41) menjadi:

Alokasi x ke π_k jika

$$\ln p_k f_k(x) = \ln p_k - \left(\frac{p}{2}\right)\ln(2\pi) - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k)$$
$$= \max_i \ln p_i f_i(x) \qquad (11-44)$$

Konstanta $\left(\frac{p}{2}\right)\ln(2\pi)$ dapat diabaikan pada (11-44) karena sama untuk semua populasi. Selanjutnya kita definisikan nilai diskriminan kuadrat untuk ke-I populasi berupa

$$d_i^Q(x) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i) + \ln p_{i,} \quad i=1,2,...,g \qquad (11\text{-}45)$$

Nilai kuadrat, $d_i^Q(x)$, terdiri dari kontribusi dari variansi umum $|\Sigma_i|$, probabilitas prior p_i, dan kuadrat jarak dari x ke populasi mean μ_i. Catatan, meskipun fungsi jarak berbeda, dengan orientasi dan ukuran konstanta jarak ellipsoid berbeda, sebaiknya lakukan pada masing-masing populasi.

Gunakan nilai diskriminasi aturan klasifikasi dari (11-44) seperti di bawah

Probabilitas Total Minimum dari Aturan Misklasifikasi untuk Populasi Normal-$|\Sigma_i|$ berbeda

Alokasi x ke π_k jika

Nilai kuadrat $d_i^Q(x) =$ maks dari $d_1^Q(x), d_2^Q(x), ..., d_g^Q(x)$
dimana $d_i^Q(x)$ diberikan pada (11-45), $i=1,2,...,g$.

Pada latihan, μ_i dan Σ_i tidak diketahui, tetapi latihan penentuan dari pengklasifikasian observasi yang benar sering berguna untuk mengestimsi. Kuantitas sampel yang relevan untuk populasi π_i adalah

$\overline{x_i} =$ vektor mean sampel
$S_i =$ matrik kovarian sampel
$n_i =$ ukuran sampel

Estimasi nilai diskriminan kuadrat $\hat{d}_i^Q(x)$ adalah

$$\hat{d}_i^Q(x) = -\frac{1}{2}\ln|S_i| - \frac{1}{2}\left(x-\bar{x}_i\right)' S_i^{-1}\left(x-\bar{x}_i\right) + \ln p_i \qquad (11\text{-}47)$$

Dan aturan klasifikasi yang sesuai dengan sampel sebelumnya.

Aturan Estimasi TPM Minimum untuk Beberapa Populasi Normal-Σ_i berbeda

Alokasi x ke π_k jika

Nilai kuadrat $\hat{d}_i^Q(x) =$ maks dari $\hat{d}_1^Q(x), \hat{d}_2^Q(x), ..., \hat{d}_g^Q(x)$ (11-48)

dimana $\hat{d}_i^Q(x)$ diberikan pada (11-47), $i = 1, 2, ..., g$.

Penyederhanaan bisa dilakukan jika matriks kovarian populasi Σ_i sama. Ketika $\Sigma_i = \Sigma$, untuk $i = 1, 2, ..., g$, nilai diskriminan pada (11-45) menjadi

$$d_i^Q(x) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}x'\Sigma^{-1}x + \mu_i'\Sigma^{-1}x - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + \ln p_i$$

Dua bentuk pertama nilainya sama untuk $d_1^Q(x), d_2^Q(x), ..., d_g^Q(x)$, dan mengakibatkan, dapat kita abaikan. Bentuk yang lainnya terdiri dari konstanta $c_i = \ln p_i - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i$ dan kombinasi linier dari komponen x.

Definisi nilai diskriminan linier

$$d_i^Q(x) = \mu_i'\Sigma^{-1}x - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + \ln p_i \qquad (11\text{-}49)$$

Estimasi $\hat{d}_i^Q(x)$, dari nilai diskriminan linier $d_i^Q(x)$ sesuai dengan estimasi gabungan Σ,

$$S_{\text{pooled}} = \frac{(n_1-1)S_1 + (n_2-1)S_2 + ... + (n_g-1)S_g}{n_1 + n_2 + ... + n_g - g} \qquad (11\text{-}50)$$

Dan diberikan oleh

$$\hat{d}_i(x) = \bar{x}_i' S_{\text{pooled}}^{-1} x - \frac{1}{2} \bar{x}_i' S_{\text{pooled}}^{-1} \bar{x}_i + \ln p_i \qquad (11\text{-}51)$$

Aturan Estimasi TPM Minimum untuk Populasi Normal dengan Kovarian sama

Alokasi nilai x ke π_k jika

Nilai diskriminan linier $\hat{d}_i^Q(x) =$ maks dari $\hat{d}_1^Q(x), \hat{d}_2^Q(x), ..., \hat{d}_g^Q(x)$ (11-52)

dimana $\hat{d}_i^Q(x)$ diberikan pada (11-51), $i = 1, 2, ..., g$.

