Centro de Investigación en Matemáticas (CIMAT). Unidad Monterrey

Tarea 5

Hairo Ulises Miranda Belmonte

April 11, 2019

1 Problema

Para datos de clasificación binaria $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, considera la siguiente función de costo:

$$\mathcal{L} = \sum_i (\theta(y_i) - \beta' \mathbf{x}_i - \beta_0)^2 \tag{1.1}$$

Definimos n_+, n_- el número de observaciones con $y_i = 1$ y $y_i = -1$, respectívamente $\mathbf{c}_+, \mathbf{c}_-$ el centroide de las observaciones con $y_i = 1$, y $y_i = -1$ y $\mathbf{c}$ el centroide de todos los datos. Como en clase, construimos las matrices:

$$\begin{aligned} \mathbf{S}_B &= (\mathbf{c}_+ - \mathbf{c}_-)(\mathbf{c}_+ - \mathbf{c}_-)' \\ \mathbf{S}_W &= \textstyle\sum_{i:y_i=1}(\mathbf{x}_i - \mathbf{c}_+)(\mathbf{x}_i - \mathbf{c}_+)' + \sum_{i:y_i=-1}(\mathbf{x}_i - \mathbf{c}_-)(\mathbf{x}_i - \mathbf{c}_-)' \end{aligned}$$

a) Verifica que

$$\mathbf{S}_W = \sum_{i:y_i=1} \mathbf{x}_i \mathbf{x}_i' + \sum_{i:y_i=-1} \mathbf{x}_i \mathbf{x}_i' - n_+ \mathbf{c}_+ \mathbf{c}_+' - n_- \mathbf{c}_- \mathbf{c}_-'$$

b) Verifica que el vector $\mathbf{S}_B \boldsymbol{\beta}$, es un múltiplo del vector $(\mathbf{c}_+ - \mathbf{c}_-)$.

c) Si definimos $\theta(1) = n/n_+$ y $\theta(-1) = -n/n_-$, verifica que en el mínimo de (1.1):

$$\beta_0 = -\boldsymbol{\beta}' \mathbf{c},$$

$$(\mathbf{S}_W + \frac{n_+ n_-}{n} \mathbf{S}_B)\boldsymbol{\beta} = n(\mathbf{c}_+ - \mathbf{c}_-) \tag{1.2}$$

d) Usando el resultado de inciso b, argumenta que (1.2) implica que en el mínimo:

$$\boldsymbol{\beta} \sim S_W^{-1}(\mathbf{c}_+ - \mathbf{c}_-),$$

es decir la solución coincide con la del Fisher Discriminant Analysis (FDA).

e) Lo anterior permite implementar FDA usando algún algoritmo de mínimos cuadrados (`lm()` en `R`, o algúna otra librería). Ilustra cómo funciona el método con algunos conjuntos de datos en 2D bien elegidos.

f) Observamos que muestra que FDA es muy robusto a datos atípicos.
Una posibilidad para hacerlo más robusto es usar mínimos cuadrados ponderados. Por ejemplo `lm()` tiene un argumento opcional `weights` donde se pueden proporcionar pesos $w_i, i = 1, \ldots, n$ para minimizar:

$$\sum_i w_i(\theta(y_i) - \boldsymbol{\beta}^t \mathbf{x}_i - \beta_0)^2.$$

¿Cómo elegirías estos pesos? Verifica tu propuesta con algunos ejemplos en 2D.

1.1 Solución inciso "a"

Se trabaja el ejercicio en dos partes, la primera:

$$\Sigma_{i:y_{i=1}}(x_i - c_+)(x_i - c_+)^{'} =$$

$$= \Sigma_{i:y_{i=1}}(x_i - c_+)x_i^{'} + \Sigma_{i:y_{i=1}}(x_i - c_+)^{'}(-c_+^{'})$$
$$= \Sigma_{i:y_{i=1}}x_i x_i^{'} - c_+\Sigma_{i:y_{i=1}}x_i^{'} - [\Sigma_{i:y_{i=1}}x_i c_+^{'} - n_+ c_+ c_+^{'}]$$

Y debido a que sabemos lo siguiente

$$n_+ c_+ = \Sigma_{i:y_{i=1}}x_i$$
$$n_+ c_+^{'} = \Sigma_{i:y_{i=1}}x_i^{'}$$

el resultado se puede expresar de la siguiente forma

$$= \Sigma_{i:y_{i=1}}x_i x_i^{'} - n_+ c_+ c_+^{'} - \Sigma_{i:y_{i=1}}x_i c_+^{'} + n_+ c_+ c_+^{'}$$
$$= \Sigma_{i:y_{i=1}}x_i x_i^{'} - \Sigma_{i:y_{i=1}}x_i c_+^{'}$$
$$= \Sigma_{i:y_{i=1}}x_i x_i^{'} - n_+ c_+ c_+^{'}$$

Asimismo se calcula la segunda parte

$$\Sigma_{i:y_{i=-1}}(x_i - c_-)(x_i - c_-)^{'} =$$

$$= \Sigma_{i:y_{i=-1}}(x_i - c_-)x_i^{'} + \Sigma_{i:y_{i=-1}}(x_i - c_-)^{'}(-c_-^{'})$$
$$= \Sigma_{i:y_{i=-1}}x_i x_i^{'} - c_-\Sigma_{i:y_{i=-1}}x_i^{'} - [\Sigma_{i:y_{i=-1}}x_i c_-^{'} - n_- c_- c_-^{'}]$$

$$n_- c_- = \Sigma_{i:y_{i=-1}} x_i$$
$$n_- c_-^{'} = \Sigma_{i:y_{i=-1}} x_i^{'}$$

el resultado se expresa como:

$$= \Sigma_{i:y_{i=-1}} x_i x_i^{'} - n_- c_- c_-^{'} - \Sigma_{i:y_{i=-1}} x_i c_-^{'} + n_- c_- c_-^{'}$$
$$= \Sigma_{i:y_{i=-1}} x_i x_i^{'} - \Sigma_{i:y_{i=-1}} x_i c_-^{'}$$
$$= \Sigma_{i:y_{i=-1}} x_i x_i^{'} - n_- c_- c_-^{'}$$

por último, se juntan los dos términos

$$= \Sigma_{i:y_{i=1}} x_i x_i^{'} - n_+ c_+ c_+^{'} + \Sigma_{i:y_{i=-1}} x_i x_i^{'} - n_- c_- c_-^{'}$$

se ordenan los términos

$$= \Sigma_{i:y_{i=1}} x_i x_i^{'} + \Sigma_{i:y_{i=-1}} x_i x_i^{'} - n_+ c_+ c_+^{'} - n_- c_- c_-^{'}$$

con esto se demuestra que

$$S_w = \Sigma_{i:y_{i=1}} x_i x_i^{'} + \Sigma_{i:y_{i=-1}} x_i x_i^{'} - n_+ c_+ c_+^{'} - n_- c_- c_-^{'}$$

