Universidad Técnica Federico Santa María

MAT281: Aplicaciones de la Matemática en Ingeniería

Tarea 1

Profesor:
Alfredo Alegría
Departamento de Matemática

Alumno:
Alejandro Villazón G.
201910009-2

15 de Octubre del 2021

Problema 1

En este problema se analiza la conexión entre la regresión logística y el análisis discriminante lineal. Considere un problema de clasificación binaria. El *log-odd* se define como

$$\log\left(\frac{P(Y=1|\boldsymbol{X}=\boldsymbol{x})}{P(Y=0|\boldsymbol{X}=\boldsymbol{x})}\right)$$

a. Calcule el *log-odd* para el modelo de regresión logística.

Solución: Recuerde que para el modelo de regresión logística tenemos que $P(Y=1|\boldsymbol{X}=\boldsymbol{x})=\frac{u}{1+u}$ donde $u=\exp\left(\beta_0+\sum_{j=1}^{d}\beta_j x_j\right)$. Así

$$\begin{aligned}
\log\left(\frac{P(Y=1|\boldsymbol{X}=\boldsymbol{x})}{P(Y=0|\boldsymbol{X}=\boldsymbol{x})}\right) &= \log\left(\frac{P(Y=1|\boldsymbol{X}=\boldsymbol{x})}{1-P(Y=1|\boldsymbol{X}=\boldsymbol{x})}\right)\\
&= \log\left(\frac{\frac{u}{1+u}}{1-\frac{u}{1+u}}\right)\\
&= \log(u)\\
&= \log(\exp\left(\beta_0+\sum_{j=1}^{d}\beta_j x_j\right))\\
&= \beta_0+\sum_{j=1}^{d}\beta_j x_j\\
&= \beta_0+\boldsymbol{\beta}^T\boldsymbol{x}
\end{aligned}$$

donde $\boldsymbol{\beta}^T=(\beta_1,\beta_2,\ldots,\beta_n)$.

b. Muestre que en el caso del análisis discriminante lineal el *log-odd* se puede escribir en el formato $\alpha_0+\boldsymbol{\alpha}^T\boldsymbol{x}$, para algún $\alpha_0\in\mathbb{R}$ y $\boldsymbol{\alpha}\in\mathbb{R}^d$. Determine explícitamente α_0 y $\boldsymbol{\alpha}$.

Solución: Recuerde que para el caso LDA se asume que

$$f_k(\boldsymbol{x})=\frac{1}{(2\pi)^d|\boldsymbol{\Sigma}_k|^{1/2}}\exp\left(-\frac{1}{2}r_k^2\right),\quad k=0,1.$$

donde $r_k^2=(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)$ con $k=0,1$, tal que $\boldsymbol{\Sigma_0}=\boldsymbol{\Sigma_1}=\boldsymbol{\Sigma}$. Así, tenemos que

$$\begin{aligned}
\log\left(\frac{P(Y=1|\boldsymbol{X}=\boldsymbol{x})}{P(Y=0|\boldsymbol{X}=\boldsymbol{x})}\right) &= \log\left(\frac{\theta_1 f_1(\boldsymbol{x})/(\theta_0 f_0(\boldsymbol{x})+\theta_1 f_1(\boldsymbol{x}))}{\theta_0 f_0(\boldsymbol{x})/(\theta_0 f_0(\boldsymbol{x})+\theta_1 f_1(\boldsymbol{x}))}\right)\\
&= \log\left(\frac{\theta_1}{\theta_0}\right)+\log\left(\frac{\exp\left(-\frac{1}{2}r_1^2\right)}{\exp\left(-\frac{1}{2}r_0^2\right)}\right)\\
&= \log\left(\frac{\theta_1}{\theta_0}\right)+\frac{1}{2}\left(r_0^2-r_1^2\right)\\
&= \log\left(\frac{\theta_1}{\theta_0}\right)+\frac{1}{2}\left((\boldsymbol{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)-(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right)\\
&= \log\left(\frac{\theta_1}{\theta_0}\right)+\frac{1}{2}(-2\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}+2\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}+\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0-\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1)\\
&= \left(\log\left(\frac{\theta_1}{\theta_0}\right)+\frac{1}{2}\left(\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0-\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1\right)\right)+\left(\boldsymbol{\mu}_1^T-\boldsymbol{\mu}_0^T\right)\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\\
&= \alpha_0+\boldsymbol{\alpha}^T\boldsymbol{x}
\end{aligned}$$