Keterangan. Persamaan (11-49) adalah fungsi linier dari x yang sesuai. Sama seperti untuk kasus kovarian sama yang didapat dari (11-45) dengan mengabaikan bentuk konstanta, $-\frac{1}{2}\ln|\Sigma|$. Hasilnya, estimasi sampel yang dimasukkan untuk kuantitas populsi yang tidak diketahui, selanjutnya dapat diinterpretasikan dalam bentuk kuadrat jarak

$$D_i^2(x) = \left(x - \bar{x}_i\right) S_{pooled}^{-1} \left(x - \bar{x}_i\right) \qquad (11\text{-}53)$$

Dari x ke vektor mean sampel $\bar{x}_i$. Aturan menempatkannya adalah:

Masukkan x ke populasi π_i yang $-\frac{1}{2} D_i^2(x) + \ln p_i$ terbesar (11-54)

Kita lihat bahwa aturan ini atau ekivalennya, (11-52)-masukkan x ke populasi terdekat. (ukuran jarak dinyatakan oleh $\ln p_i$.)

Jika probabilitas priornya tidak diketahui, prosedur yang berguna adalah menentukan $p_1 = p_2 = ... = p_g = \frac{1}{g}$. Observasi selanjutnya dimasukkan ke populasi yang terdekat.

Contoh 11.10

Mari kita hitung nilai diskriminan linier yang berasal dari data dengan $g = 3$. Populasi diasumsikan sebagai normar bivariat dengan matrik kovarian biasa.

Sampel acak dari populasi π_1, π_2, dan π_3 disebutkan di bawah, beserta mean sampel dan matriks kovariannya.

$$\pi_1 : X_1 = \begin{bmatrix} -2 & 0 & -1 \\ 5 & 3 & 1 \end{bmatrix} \text{ sehingga } n_1 = 3, \quad \bar{x}_1 = \begin{bmatrix} -1 \\ 3 \end{bmatrix}, \text{ dan } S_1 = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

$$\pi_2 : X_2 = \begin{bmatrix} 0 & 2 & 1 \\ 5 & 3 & 1 \end{bmatrix} \quad \text{sehingga } n_2 = 3, \quad \bar{x}_2 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \text{ dan } S_2 = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

$$\pi_3 : X_3 = \begin{bmatrix} 1 & 0 & -1 \\ -2 & 0 & -4 \end{bmatrix} \text{sehingga } n_3 = 3, \quad \bar{x}_3 = \begin{bmatrix} 0 \\ -2 \end{bmatrix}, \text{ dan } S_3 = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$$

Diberikan $p_1 = p_2 = 0,25,$ dan $p_3 = 0,50$, mari kita klasifikasi observasi $x_0' = [x_{01}, x_{02}] = [-2, -1]$ berdasarka (11-52). Dari (11-50)

$$S_{\text{pooled}} = \frac{(3-1)\begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix} + (3-1)\begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix} + (3-1)\begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}}{9-3}$$

$$= \frac{2}{6}\begin{bmatrix} 1+1+1 & -1-1+1 \\ -1-1+1 & 4+4+4 \end{bmatrix} = \begin{bmatrix} 1 & -\frac{1}{3} \\ -\frac{1}{3} & 4 \end{bmatrix}$$

sehingga

$$S_{pooled}^{-1} = \frac{9}{35}\begin{bmatrix} 4 & \frac{1}{3} \\ \frac{1}{3} & 1 \end{bmatrix} = \frac{1}{35}\begin{bmatrix} 36 & 3 \\ 3 & 9 \end{bmatrix}$$

selanjutnya,

$$\bar{x}_1' S_{pooled}^{-1} = [-1 \quad 3]\frac{1}{35}\begin{bmatrix} 36 & 3 \\ 3 & 9 \end{bmatrix} = \frac{1}{35}[-27 \quad 24]$$

dan

$$\bar{x}_2' S_{pooled}^{-1} \bar{x}_2 = \frac{1}{35}[48 \quad 39]\begin{bmatrix} 1 \\ 4 \end{bmatrix} = \frac{204}{35}$$

