Minería de datos

Trabajo análisis discriminante de bd iris

Kevin Heberth Haquehua Apaza

30 de julio del 2025

Table of Contents

[Análisis de base de datos iris mediante análisis discriminante 1](#_Toc206688311)

[Con el análisis realizado en clases se observo algunos resultados que sugerían realizar una transformación o buscar alguna otra solución. Desarrolle la solución e interprete los resultados 1](#_Toc206688312)

[Solución 4](#_Toc206688313)

# Análisis de base de datos iris mediante análisis discriminante

## Con el análisis realizado en clases se observo algunos resultados que sugerían realizar una transformación o buscar alguna otra solución. Desarrolle la solución e interprete los resultados

Primeramente veamos el resultado original

Librerias a utilizar

Cargamos los datos

data("iris")

Realizamos la partición de la data

#dividir la data  
set.seed(12345)  
muestra = createDataPartition(iris$Species, p =0.8, list =F)  
train = iris[muestra,]  
test = iris[-muestra,]

Ahora ejecutemos el modelo lineal discriminante

discrim\_l = lda(Species ~ Sepal.Length + Sepal.Width +Petal.Length + Petal.Width,  
 data =train)  
discrim\_l

## Call:  
## lda(Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width,   
## data = train)  
##   
## Prior probabilities of groups:  
## setosa versicolor virginica   
## 0.3333333 0.3333333 0.3333333   
##   
## Group means:  
## Sepal.Length Sepal.Width Petal.Length Petal.Width  
## setosa 4.9925 3.4050 1.4775 0.245  
## versicolor 5.9675 2.7625 4.2575 1.345  
## virginica 6.6000 3.0050 5.5700 2.060  
##   
## Coefficients of linear discriminants:  
## LD1 LD2  
## Sepal.Length 0.8090913 0.2925911  
## Sepal.Width 1.9393787 -2.4494212  
## Petal.Length -2.2164336 0.6391380  
## Petal.Width -3.1630675 -2.4130658  
##   
## Proportion of trace:  
## LD1 LD2   
## 0.9916 0.0084

Evaluemos y veamos la matriz de confusión

#evaluacion  
prediccion = predict(discrim\_l, test)  
prediccion$class

## [1] setosa setosa setosa setosa setosa setosa   
## [7] setosa setosa setosa setosa versicolor versicolor  
## [13] versicolor versicolor versicolor versicolor versicolor versicolor  
## [19] versicolor versicolor virginica virginica virginica virginica   
## [25] virginica virginica virginica virginica virginica virginica   
## Levels: setosa versicolor virginica

prediccion$posterior

## setosa versicolor virginica  
## 5 1.000000e+00 2.553396e-26 2.395489e-49  
## 7 1.000000e+00 9.349809e-22 5.479903e-43  
## 15 1.000000e+00 7.077285e-35 8.290507e-61  
## 18 1.000000e+00 2.013902e-24 1.093475e-46  
## 21 1.000000e+00 8.818530e-23 6.234255e-45  
## 23 1.000000e+00 6.017800e-29 8.656654e-53  
## 27 1.000000e+00 7.660503e-20 2.775055e-40  
## 34 1.000000e+00 9.239481e-34 6.757276e-59  
## 42 1.000000e+00 6.580155e-12 5.591279e-31  
## 47 1.000000e+00 1.423772e-26 1.847393e-49  
## 51 1.705622e-20 9.999837e-01 1.628641e-05  
## 52 7.795853e-22 9.998520e-01 1.480127e-04  
## 56 1.754090e-25 9.994334e-01 5.666046e-04  
## 58 3.388938e-16 1.000000e+00 2.412084e-08  
## 68 3.033131e-18 9.999998e-01 1.594068e-07  
## 70 3.994890e-20 9.999991e-01 8.638759e-07  
## 74 7.933651e-25 9.998675e-01 1.325017e-04  
## 85 3.556083e-27 9.830436e-01 1.695644e-02  
## 91 8.574268e-26 9.997294e-01 2.705795e-04  
## 100 1.743526e-21 9.999807e-01 1.925693e-05  
## 111 1.105697e-35 1.685661e-02 9.831434e-01  
## 113 5.855433e-44 1.303059e-04 9.998697e-01  
## 119 3.309200e-67 1.459771e-10 1.000000e+00  
## 124 7.464250e-36 8.877668e-02 9.112233e-01  
## 127 6.998661e-34 2.034541e-01 7.965459e-01  
## 130 4.002543e-36 1.433569e-01 8.566431e-01  
## 133 1.100827e-51 8.955581e-07 9.999991e-01  
## 135 1.335121e-39 4.990247e-02 9.500975e-01  
## 138 7.892275e-39 6.398729e-03 9.936013e-01  
## 143 4.514135e-43 5.304245e-04 9.994696e-01