note que $\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k\in\mathbb{R}$ para $k=1,0$, por lo que $\alpha_0:=\log\left(\frac{\theta_1}{\theta_0}\right)+\frac{1}{2}\left(\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0-\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1\right)$ está bien definido y pertenece a $\mathbb{R}$. Por otro lado, tenemos que $\boldsymbol{\alpha}:=\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0\right)$ el cual pertenece a $\mathbb{R}^d$ pues $\boldsymbol{\mu}_k\in\mathbb{R}^d$ para $k=0,1$ y $\boldsymbol{\Sigma}^{-1}\in\mathbb{R}^{d\times d}$.

c. ¿Qué se puede concluír? ¿Cuáles son las diferencias fundamentales entre ambos métodos?

Solución: Podemos concluir que en ambos casos el *log-odd* se puede escribir en el mismo formato. La diferencia fundamental entre ambos métodos es que en el Análisis discriminante se asume una densidad gaussiana para f_0 y f_1 para así calcular $r(\boldsymbol{x})$ mediante el teorema de Bayes, en cambio en el método de regresión logística se asume directamente una forma para $r(\boldsymbol{x})$ sin pasar por las funciones de densidad f_k. Por lo que para la regresión logística se estiman los parámetros del vector $\boldsymbol{\beta}$ y para el caso del Análisis discrimimante lineal se estiman las medias $\boldsymbol{\mu}_k$ y la matriz de covarianza $\boldsymbol{\Sigma}$.

Problema 2

La siguiente tabla muestra un conjunto de datos que contiene 6 observaciones, 3 covariables y 1 variable respuesta cualitativa:

X_1	X_2	X_3	Y
0	3	0	Rojo
2	0	0	Rojo
0	1	3	Rojo
0	1	2	Verde
-1	0	1	Verde
1	1	1	Rojo

Suponga que queremos usar estos datos para predecir Y cuando $X_1 = X_2 = X_3 = 0$ usando el método de K vecinos cercanos.

a. Calcule la distancia Euclideana entre cada observación y el punto $X_1 = X_2 = X_3 = 0$.

Solución: Sea $\boldsymbol{X}_i$ el vector de covariables de la observación i-ésima y $\boldsymbol{x} = (0, 0, 0)$ así tenemos las siguientes distancias:

$$\begin{aligned}
\|\boldsymbol{X}_1 - \boldsymbol{x}\|_2 &= 3\\
\|\boldsymbol{X}_2 - \boldsymbol{x}\|_2 &= 2\\
\|\boldsymbol{X}_3 - \boldsymbol{x}\|_2 &= \sqrt{10}\\
\|\boldsymbol{X}_4 - \boldsymbol{x}\|_2 &= \sqrt{5}\\
\|\boldsymbol{X}_5 - \boldsymbol{x}\|_2 &= \sqrt{2}\\
\|\boldsymbol{X}_6 - \boldsymbol{x}\|_2 &= \sqrt{3}
\end{aligned}$$

b. ¿Cuál es la predicción con $K = 1$?

Solución: Verde, pues la observación más cercana a $\boldsymbol{x}$ es $\boldsymbol{X}_5$, por lo que clasificamos a Y en la clase de Y_5 que es Verde.

c. ¿Cuál es la predicción con $K = 3$?

Solución: Note que las 3 observaciones más cercanas a $\boldsymbol{x}$ son $\boldsymbol{X}_5, \boldsymbol{X}_6$ y $\boldsymbol{X}_2$. Como $Y_5 = \text{Verde}, Y_6 = \text{Rojo}$ y $Y_2 = \text{Rojo}$, tenemos que

$$P(Y = \text{Rojo} \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{2}{3} \quad \text{y} \quad P(Y = \text{Verde} \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{1}{3}.$$

Luego, Y es clasificado como Rojo.

d. Si en este problema la frontera de decisión del clasificador de Bayes es altamente no-lineal, ¿se espera que el *mejor* valor de K sea grande o pequeño?

