REPORTE FINAL APLICANDO UN ANALISIS DISCRIMINANTE LINEAL (LDA)

Arleth Michell Morales García

2022-06-05

INTRODUCCIÓN

Los halcones son aves de rapiña, por lo que son conocidos por sus increíbles habilidades de caza y por ser un depredador despiadado, dominante dentro de su entorno. Existen más de 40 especies diferentes de halcones, que se pueden encontrar en todo el mundo. En este estudio nos concentraremos en 3 tipos de especie.

OBJETIVO:

Determinar un modelo matemático que permita diferenciar a qué tipo de las 3 especies pertenece un Halcón.

DESCRIPCIÓN DE LA MATRIZ DE DATOS

Elegí una base de datos en el repositorio de *KEY2STATS* acerca de 3 especies de Halcones con 72 observaciones y 20 variables. Pero eliminé 6 columnas de variables que no eran necesarias.

DESCRIPCIÓN DE LAS VARIABLES UTILIZADAS

Species:

CH = Cooper

RT = Cola roja

SS = Espinillas afiladas

Wing: Longitud (en mm) de la pluma principal del ala desde la punta hasta la muñeca a la que se une

Weight: Peso corporal (en g)

Culmen: Longitud (en mm) del pico superior desde la punta hasta donde choca con la parte carnosa del ave

Hallux: Longitud (en mm) de la garra asesina

Tail: Medida (en mm) relacionada con la longitud de la cola (inventada en el MacBride Raptor Center)

StandardTail: Medida estándar de la longitud de la cola (en mm)

Tarsus: Longitud del hueso básico del pie (en mm)

ELECCIÓN DE LA MATRIZ DE DATOS

```
library(readr)
H <- read_csv("Halcones.csv")
Ha<- as.data.frame(H)
```

Covertí la variable Species en factor.

EXPLORACIÓN DE LA MATRIZ

Dimensión

```
dim(Ha)
```

```
## [1] 72 14
```

Nombre de las variables

```
colnames(Ha)
```

```
##  [1] "Month"        "Day"          "Year"         "Species"      "Wing"
##  [6] "Weight"       "Culmen"       "Hallux"       "Tail"         "StandardTail"
## [11] "Tarsus"       "WingPitFat"   "KeelFat"      "Crop"
```

Estructura

```
str(Ha)
```

```
## 'data.frame':    72 obs. of  14 variables:
##  $ Month       : num  9 9 9 9 9 9 9 9 9 9 ...
##  $ Day         : num  9 9 10 10 11 12 14 14 17 17 ...
##  $ Year        : num  1997 1997 1997 1997 1997 ...
##  $ Species     : chr  "RT" "SS" "SS" "RT" ...
##  $ Wing        : num  392 191 161 365 156 191 198 160 164 352 ...
##  $ Weight      : num  1142 157 98 813 94 ...
##  $ Culmen      : num  27.2 12 10.3 26.2 9.9 16.5 12.5 9.8 10.1 26.6 ...
##  $ Hallux      : num  33 13.8 11.7 30.1 11.4 14.4 14.1 11 11.4 30.3 ...
##  $ Tail        : num  235 158 133 221 125 158 159 132 136 216 ...
##  $ StandardTail: num  244 162 137 224 129 161 164 135 138 225 ...
##  $ Tarsus      : num  89.3 57 48.4 86.6 48.6 54.3 56 48.3 50.5 90.2 ...
##  $ WingPitFat  : num  0 2 1 0 1 0 2 2 2 0 ...
##  $ KeelFat     : num  1 2 2 0 2 1 2 1 1 0 ...
##  $ Crop        : num  0 0 0.25 0 0.25 0 0 0 0 0 ...
```

Detección de valores NULOS (NA)

```
anyNA(Ha)
```

```
## [1] FALSE
```

METODOLOGÍA DE ANÁLISIS

El ***Análisis Discriminante Lineal o Linear Discrimiant Analysis (LDA)*** es un método de clasificación supervisado de variables cualitativas en el que dos o más grupos son conocidos a priori y nuevas observaciones se clasifican en uno de ellos en función de sus características. Haciendo uso del teorema de Bayes, LDA estima la probabilidad de que una observación, dado un determinado valor de los predictores, pertenezca a cada una de las clases de la variable cualitativa, $P(Y = k|X = x)$. Finalmente se asigna la observación a la clase k para la que la probabilidad predicha es mayor. El algortimo LDA comienza por encontrar direcciones que maximizan la separación entre clases, luego utiliza estas direcciones para predecir la clase de individuos. Estas direcciones llamadaa discriminantes lineales, son combinaciones lineales de variables predictoras.

RESULTADOS

Se define la matriz de datos y la variable respuesta con las clasificaciones.

```
x<-Ha[,5:11]
y<-Ha[,4]
```

Se optó por trabajar con 8 variables que son: Especies (que es la variable y), Wing, Weight, Culmen, Hallux, Tail, StandardTail, Tarsus.