prediccion$x

## LD1 LD2  
## 5 8.8267271 -0.63889692  
## 7 7.7989081 -0.50735568  
## 15 10.6930384 -1.51242015  
## 18 8.3973917 -0.60600227  
## 21 8.0975579 0.15976515  
## 23 9.3896640 -1.01158854  
## 27 7.3629512 -0.50379824  
## 34 10.3949000 -1.96225412  
## 42 5.8063258 2.09383477  
## 47 8.8522253 -0.97169446  
## 51 -1.4407535 0.13953005  
## 52 -1.7992283 -0.40515877  
## 56 -2.5087302 0.85240913  
## 58 -0.3231142 1.55505895  
## 68 -0.7862653 1.59487492  
## 70 -1.2089793 1.65710678  
## 74 -2.3120736 1.33857974  
## 85 -2.9961954 -0.20786560  
## 91 -2.5204741 1.46116795  
## 100 -1.6221568 0.59675395  
## 111 -4.6297131 -1.19894979  
## 113 -5.9777416 -0.60693961  
## 119 -9.7609314 1.04834084  
## 124 -4.6853205 0.32202818  
## 127 -4.3506484 -0.01608684  
## 130 -4.7375014 0.90837111  
## 133 -7.2272040 -0.41148458  
## 135 -5.3273531 1.92107499  
## 138 -5.1585200 -0.24499842  
## 143 -5.8494596 0.06225366

#matriz de confusion  
confusionMatrix(test$Species, prediccion$class)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction setosa versicolor virginica  
## setosa 10 0 0  
## versicolor 0 10 0  
## virginica 0 0 10  
##   
## Overall Statistics  
##   
## Accuracy : 1   
## 95% CI : (0.8843, 1)  
## No Information Rate : 0.3333   
## P-Value [Acc > NIR] : 4.857e-15   
##   
## Kappa : 1   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: setosa Class: versicolor Class: virginica  
## Sensitivity 1.0000 1.0000 1.0000  
## Specificity 1.0000 1.0000 1.0000  
## Pos Pred Value 1.0000 1.0000 1.0000  
## Neg Pred Value 1.0000 1.0000 1.0000  
## Prevalence 0.3333 0.3333 0.3333  
## Detection Rate 0.3333 0.3333 0.3333  
## Detection Prevalence 0.3333 0.3333 0.3333  
## Balanced Accuracy 1.0000 1.0000 1.0000

### Solución

El caso es que las variables o clases estan bien separadas y debido a las pocas observaciones 120, puede ser el caso de necesitar una tranformación de datos (normal o log) o tambien realizar una validacion cruzada

#### Tranformación de datos

**Por la normal**

#dividir la data  
iris\_tran\_norm <- scale(iris[,1:4])  
iris\_st <- data.frame(cbind(iris\_tran\_norm, iris[,5]))  
muestra = createDataPartition(iris\_st$V5, p =0.8, list =F)  
train = iris[muestra,]  
test = iris[-muestra,]

Ahora ejecutemos el modelo lineal discriminante

discrim\_l = lda(Species ~ Sepal.Length + Sepal.Width +Petal.Length + Petal.Width,  
 data =train)  
discrim\_l

## Call:  
## lda(Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width,   
## data = train)  
##   
## Prior probabilities of groups:  
## setosa versicolor virginica   
## 0.3416667 0.3250000 0.3333333   
##   
## Group means:  
## Sepal.Length Sepal.Width Petal.Length Petal.Width  
## setosa 4.982927 3.382927 1.426829 0.2317073  
## versicolor 5.928205 2.756410 4.276923 1.3230769  
## virginica 6.632500 2.987500 5.600000 2.0675000  
##   
## Coefficients of linear discriminants:  
## LD1 LD2  
## Sepal.Length 0.8894355 -0.5731198  
## Sepal.Width 1.4058731 -1.6483956  
## Petal.Length -2.2963444 1.5021688  
## Petal.Width -2.7175171 -3.4276176  
##   
## Proportion of trace:  
## LD1 LD2   
## 0.9903 0.0097

