Estadística Aplicada 3 - Examen 1

Marcelino 191654

23/10/23

Ejercicio 1

En este reporte se creó un clasificador con la base de datos `MNIST` (Modified National Institute of Standards and Technology). 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(28x28) de entrenamiento y 10,000 imágenes de prueba.

El objetivo de este clasificador es 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`. 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 que se encuentra en los anexos de este reporte.

Después de entrenar cada modelo y validar métricas importantes de clasificación multiclase obtuvimos lo mostrado en Tabla 1 (que se encuentra en los anexos). Donde cada métrica entre más alta sea, es mejor. Para elegir el mejor modelo nos guiamos mucho más por la métrica de macro-Accuracy, la cual mide la proporción de predicciones correctas con respecto a todos los datos. Esta sería el complemento de lo que nos daría el error de clasificación.

Con lo cual el mejor modelo de clasificación fue el modelo QDA, superior en todas las métricas. Este resultado tiene mucho sentido por lo siguiente. En primer lugar, sabemos que la regresión logística es un modelo pésimo para clasificaciones multiclase, por lo que era de esperarse que tuviera peor desempeño. En segundo, lugar notamos que estamos entrenando más de 10,000 datos los cuales sabemos que están acotados en cierto rango y con lo cual si suponemos que vienen de alguna distribución entonces tendrían todos sus momentos y por TCL cualquier estadística que sea función de la suma de estos datos se distrubirá normal, en especial tenemos que el discriminante utilizado en LDA y QDA se distribuirá normal, por lo que es de esperarse que los modelos LDA y QDA tengan cierto desempeño aunque los datos no sean normales, superando al NB porque aprovecha mejor la normalidad asintótica (recordando que en el NB que utilizamos está aproximando las densidades condicionales con una normal). Y por último, notemos que ligeramente QDA es mejor a LDA debido a que la forma en que se escribe 3 y 5 puede tener una mayor variabilidad en los pixeles de las imágenes que la forma en que se ecribe 1, por lo cual QDA describiría un poco mejor los datos que LDA.

Por último, obviamente elegimos el QDA porque nos interesa más predicción que interpretabilidad o simplicidad.

El desempeño final de este clasificador con el conjunto de prueba se muestra en la Tabla 2 y la matriz de confusión en la Figura 2.

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 tomar en cuenta el dato de a qué especie pertenecía cada registro. Posteriormente se corrieron muchos algoritmos de clustering pero se decidieron analizar los siguientes `k-means clustering`, `average-linkage clustering` y `complete-linkage clustering` . Para el método de k-means se supuso que habían 3 clusters dado que se sabe que hay 3 especies de flores en esa base. Aunque en la vida real no sabremos cuántos clusters hay, de hecho al comparar gráficas de pares de columnas de los datos (véase la Figura 4 en los anexos), solo se podían apreciar bien 2 tipos de clusters bien definidos, por lo que hay que tener cuidado al elegir el número de clusters.

Los resultados de los dendogramas de los clusters jerárquicos se pueden apreciar en la Figura 3 que se encuentra en los anexos de este reporte. Notamos que en los dos métodos hay dos clusters bien definidos, por lo que necesariamente deben haber al menos dos clases. Sin embargo, cuando analizamos los dendogramas para medir una tercera clase, observamos que realmente no pareciera que esté bien definida por lo que podría ser que no hayan podido clusterizar bien estos métodos, sabiendo que deben haber 3 especies. Con lo cual procedimos a analizar los 3 métodos para conocer el grado de clasificación que tienen con la métrica de `accuracy`, dado que tenemos las etiquetas de los datos, y suponiendo que ciertas clusters se refierían a ciertas etiquetas. Y obtuvimos los datos que se muestran en la tabla.

Method	Accuracy
Average Linkage	0.9066667
Complete Linkage	0.8400000
K-Means Clustering	0.8933333

Por lo tanto, notamos que el mejor método fue `average-linkage` y ligeramente mejor que el `knn`. Esto tiene sentido, porque aunque se vio que era difícil de clasificar tanto por lo visto en los dendogramas como por lo visto en las imágenes de datos, estos métodos podrían sortear mejor este problema, a comparación del `single-linkage` que podría desviarse fácilmente al irse a los extremos en sus criterior de separación. Además, hay que reconocer que el `average-linkage` fuera ligeramente mejor sin conocer el número de clusters previamente.