1.2 Solución inciso "b"

Sea β un vector con los coeficientes β_i de $i = 1, ..., n$ de la forma $\beta = [\beta_1, \beta_2, ..., \beta_n]^{'}$, sustituyendo S_B se tiene:

$$(c_+ - c-)(c_+ - c-)^{'} \beta$$

donde $(c_+ - c-)^{'} \beta$ es un producto punto, entonces sea n una constante tal que

$$(c_+ - c-)^{'} \beta = n$$

se tiene que

$$(c_+ - c-)(c_+ - c-)^{'} \beta =$$
$$= (c_+ - c-)n$$

por lo tanto se demuestra que $S_B \beta$ es un múltiplo de $(c_+ - c-)$

1.3 Solución inciso "c"

Separamos la función de costos en aquellas observaciones clasificadas en clase $y = 1$ y $y = -1$ de la siguiente manera:

$$\Sigma_{i=y=1}(\frac{n}{n_+} - \beta' X_i - \beta_0)^2 + \Sigma_{i=y=-1}(-\frac{n}{n_-} - \beta' X_i - \beta_0)^2$$

primero se toma la derivada de la función de costo respecto a β_0

$$\frac{d}{d\beta_0} = 2\Sigma_{i=y=1}(\frac{n}{n_+} - \beta' X_i - \beta_0)(-1) + 2\Sigma_{i=y=-1}(-\frac{n}{n_-} - \beta' X_i - \beta_0)(-1)$$

Se reducen términos y se iguala a cero.

$$-2\Sigma_{i=y=1}(\frac{n}{n_+} - \beta' X_i - \beta_0) - 2\Sigma_{i=y=-1}(-\frac{n}{n_-} - \beta' X_i - \beta_0) = 0$$

$$-2[\Sigma_{i=y=1}(\frac{n}{n_+} - \beta' X_i - \beta_0) + \Sigma_{i=y=-1}(-\frac{n}{n_-} - \beta' X_i - \beta_0)] = 0$$

$$\Sigma_{i=y=1}(\frac{n}{n_+} - \beta' X_i - \beta_0) + \Sigma_{i=y=-1}(-\frac{n}{n_-} - \beta' X_i - \beta_0) = 0$$

$$\frac{nn_+}{n_+} - \beta' \Sigma_{i:y=i} X_i - n_+\beta_0 - \frac{nn_-}{n_-} - \beta' \Sigma_{i:y=-1} X_i - n_+\beta_0 = 0$$

$$n - \beta' \Sigma_{i:y=i} X_i - n_+\beta_0 - n - \beta' \Sigma_{i:y=-1} X_i - n_+\beta_0 = 0$$

$$-\beta' \Sigma_{i:y=i} X_i - n_+\beta_0 - \beta' \Sigma_{i:y=-1} X_i - n_+\beta_0 = 0$$

$$-\beta' \Sigma_{i:y=i} X_i - \beta' \Sigma_{i:y=-1} X_i = n_+\beta_0 + n_+\beta_0$$

$$-\beta'(\Sigma_{i:y=i} X_i + \Sigma_{i:y=-1} X_i) = (n_+ + n_+)\beta_0$$

$$-\beta'(\Sigma_{i:y=i} X_i + \Sigma_{i:y=-1} X_i) = (n_+ + n_+)\beta_0$$

en el siguiente paso se realiza la suma de todas las observaciones de tal forma que $\Sigma_{i:y=i} X_i + \Sigma_{i:y=-1} X_i = \Sigma_i X_i$ y $(n_+ + n_+) = n$, de esta manera:

$$-\beta'(\Sigma_i X_i) = n\beta_0$$

Cabe mencionar que se utiliza el siguiente resultado; $\Sigma_i X_i = nc$ con c como el centroide total.

$$-\beta'(nc) = n\beta_0$$

$$-\beta' c = \beta_0$$

A continuación derivamos respecto β, en muchos de los casos se utiliza lo siguiente.

$$n_- c_- = \Sigma_{i:y=-1} X_i$$
$$n_+ c_+ = \Sigma_{i:y=1} X_i$$
$$n_- c_-^{'} = \Sigma_{i:y=-1} X_i^{'}$$
$$n_+ c_+^{'} = \Sigma_{i:y=1} X_i^{'}$$

Se realiza la primera derivada

$$\frac{d}{d\beta} = 2\Sigma_{i=y=1}(\frac{n}{n_+} - \beta' X_i - \beta_0)(-X_i) + 2\Sigma_{i=y=-1}(-\frac{n}{n_-} - \beta' X_i - \beta_0)(-X_i)$$

$$-2[\Sigma_{i=y=1}(\frac{n}{n_+} - \beta' X_i - \beta_0)(-X_i) + \Sigma_{i=y=-1}(-\frac{n}{n_-} - \beta' X_i - \beta_0)(-X_i)] = 0$$

$$\Sigma_{i=y=1}(\frac{n}{n_+} - \beta' X_i - \beta_0)(X_i) + \Sigma_{i=y=-1}(-\frac{n}{n_-} - \beta' X_i - \beta_0)(X_i) = 0$$

$$\frac{n\Sigma_{i=y=1} X_i^{'}}{n_+} - \beta'\Sigma_{i=y=1} X_i X_i^{'} - \beta_0\Sigma_{i=y=1} X_i^{'}) - \frac{n\Sigma_{i=y=-1} X_i^{'}}{n_-} - \beta'\Sigma_{i=y=-1} X_i X_i^{'} - \beta_0\Sigma_{i=y=-1} X_i^{'} = 0$$

$$-\beta'\Sigma_{i=y=1} X_i X_i^{'} - \beta'\Sigma_{i=y=-1} X_i X_i^{'} = -\frac{n\Sigma_{i=y=1} X_i^{'}}{n_+} + \beta_0\Sigma_{i=y=1} X_i^{'} + \frac{n\Sigma_{i=y=-1} X_i^{'}}{n_-} + \beta_0\Sigma_{i=y=-1} X_i^{'}$$

$$-\beta'\Sigma_{i=y=1} X_i X_i^{'} - \beta'\Sigma_{i=y=-1} X_i X_i^{'} = -\frac{nn_+ c_+^{'}}{n_+} + \beta_0\Sigma_{i=y=1} X_i^{'} + \frac{nn_- c_-^{'}}{n_-} + \beta_0\Sigma_{i=y=-1} X_i^{'}$$

$$-\beta'\Sigma_{i=y=1} X_i X_i^{'} - \beta'\Sigma_{i=y=-1} X_i X_i^{'} = -nc_+^{'} + \beta_0\Sigma_{i=y=1} X_i^{'} + nc_-^{'} + \beta_0\Sigma_{i=y=-1} X_i^{'}$$

agrupando términos

$$-\beta'\Sigma_{i=y=1} X_i X_i^{'} - \beta'\Sigma_{i=y=-1} X_i X_i^{'} = n(c_+^{'} - c_-^{'}) + \beta_0 n_+ c_+^{'} + \beta_0 n_- c_-^{'}$$