$$\hat{d}_1(x_0) = -1,386 + \left(\frac{-27}{35}\right)(-2) + \left(\frac{24}{35}\right)(-1) - \frac{1}{2}\left(\frac{99}{35}\right) = -1,943$$

$$\hat{d}_2(x_0) = -1,386 + \left(\frac{48}{35}\right)(-2) + \left(\frac{39}{35}\right)(-1) - \frac{1}{2}\left(\frac{204}{35}\right) = -1,8158$$

$$\hat{d}_3(x_0) = -0,693 + \left(\frac{-6}{35}\right)(-2) + \left(\frac{-18}{35}\right)(-1) - \frac{1}{2}\left(\frac{36}{35}\right) = -3,50$$

Perhatikan bentuk linier dari $\hat{d}_1(x_0) =$ konstanta + (konstanta)x_{01} +(konstanta)x_{02} . Dengan cara yang sama

$$\bar{x}_2' S_{pooled}^{-1} = [1 \quad 4]\frac{1}{35}\begin{bmatrix} 36 & 3 \\ 3 & 9 \end{bmatrix} = \frac{1}{35}[48 \quad 39]$$

$$\bar{x}_2' S_{pooled}^{-1} \bar{x}_2 = \frac{1}{35}[48 \quad 39]\begin{bmatrix} 1 \\ 4 \end{bmatrix} = \frac{204}{35}$$

dan

$$\hat{d}_2(x_0) = \ln(0,25) + \left(\frac{48}{35}\right)x_{01} + \left(\frac{39}{35}\right)x_{02} - \frac{1}{2}\left(\frac{204}{35}\right)$$

Terakhir,

$$\bar{x}_3' S_{pooled}^{-1} = [0 \quad -2]\frac{1}{35}\begin{bmatrix} 36 & 3 \\ 3 & 9 \end{bmatrix} = \frac{1}{35}[-6 \quad -18]$$

$$\bar{x}_3' S_{pooled}^{-1} \bar{x}_3 = \frac{1}{35}[-6 \quad -18]\begin{bmatrix} 0 \\ -2 \end{bmatrix} = \frac{36}{35}$$

dan

$$\hat{d}_3(x_0) = \ln(0,50) + \left(\frac{-6}{35}\right)x_{01} + \left(\frac{-18}{35}\right)x_{02} - \frac{1}{2}\left(\frac{36}{35}\right)$$

Substitusikan nilai $x_{01} = -2$ dan $x_{02} = -1$ didapat

$$\hat{d}_1(x_0) = -1,386 + \left(\frac{-27}{35}\right)(-2) + \left(\frac{24}{35}\right)(-1) - \frac{1}{2}\left(\frac{99}{35}\right) = -1,943$$

$$\hat{d}_2(x_0) = -1,386 + \left(\frac{48}{35}\right)(-2) + \left(\frac{39}{35}\right)(-1) - \frac{1}{2}\left(\frac{204}{35}\right) = -8,158$$

$$\hat{d}_3(x_0) = -0,693 + \left(\frac{-6}{35}\right)(-2) + \left(\frac{-18}{35}\right)(-1) - \frac{1}{2}\left(\frac{36}{35}\right) = -0,350$$

Karena $\hat{d}_3(x_0) = -0,350$ nilai diskriminan yang terbesar, maka kita alokasikan x ke π_3.

Contoh 11.11

Pengakuan pekerja dari sekolah bisnis yang mempunya indek kelulusannya berupa nilai grade point average (GPA) dan graduate manegement aptitude test (GMAT) yang dapat membantu memutuskan pelamar mana yang diakui program lulusan sekolahnya. Gambar 11.9 menunjukkan pasangan nilai $x_1 =$ GPA, $x_2 =$ GMAT untuk grup dari pelamar yang tersedia dikategorikan seperti π_1: admit; π_2: not admit; dan π_3 borderline. Data gambar tersebut disajikan pada tabel 11.6 (lihat latihan 11.29). hasil dari data tersebut (lihat SAS statistical software output pada panel 11.1 pada halaman 534-535)

$$n_1 = 31 \qquad n_2 = 28 \qquad n_3 = 26$$

$$\bar{x}_1 = \begin{bmatrix} 3.40 \\ 561.23 \end{bmatrix} \qquad \bar{x}_2 = \begin{bmatrix} 2.48 \\ 447.07 \end{bmatrix} \qquad \bar{x}_3 = \begin{bmatrix} 2.99 \\ 446.23 \end{bmatrix}$$

$$\bar{x} = \begin{bmatrix} 2.97 \\ 488.45 \end{bmatrix} \qquad S_{pooled} = \begin{bmatrix} 0.0361 & -2.0188 \\ -2.0188 & 3655.9011 \end{bmatrix}$$

Misalkan ada pelamar baru dengan nilai GPA dari $x_1 = 3,21$ dan GAMAT dari $x_2 = 497$. Mari kita klasifikasikan pelamar ini menggunakan aturan pada (11-54) dengan probabilitas prior sama.

