Estadística Aplicada 3 - Examen 1

Marcelino

21/10/23

Ejercicio 1

En este reporte se creó un clasificador con la base de datos `MNIST` (Modified National Institute of Standards and Technology), esta base es uno de los conjuntos de datos más icónicos en el campo del aprendizaje automático y la visión por computadora. Se compone de un conjunto de imágenes en escala de grises de dígitos escritos a mano, del 0 al 9, y ha sido ampliamente utilizada para entrenar diversos modelos de reconocimiento de imágenes. MNIST contiene 70,000 imágenes en total, divididas en 60,000 imágenes de entrenamiento y 10,000 imágenes de prueba. Cada imagen tiene un tamaño de 28x28 píxeles, lo que da un total de 784 píxeles por imagen.

El objetivo de este clasificador era predecir si una imagen contenía un 1, 3 o 5. Para lograr encontrar el mejor clasificador primero se prepararon los datos de tal forma que la variable respuesta fuera de tipo factor y los regresores fueran los pixeles de la imagen. Posteriormente, se procedió a dividir la base de datos en un conjunto de entrenamiento, validación y prueba. El conjunto de entrenamiento y validación se utilizó para encontrar el mejor modelo y el conjunto de prueba para evaluar el desempeño del modelo. Los modelos que se utilizaron para encontrar el mejor clasificador fueron: LDA, QDA, Naive Bayes y Regresión Logística.

Además, como la base de datos era muy grande en cuanto al número de regresores potenciales, se decidió reducir la dimensionalidad de los datos con PCA, antes de entrenar los modelos. Para esto, se realizó PCA sobre el conjunto de entrenamiento y se seleccionaron los primeros 50 componentes principales, ya que estos explicaban la mayor parte de la varianza de los datos, utilizamos Scree plot para comprobarlo, como se puede apreciar en la siguiente imagen.

Después de correr cada modelo y validar sus métricas obtuvimos lo siguiente:

Con lo cual el mejor modelo de clasificación para este es el modelo _____. Esta es su matriz de confusión en el grupo de testeo, y sus métricas de desempeño.

Ejercicio 2

```
library(cluster)

#Iris DB
data <- as.matrix(iris[,1:4])
dist_mat <- dist(data, method = 'euclidean')
```

```
hclust_single <- hclust(dist_mat, method = 'single')
hclust_average <- hclust(dist_mat, method='average')
hclust_complete <- hclust(dist_mat, method= 'complete')
divisive_model <- agnes(dist_mat, method = "single")
```