$$-\beta'\Sigma_{i=y=1} X_i X_i^{'} - \beta'\Sigma_{i=y=-1} X_i X_i^{'} - \beta_0 n_+ c_+^{'} - \beta_0 n_- c_-^{'} = -n(c_+^{'} - c_-^{'})$$

se sustituye $\beta 0 = -\beta^{'} c$ y tenemos:

$$\beta'\Sigma_{i=y=1} X_i X_i^{'} - \beta'\Sigma_{i=y=-1} X_i X_i^{'} - (-\beta^{'} c) n_+ c_+^{'} - (-\beta^{'} c) n_- c_-^{'} = n(c_+^{'} - c_-^{'})$$

$$\beta'\Sigma_{i=y=1} X_i X_i^{'} - \beta'\Sigma_{i=y=-1} X_i X_i^{'} + \beta^{'} c n_+ c_+^{'} + \beta^{'} c n_- c_-^{'} = n(c_+ - c_-)^{'}$$

se realiza factorización respecto $\beta^{'}$

$$\beta'(\Sigma_{i=y=1}X_iX_i^{'} + \Sigma_{i=y=-1}X_iX_i^{'} - cn_+c_+^{'} - cn_-c_-^{'}) = n(c_+ - c_-)^{'}$$

separando c en $c_+ + c_-$

$$\beta'(\Sigma_{i=y=1}X_iX_i^{'} + \Sigma_{i=y=-1}X_iX_i^{'} - (c_+ + c_-)n_+c_+^{'} - (c_+ + c_-)n_-c_-^{'}) = n(c_+ - c_-)^{'}$$

se toma transpuesta a la ecuación por ambos lados, donde los términos del lado izquierdo que multiplican a β son simétricos; entonces, por definición se tiene:

$$\beta(\Sigma_{i=y=1}X_iX_i^{'} + \Sigma_{i=y=-1}X_iX_i^{'} - (c_+ + c_-)n_+c_+^{'} - (c_+ + c_-)n_-c_-^{'}) = n(c_+ - c_-)$$

$$\beta(\Sigma_{i=y=1}X_iX_i^{'} + \Sigma_{i=y=-1}X_iX_i^{'} - c_+n_+c_+^{'} + c_-n_-c_-^{'} - c_+n_-c_+^{'} + c_-n_+c_-^{'}) = n(c_+ - c_-)$$

$$\beta(\Sigma_{i=y=1}X_iX_i^{'} + \Sigma_{i=y=-1}X_iX_i^{'} - n_+c_+c_+^{'} + n_+c_-c_+^{'} - n_-c_+c_-^{'} + n_-c_-c_-^{'}) = n(c_+ - c_-)$$

$$\beta(S_W + n_+c_-c_+^{'} - n_-c_+c_-^{'}) = n(c_+ - c_-)$$

el cual se expresa de forma equivalente como:

$$\beta(S_W + n_+n_-(c_+ - c_-)(c_+ - c_-)^{'}/n) = n(c_+ - c_-)$$

1.4 Solución inciso "d"

Se parte de lo siguiente

$$(S_W + \frac{n_+n_-S_B}{n})\beta = n(c_+ - c_-)$$

Se sustituye y se obtiene

$$(S_W + \frac{n_+n_-(c_+ - c-)(c_+ - c-)^{'}}{n})\beta = n(c_+ - c_-)$$

por el resultado del inciso "a", reducimos el término de la izquierda

$$S_W\beta + \frac{n_+n_-(c_+ - c-)(c_+ - c-)^{'}\beta =}{n} = n(c_+ - c_-)$$

$$S_W\beta + \frac{n_+n_-(c_+ - c-)n}{n} = n(c_+ - c_-)$$

$$S_W\beta + n_+n_-(c_+ - c-) = n(c_+ - c_-)$$

despejando

$$S_W\beta = n(c_+ - c_-) - n_+n_-(c_+ - c-)$$

$$S_W\beta = (c_+ - c_-)(n - n_+n_-)$$

multiplicando ambos lados de la ecuación por S_W^{-1}

$$\beta = S_W^{-1}(c_+ - c_-)(n - n_+n_-)$$

donde $(n - n_+n_-)$ es una constante, y es por eso que se tiene que β aproxima al siguiente resultado:

$$\beta \approx S_W^{-1}(c_+ - c_-)$$

1.5 Solución inciso "e" y "f"

Para ilustrar el método se generan dos conjuntos de datos de una muestra de 100 observaciones; el primer conjunto es una mezcla de gaussianas con $\sigma = 1$ y $\mu = 2$; el segundo conjunto de observaciones tiene de parámetros $\sigma = 1$ y $\mu = 6$.

En el inciso "f", se menciona que el método FDA no detecta datos extremos y es necesario pesar las observaciones, los pesos se deciden asignando a cada observación su inverso proporcional de la distancia euclidiana al cuadrado respecto a su centroide, esto le dará importancia a las observaciones con distancia lejanas a su centroide.

$$d = \sqrt{(\Sigma_{i=1}^{n}(\mu - X_i)^2)}$$
$$w = 1/d$$

En la figura 1.1, se presenta el primer conjunto de datos, se realiza mínimos cuadrados con la función **lm** en R para ajustar el modelo y generar el hiperplano separador. La figura del centro es el hiperplano al utilizar FDA, se observa que dado a que las observaciones son linealmente separables, se obtiene buena clasificación en las observaciones. En la figura derecha, se utilizan los pesos de la forma ya mencionada, en este caso no se presentan diferencias interesantes, debido a que el conjunto de datos no cuenta valores extremos respecto a sus centroides.