Dengan $x_0' = [3.21, 497]$ jarak sampelnya

$$\begin{aligned} D_1^2(x_0) &= (x_0 - \bar{x}_1) S_{pooled}^{-1} (x_0 - \bar{x}_1) \\ &= [3.21 - 3.40, 497 - 561.23] \begin{bmatrix} 28.6096 & 0.0158 \\ 0.0158 & 0.0003 \end{bmatrix} \begin{bmatrix} 3.21 & -3.40 \\ 497 & -561.23 \end{bmatrix} \\ &= 2.58 \end{aligned}$$

$$D_2^2(x_0) = (x_0 - \bar{x}_2) S_{pooled}^{-1} (x_0 - \bar{x}_2) = 17.10$$

$$D_3^2(x_0) = (x_0 - \bar{x}_3) S_{pooled}^{-1} (x_0 - \bar{x}_3) = 2.47$$

Karena jarak dari $x_0' = [3.21, 497]$ ke mean grup $\bar{x}_3$ yang terkecil, kita masukkan pelamar ini ke π_3: borderline.

2.6 Metode Fisher Untuk Pendiskriminasian Diantara Beberapa Populasi

Tujuan utama dari analisis diskriminan Fisher adalah untuk memisahkan populasi. Dalam hal ini tidak perlu diasumsukan bahwa g buah populasi berdistribusi normal multivariat. Akan tetapi kita mengasumsikan matriks kovarian populasi $p \times p$ adalah sama, yaitu $\sum_1 = \sum_2 = \cdots = \sum_g = \sum$.

Misal $\bar{\mu}$ = vektor rata-rata dari kombinasi poopulasi

B_0= jumlah dari *cross-products* diantara grup

maka $B_0 = \sum_{i=1}^{g}(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})' \; dengan \; \bar{\mu} = \frac{1}{g}\sum_{i=1}^{g}\mu_i$ (11-58)

Perhatikan kombinasi linear : $Y = \ell'X$

yang mempunyai nilai ekspektasi :

$E(Y) = E(\ell'X) = \ell E(X|\pi_i) = \ell\mu_i$ untuk populasi π_i.

dan variansi :

$Var(Y) = \ell' Cov(X)\ell = \ell'\Sigma\ell$ untuk semua populasi.

Akibatnya, nilai ekspektasi $\mu_{iY} = \ell'\mu_i$ berubah sebagai populasi darimana X yang dipilih berubah.

Definisikan rata-rata keseluruhan, adalah :

$$\bar{\mu}_Y = \frac{1}{g}\sum_{i=1}^{g}\mu_{iY} = \frac{1}{g}\sum_{i=1}^{g}\ell'\mu_i = \ell'\left(\frac{1}{g}\sum_{i=1}^{g}\mu_i\right) = \ell'\bar{\mu}$$

dan membentuk rasio :

$$\frac{\begin{pmatrix} jumlah\ kuadrat\ jarak\ dari \\ populasi\ terhadap\ rata-rata \\ keseluruhan\ dari\ Y \end{pmatrix}}{variansi\ Y} = \frac{\sum_{i=1}^{g}(\mu_{iY} - \bar{\mu}_Y)^2}{\sigma_Y^2} = \frac{\sum_{i=1}^{g}(\ell'\mu_i - \ell'\bar{\mu})^2}{\ell'\Sigma\ell}$$

$$= \frac{\ell'\left(\sum_{i=1}^{g}(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})'\right)\ell}{\ell'\Sigma\ell}$$

atau

$$\frac{\sum_{i=1}^{g}(\mu_{iY} - \bar{\mu}_Y)^2}{\sigma_Y^2} = \frac{\ell' B_0 \ell}{\ell'\Sigma\ell} \qquad (11-59)$$

Rasio (11-59) mengukur variabilitas antar grup dari nilai Y relative terhadap variabilitas didalam grup. Kita dapat memilih ℓ untuk memaksimalkan rasio.

