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. Para esta división se utilizó el conjunto de prueba que viene con los datos y para el conjunto de validación se utilizó muestreo alteatorio estratificado con respecto a la variable `label` con la semilla de `191654`. 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 Figura 1.

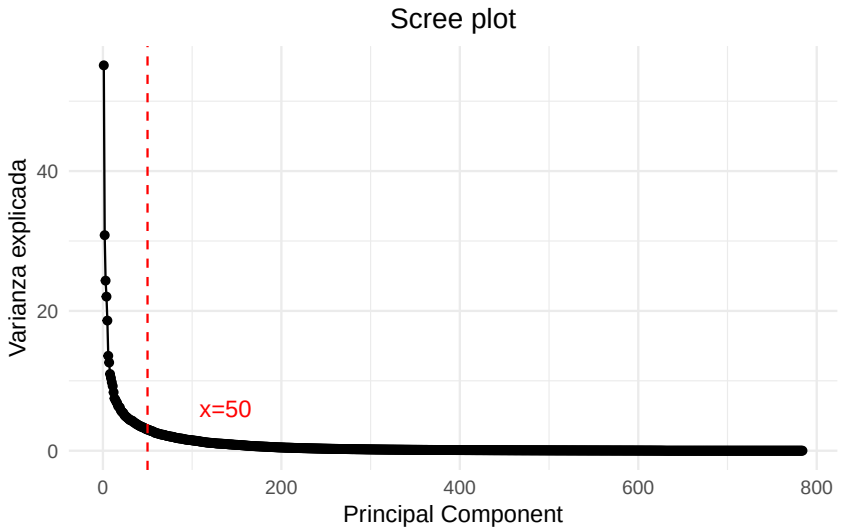


Figura 1: Scree plot para seleccionar el número de componentes principales a utilizar en PCA.

Después corrimos cada modelo y validamos métricas importantes de clasificación multiclase mostrados en la Tabla 1.

Tabla 1: Resumen de resultados

Modelo	Accuracy	F_Measure	Recall	Precision	MCC	Promedio
QDA	0.9614859	0.9607660	0.9609410	0.9609523	0.9422218	0.9572734
LDA	0.9475553	0.9460192	0.9464634	0.9461422	0.9213084	0.9414977
LR	0.6929801	0.8336565	0.6566859	0.7535628	0.6217909	0.7117352
NB	0.8929254	0.8902490	0.8897470	0.8931629	0.8397478	0.8811664

Con lo cual el mejor modelo de clasificación para este es el modelo QDA, superior en todas las métricas. Estas fueron las macro de Accuracy, Precision, Recall, F-meas, MCC. Donde cada uno entre más alta sea es mejor. La Macro-Accuracy mide la proporción de predicciones correctas para cada clase y luego calcula el promedio. Macro-Precision refleja la cantidad de predicciones positivas correctas para cada clase, frente a todas las predicciones positivas, siendo especialmente relevante cuando las falsas alarmas (falsos positivos) son costosas. Macro-Recall se centra en cuántas observaciones positivas reales de cada clase fueron capturadas por el modelo. Macro-F-meas (Medida F1 macro) ofrece un equilibrio entre Macro-Precision y Macro-Recall, siendo útil en contextos donde ambas métricas son igualmente importantes para todas las clases. Finalmente, Macro-MCC (Coeficiente de correlación de Matthews macro) proporciona una medida de la calidad de las clasificaciones, tomando en cuenta verdaderos y falsos positivos y negativos para cada clase y calculando el promedio.

Y el desempeño final de este clasificador con el conjunto de prueba fue el siguiente:

Tabla 2: Resumen de resultados QDA

Modelo	Accuracy	F_Measure	Recall	Precision	MCC
QDA	0.9716826	0.970611	0.9707141	0.9706929	0.9574273

Y su respectiva matriz de confusión.

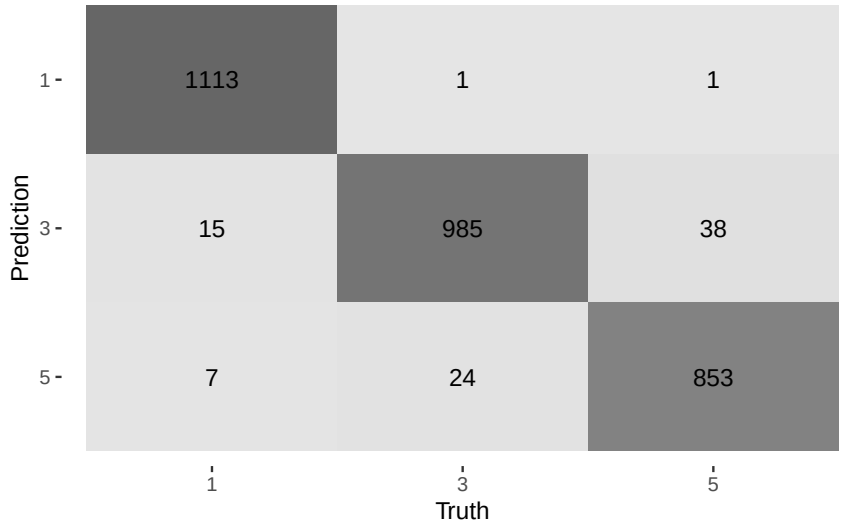


Figura 2: Matriz de confusión para el modelo QDA.