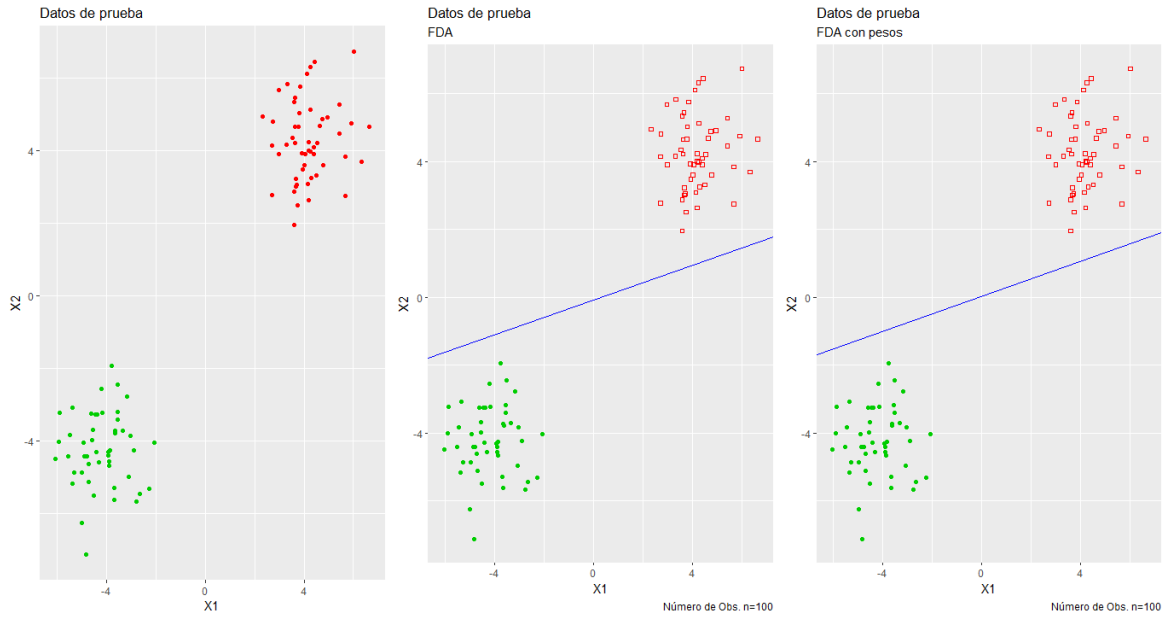

Figure 1.1: Ajuste con FDA para Datos de prueba

En la figura 1.2, se presentan los resultados para el segundo conjunto de datos, los cuales se observan más separados de sus centroides con algunas observaciones que se mezclan entre clases. En este caso, FDA falla, trazando un hiperplano que cruza las observaciones; pesando las observaciones como se realizó con anterioridad, los resultados siguen siendo malos.

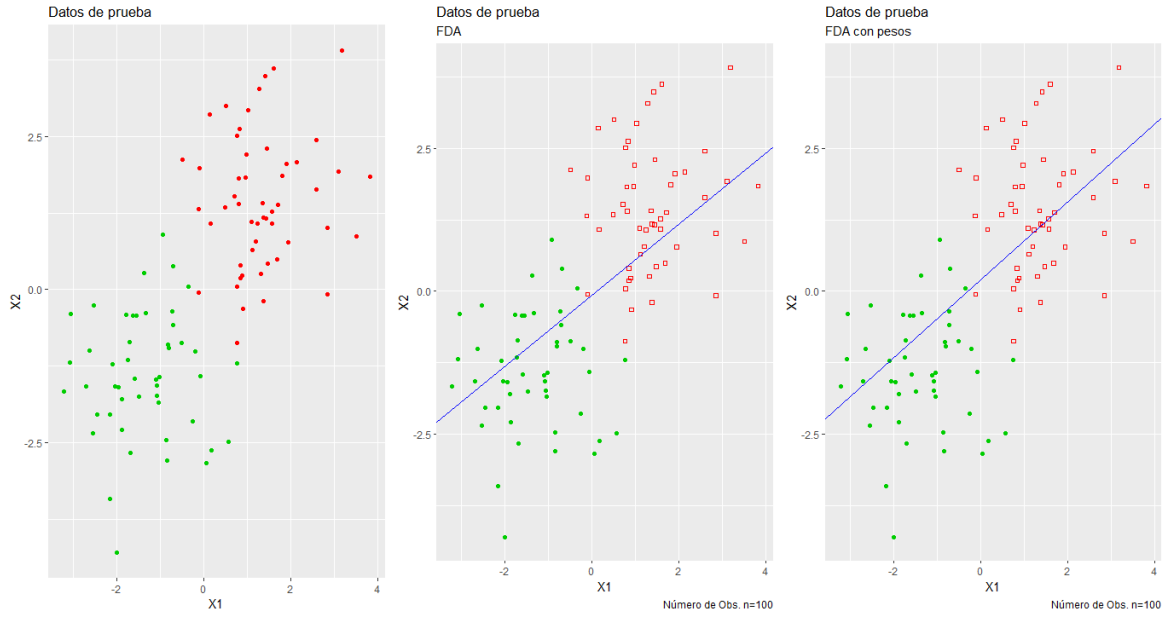

Figure 1.2: Ajuste con FDA para Datos de prueba

En conclusión el último conjunto de datos bajo FDA no es linealmente separable, y no es

robusto bajo datos atípicos; asimismo, los pesos que se les dan no contribuyen de manera significativa, pero la idea de dar pesos inversamente proporcional a la distancia de sus centroides, parece ser una buena medida de asignar importancia a las observaciones atípicas.

2 Problema

Este ejercicio es sobre el método de clasificación binaria perceptron.

a) Implementa el modelo clásico de perceptrón (versión que trabaja en línea). Aplícalo primero a un conjunto de datos artificiales en dos dimensiones y con dos categorias. Discute tus resultados. Incluye gráficas del ajuste.

b) Aplica el clasificador al conjunto de datos `pima` que vimos en clase. Usa `pima.tr` para ajustar el modelo y `pima.te` para verificar su calidad predictiva. ¿ Qué puedes decir sobre su desempeño? Comenta tus hallazgos.

2.1 Solución inciso "a"

Para realizar este ejercicio el perceptrón se programa en base al pseudocódigo que se obtiene en el libro de Cristianini N y Shawe-Taylor J(2000)[1]. Se generan dos conjuntos de datos en dos dimensiones para probar el perceptrón; el primer conjunto es una mezcla de normales con $\sigma = 1$ y $\mu = 5$; el segundo cuenta con los parámetros $\sigma = 1$ y $\mu = 2$[2].

En la figura 2.1, se presenta el primer conjunto de datos, en la figura derecha se presenta el hiperplano separador que realiza el perceptrón, se observa que no se tiene problema alguno para separar las observaciones por sus respectivas regiones.

[1] El presudocódigo se encuentra en la sección Presudocódigo de este ejercicio

[2] Nota: se le deja al usuario el código en el archivo **Tarea 5 Ejercicio 2**, la opción de jugar con el perceptrón y realizar cambios en los parámetros de la distribución de los datos sintéticos.