Solución: Se espera que sea pequeño, pues si K fuera grande por lo visto en clases se esperaría una división lineal.

Problema 3

Una estación de radio clasificará a sus auditores en jóvenes (0) o adultos (1) a partir de sus gustos musicales. Para llevar a cabo este proceso, se realizó una encuesta a 10 auditores, donde cada uno ha manifestado si le agradan o no ciertos grupos. Los datos se encuentran disponibles en el archivo `gustos_musicales.txt` en AULA.

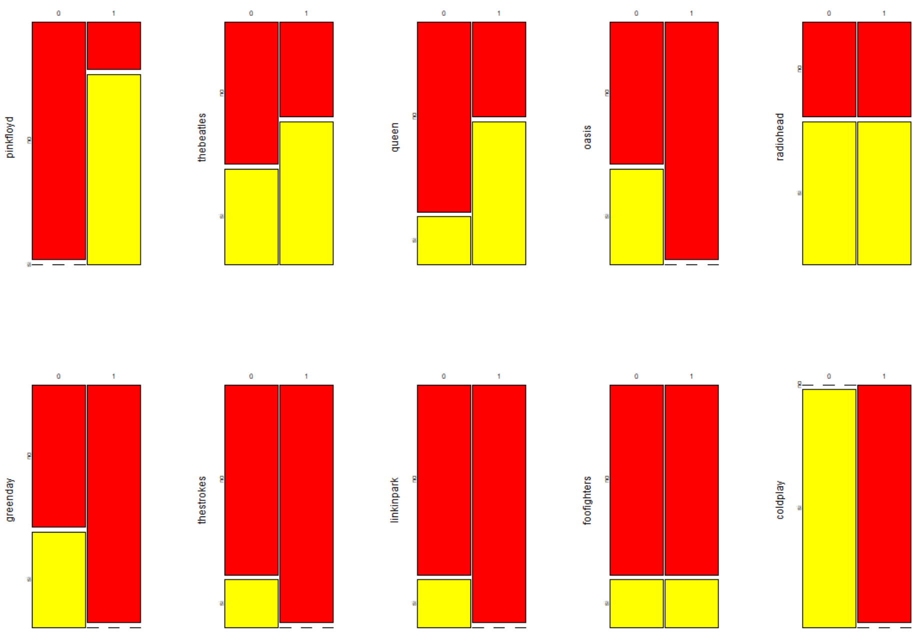
a. Ajuste un modelo de Bayes ingenuo a este conjunto de datos. Reporte las distribuciones marginales estimadas para cada covariable. Comente los resultados.

Solución: Se ejecutó el siguiente código en R:

```
library("naivebayes")
path = "/Users/Aleja/Documents/6to Sem/MAT281/T1/gustos_musicales.txt"
datos = read.table(file=path, header=TRUE, sep=",")
datos$etiqueta <- as.factor(datos$etiqueta)
attach(datos)

modelo <-naive_bayes(etiqueta ~ pinkfloyd + thebeatles + queen + oasis
                     + radiohead + greenday + thestrokes + linkinpark
                     + foofighters + coldplay, data = datos)

par(mfrow=c(2,5))
plot(modelo)
```



Note que la covariable “coldplay” es decisiva, puesto que a los que les gusta esta banda en su totalidad son jovenes y a los que no son adultos.

Si nos fijamos en las covariables “foofighters” y “radiohead” tenemos una proporción similar tanto en jovenes como adultos, con una mayoría de personas que no les gusta “foofighters” y que les gusta “radiohead”, podriamos decir que estas variables no son muy decisivas dada la proporción similar.

Por otro lado, podemos notar que las covariables “linkinpark”, “oasis”, “greenday” y “thestrokes” tienen proporciones muy similares en ambas etiquetas, se puede ver que la totalidad de adultos no sienten agrado por estas cuatro bandas y una pequeña proporción de jovenes les gusta estas bandas. Si una persona siente agrado por una de estas bandas podría tender a ser joven dados los datos.

Note que, las covariables “thebeatles” y “queen” tienen proporciones similares en ambas etiquetas, la mayoría de los jovenes no sienten agrado por estas bandas y la mayoría de los adultos si sienten agrado.

