Diskriminantna analiza

Prof.dr.sc. Bojana Dalbelo Bašić

Fakultet elektrotehnike i računarstva

Uvod

Primjer:

Kako odrediti sigurnu vrstu anestetika za pacijenta koji mora na operaciju ?

O pacijentu su poznati određeni podaci, npr.: dob, spol, krvni tlak, težina, krvna slika, alergijska reakcija, itd.

Temeljem ovakvog znanja anesteziolog mora odrediti da li i koji anestetik smije dati pacijentu.

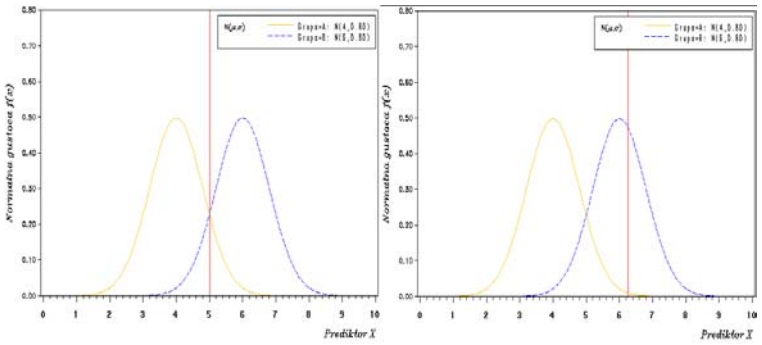
Temeljem ovakvog znanja anesteziolog želi znati:

- Može li se oblikovati pravilo koje bi nove pacijente ispravno klasificiralo u dvije grupe: grupu onih kojima se smije dati anestetik, te grupu onih kojima se ne smije dati anestetik ?
- Ako se može oblikovati pravilo, koje je to pravilo ?
- Kolika je vjerojatnost krive klasifikacije?
- Koje su posljedice krive klasifikacije?
- Što najviše utječe na klasifikaciju?

Cilj diskriminativne analize

Naći funkcije prediktorskih varijabli koje maksimalno separiraju grupe opservacija

Klasifikacija novih opservacija u grupe

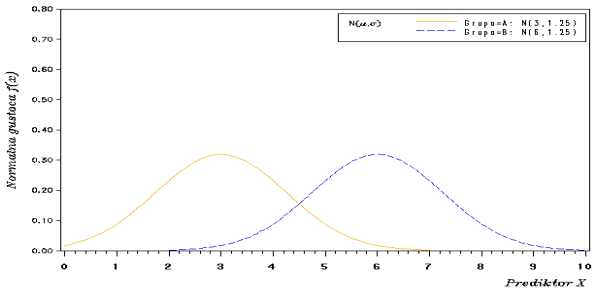


Osnovni pojmovi

- opservacija: objekt iz populacije (grupe)
- **prediktorska varijabla** (nezavisna, prediktor):
  - kvantitativna, kontinuirana varijabla
  - njene su vrijednosti dobivene mjerenjem objekata
  - temeljem vrijednosti vrši se diskriminacija
- **kriterijska varijabla** (zavisna varijabla, kriterij, varijabla grupe):
  - kvalitativna, diskretna varijabla
  - određuje pripadnost objekta grupi
- separacija
- klasifikacija

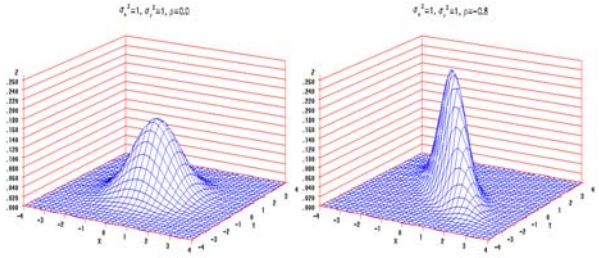
Potrebni preduvjeti

1. Prediktorske varijable moraju biti multivarijatno normalno distribuirane u populacijama Π_i definiranim varijablom grupe i=1,2



Potrebni preduvjeti

2. Martice varijanci-kovarijanci moraju biti homogene u populacijama Π_i definiranim varijablom grupe.



- Ako su uvjeti ispunjeni primjenjuje se **linearna diskriminativna funkcija**
- Ako uvjeti nisu ispunjeni primjenjuje se kvadratna ili neparametarska diskriminativna funkcija

Pravila za diskriminaciju grupa

- Pravilo vjerodostojnosti
  - funkcija gustoće
- Pravilo linearne diskriminativne funkcije
  - Fisherovo pravilo
- Pravilo Mahalanobisove udaljenosti
  - opisuje kvadratnu udaljenost opservacije od pripadne sredine grupe

Broj diskriminativnih funkcija

- broj diskriminativnih funkcija za separaciju g grupa, $g \geq 2$, na osnovu p prediktorskih varijabli, jednak je:

$$\min(g\text{-}1, p)$$

Iris data set min(3-1,4) = 2

Temeljna ideja

Maksimizirati

(SS između grupa)/(SS unutar grupa)

Pravilo linearne diskriminativne funkcije

- Osnovni Fisherov cilj je separacija grupa (primjer dvije populacije) hiperravnina

$$Y = \mathbf{b}^T\boldsymbol{X} = \boldsymbol{b}_1\mathbf{x}_1 + \boldsymbol{b}_2\mathbf{x}_2 + \ldots + \boldsymbol{b}_p\mathbf{x}_p = (\mu^{(1)} - \mu^{(2)})^T \Sigma^{-1}\boldsymbol{X}$$

  Cilj je linearne funkcije Y maksimizirati (kvadriranu) udaljenost $\mu_Y^{(1)}$ i $\mu_Y^{(2)}$ u odnosu na varijabilitet od Y.

