INTRODUCCION AL ANALISIS MULTIVARIADO

Lab. No.4 - LOGISTICA MULTINOMIAL Y DISCRIMINANTE

IRIS

La base de datos Iris corresponde a Fisher o a Anderson y es famosa. Se cuenta con las medidas en centímetros de un conjunto de 150 flores: largo y ancho del sépalo, largo y ancho del pétalo. Se cuenta con 50 flores de cada una de 3 especies: Iris setosa, versicolor, y virginica. Los datos se encuentran en el archivo `iris.Rdata`.

1. Haga una base con una muestra de 100 flores. Use RNGkind(sample.kind = “Rounding”) y set.seed(10). La base de aprendizaje se llamará `basea` y la de validaciónse llamará `basev`.

- Observe cuántos datos quedaron en la base de entrenamiento de cada especie.
- Cómo podría hacerse para tener el mismo número de datos de cada especie en la base de entrenamiento? (no tiene que hacerlo)

2. Visualice los datos de la base de aprendizaje por pares de variables poniendo colores por especie.

- Observe el comportamiento de los dos tipos para las diferentes combinaciones de variables. Vea en particular si algunas de ellas serían suficientes para clasificar.

3. Proponga un modelo logístico multinomial y estime sus parámetros usando la base de entrenamiento. Use la función `multinom` de la librería `nnet`.

- Observe si hubo convergencia. Qué se entiende con que hay convergencia? Qué se podría hacer si no hubiera convergencia?
- Tome el individuo 32 de la base de validación. Cuáles son los valores de las 4 variables para ese individuo? A cuál especie pertenece?
- Obtenga la parte lineal para ese individuo en cada ecuación (versicolor y virginica).
- Obtenga la probabilidad de pertenencia a la especie versicolor para ese individuo. Use 3 decimales.
- Obtenga la probabilidad de pertenencia a la especie virginica para ese individuo. Use 3 decimales.
- Obtenga la probabilidad de pertenencia a la especie setosa para ese individuo. Use 3 decimales.
- Obtenga las probabilidades de pertenencia a cada especie para los individuos de la base de validación usando `predict`, indique `type="probs"` para obtener las probabilidades. Extraiga las probabilidades para el individuo 32 (use 3 decimales).
- Haga la tabla de confusión al clasificar los datos de la base de validación. Para obtener la clasificación use `predict` pero elimine `type="probs"`.
- Realice el proceso de selección de variables hacia atrás. ¿Cuál resultado es más conveniente? Use la función `step(mod1)` que usa el criterio de Akaike.
- Obtenga la tabla de confusión y compárela con la obtenida con el modelo completo. Para almacenar el resultado del step en mod2, haga `mod2=step(mod1)`

SECUNDARIA

El conjunto de datos contiene variables sobre 200 estudiantes. Los estudiantes que ingresan a la escuela secundaria hacen la elección de un programa de tres posibles: general, vocacional y académico. Su elección puede ser modelada usando algunas variables predictoras. A continuación, se describen las variables:

genero: género del estudiante (femenino, masculino).

nivelsocio: estrato socioeconómico (bajo, medio, alto).

tipo: tipo de escuela (privada, publica).

programa: tipo de programa elegido por el estudiante (general, vocacional, académico).

lectura, escritura, mate, ciencias y sociales son variables continuas que representan los puntajes en cada una de esas materias. Los datos se encuentran en el archivo `secundaria.Rdata` .

1. Cargue la base y observe cuántos estudiantes hay de cada programa.
2. Corra un modelo de regresión logística multinomial para el programa como respuesta. Use como predictores el género, el estrato socioeconómico y el tipo de colegio como predictores.

- Observe cuántos individuos hay en cada combinación de la variable respuesta y cada predictor.
- Note que solo hay dos estudiantes de escuela privada en vocacional. Busque esos estudiantes y elimínelos de la base.
- Corra nuevamente el modelo usando la base donde se eliminaron esos dos estudiantes.
- Compare los errores estándar de los dos modelos.

CRANEOS

- Se tienen datos de 32 cráneos recogidos en el Tibet los cuales han sido clasificados en 2 tipos raciales.
- Se cuenta con 5 medidas antropométricas de longitudes y anchuras de cráneo y cara las cuales se van a utilizar para construir una función discriminante.
- Los datos se encuentran en el archivo “Tibet.Rdata”

1. Visualice los datos por pares de variables poniendo colores por tipo de cráneo.

- Observe el comportamiento de los dos tipos para las diferentes combinaciones de variables. Vea en particular si algunas de ellas serían suficientes para clasificar.

2. Obtenga las matrices de covariancias para cada tipo de cráneo.

- Compárelas visualmente.
- Haga la prueba multivariada de Box (M de Box) para verificar que las dos matrices de covariancias son iguales. Use la función boxM de la librería biotools: boxM(basex,tipo). Recuerde que en basex sólo se incluyen las variables métricas que se usarán para hacer la función de clasificación.
- Obtenga la matriz de covariancias combinada.

3. Verifique el supuesto de normalidad:

- Haga el qqplot multivariado para cada tipo de cráneo (tiene que hacer dos gráficos). Para hacer esto defina una sub-base para cada tipo de cráneo llamada b; n y p son el número de filas y columnas de b,

respectivamente. Primero se calculan las distancias de Mahalanobis cuadrática de cada punto a su centroide. Debe especificarse la matriz de covariancias:

4. Asuma probabilidades a priori iguales para cada tipo de cráneo. Calcule el valor de las dos funciones discriminantes lineales para cada cráneo mediante:

$$L_i(x) = \bar{x}_i' S_p^{-1} \left(x - \frac{1}{2} \bar{x}_i \right)$$

- Calcule las probabilidades a posteriori.
- Decida a cuál tipo asigna cada cráneo usando los resultados de las dos funciones anteriores.
- Haga una tabla de confusión y calcule los porcentajes de clasificación errónea para cada tipo.
- Use la función lda de la librería MASS. Debe escribir el modelo de la misma forma que se hace en lm. Llámelo mod1. Se pueden indicar probabilidades a priori con prior=c(0.5,0.5), si no se hace, el default son las proporciones observadas.
- Haga la clasificación basada en las funciones discriminantes obtenidas anteriormente en mod1. Para esto use predict(mod) y observe el resultado. Compare estos resultados con los obtenidos manualmente en los puntos anteriores.
- Clasifique dos cráneos que tienen los siguientes valores para las variables utilizadas: 171,140.5,127.0,69.5,137.0 y 179.0,132.0,140.0,72.0,138.5. Use la función predict de forma similar a un modelo de regresión.
- Realice el proceso de selección de variables con el stepwise hacia adelante para clasificación mediante la función greedy.wilks de la librería klaR: greedy.wilks(tipo ~ .,data=base, prior=c(0.5,0.5), “lda”, niveau = 0.05)
- Escriba el modelo sugerido por el resultado del proceso de selección de variables y obtenga nuevamente la tabla de confusión. Compárela con la que obtuvo anteriormente. Es conveniente este resultado?

5. Proponga un modelo logístico y estime sus parámetros.

- Haga la tabla de confusión y compárela con la obtenida anteriormente con LDA y con todas las variables.
- Realice el proceso de selección de variables. Compárelo con el obtenido con el que sugiere el LDA.