Finalmente, para la covariable “pinkfloyd” podemos notar que a la mayoría de las personas que no les gusta esta banda son jovenes y a las que si en su totalidad son adultos, lo cual se puede deber a la antiguedad de la banda y la edad de los encuestados, esta variable puede ser considerada influyente.

b. Responda la encuesta con sus gustos musicales. De acuerdo al modelo ajustado en el inciso **a.**, ¿a cuál de las dos clases pertenece?

Solución: De la lista de grupos/bandas musicales solo me gustan Queen y Coldplay. Al aplicar la función `predict()` en R como sigue

```
predict(modelo, newdata = data.frame(pinkfloyd = "no", thebeatles = "no",
    queen = "si", oasis = "no", radiohead = "no", greenday = "no",
    thestrokes = "no", linkinpark = "no", foofighters = "no", coldplay = "si"),
    type = "prob")
```

```
        0           1
0.9982669 0.001733102
```

tenemos que el resultado es 0.0017 para la etiqueta 1. Por lo tanto, se me clasifica como Joven (0).

Problema 4

Considere el conjunto de datos de **factores de riesgo coronario** en tres regiones rurales de Sudáfrica (ver el archivo `datos_heart_disease.txt` disponible en AULA). Lleve a cabo un proceso de selección de variables siguiendo los siguientes pasos:

i. Ajuste un modelo de regresión logística con las 7 covariables que se señalan a continuación: presión sanguínea sistólica (`sbp`), tabaco acumulado (`tobacco`), colesterol unido a lipoproteínas de baja densidad (`ldl`), historial familiar de problemas cardiacos (`famhist`), obesidad (`obesity`), consumo actual de alcohol (`alcohol`) y edad (`age`).

ii. Elimine la covariable menos significativa basándose en el test de Wald.

iii. Vuelva a ajustar un modelo de regresión logística con las covariables restantes y nuevamente elimine la menos significativa.

iv. Repita este procedimiento hasta que no se puedan quitar más covariables del modelo.

Solución: Se ajustó el siguiente modelo en R:

```
path = "/Users/Aleja/Documents/6to Sem/MAT281/T1/datos_heart_disease.txt"
datos = read.table(file=path, header=TRUE, sep=",")
attach(datos)

modelo <- glm(chd ~ sbp+tobacco+ldl+famhist+obesity+alcohol+age,
              family=binomial(link="logit"))
summary(modelo)
```

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.1295997  0.9641558  -4.283 1.84e-05 ***
sbp             0.0057607  0.0056326   1.023  0.30643
tobacco         0.0795256  0.0262150   3.034  0.00242 **
ldl             0.1847793  0.0574115   3.219  0.00129 **
famhistPresent  0.9391855  0.2248691   4.177 2.96e-05 ***
obesity        -0.0345434  0.0291053  -1.187  0.23529
alcohol         0.0006065  0.0044550   0.136  0.89171
age             0.0425412  0.0101749   4.181 2.90e-05 ***
```

```
#Se elimina la covariable "alcohol"
#Se ajusta nuevamente el modelo:
modelo <- glm(chd ~ sbp+tobacco+ldl+famhist+obesity+age,
              family=binomial(link="logit"))
summary(modelo)
```

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.127753   0.963799  -4.283 1.85e-05 ***
sbp             0.005862   0.005585   1.050  0.29389
tobacco         0.080220   0.025736   3.117  0.00183 **
ldl             0.184151   0.057224   3.218  0.00129 **
famhistPresent  0.941306   0.224342   4.196 2.72e-05 ***
obesity        -0.034546   0.029100  -1.187  0.23517
age             0.042421   0.010131   4.187 2.82e-05 ***
```

```
#Se elimina la covariable "sbp"
#Se ajusta nuevamente el modelo:
modelo <- glm(chd ~ tobacco+ldl+famhist+obesity+age,
              family=binomial(link="logit"))
summary(modelo)
```

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.546637   0.786171  -4.511 6.44e-06 ***
tobacco         0.080760   0.025604   3.154  0.00161 **
ldl             0.185226   0.057212   3.238  0.00121 **
famhistPresent  0.933002   0.223774   4.169 3.05e-05 ***
obesity        -0.030526   0.028729  -1.063  0.28799
age             0.045256   0.009774   4.630 3.65e-06 ***
```