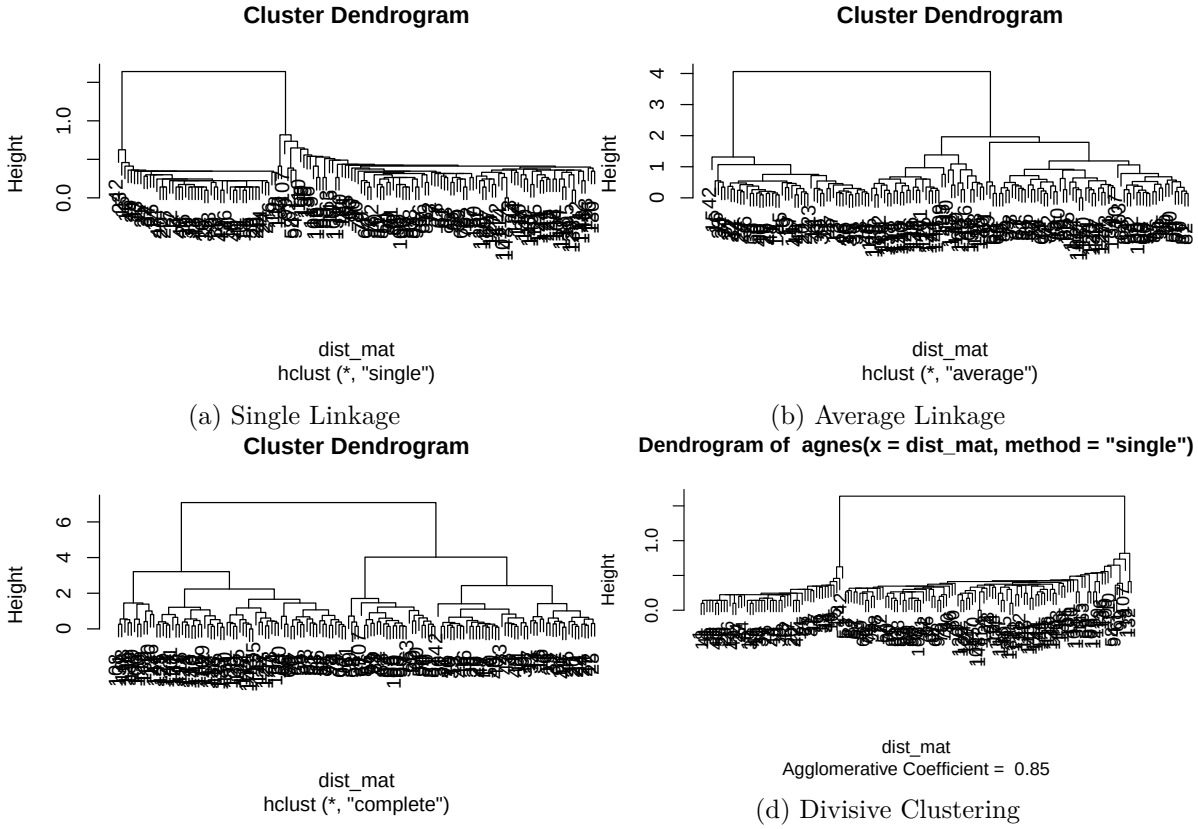
Con lo cual es excelente para poder clasificar correctamente los dígitos 1, 3 y 5.

El modelo elegido tiene sentido porque…

Ejercicio 2

En esta segunda parte del reporte se realizó un análisis de clustering sobre la base de datos `iris`. El conjunto de datos `iris`, introducido por el biólogo Ronald Fisher en 1936, consiste en mediciones de cuatro variables (longitud y ancho de sépalos y pétalos) de tres especies de flores iris (setosa, versicolor y virginica). Con 50 observaciones por especie, este conjunto ha sido ampliamente utilizado en estadística y aprendizaje automático como ejemplo para técnicas de análisis y clasificación debido a su claridad y tamaño manejable.

Para poder hacer cluster se calculó primero la matriz de distancias euclidianas entre las observaciones sin toma en cuenta el dato de a qué especie pertenecía cada registro, posteriormente se corrieron los algoritmos de clustering `single`, `average`, `complete` y `divisive`. Los resultados de estos algoritmos se muestran a continuación.



(a) Single Linkage

(b) Average Linkage

(c) Complete Linkage

(d) Divisive Clustering

Figura 3: Dendogramas de los algoritmos de clustering

Tomando en cuenta la base original con etiquetas sobre las especies a las que pertenecía cada registro notamos lo siguiente…

```
Loading required package: lattice


Attaching package: 'caret'

The following objects are masked from 'package:yardstick':

    precision, recall, sensitivity, specificity

The following object is masked from 'package:purrr':

    lift

  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [75] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[112] 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[149] 2 2

  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [75] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 3 3 3 3 2 3 3 3 3
[112] 3 3 2 2 3 3 3 3 2 3 2 3 2 3 3 2 2 3 3 3 3 3 2 3 3 3 3 2 3 3 3 2 3 3 3 2 3
[149] 3 2

  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 3 2 3 2 3 2 3 3 3 3 2 3 2 3 3 2 3 2 3 2 2
 [75] 2 2 2 2 2 3 3 3 3 2 3 2 2 2 3 3 3 2 3 3 3 3 3 2 3 3 2 2 2 2 2 2 3 2 2 2 2
[112] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[149] 2 2

  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [75] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[112] 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[149] 2 2
```