Tarea 5

Rodrigo Carvajal Pizarro (rcarvaja@astro.puc.cl)
https://github.com/racarvajal/Tarea5_Astro_Stats.git

Resolución

Problema 1

Los datos a clasificar, sin ninguna modificación, pueden verse en la Figura 1

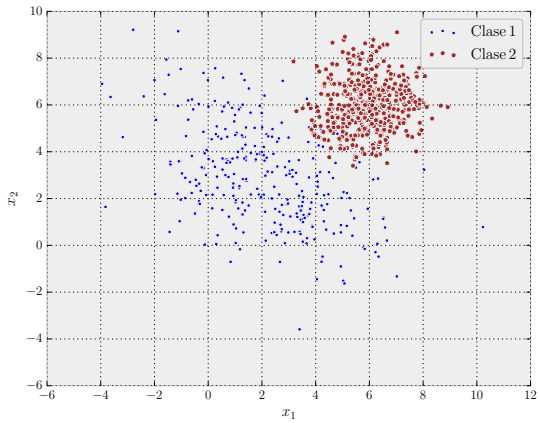

Figura 1: Datos entregados

a) Para implementar la regresión lineal para clasificar los datos, se busca una recta (o un plano, al tener más de una dimensión) que minimice la distancia con los puntos escogidos. Esta minimización se realizó con el método de los mínimos cuadrados y se llega a los parámetros para el plano. Estos parámetros son:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Con

$$\beta_0 = 0{,}58555898,\ \beta_1 = 0{,}1159599,\ \beta_2 = 0{,}10686815$$

Como la pertenencia a una u otra clase puede expresarse numéricamente (1 o 2), es sencillo mostrar la línea divisoria entre ambas clases a partir del plano obtenido con la regresión lineal. Se determina un valor que da cuentra de la división (1,5) y se convierte la ecuación del plano en una ecuación de una recta.

$$\hat{y} = 1{,}5 = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Esta ecuación puede graficarse junto con los puntos a clasificar y puede verse en la figura 2.

Para utilizar el método de Linear Discriminant Analysis, se utiliza un discriminante lineal, tal como se mostró en clases:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

En que Σ corresponde a la matriz de covarianza (que será constante para todas las clases), μ_k es el valor medio de los datos de cada clase y π_k corresponde al prior de cada clase.

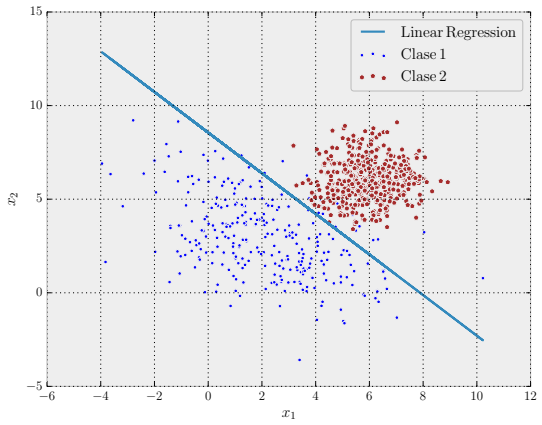

Figura 2: Datos entregados junto con la frontera de decisión de la regresión lineal

Se calcula $\delta_k(x)$ para cada clase y se comparan sus valores. La frontera de decisión se encuentra en la zona en que los discriminantes se igualan ($\delta_k(x) = \delta_l(x)$). El resultado de este método se muestra en la Figura 3.