Anexos

Tablas

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

Tabla 2: Resumen de resultados QDA

Modelo	Accuracy	F_Measure	Recall	Precision	MCC
QDA	0.9716826	0.9705247	0.9706997	0.9704187	0.9573689

Figuras

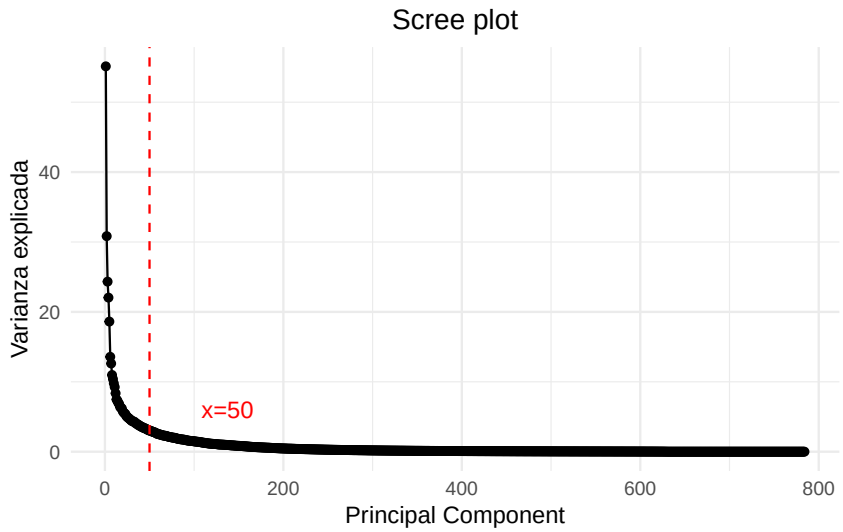


Figura 1: Scree plot para seleccionar el número de componentes principales a utilizar en PCA.

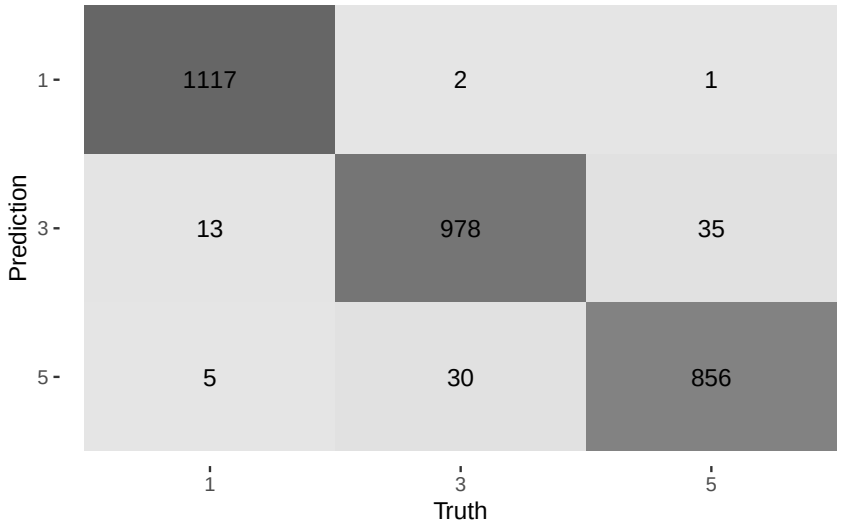
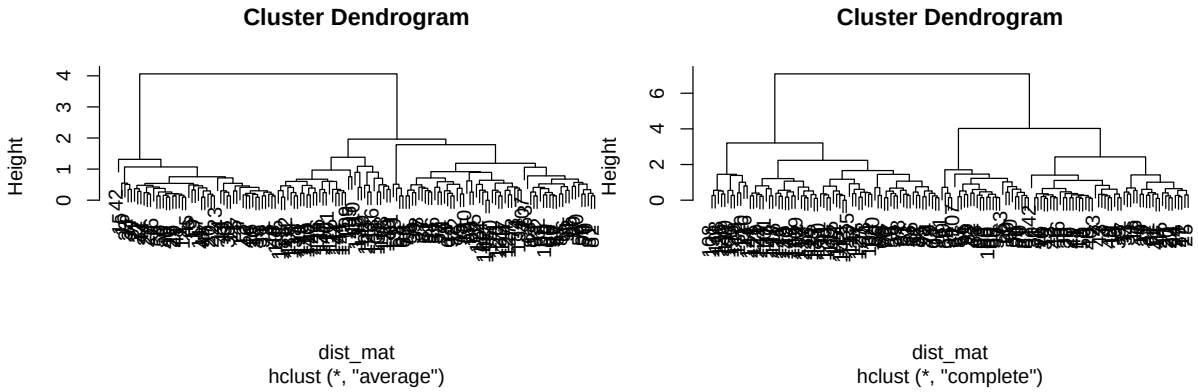