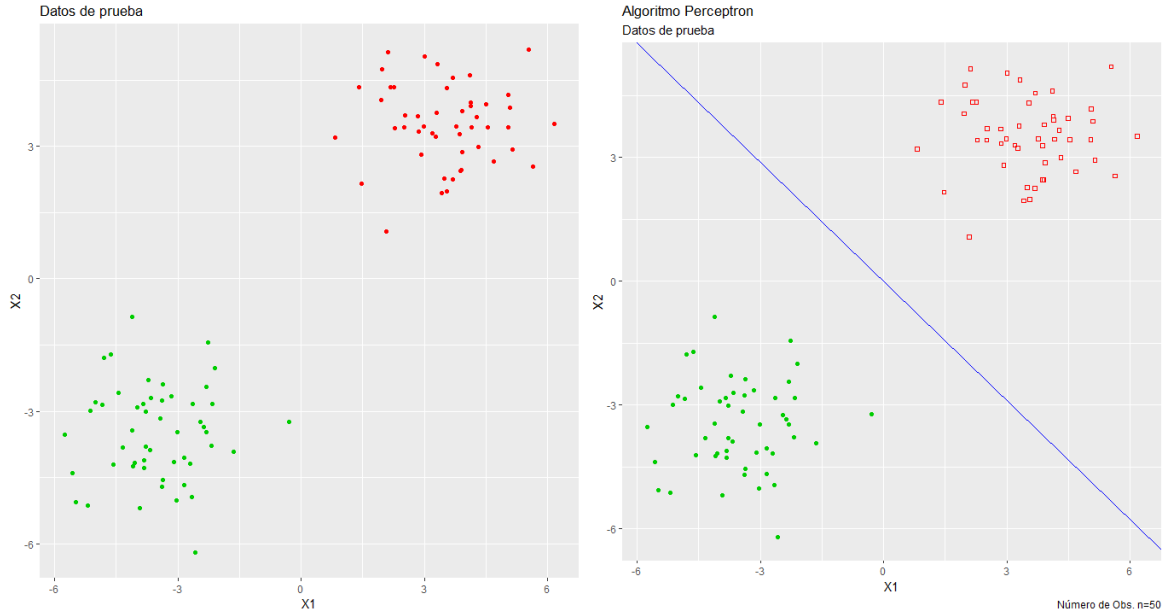

Figure 2.1: Perceptrón para Datos de prueba

En la figura 2.2, se presenta el segundo conjunto de datos, en este se pretende tener observaciones que compliquen su separación de forma lineal; no obstante, en la figura derecha se observa que el hiperplano separador, realiza bien su función, solo dejando a tres observaciones en clases a las cuales no pertenecen.

Figure 2.2: Perceptrón para Datos de prueba

En conclusión, la versión más sencilla de una red neuronal, al generar el hiperplano separador,

logra reasignar a las observaciones en sus regiones que petenecen de forma correcta, siempre y cuando el conjunto de datos sean lineal mente separables, ya que de no ser así el algoritmo no converge.

2.2 Solución inciso "b"

Se aplica el clasificador al conjunto de datos **pima**, se entrena el peceptrón con los datos **pima.tr** y se verifica su calidad predictiva con los datos del archivo **pima.te**. Se utiliza el algoritmo implementado en el inciso anterior para obtener el hiperplano separador y entrenar el modelo. Los datos de entrenamiento se escalan para homogeneizar la unidad de medida, y como consecuencia el sesgo del modelo es cero; una vez implementando el perceptrón, realizamos PCA sobre las observaciones de entrenamiento para obtener una representación visual del las observaciones y el hiperplano separador.

En la figura 2.3, se observa del lado izquierdo el **screeplot**, el cual sugiere utilizar las dos primeras componentes. En el lado derecho, se observan las proyecciones de los datos en las dos primeras componentes. El problema de este conjunto de datos se presenta en que no parecen ser linealmente separables respecto a las dos clases que lo conforman (Yes:1, No:-1), con consecuencia de que el perceptrón presenta problemas para converger.

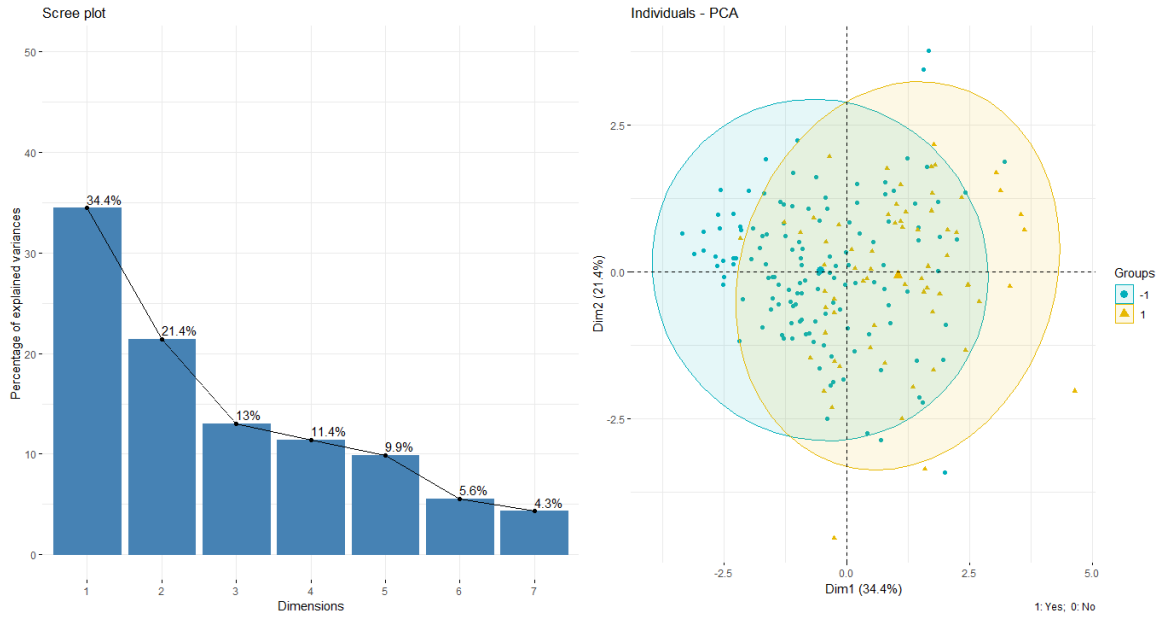

Figure 2.3: Resultado PCA en las observaciones de entrenamiento

Previo a evaluar la eficiencia predictiva, se toma el espacio de las dos componentes y se traza el hiperplano separador obtenido con los datos de entrenamiento. En la figura 2.4, se observan los resultados, vemos que no logra realizar la separación de las observaciones por sus regiones correspondientes, esto por que el conjunto de datos no son del todo linealmente separables.

