Capítulo 12

Análise discriminante de Mahalanobis

12.1. Sobre análise discriminante

O objetivo principal da análise discriminante é discriminar indivíduos alocados em grupos previamente estabelecidos. Perceba então que, diferentemente da análise de agrupamento, neste caso conhecemos os grupos, ao menos inicialmente.

A análise discriminante pode ser utilizada para classificação de novos indivíduos, isto é, observações que não foram utilizadas na construção da regra de discriminação ou classificação.

Existem alguns métodos para a construção da regra de discriminação, tais como as funções discriminantes canônicas, como visto no capítulo sobre MANOVA. Neste capítulo, utilizaremos um método bastante simples de discriminação, baseado na distância de Mahalanobis, apresentada no capítulo sobre análise de agrupamento.

12.2. Utilizando a distância de Mahalanobis para discriminação

A nossa abordagem para discriminação de grupos aqui consiste em calcular as distâncias quadradas generalizadas de Mahalanobis de cada observação multivariada ao centro de cada um dos grupos. Considere então a i-ésima $(i = 1, 2, ..., n_j)$ observação p-variada pertencente ao j-ésimo $(j = 1, 2, ..., k)$ grupo, $\mathbf{x}_{ij}$. Seja $\bar{\mathbf{x}}_{j'}$ o vetor de médias do j'-ésimo $(j' = 1, 2, ..., k)$ grupo. A distância de Mahalanobis dessa observação ao centro desse grupo é dada por:

$$D^2_{ij,j'} = (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{j'})^T \hat{\mathbf{\Sigma}}_c^{-1} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{j'})$$

em que $\hat{\mathbf{\Sigma}}_c^{-1}$ é a estimativa da matriz de covariâncias combinadas para grupos.

Agora considere C_j a variável aleatória que representa o grupo ou classe ao qual está alocada a obsevação $\mathbf{x}_{ij}$. A classe predita pelo método $\hat{C}_{j'}$ para essa observação é tal que

$$j' \Rightarrow \min_{j'=1}^{k} (D^2_{ij,j'})$$

isto é, o indivíduo é alocado ao grupo cuja distância do seu centro é a menor.

12.3. Exemplo

Considere novamente os dados 'iris', que contêm cinquenta observações tomadas em quatro variáveis para cada um dos três grupos (fator *Species*). Uma análise discriminante com base nas distâncias quadradas generalizadas de Mahalanobis pode ser realizada usando a função `D2.disc()` do pacote **biotools** (Silva, 2016), da seguinte forma:

```
R> library(biotools)
R> adm <- D2.disc(data= iris[,-5], grouping= iris[,5])
R> adm

Call:
D2.disc.default(data= iris[, -5], grouping= iris[,5])

Mahalanobis distances from each class and class
prediction (first 6 rows):
     setosa versicolor virginica grouping   pred
1 0.2910898   98.88475  191.7886   setosa setosa
2 2.0313451   80.97126  169.1870   setosa setosa
3 0.5532814   87.28938  177.0701   setosa setosa
4 2.0866979   75.29369  160.7244   setosa setosa
5 0.5956300  100.92317  193.8540   setosa setosa
6 1.9447533   95.93997  183.1140   setosa setosa
  misclass
1
2
3
4
5
6

Class means:
           Sepal.Length Sepal.Width Petal.Length
setosa            5.006       3.428        1.462
versicolor        5.936       2.770        4.260
virginica         6.588       2.974        5.552
           Petal.Width
setosa           0.246
versicolor       1.326
virginica        2.026
```

```
Confusion matrix:
           new setosa new versicolor new virginica
setosa             50              0             0
versicolor          0             48             2
virginica           0              1            49
```

O *output* traz as distâncias (`D2`) das observações[1] ao centro de cada grupo (`means`), a matriz de covariâncias combinadas (`pooled`) e a matriz de confusão (`confusion.matrix`) contendo o número de classificações corretas na diagonal e incorretas fora dela. No objeto (`D2`) encontramos uma coluna (`misclass`) que indica (com um asterisco) onde houve divergência entre a classificação original (`grouping`) e a predita pelo método (`pred`).

Se analisarmos o objeto `D2` (digite `adm$D2`), verificaremos então que a primeira observação apresentará a menor distância de Mahalanobis ao centro do seu próprio grupo (*setosa*), sendo assim corretamente classificada. Por outro lado, a observação encontrada na linha 71 apresenta distância para o centro do grupo *virginica* ainda menor em relação ao seu próprio grupo (*versicolor*). Assim, essa observação foi realocada ao grupo *virginica* (veja a coluna `pred`). O mesmo ocorre com a observação 84. Com a observação 134 uma reclassificação também foi feita.

Na matriz de confusão, observamos que, dos 50 indivíduos pertencentes à classe *setosa*, todos foram corretamente classificados. Porém, na classe *versicolor*, dos 50 indivíduos, 2 deles (observações 71 e 84) foram realocados à classe *virginica* e os 48 restantes foram corretamente classificados. A classe *virginica* também teve um dos seus indivíduos realocado (observação 134).

Finalmente, os grupos finais obtidos são: *setosa*, permanecendo inalterado; *versicolor*, agora com 49 indivíduos; e *virginica*, agora com 51 indivíduos.

[1] Um *print method* foi criado para que o output fosse resumido às seis primeiras observações multivariadas, evitando assim excesso de informação ao executar a função com data frames extensos.

12.4. Exercício

Utilize os resultados obtidos no exercício do capítulo sobre MANOVA com os dados 'Wolves', disponíveis no pacote **candisc** (Friendly & Fox, 2013), para realizar uma análise discriminante com base na distância de Mahalanobis. Para tal, utilize a matriz de covariâncias residuais. Verifique quantas reclassificações foram feitas.

12.5. Bibliografia

Friendly, M.; Fox, J. (2013) **candisc**: Visualizing generalized canonical discriminant and canonical correlation analysis. R package version 0.6-5. Disponível em:
`http://CRAN.R-project.org/package=candisc`

Mahalanobis, P. C. (1936) On the generalized distance in statistics. **Proceedings of The National Institute of Sciences of India**, 12:49-55.

Manly, B. F. J. (2005) **Multivariate statistical methods**: a primer. 3. ed. Boca Raton, FL: Chapman and Hall/CRC Press.

Silva, A. R. (2016) **biotools**: Tools for biometry and applied statistics in agricultural science. R package version 3.0. Disponível em
`http://CRAN.R-project.org/package=biotools`.