Evaluemos y veamos la matriz de confusión

#evaluacion  
prediccion = predict(discrim\_l, test)  
prediccion$class

## [1] setosa setosa setosa setosa setosa setosa   
## [7] setosa setosa setosa versicolor versicolor versicolor  
## [13] versicolor versicolor versicolor versicolor versicolor versicolor  
## [19] versicolor versicolor virginica virginica virginica virginica   
## [25] virginica versicolor virginica virginica virginica virginica   
## Levels: setosa versicolor virginica

prediccion$posterior

## setosa versicolor virginica  
## 6 1.000000e+00 2.097159e-21 3.466670e-41  
## 7 1.000000e+00 5.374213e-19 1.176329e-38  
## 16 1.000000e+00 6.507967e-28 1.739622e-49  
## 19 1.000000e+00 6.625233e-23 1.356711e-43  
## 25 1.000000e+00 1.165094e-15 1.201821e-34  
## 27 1.000000e+00 1.462511e-17 2.158562e-36  
## 40 1.000000e+00 6.360550e-21 1.425099e-41  
## 45 1.000000e+00 1.591946e-17 3.925905e-36  
## 48 1.000000e+00 1.176630e-18 1.080951e-38  
## 52 5.780022e-20 9.995983e-01 4.016895e-04  
## 60 6.827737e-21 9.998176e-01 1.823868e-04  
## 64 3.470490e-24 9.974901e-01 2.509901e-03  
## 65 5.749231e-14 9.999994e-01 6.310440e-07  
## 73 4.349109e-29 8.887511e-01 1.112489e-01  
## 75 5.997405e-18 9.999887e-01 1.126670e-05  
## 82 9.577287e-16 9.999999e-01 9.154007e-08  
## 87 1.773827e-21 9.989378e-01 1.062208e-03  
## 92 1.651432e-22 9.991603e-01 8.396657e-04  
## 98 9.744756e-19 9.999799e-01 2.005247e-05  
## 99 1.677313e-10 1.000000e+00 5.502406e-09  
## 112 5.361968e-38 2.017101e-03 9.979829e-01  
## 113 1.742297e-39 1.794787e-04 9.998205e-01  
## 125 2.965705e-40 8.876413e-05 9.999112e-01  
## 127 5.613504e-30 2.423146e-01 7.576854e-01  
## 129 1.576487e-44 1.326397e-05 9.999867e-01  
## 134 3.511930e-29 8.405248e-01 1.594752e-01  
## 135 3.649642e-36 1.477824e-01 8.522176e-01  
## 138 8.866543e-36 9.145141e-03 9.908549e-01  
## 140 8.197673e-37 7.165453e-04 9.992835e-01  
## 150 7.860080e-34 2.882195e-02 9.711781e-01

prediccion$x

## LD1 LD2  
## 6 7.6707503 -1.48738652  
## 7 7.2169204 -0.31258177  
## 16 9.0997864 -2.78395401  
## 19 8.0687453 -1.15172115  
## 25 6.5183870 0.66664042  
## 27 6.8416740 -0.58415768  
## 40 7.7037554 -0.10616303  
## 45 6.8040635 -0.85017727  
## 48 7.2074975 0.35985911  
## 52 -1.8429584 -0.47093614  
## 60 -1.9636592 0.48246590  
## 64 -2.7190682 0.83871398  
## 65 -0.3660554 -0.18435002  
## 73 -3.8345509 1.34112026  
## 75 -1.2619481 0.40867229  
## 82 -0.5723148 1.87566190  
## 87 -2.1759840 -0.17759877  
## 92 -2.3488464 0.52365755  
## 98 -1.4398352 0.52329625  
## 99 0.5481876 0.54579036  
## 112 -5.4699773 0.18394965  
## 113 -5.6952135 -0.92490670  
## 125 -5.8216639 -1.06167964  
## 127 -4.0873532 -0.27458858  
## 129 -6.5617967 -0.21576279  
## 134 -3.8720578 1.14703534  
## 135 -5.2075400 2.68518456  
## 138 -5.0951452 0.16778693  
## 140 -5.2360482 -1.29727512  
## 150 -4.7619125 0.01831887

#matriz de confusion  
confusionMatrix(test$Species, prediccion$class)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction setosa versicolor virginica  
## setosa 9 0 0  
## versicolor 0 11 0  
## virginica 0 1 9  
##   
## Overall Statistics  
##   
## Accuracy : 0.9667   
## 95% CI : (0.8278, 0.9992)  
## No Information Rate : 0.4   
## P-Value [Acc > NIR] : 5.303e-11   
##   
## Kappa : 0.9497   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: setosa Class: versicolor Class: virginica  
## Sensitivity 1.0 0.9167 1.0000  
## Specificity 1.0 1.0000 0.9524  
## Pos Pred Value 1.0 1.0000 0.9000  
## Neg Pred Value 1.0 0.9474 1.0000  
## Prevalence 0.3 0.4000 0.3000  
## Detection Rate 0.3 0.3667 0.3000  
## Detection Prevalence 0.3 0.3667 0.3333  
## Balanced Accuracy 1.0 0.9583 0.9762