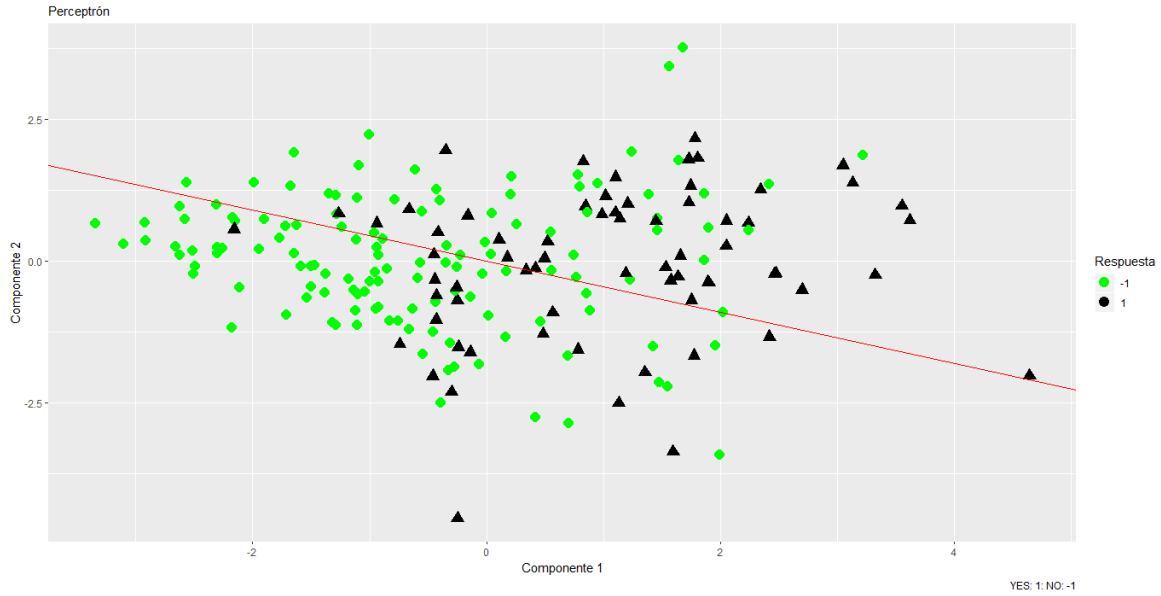

Figure 2.4: Perceptrón para datos de entrenamiento

Se utilizan los coeficientes obtenidos de los datos de entrenamiento con el algoritmo perceptrón $\beta = [\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7]'$, y se evalúa el ajuste del modelo con los datos de prueba, dando el mismo trato que al que se les da a los de entrenamiento.

$$g = X_{prueba}\beta$$

Al tener los valores de g, se utiliza la función de activación signo para obtener $\hat{y} = sign(g(x))$, y poder contrastar contra las clases originales de las observaciones de prueba. Para ver la eficiencia en la predicción se realiza una tabla de confusión. En la tabla 2.1, se presentan que tan bien predice el modelo perceptrón, las filas son las etiquetas originales en las respuestas y las columnas son las predicciones que se obtienen al evaluar con la función de activación los resultados del ajuste del perceptrón a los datos de prueba; se puede observar, que las respuestas originales YES son 109, y con el algoritmo perceptrón se predicen 155; las respuestas originales NO son 223, con el algoritmo perceptrón se predicen 177

	NO-predicción	YES-predicción	Total
NO-original	158	65	223
YES-original	19	90	109
Total	177	155	332

Table 2.1: Tabla de confusión

En conclusión, el algoritmo perceptrón es eficiente cuando el conjunto de datos es linealmente separable, esto sucede si las clases en las observación no suelen traslapare; asimismo, en base a la tabla de confusión, el modelo ajustado del perceoptrón no arroja resultados tan equivocados.

2.3 Pseudocódigo

Este es el pseudocódigo que se presenta en el libro de Cristianini N y Shawe-Taylor J(2000), se toma la imagen para respetar su notación.

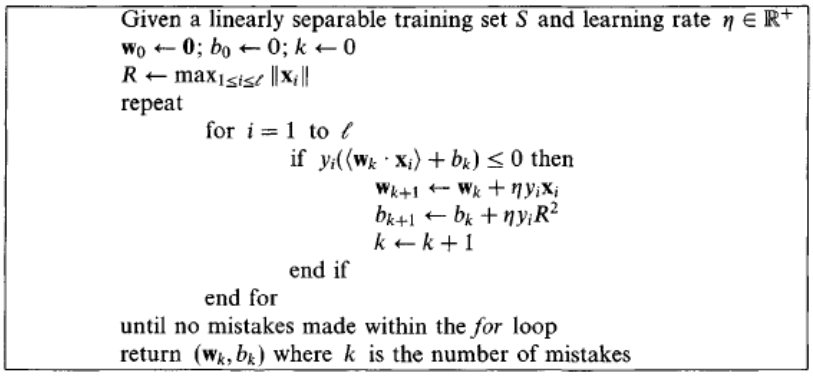

Figure 2.5: Pseudocódigo Perceptron versión en linea de Cristianini N y Shawe-Taylor J(2000)

2.4 Referencia

Cristianini N y Shawe-Taylor J(2000). An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press; 2000.

3 Problema

Los archivos contenidos en las carpetas `email_train` e `email_test` corresponden a correos electrónicos en inglés clasificados como Spam y No-Spam.

a) Implementa clasificadores de Spam usando regresión logística, LDA, QDA y FDA. Usa los datos `email_train` e `email_test` para ajustar y probar los métodos, respectívamente. Compara su desempeño.

b) Las curvas ROC (Receiver Operating Characteristics) es un método muy común para comparar algoritmos de clasificación binarios basado en la tabla de errores (falsos positivos y falsos negativos) que se cometen. Usa los resultados del inciso anterior para comparar los clasificadores usando este criterio. ¿ Cuál método elegirias? Usa el criterio del área bajo la curva (AUC).

3.1 Solución inciso "a"

En este ejercicio se implementa la librería **tm.pluging.mail**, para la lectura de los archivos. Se utilizan diversas funciones en el procesamiento de texto de la librería **tm**. El proceso de la limpieza en los e-mails consiste en eliminar espacios en blanco, números, *stop words*, *stems* y convertir palabras en mayúsculas a minúsculas; ya por último, una vez que el corpus de

textos se encuentre "limpio", se obtiene la matriz de términos.

Lo anterior se realiza a la base de entrenamiento y a la de prueba; asimismo, se decide utilizar los 25 términos más frecuentes para la construcción del clasificador. Las palabras que se seleccionan para el entrenamiento del modelo se utilizan con los datos de prueba; i.e., a raíz de que el modelo se entrena con esos términos, los coeficientes que se estiman correspondes a esas covariables, por lo tanto, es necesario que se tengan las mismas variables (términos) de respuesta para la base de prueba.

La figura 3.1, muestra las 25 palabras que aparecen con más frecuencia en los e-mails, siendo las que se utilizan para generar el clasificador de mensajes que son Spam.

Figure 3.1: 25 palabras más frecuentes

Se ajustan cuatro modelos; Logístico, FDA, LDA y LDQ. La tabla 3.1, muestra el desempeño del entrenamiento en los modelos[3], el cual se evalúa por medio del criterio AIC, la tasa de precision del modelo y su respectiva tasa de error al realizar el ajuste en las observaciones, cabe mencionar que el modelo LDQ no reporta el estadístico AIC; de esta manera, se observa que el modelo Logístico tiene mejor tasa de precisión en los datos de entrenamiento y por ende la menor tasa de error, seguido del clasificador LDA.

Modelos	AIC	Tasa de Precisión	Tasa de Error
FDA	2437.415	37%	63%
Logit	1915.4	83%	17%
LDA	1915.4	80%	20%
LDQ	-	60%	40%

Table 3.1: Desempeño de los clasificadores; Datos de Entrenamiento

Después de entrena el modelo, se utilizan los datos de prueba para probar la eficiencia en

[3]Para no insertar tablas largas, no se presentan las salidas de las regresiones; no obstante, en el archivo **Tarea 5 Ejercicio 3**, se le deja al lector el código con los coeficientes.

el clasificador. En la tabla 3.2, se mide el desempeño de los clasificadores en base a la tasa de precisión y la tasa de error en la clasificación. Se observa que el clasificador con mayor precisión para asignar la categoría correcta a los datos de prueba fue el LDQ, con tasa de precisión del 73% y tasa de error del 25%; respecto a los otros clasificadores, su tasa de error es cercana al 50%, provocando equivocación de forma constante en la detección de los mensajes Spam.