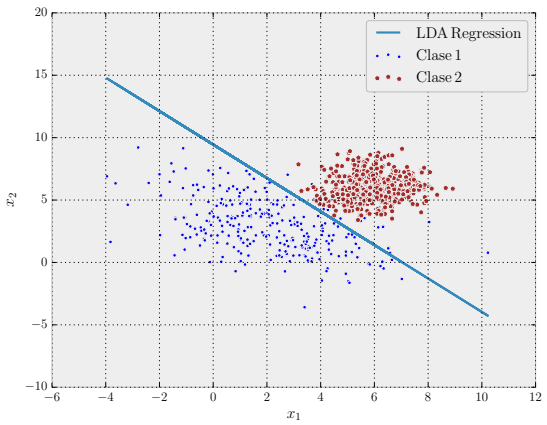

Figura 3: Datos entregados junto con la frontera de decisión del método LDA (Linear Discriminant Analysis)

Por último, se realiza la clasificación utilizando el método Quadratic Discriminant Analysis (QDA) que se basa en la misma idea de LDA, pero ahora cada clase tiene una matriz de covarianza diferente.

$$\delta_k(x) = -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k + \log\pi_k + \log|\Sigma_k|$$

Ahora, el discriminante será cuadrático en x por lo que la frontera de decisición corresponderá a una curva (elipse, parábola, etc.). Esto puede verse en la Figura 4.

A fin de comparar los tres métodos y notar que sus resultados son diferentes, se graficaron las tres fronteras de decisión juntas en la Figura 5.

Si solo se consideran los puntos a clasificar y sin ver las fronteras de decisión ya obtenidas, puede esperarse que el método que mejor toma en cuenta los puntos es QDA. La distribución de los puntos indica que

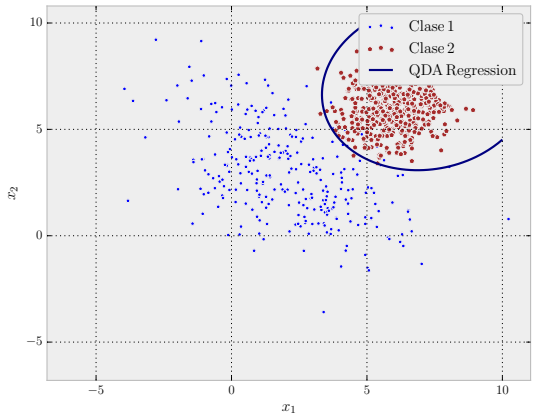

Figura 4: Datos entregados junto con la frontera de decisión del método QDA (Quadratic Discriminant Analysis)

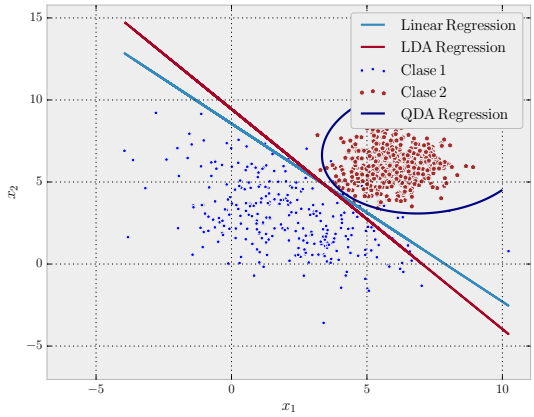

Figura 5: Datos entregados junto con las fronteras de decisión de Regresión Lineal, LDA y QDA

debería haber una frontera que, en el mejor de los casos, sea curva. Entre los tres métodos, aquel que entrega una frontera curva es QDA.

b) Además de los métodos de Regresión Lineal, Linear Discriminant Analysis (LDA) y Quadratic Discriminant Analysis (QDA), se analizarán los métodos de Nearest-Neighbor y AdaBoost.

Al utilizar el paquete `scikit-learn` los parámetros necesarios para ejecutar cada una de las clasificaciones son menos y más sencillos. Para el método de Nearest-neighbors, se requiere un número de vecinos a considerar. En este caso, se utilizaron 20 vecinos. Con AdaBoost, se requiere, como parámetro, el número de estimadores. Para este cálculo, se utilizaron 100. Con estos parámetros, los resultados pueden verse en las Figuras 6 y 7.