De la misma forma manda una clasificación perfecta

**Por una tranformación log**

#dividir la data  
iris\_tran\_norm <- log(iris[,1:4])  
iris\_st <- data.frame(cbind(iris\_tran\_norm, iris[,5]))  
muestra = createDataPartition(iris\_st$iris...5., p =0.8, list =F)  
train = iris[muestra,]  
test = iris[-muestra,]

Ahora ejecutemos el modelo lineal discriminante

discrim\_l = lda(Species ~ Sepal.Length + Sepal.Width +Petal.Length + Petal.Width,  
 data =train)  
discrim\_l

## Call:  
## lda(Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width,   
## data = train)  
##   
## Prior probabilities of groups:  
## setosa versicolor virginica   
## 0.3333333 0.3333333 0.3333333   
##   
## Group means:  
## Sepal.Length Sepal.Width Petal.Length Petal.Width  
## setosa 5.0550 3.4575 1.485 0.2500  
## versicolor 5.9325 2.7550 4.225 1.3100  
## virginica 6.6425 2.9675 5.580 2.0125  
##   
## Coefficients of linear discriminants:  
## LD1 LD2  
## Sepal.Length 0.9460727 0.2486590  
## Sepal.Width 1.4378567 -2.4054054  
## Petal.Length -2.1536775 0.5667747  
## Petal.Width -2.9228834 -2.3718011  
##   
## Proportion of trace:  
## LD1 LD2   
## 0.991 0.009

Evaluemos y veamos la matriz de confusión

#evaluacion  
prediccion = predict(discrim\_l, test)  
prediccion$class

## [1] setosa setosa setosa setosa setosa setosa   
## [7] setosa setosa setosa setosa versicolor versicolor  
## [13] versicolor versicolor versicolor versicolor virginica versicolor  
## [19] versicolor versicolor virginica virginica virginica virginica   
## [25] virginica virginica virginica virginica virginica virginica   
## Levels: setosa versicolor virginica

prediccion$posterior

## setosa versicolor virginica  
## 3 1.000000e+00 2.381151e-18 1.427106e-37  
## 14 1.000000e+00 2.276706e-18 5.826960e-38  
## 23 1.000000e+00 2.568319e-23 7.119152e-44  
## 28 1.000000e+00 1.947829e-20 3.196435e-40  
## 29 1.000000e+00 1.201699e-20 1.181126e-40  
## 31 1.000000e+00 2.126094e-15 1.197358e-33  
## 38 1.000000e+00 5.870451e-22 2.619143e-42  
## 44 1.000000e+00 1.499597e-14 4.219721e-31  
## 46 1.000000e+00 2.012025e-15 1.325645e-33  
## 48 1.000000e+00 4.222626e-17 8.393340e-36  
## 55 3.598246e-22 9.956920e-01 4.308033e-03  
## 57 1.679528e-21 9.773464e-01 2.265364e-02  
## 63 2.090457e-17 9.999990e-01 1.041136e-06  
## 65 7.617010e-14 9.999974e-01 2.581030e-06  
## 67 2.902764e-23 9.573845e-01 4.261549e-02  
## 68 1.161688e-15 9.999987e-01 1.295455e-06  
## 84 3.451829e-31 9.201352e-02 9.079865e-01  
## 85 5.083644e-24 9.084536e-01 9.154645e-02  
## 86 3.671968e-20 9.878405e-01 1.215950e-02  
## 88 2.303788e-22 9.995640e-01 4.359835e-04  
## 103 4.331736e-41 2.530827e-05 9.999747e-01  
## 111 2.849617e-31 9.038320e-03 9.909617e-01  
## 114 6.021280e-40 1.094791e-04 9.998905e-01  
## 115 4.682142e-45 5.004826e-07 9.999995e-01  
## 122 6.896112e-37 3.689448e-04 9.996311e-01  
## 126 4.420358e-35 2.397433e-03 9.976026e-01  
## 129 8.242230e-43 8.907395e-06 9.999911e-01  
## 138 4.197174e-34 3.592734e-03 9.964073e-01  
## 144 1.533634e-44 6.985928e-07 9.999993e-01  
## 149 4.272782e-40 5.460057e-06 9.999945e-01