Figura 2: Matriz de confusión para el modelo QDA.



(a) Average Linkage (b) Complete Linkage

Figura 3: Dendogramas de los algoritmos de clustering

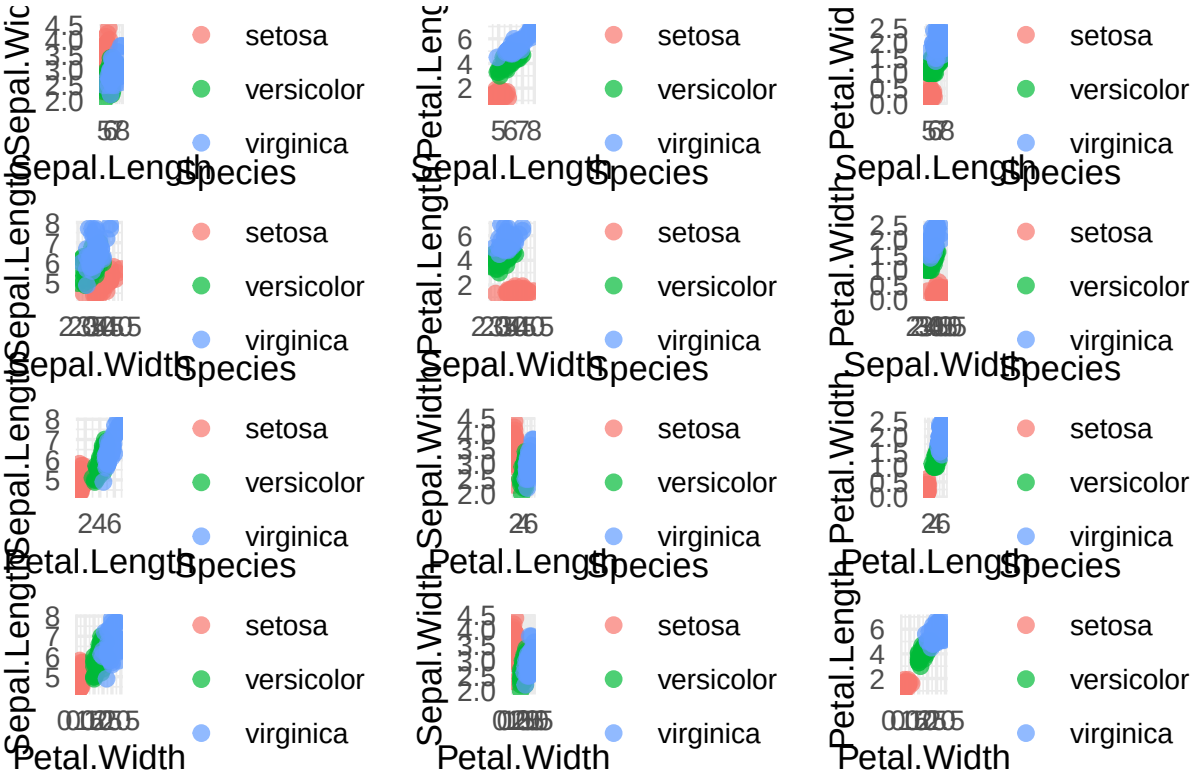


Figura 4: Comparación de columnas a pares de los datos de Iris