Con los cinco métodos utilizados, pueden construirse matrices de confusión. Estas matrices permiten conocer el grado de asertividad de cada uno de los métodos, es decir, cuán eficientes son en clasificar nuevos datos a partir de lo aprendido con datos de entrenamiento.

El paquete `scikit-learn` cuenta con un método para obtener matrices de confusión para cada método una vez que se han dividido los datos entre subconjuntos de entrenamiento y de validación. Para comparar los resultados, los métodos de Regresión Lineal, LDA y QDA fueron reimplementados con `scikit-learn`.

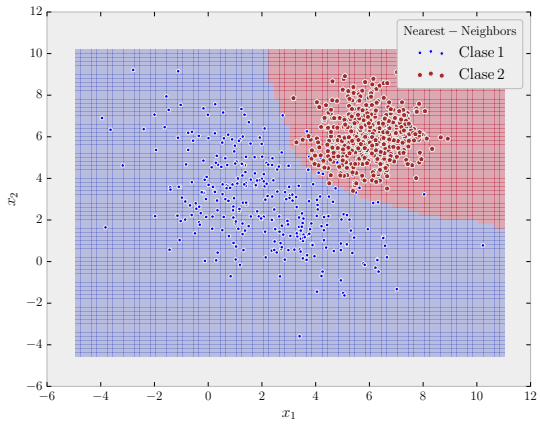

Figura 6: Datos entregados junto con la clasificación de Nearest-Neighbors (20 vecinos)

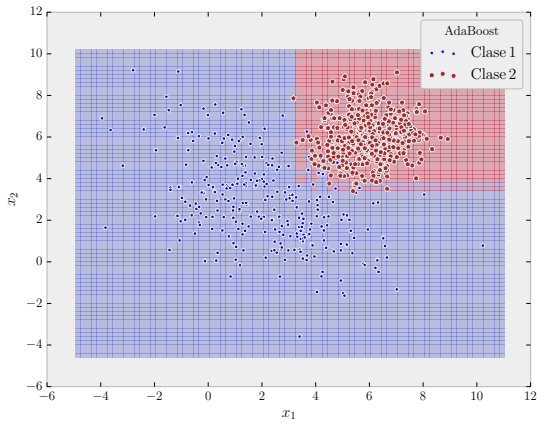

Figura 7: Datos entregados junto con la clasificación de AdaBoost (n estimators = 100)

Se logró calcular las matrices de confusión para cuatro de los métodos. Para regresión lineal, no fue posible implementar este cálculo. Pero se puede asumir que sus resultados son similares a los de Linear Discriminant Analysis. De este modo, las matrices (promediadas sobre 300 ejecuciones con diferentes conjuntos de prueba) son las siguientes.

$$\mathbb{M}_{\text{Nearest}-\text{Neighbors}} = \begin{pmatrix} 70{,}016 & 4{,}615 \\ 1{,}003 & 125{,}21 \end{pmatrix}$$

$$\mathbb{M}_{\text{AdaBoost}} = \begin{pmatrix} 71{,}812 & 2{,}966 \\ 2{,}096 & 123{,}123 \end{pmatrix}$$

$$\mathbb{M}_{\text{LDA}} = \begin{pmatrix} 69{,}555 & 5{,}224 \\ 0{,}023 & 125{,}197 \end{pmatrix}$$

$$\mathbb{M}_{\text{QDA}} = \begin{pmatrix} 71{,}337 & 3{,}441 \\ 0{,}531 & 124{,}688 \end{pmatrix}$$

Al analizar una matriz de confusión, se busca que la mayor parte de los valores se encuentre en la diagonal. Esto significa que los puntos han sido clasificados como lo que, realmente, son. Si hay valores fuera de la diagonal, hay puntos que han sido clasificados erróneamente (confusión).