Biasanya $\sum$ dan μ_i tidak tersedia, tapi kita mempunyai *training set* yang membuat observasi yang telah diklasidikasikan dengan benar. Misal *training set* memuat sampel acak berukuran n_i dari populasi π_i, i = 1, 2, …, g. Misal dari p x n_i himpunan data, dari populasi π_i, dengan X_i dan kolom ke-j nya dinotasikan sebagai x_{ij}. Setelah mengkonstruksi vector rata-rata sampel :

$$\bar{x}_i = \frac{1}{n_i}\sum_{i=1}^{n_i} x_{ij}$$

dan matriks kovarian S_i, i = 1, 2, …, g, kita mendefinisikan vektor rata-rata keseluruhan sebagai :

$$\bar{x} = \frac{\sum_{i=1}^{g} n_i\,\bar{x}_i}{\sum_{i=1}^{g} n_i} = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{g} n_i}$$

yang merupakan vektor rata-rata p x i yang diambil dari semua sampel observasi didalam training set.

Bersesuaian dengan matriks B_0 (11-58), kita mendefinisikan matriks *sample between groups*

$$\hat{B}_0 = \sum_{i=1}^{g}(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \qquad (11-60)$$

Demikian juga, penaksiran $\sum$ berdasarkan pada matriks samel didalam grup.

$$W = \sum_{i=1}^{g}(n_i - 1)S_i = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_i\right)\left(x_{ij} - \bar{x}_i\right)' \qquad (11.61)$$

Akibatnya, $W/\left(n_1 + n_2 + \cdots + n_g - g\right) = S_{pooled}$ adalah estimasi dari $\boldsymbol{\Sigma}$.

Diskriminan Linear Sampel Fisher

Misalkan $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_s > 0$ menunjukkan $s \le min(g - 1., p)$ buah nilai eigen tak nol dari $W^{-1}\hat{B}_0$ dan $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_s$ adalah nilai vektor eigen yang berkorespondensi (sehingga $\hat{e}'S_{pooled}\hat{e} = 1$). Maka vektor koefisien $\hat{\ell}$ yang

memaksimalkan rasio

$$\frac{\hat{\ell}'\hat{\mathrm{B}}_0\hat{\ell}}{\hat{\ell}'\hat{\mathrm{W}}\hat{\ell}} = \frac{\hat{\ell}'\left(\sum_{i=1}^{g}(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'\right)\ell}{\hat{\ell}'\left[\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)'\right]\hat{\ell}}$$

Diberikan oleh $\hat{\ell}_1 = \hat{e}_1$. Kombinasi linear $\hat{\ell}_1'x = \hat{e}_1'x$ disebut diskriminan pertama sampel. Sehingga $\hat{\ell}_k'x = \hat{e}_k'x$ adalah diskriminan ke-*k* dari sampel (diskriminan sampel yang ke-*k*), $k \leq s$.

2.6.1 Penggunaan Diskriminan Fisher untuk Klasifikasi

Diskriminan Fisher dipakai untuk mendapatkan representasi data dalam dimensi yang lebih rendah, yang memisahkan populasi sebanyak mungkin. Diskriminan juga merupakan dasar dari aturan klasifikasi.

Misalkan diberikan $Y_k = \ell_k'X = diskriminan\ ke-k, \quad k \leq s$ (11-64)

kita simpulkan bahwa

$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_s \end{bmatrix}$ mempunyai vector rata-rata $\mu_{iY} = \begin{bmatrix} \mu_{iY1} \\ \vdots \\ \mu_{iYs} \end{bmatrix} = \begin{bmatrix} \ell_1'\mu_i \\ \vdots \\ \ell_s'\mu_i \end{bmatrix}$

dibawah populasi π_i dan matriks kovarian **I** untuk semua populasi.

Jika jumlah kuadrat dari Y = *y* ke μ_{iY} adalah

$$(y - \mu_{iY})'(y - \mu_{iY}) = \sum_{j=1}^{s}(y_j - \mu_{iYj})^2$$

maka atura klasifikasi yang layak adalah menempatkan *y* ke populasi π_k jika jarak kuadrat dari *y* ke □□□ lebih kecil daripada jarak kuadrat dari *y* ke □□□ untuk □ ≠ □.

Jika hanya *r* buah diskriminan yang digunakan untuk pengalokasian, maka aturannya menjadi :

Alokasikan *x* ke □□ jika

$$\sum_{□=1}^{□}(□□ - □□□□)^2 = \sum_{□=1}^{□}\left[\ell_□'(□ - □□)\right]^2$$

$$\leq \sum_{□=1}^{□}\left[\ell_□'(□ - □□)\right]^2 \qquad (11-65)$$

untuk setiap $\square \neq \square$.