Modelos	Tasa de Precisión	Tasa de Error
FDA	51%	49%
Logit	57%	43%
LDA	53%	47%
LDQ	73%	27%

Table 3.2: Desempeño de los clasificadores; Datos de Prueba

Para evaluar el desempeño individual de cada modelo y argumentar la tasa de error mostrado previamente, se construyen tablas de confusión para cada clasificador utilizando los datos de prueba. Se debe mencionar que las filas son las etiquetas originales (Spam: 1, No-Spam:0), y las columnas son las asignaciones realizadas por cada clasificador.

Los siguientes cuadros muestra la tabla de confusión de los clasificadores; en estas tablas lo que uno espera son representaciones simétricas en la estructura. Para implementar FDA, se utiliza la función **lm** en R; se observa en la tabla 3.3, mal desempeño en la clasificación de los datos de prueba utilizando FDA, de 500 e-mails etiquetados como Spam y 500 como No-Spam, FDA clasifica 4 como No-Spam, y 996 como Spam; la tasa de error de este clasificador es del 49%.

	No-Spam	Spam
No-Spam	4	496
Spam	0	500

Table 3.3: Tabla de confusión; clasificador FDA datos de Prueba

En la tabla 3.4, se presenta la tabla de confusión del clasificador logístico utilizando los datos de prueba; con este clasificador se tiene una tasa de error del 43%, de 500 e-mails etiquetados como Spam y 500 como No-Spam, clasifica 123, como Spam y 877 como No-spam.

	No-Spam	Spam
No-Spam	472	28
Spam	405	95

Table 3.4: Tabla de confusión; clasificador Logit datos de Prueba

En la tabla 3.6, se presenta la tabla de confusión del clasificador LDA utilizando los datos de prueba; su tasa de error es del 47%, de 500 e-mails etiquetados como Spam y 500 como No-Spam, clasifica 97 mensajes como Spam y 903 como No-spam.

	No-Spam	Spam
No-Spam	464	36
Spam	439	61

Table 3.5: Tabla de confusión; clasificador LDA datos de Prueba

La figura 3.2, se presenta de forma gráfica la tabla de confusión, puntos negros son e-mails etiquetados como No-Spam, puntos rojos son etiquetados como Spam. Se puede observar que los puntos que se posicionan en el valor uno del eje de las ordenadas, tendrían que ser aquellos que son Spam, no obstante el clasificador manda e-mails No-Spam como Spam.

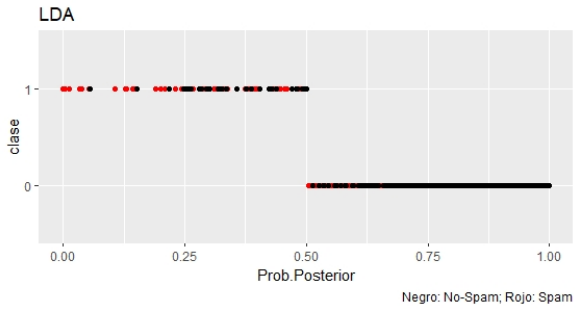

Figure 3.2: Clasificación datos de prueba LDA

En la tabla 3.7, se muestra la tabla de confusión del clasificador LDQ utilizando los datos de prueba; su tasa de error es del 27%, de 500 e-mails etiquetados como Spam y 500 como No-Spam, clasifica 687, como Spam y 313 como No-spam.

	No-Spam	Spam
No-Spam	275	225
Spam	38	462

Table 3.6: Tabla de confusión; clasificador LDQ datos de Prueba

En la figura 3.3, se presenta de forma visual la clasificación de los e-mails con el modelo LDQ. Se puede observar, que tiende a clasificar e-mails Spam como No-Spam, pero en menor medida que el método LDA.

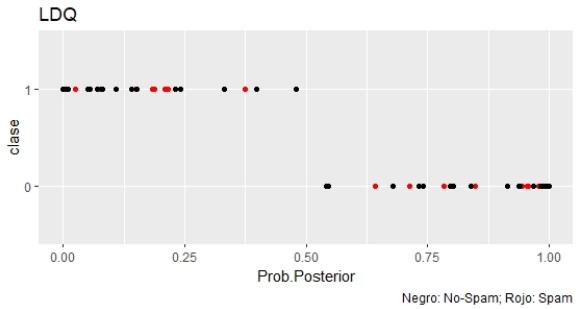

Figure 3.3: Clasificación datos de prueba LDQ

En conclusión, bajo los criterios de la tasa de error y la tabla de confusión, el clasificador LDQ, realiza mejor clasificación en los datos de prueba respecto a los demás.

3.2 Solución inciso "b"

En este ejercicio se utilizan las curvas ROC para comparar los algoritmos de clasificación binarios utilizados en el inciso anterior en base a la tabla de errores conocidos como; **falsos positivos, y falsos negativos**; asimismo, se utiliza el criterio del área bajo la curva para la elección del mejor clasificador binario.

A manera general, los gráficos de la curva ROC tienen en su eje de las ordenadas el porcentaje de verdaderos positivos; i.e., las veces que el clasificador etiqueta como Spam y la etiqueta es Spam. En el eje de las abscisas se encuentra el porcentaje de falsos positivos; i.e., el porcentaje de veces que el clasificador etiqueta de la forma incorrecta. Lo que se espera de la curva ROC, es que la linea del gráfico se aleje lo más posible de la line recta que sale del origen; también lo que uno desea es una curva que tienda a uno lo más rápido posible; i.e., que la curva crezca y se estabilice en el porcentaje uno, en una cantidad baja de falsos negativos.

La figura 3.4, presenta la curva ROC de los clasificadores implementados en el inciso anterior. De forma implícita uno recupera la tasa de error observando el corte donde la curva tiene a estabilizarse y comienza a converger a uno.

En la figura superior izquierda se encuentra la curva ROC del clasificador FDA, el cual tiende a aproximarse a la linea horizontal; a su vez, se observa que despues del 50% de falsos positivos, mejora sus porcentajes de verdaderos positivos. La figura superior derecha es la curva ROC del clasificador logístico, la cual se separa de la linea horizontal, y después del 30% de la tasa de falsos positivos mejora en el porcentaje de los verdaderos positivos. La curva ROC del clasificador LDA se presenta en la figura inferior izquierda, cuya representación es la misma que la del modelo FDA, con porcentaje de falsos positivos que son prácticamente los mismos al modelo FDA. En la figura inferior derecha se observa la curva ROC del clasificador LDQ; su desempeño se asemeja al del clasificador logístico, con una curva que se separa de la linea horizontal y porcentajes de verdaderos positivos que incrementan a niveles de falsos negativos pequeños.