Definir como n y p el número de species y variables

```
n<-nrow(x)
p<-ncol(x)
```

SE APLICA EL ANÁLISIS DISCRIMINATE LINEAL

```
library(MASS)
```

```
modelo_lda <- lda(formula = y ~ Wing + Weight +
                  Culmen + Hallux + Tail +
                  StandardTail + Tarsus, data = Ha)
```

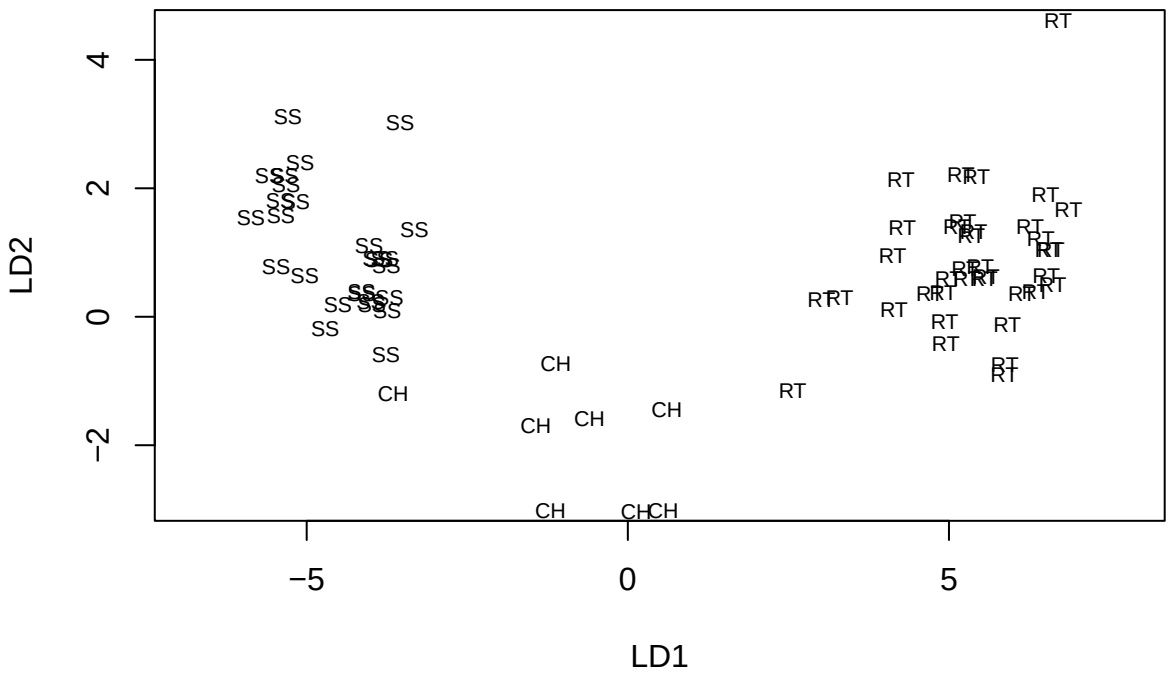
De acuerdo con las probabilidades previas de grupos, es más probable a pertenecer al grupo de especies Espinillas afiladas (SS)

Ecuaciones:

$$LD1 = 0.022(Wing) - 0.0004(Weight) - 0.11(Culmen) + 0.37(Hallux) - 0.02(Tail) - 0.009(StandardTail) + 0.06(Tarsus)$$

$$LD2 = 0.001(Wing) + 0.003(Weight) - 0.27(Culmen) + 0.33(Hallux) - 0.02(Tail) - 0.10(StandardTail) + 0.05(Tarsus)$$

```
plot(modelo_lda, main = "Gráfico de discriminantes lineales")
```



Agrupamiento de cada una de las variables por especie.

Despúes de haber obtenido las funciones discriminantes, ya se puede clasificar un nuevo Halcón dependiendo sus medidas.

Ponemos un ejemplo con los siguientes datos:

Wing = 250, Weight = 715, Culmen = 18.2, Hallux = 11 Tail = 218,

StandardTail = 176, Tarsus = 67.8

NUEVAS OBSERVACIONES

```
nuevas_observaciones <- data.frame(Wing = 250, Weight = 715,
                                   Culmen = 18.2, Hallux = 11, Tail = 218,
                                   StandardTail = 176, Tarsus = 67.8)
predict(object = modelo_lda, newdata = nuevas_observaciones)
```

```
## $class
## [1] SS
## Levels: CH RT SS
##
## $posterior
##              CH           RT        SS
## 1 0.0005639214 2.093009e-23 0.9994361
##
## $x
```

```
##         LD1       LD2
## 1 -5.925058 -1.746037
```

El resultado muestra que, de acuerdo con la función discriminante, la probabilidad posterior de que un Halcón pertenezca a la especie SS es de 99.9%

Evaluación de los errores de clasificación.

```
predicciones <- predict(object = modelo_lda, newdata = Ha[, 5:11],
                        method = "predictive")
```

```
table(Ha$Species, predicciones$class,
      dnn = c("Clase real", "Clase predicha"))
```

```
##           Clase predicha
## Clase real CH RT SS
##         CH  7  0  1
##         RT  0 37  0
##         SS  0  0 27
```

```
trainig_error <- mean(Ha$Species != predicciones$class) * 100
paste("trainig_error=", trainig_error, "%")
```

```
## [1] "trainig_error= 1.38888888888889 %"
```

La precisión de clasficación del modelo discriminante es del 98.62% ya que solo tiene 1.38% de error de clasificación

CONCLUSIONES

Cómo se mencionó anteriormente, la probabilidad más alta de que pertenezcan a un grupo es al grupo SS, se puede comprobar con nuestro ejemplo de nuevas predicciones. El margen de error de este modelo es muy mínimo por lo que podemos concluir que estamos ante un buen modelo matemático de clasificación para los tres tipos de especies de los Halcones.

BIBLIOGRAFÍA

https://www.key2stats.com/data-set/view/840

http://www.halconpedia.com/

https://rpubs.com/Joaquin_AR/233932