Result 11.6

Misalkan $\square_{\square} = \ell_{\square}'\square$ dimana $\ell_{\square} = \Sigma^{-1/2}e_j$ dan e_j adalah sebuah vector eigen dari $\Sigma^{-1/2}\square_0\Sigma^{-1/2}$. Maka

$$\sum_{\square=1}^{\square}(\square_{\square} - \square_{\square\square\square})^2 = \sum_{\square=1}^{\square}\left[\ell_{\square}'(\square - \square_{\square})\right]^2 = (\square - \square_{\square})'\Sigma^{-1}(\square - \square_{\square})$$

$$= -2\square_{\square}(\square) + \square'\Sigma^{-1}\square + 2\square\square\ \square_{\square}$$

Jika $\square_1 \geq \cdots \geq \square_{\square} > 0 = \square_{\square-1} = \cdots = \square_{\square}$, $\sum_{\square=\square+1}^{\square}(\square_{\square} - \square_{\square\square\square})^2$ bernilai konstan untuk semua populasi untuk i = 1, 2, . . ., g, jadi hanya *s* diskriminan y_j pertama, atau $\sum_{\square=1}^{\square}(\square_{\square} - \square_{\square\square\square})^2$, i = 1, 2, . . ., g yang memberikan konstribusi pada klasifikasi.

Sekarang akan ditunjukkan aturan klasifikasi berdasarkan $r \leq s$ diskriminan sampel yang pertama.

Prosedur Klasifikasi Fisher Berdasarkan Diskriminan Sampel

Alikasikan x pada $\square_{\square}$ jika

$$\sum_{\square=1}^{\square}(\acute{\square}_{\square} - \bar{\square}_{\square\square})^2 = \sum_{\square=1}^{\square}\left[\hat{\ell}_{\square}(\square - \bar{\square}_{\square})\right]^2 \leq \sum_{\square=1}^{\square}\left[\hat{\ell}_{\square}(\square - \bar{\square}_{\square})\right]^2 \qquad (11-67)$$

untuk $\square \neq \square$

dimana $\hat{\ell}_{\square}$ didefinisikan dalam (11-62) dan $r \leq s$.

BAB III
KESIMPULAN DAN SARAN

3.1 Kesimpulan

1. AER mengindikasikan fungsi klasifikasi sampel yang akan datang. Nilai error optimum tergantung pada fungsi densitas yang tidak diketahui $f_1(x)$ dan $f_2(x)$.
2. Metode Fisher mengubah bentuk pengamatan-pengamatan multivariat x ke pengamatan-pengamatan univariat y bahwa y berasal dari populasi-populasi π_1 dan π_2 yang telah dipisahkan.
3. Satu aturan pengelompokan yang didasarkan pada Fungsi Diskriminan Fisher. Kelompokan x_o ke π_1 jika $y_0 - \hat{m} \geq 0$ atau kelompokan x_o ke π_2 jika $y_0 - \hat{m} < 0$.
4. Untuk menglaokasikan nilai x ke $\pi_1, \pi_2, ...,$ atau π_k dapat kita gunakan aturan ECM minimum dengan nilai misklasifikasi sama sesuai dengan persaman (11-40) dan (11-41) atau dengan menggunakan probabilitas posteriornya.
5. Kita juga dapat menggunakan persamaan (11-45) untuk kovarian berbeda, namun untuk nilai estimasinya kita bisa menggunakan persamaan (11-48), tetapi untuk kovarian sama, kita dapat menggunakan persamaan (11-52).
6. Untuk menghitung skor diskriminan diantara beberapa populasi, kita dapat menggunakan Diskriminan Linear Sampel Fisher dengan memaksimalkan rasio pada (11-62).
7. Untuk pengklasifikasiannya, kita dapat menggunakan Metode Fisher yang sesuai dengan persamaan (11-67).

3.2 Saran

1. Untuk kasus populasi yang jumlahnya > 4, sebaiknya dalam menyelesaikan pengklasifikasian menggunakan suatu program computer. Dengan begitu, waktu, tenaga, dan materi termanfaatkan dengan efektif.
2. Selain menggunakan referensi dari makalah ini, pembaca dapat memanfaatkan referensi lain agar dapat melengkapi informasi yang belum tersedia di makalah ini.