```
#Se elimina la covariable "obesity"
#Se ajusta nuevamente el modelo:
modelo <- glm(chd ~ tobacco+ldl+famhist+age,family=binomial(link="logit"))
summary(modelo)
```

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.204275   0.498315  -8.437  < 2e-16 ***
tobacco         0.080701   0.025514   3.163  0.00156 **
ldl             0.167584   0.054189   3.093  0.00198 **
famhistPresent  0.924117   0.223178   4.141 3.46e-05 ***
age             0.044042   0.009743   4.521 6.17e-06 ***
```

Responda las siguientes preguntas:

a. ¿Cuales son las variables que sobrevivieron a este proceso?

Solución: Las variables `ldl`, `tobbaco`, `famhist` y `age` sobrevivieron al proceso.

b. ¿De qué otra manera se podría realizar un proceso de selección de variables?

Solución: Se puede aplicar un método similar al hecho en el punto anterior, fijandonos en un criterio para añadir o eliminar variables, este método es conocido como "stepwise.º "paso a paso", en el curso Análisis de Regresión se ven algunos criterios como C_p de Mallows, $R^2_{ajustado}$ o AIC para encontrar el mejor subconjunto de covariables para el modelo. El método stepwise tiene tres modalidades:

- Dirección *forward*: consiste en comenzar el modelo sin covariables e ir añadiendo variable a variable quedandonos con una si es que el modelo mejora según el criterio que estemos utilizando. Se repite el proceso agregando una a una las covariables restantes quedandonos con la que mejore el modelo y así sucesivamente hasta que el modelo no pueda ser mejorado por ninguna covariable restante
- Dirección *backward*: el modelo comienza con todas las covariables, se va eliminando una a una, si el modelo mejora según el criterio, entonces esa covariable se elimina. Luego, se repite el proceso con las covariables restantes.
- Dirección *both*: es una combinación de las modalidades anteriores, es decir, permite agregar o eliminar covariables en cada paso si es que el modelo mejora según el criterio.

En R está implementada la función `step()` que aplica el método descrito anteriormente basandose en el Criterio de Información de Akaike (AIC) el cual se busca minimizar. Si aplicamos la función `step()` para nuestro modelo obtenemos que el mejor subconjunto de covariables es el mismo obtenido en **a.** y se van eliminando en el mismo orden, como se puede ver a continuación:

```
modelo <- glm(chd ~ sbp+tobacco+ldl+famhist+obesity+alcohol+age,
              family=binomial(link="logit"))
z <- step(modelo, direction="backward")


Start:  AIC=499.17
chd ~ sbp + tobacco + ldl + famhist + obesity + alcohol + age

          Df Deviance    AIC
- alcohol  1   483.19 497.19
- sbp      1   484.22 498.22
- obesity  1   484.61 498.61
<none>         483.17 499.17
- tobacco  1   493.05 507.05
- ldl      1   494.09 508.09
- famhist  1   500.89 514.89
- age      1   501.51 515.51


Step:  AIC=497.19
chd ~ sbp + tobacco + ldl + famhist + obesity + age

          Df Deviance    AIC
- sbp      1   484.30 496.30
- obesity  1   484.63 496.63
<none>         483.19 497.19
- tobacco  1   493.62 505.62
- ldl      1   494.12 506.12
- famhist  1   501.07 513.07
- age      1   501.54 513.54

Step:  AIC=496.3
chd ~ tobacco + ldl + famhist + obesity + age

          Df Deviance    AIC
- obesity  1   485.44 495.44
<none>         484.30 496.30
- tobacco  1   494.99 504.99
- ldl      1   495.36 505.36
- famhist  1   501.93 511.93
- age      1   507.07 517.07


Step:  AIC=495.44
chd ~ tobacco + ldl + famhist + age

          Df Deviance    AIC
<none>         485.44 495.44
- ldl      1   495.39 503.39
- tobacco  1   496.18 504.18
- famhist  1   502.82 510.82
- age      1   507.24 515.24
```