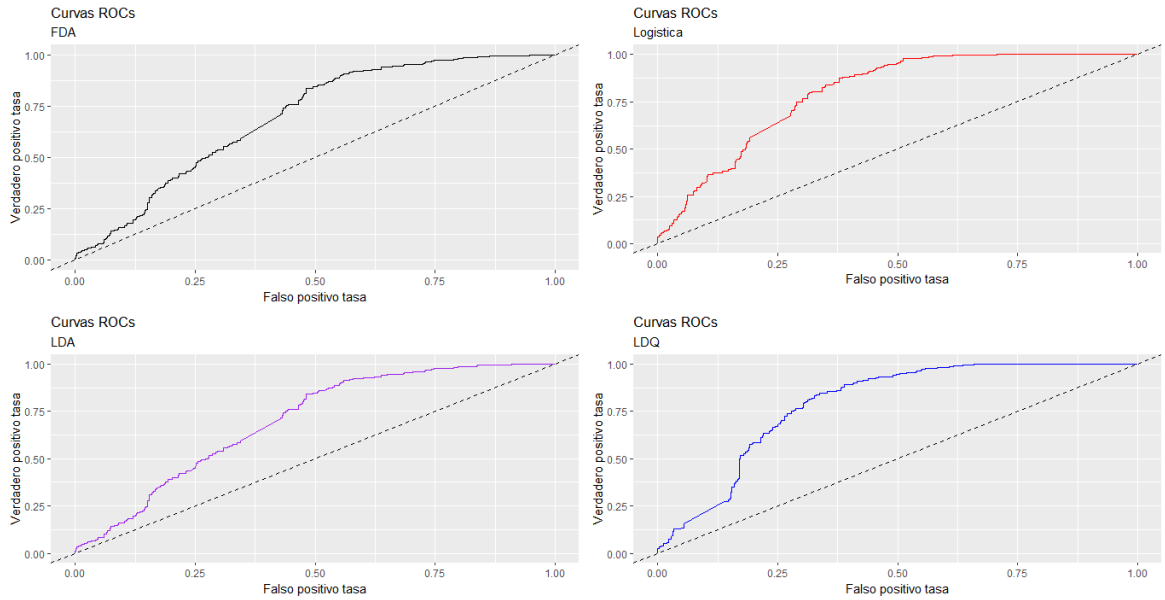

Figure 3.4: Curvas ROC's

La figura 3.5, se presentan la curva ROC de los clasificadores logístico, LDA, LDQ[4] que se observaron en la figura 3.4, pero ahora en un sólo gráfico; la linea roja es la curva ROC del clasificador logístico, la morada del LDA, y el azul del LDQ; bajo esta representación se aprecia que curva se aleja más de la linea horizontal, que en este caso es la del clasificador LDQ; no obstante , se puede observar que al inicio de la curva ROC de LDQ, se marca un cambio de pendiente, lo cual tendrá efecto en el criterio del área bajo la curva.

[4]La curva ROC del clasificador FDA se traslapa con la del LDA.

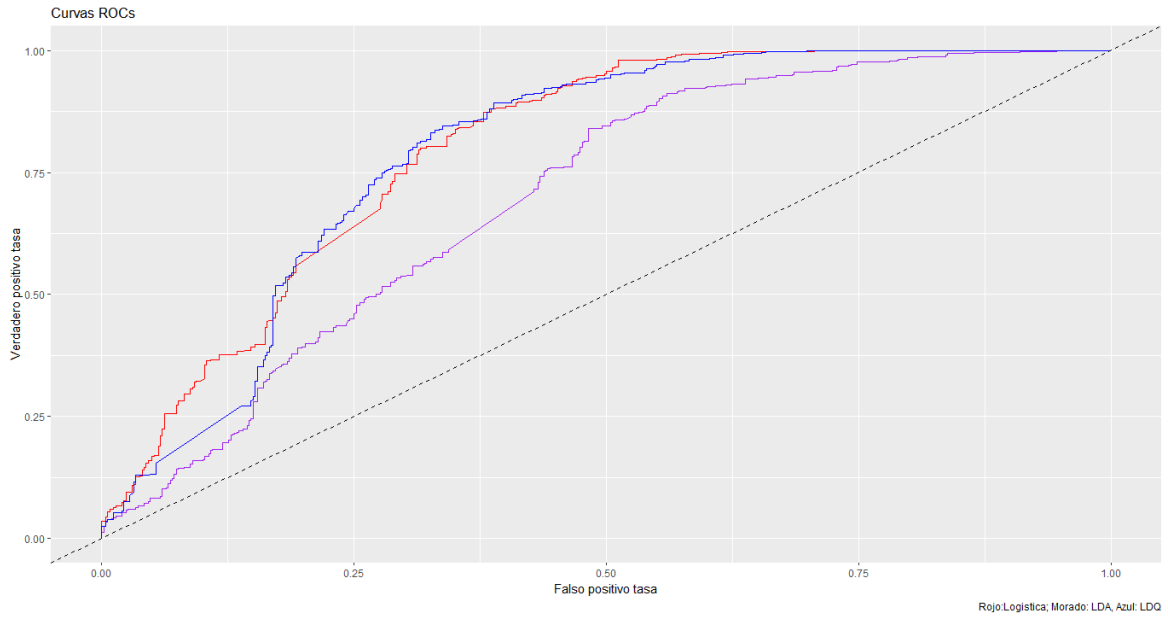

Figure 3.5: Curvas ROC's

Lo anterior da información sobre que clasificador es mejor en la clasificación de e-mails Spam; no obstante, en este ejercicio se utiliza el criterio del área bajo la curva para seleccionar el mejor método. Se utiliza la función **auc** de la librería **ROC** en R, para calcular el valor del área de la curva de la curva ROC de cada clasificador. En la tabla 3.7, se presenta el área bajo la curva (AUC) para cada clasificador, cabe mencionar que el mejor resultado es una área cercana a uno. El clasificador FDA y LDA reportan el mismo AUC de 0.6981; el clasificador logístico reporta un AUC de 0.7977 ; el AUC de modelo LDQ es de 0.7896.

Modelos	Área bajo la curva
FDA	0.6981
Logit	0.7977
LDA	0.6981
LDQ	0.7896

Table 3.7: Desempeño de los clasificadores; Datos de Prueba

En conclusión, el mejor modelo en base al AUC es el logístico y el LDQ, pero dado a las tasas de error que presentan y al porcentaje de verdaderos positivos que muestra la curva ROC, se concluye que el clasificador LDQ es mejor para clasificar los e-mails que son Spam[5].

[5]Se tiene presente que al incrementar el número de variables; i.e., términos, los resultados pueden mejorar para todos los clasificadores; no obstante, para realizar un reporte conciso y corto, solo se muestran los resultados utilizando los 25 términos más frecuentes en el corpus; sin embargo, en el archivo **Tarea 5 Ejercicio 3.R**, se trabaja con más términos, el cual queda a disposición para realizar ajustes en el número de términos.