1. La base de datos para regresión es tomada de IRIS, un conjunto de base de datos reales de acceso abierto que puede ser usado como problemas juguete (toy problems) con fines académicos o evaluativos. La base de datos utilizada para el problema de regresión cuenta con 8 características y 1030 muestras, aunque solo se trabaja con 6 de esas características.

Se utiliza el 70% del total de las muestras para el entrenamiento y el 30% restante para la evaluación, por lo tanto, 721 muestras corresponden al conjunto de entrenamiento y 309 muestras al conjunto de evaluación. Las muestras de ambos conjuntos se randomizan con la funcion de matlab randperm que randomiza los indices de la matriz de muestras original.

Al conjunto de entrenamiento se le aplica una normalización zscore dejando una distribucion de datos de media 0 y desviación 1. Por último, se obtienen la media y la desviación utilizada para el centrado y el escalado del conjunto de entrenamiento y se usan estos valores para normalizar el conjunto de evaluación.

**3.** Se puede notar que con 100 vecinos el ECM es menor que cuando hay menos vecinos, esto se debe a que entre más grande sea el k, mas robusta es la clasificación ante el ruido. Aunque hay un limite, un k muy grande puede llevar a que se etiqueten nuevas muestras a la clase con más muestras en el conjunto de entrenamiento. Con un k moderadamente alto hay mas criterios para asignar una muestra nueva a una clase, ya que cada vecino “da un voto” de la clase a la que debería pertenecer la nueva muestra x\*

**4.** La base de datos para clasificación es tomada de IRIS, un conjunto de base de datos reales de acceso abierto que pueden ser usadas como problemas juguete (toy problems) con fines académicos o evaluativos. Esta base de datos en particular contiene 150 muestras, 4 características y 3 clases definidas (en el vector Yclass). Para este caso se utilizaron únicamente las 3 primeras características.

Se utiliza el 70% del total de las muestras para el entrenamiento y el 30% restante para la evaluación, por lo tanto, 105 muestras corresponden al conjunto de entrenamiento y 45 muestras al conjunto de evaluación. Las muestras de ambos conjuntos se randomizan con la funcion de matlab randperm que randomiza los indices de la matriz de muestras original.

Al conjunto de entrenamiento se le aplica una normalización zscore dejando una distribucion de datos de media 0 y desviación 1. Por último, se obtiene la media y la desviación utilizada en el centrado y escalado del conjunto de entrenamiento y se usan estos valores para normalizar el conjunto de evaluación.

Como se mencionó anteriormente, el problema cuenta con 3 clases: Clase 1, 2 y 3. Cada clase tiene igual número de muestras (50), por lo que se descartan riesgos de desbalance de clases.

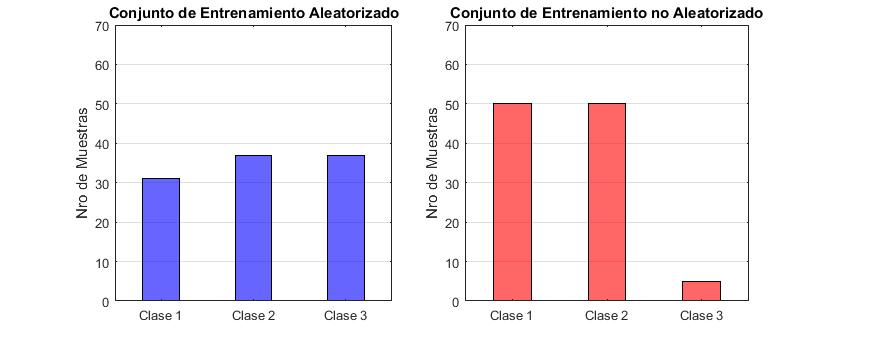
|  |  |  |
| --- | --- | --- |
| Numero de Vecinos | Eficiencia | Error de Clasificación |
| 1 | 0.6 | 0.4 |
| 2 | 0.6 | 0.4 |
| 3 | 0.37778 | 0.62222 |
| 4 | 0.64444 | 0.35556 |
| 5 | 0.53333 | 0.46667 |
| 6 | 0.57778 | 0.42222 |
| 7 | 0.57778 | 0.42222 |

**¿Por qué se usa la moda en el caso de clasificación con el método de los K vecinos?**

En el problema de clasificación sus variables forman valores discretos, es decir, es Categórica, algunos ejemplos de problemas de clasificación son: rostros, voz o estudio de crédito. La moda, a diferencia de la media, puede calcularse para variables de este tipo ya que su valor siempre existe en el conjunto de datos. Si tuviéramos un conjunto de datos [2 2 2 2 2 7] la moda sería 2 y ese sería el valor esperado, como en la mayoría de los vecinos la clase es 2, se selecciona la clase 2 para la muestra x\*, sin embargo, si utilizaramos la media, el resultado sería 2.8333, este valor no corresponde a ninguna de las clases definidas y podría provocar errores en la estimación de la clase para la muestra x\*. Es por esta razón que se considera la moda como la medida estadística más conveniente para el problema de clasificación.

**¿Por qué cree que se deben armar los conjuntos de entrenamiento y prueba de forma aleatoria?**

La base de datos tiene los valores de Yclas organizados de menor a mayor, de esta manera, al hacer los conjuntos sin aleatorización, el conjunto de entrenamiento contara solo con 5 muestras de la clase 3 y el resto serían de la clase 1 y 2, esto puede perjudicar la clasificación para las muestras de clase 3 en el conjunto de validación, ya que para una muestra x\* de clase 3, los k vecinos tendrán en su mayoría la clase 1 o 2. Con la aleatorización se garantiza que las clases en los conjuntos de entrenamiento y validación están mezcladas, y no hay un desbalance de clases en estos conjuntos.



6. Se utilizó el método de matlab classify configurando el tipo de funcion discriminante como mahalanobis, asi: class = classify(Xtest,Xtrain,Ytrain,'mahalanobis'); Con esto se encontro una eficiencia de 93% y en error del 7%

|  |  |  |
| --- | --- | --- |
| tipo de función discriminante | eficiencia | error |
| linear |  |  |
| diaglinear |  |  |
| quadratic |  |  |
| diagquadratic |  |  |
| mahalanobis | 93% | 7% |