prediccion$x

## LD1 LD2  
## 3 7.2894036 0.32455916  
## 14 7.3464271 0.83000180  
## 23 8.4160423 -0.83250129  
## 28 7.7630615 -0.15937801  
## 29 7.8346436 0.02448506  
## 31 6.5941220 0.75999799  
## 38 8.1306814 -0.29401361  
## 44 6.1893259 -1.10115280  
## 46 6.5887835 0.65000349  
## 48 6.9794286 0.35637072  
## 55 -2.4896924 0.52132248  
## 57 -2.4676347 -0.91161466  
## 63 -1.0717946 2.68607195  
## 65 -0.4591180 -0.03542562  
## 67 -2.8382187 -0.24022919  
## 68 -0.7574485 1.49031492  
## 84 -4.4756415 0.68374072  
## 85 -3.0274333 -0.28996100  
## 86 -2.1769353 -1.34010784  
## 88 -2.3825231 2.03529865  
## 103 -6.1879882 -0.49683679  
## 111 -4.4528302 -1.34335290  
## 114 -6.0008203 0.08482617  
## 115 -6.8593771 -1.50397252  
## 122 -5.4487028 -0.71833881  
## 126 -5.1443124 -0.18483417  
## 129 -6.4917072 -0.35984944  
## 138 -4.9681174 -0.42660818  
## 144 -6.7688154 -1.52687580  
## 149 -5.9720489 -2.44053962

#matriz de confusion  
confusionMatrix(test$Species, prediccion$class)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction setosa versicolor virginica  
## setosa 10 0 0  
## versicolor 0 9 1  
## virginica 0 0 10  
##   
## Overall Statistics  
##   
## Accuracy : 0.9667   
## 95% CI : (0.8278, 0.9992)  
## No Information Rate : 0.3667   
## P-Value [Acc > NIR] : 4.476e-12   
##   
## Kappa : 0.95   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: setosa Class: versicolor Class: virginica  
## Sensitivity 1.0000 1.0000 0.9091  
## Specificity 1.0000 0.9524 1.0000  
## Pos Pred Value 1.0000 0.9000 1.0000  
## Neg Pred Value 1.0000 1.0000 0.9500  
## Prevalence 0.3333 0.3000 0.3667  
## Detection Rate 0.3333 0.3000 0.3333  
## Detection Prevalence 0.3333 0.3333 0.3333  
## Balanced Accuracy 1.0000 0.9762 0.9545

En este caso ya muestra un mejor accuracy, y la especificidad del modelo bajo, siendo un resultado más real

#### Realizando validación cruzada

Definir el control de validación cruzada de 10

ctrl <- trainControl(method = "cv", number = 10)

Ahora entremos los datos con validaciones cruzadas

set.seed(12345)  
muestra = createDataPartition(iris$Species, p =0.8, list =F)  
train = iris[muestra,]  
test = iris[-muestra,]

Ahora ejecutar el modelo lineal con validación cruzada

discrim\_l\_cv <- train(Species ~ Sepal.Length + Sepal.Width +Petal.Length + Petal.Width,   
 data = train,   
 method = "lda",   
 trControl = ctrl)

Veamos los resultados

discrim\_l\_cv

## Linear Discriminant Analysis   
##   
## 120 samples  
## 4 predictor  
## 3 classes: 'setosa', 'versicolor', 'virginica'   
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 108, 108, 108, 108, 108, 108, ...   
## Resampling results:  
##   
## Accuracy Kappa   
## 0.975 0.9625

Se observa un accuracy no exacto. Veamos ahora la predicción

#evaluacion  
prediccion = predict(discrim\_l\_cv, test)  
confusionMatrix(prediccion, test$Species)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction setosa versicolor virginica  
## setosa 10 0 0  
## versicolor 0 10 0  
## virginica 0 0 10  
##   
## Overall Statistics  
##   
## Accuracy : 1   
## 95% CI : (0.8843, 1)  
## No Information Rate : 0.3333   
## P-Value [Acc > NIR] : 4.857e-15   
##   
## Kappa : 1   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: setosa Class: versicolor Class: virginica  
## Sensitivity 1.0000 1.0000 1.0000  
## Specificity 1.0000 1.0000 1.0000  
## Pos Pred Value 1.0000 1.0000 1.0000  
## Neg Pred Value 1.0000 1.0000 1.0000  
## Prevalence 0.3333 0.3333 0.3333  
## Detection Rate 0.3333 0.3333 0.3333  
## Detection Prevalence 0.3333 0.3333 0.3333  
## Balanced Accuracy 1.0000 1.0000 1.0000

De la misma forma manda una clasificación perfecta a pesar de que el modelo una clasificación exacta con la data de prueba a pesar de que con la data de entrenamiento arroja un accuracy menor.

#### CONCLUSIÓN

La mejor forma de evitar los errores de ajuste y clasificación es con la tranformación