- klasifikacija

$$\mathbf{b}^T\mathbf{x} - k > 0$$

  uz $k = 1/2\ (\mu^{(1)} - \mu^{(2)})^T\Sigma^{-1}(\mu^{(1)} - \mu^{(2)})$

Testiranje pouzdanosti klasifikacije

Analiza pogrešaka klasifikacije po tipu i broju, uz vrednovanje različitog značenja pogrešaka

Odabir značajnih prediktora

Pitanje:

- da li su za efikasnu diskriminaciju potrebni svi prediktori
- koji su prediktori najbolji

Testira se statistička značajnost pojedine prediktorske varijable (statistički testovi);

metode:

- selekcija unaprijed
- selekcija unatrag
- stepwise selekcija

Primjer (Iris dataset)

- Kriterijska varijabla (3 grupe)
  - Iris Setosa, Iris Versicolor, Iris Virginica



- Prediktorske varijable (4)
  - SepalLength, SepalWidth, PetalLength, PetalWidth

Test (Kolmogorovljev D) normaliteta prediktora po grupama irisa

Prediktorske varijable	Vrsta irisa (grupa)		
	Setosa	Versicolor	Virginica
petal length	< 0.01	0.08	0.11
petal width	< 0.01	< 0.01	0.07
sepal length	0.10	> 0.15	0.09
sepal width	> 0.15	0.07	0.04

Uz pretpostavku da su uvjeti diskriminativne analize zadovoljeni koeficijenti za formiranje linearnih funkcija bili bi:

$\mathbf{b}_1$ = [-0.083, -0.15, 0.22, 0.28]

$\mathbf{b}_2$ = [-0.002, 0.21, -0.09, 0.28]

Obje funkcije su statistički značajne pri čemu je prva statistički značajnija od druge

(tj. prva funkcija jače diskriminira grupe od druge)

Koeficijenti za formiranje kanoničkih funkcija

Raw Canonical Coefficients		
Variable	Can1	Can2
SepalLength	-.082	0.002
SepalWidth	-0.15	0.21
PetalLength	0.22	-0.09
PetalWidth	0.28	0.28

Provjera točnosti

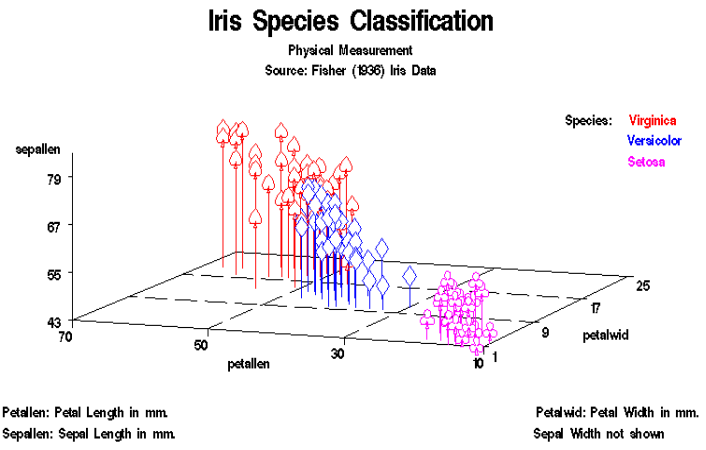
Rezultati pokazuju da je krivo klasificirano 3% primjeraka (metoda krosvalidacije) odnosno 2% primjeraka (metoda resupstitucije)

Odabir značajnih prediktora za diskriminaciju

– pokazalo se da su prediktori PetalLength, SepalWidth, PetalWidth, statistički značajni dok je SepalLength na granici značajnosti

Rezime rezultata metode u koracima (stepwise)

Stepwise Selection Summary								
Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda
1	1	PetalLength		0.94	1180.16	<.0001	0.05	<.0001
2	2	SepalWidth		0.37	43.04	<.0001	0.03	<.0001
3	3	PetalWidth		0.32	34.57	<.0001	0.02	<.0001
4	4	SepalLength		0.06	4.72	0.0103	0.02	<.0001



Literatura

Sharma, S. (1996), "Applied Multivariate Techniques", John Wiley & Sons, Inc., pp. 237 – 316.

Johnson, R. A., Wichern, D. W. (1988), "Applied Multivariate Statistical Analysis", Prentice Hall, International Edition, pp. 470 – 542.