Para poder comparar estas matrices, se recurre a otro valor, el *Misclassification Rate*, que indica la fracción de elementos que fueron clasificados erróneamente sobre todas las clasificaciones utilizando un método de los ya mencionados. Para obtener valores adecuados, también se ejecutó la clasificación 300 veces para cada método y los resultados son los siguientes.

$$\text{Misclassification rate}_{\text{Nearest}-\text{Neighbors}} = 0{,}02312$$
$$\text{Misclassification rate}_{\text{AdaBoost}} = 0{,}02531$$
$$\text{Misclassification rate}_{\text{LDA}} = 0{,}02623$$
$$\text{Misclassification rate}_{\text{QDA}} = 0{,}01986$$

Se busca que la tasa de mala clasificación sea mínima con un método. En este caso, el método que menor valor entrega de la tasa de mala clasificación es QDA con un valor = 0,01986. Esto significa que se equivoca en la clasificación un 1,9 % de las veces.

Con lo anterior, se puede decir que, considerando la estructura de los datos, el mejor clasificador es Quadratic Discriminant Analysis (QDA). Este resultado es congruente con lo expresado anteriormente en esta tarea en que se busca el mejor clasificador sobre la base de las frontera de decisión que entregará.

c) Ahora, ya conociendo los valores reales de las matrices de covarianza y los valores medios de las distribuciones que generan los puntos de las Clases 1 y 2, se puede estudiar la frontera de decisión entre ambos conjuntos.

Lo primero que puede hacerse es graficar cada una de las distribuciones. Mediante un gráfico de isocurvas tal como el de la Figura 8

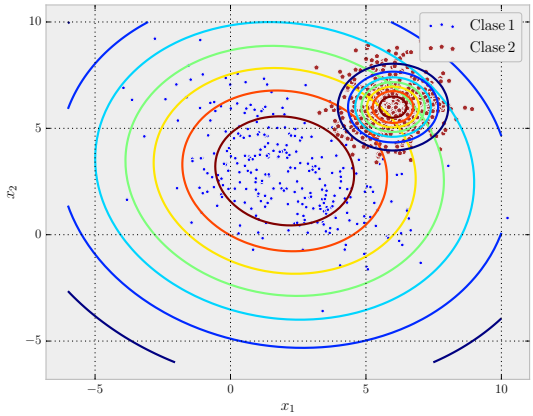

Figura 8: Datos entregados junto con las isocurvas de ambas distribuciones

Si, ahora, se grafica la zona en que ambas distribuciones se igualan, se tendrá el clasificador de Bayes correspondiente a los puntos entregados. Esto puede notarse en la Figura 9.

Lo primero que puede notarse es la similitud entre la *verdad* de la frontera de decisión y aquella generada a través de QDA. Esto puede explicarse, entre otras razones, por la simpleza de ambas distribuciones. Las dos corresponden a Normales Multivariadas y no a otros tipos más complejos de funciones. La intersección entre ambas funciones tendrá la forma de una elipse, algo similar a lo que entrega un método como QDA.

De este modo, se confirma que el mejor método para clasificar los datos entregados en esta tarea es el de QDA que entrega una frontera de decisión curva muy similar a la original.

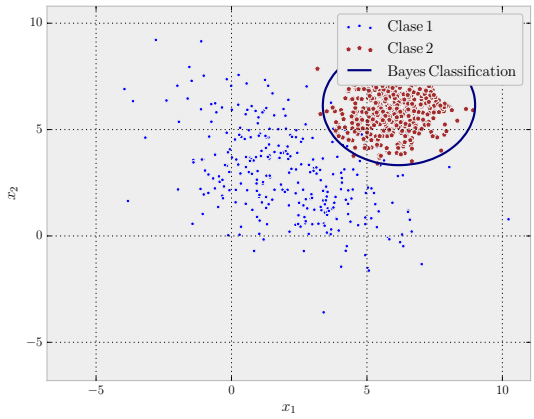

Figura 9: Datos entregados junto con el casificador de Bayes asumiendo como verdaderas las distribuciones entregadas