Centro de Investigación en Matemáticas (CIMAT). Unidad Monterrey

Tarea 6

Victor Manuel Gómez Espinosa

5 de junio de 2020

1. Problema 1

Para este problema se complementó el ejercicio 2 de la tarea 4, es decir, se utilizó el conjunto de datos MNIST que es una base de dígitos del 0 al 9 a mano, que son imágenes de 28x28 pixeles, en total se cuenta con 70,000 imágenes.

Se tomó un conjunto de entrenamiento (X_train, y_train) del 75% y otro de prueba del 25% restante (X_test, y_test), ambos balanceados (Figura 1.1). El conjunto de entrenamiento se utilizó para ajustar diferentes clasificadores (regresión logística, redes neuronales, máquinas de soporte vectorial, arboles de clasificación y AdaBoost) y de estos seleccionar el mejor modelo para predecir la respuesta o clasificación de las imágenes, es decir a la imagen asignarle una etiqueta del digito correspondiente.

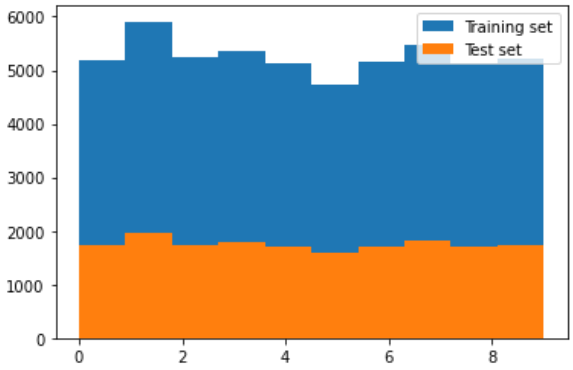

Figura 1.1: Conjuntos de datos, de entrenamiento y prueba.

El ejercicio se continuó bajo las mismas condiciones de la tera 4 para poder comparar el desempeño con el mejor modelo pasado, que fue el de QDA con un *accuracy* de 0.89 para el conjunto de prueba. Debido a las razones anteriores, se empleó la misma representación en menor dimensión que permite, separar las clases y por lo tanto podría mejorar el modelo y la predicción. Para esto, se estandarizaron los conjuntos de datos, se aplicó la transformación asociada al método LDA (identifica atributos que maximizan la varianza entre clases y minimiza la varianza entre clases) que reduce a 9 dimensiones (Figura 1.2) y a esto se ajustó nuevamente los modelos.

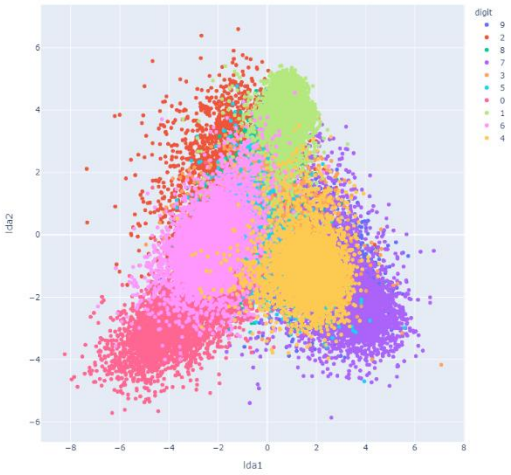

Figura 1.2: Representaciones de los dígitos en los primeros dos componentes LDA.

Para seleccionar el mejor modelo, para todos se probaron diferentes configuraciones, es decir, con diferentes parámetros de cada modelo utilizando la función GridSearchCV de scikit-learn y como criterio para seleccionar el mejor se utilizó el *accuracy* para 5-folds.

Observe la Figura 1.3, que representa para cada clasificador el desempeño o *accuracy* para el conjunto de entrenamiento como de prueba de cada modelo evaluado modificando los parámetros y el círculo rojo representa el mejor modelo seleccionado para cada clasificador y en el titulo los correspondientes parámetros de este. Observe que los 2 modelos que superan 0.9 tanto para el conjunto de prueba como para el de entrenamiento son los de redes neuronales (MLP) con dos capas ocultas con 50 unidades cada una y parámetro de Alpha de 0.01, así como también el de maquinas de soporte vectorial (SVM) con kernel gaussiano y parámetros gamma=0.1 y C=5. Los desempeños de ambos modelos son muy similares, pero finalmente se seleccionó como el mejor el de maquina de soporte vectorial debido al menor tiempo que le tomó ajustarse al conjunto de entrenamiento.

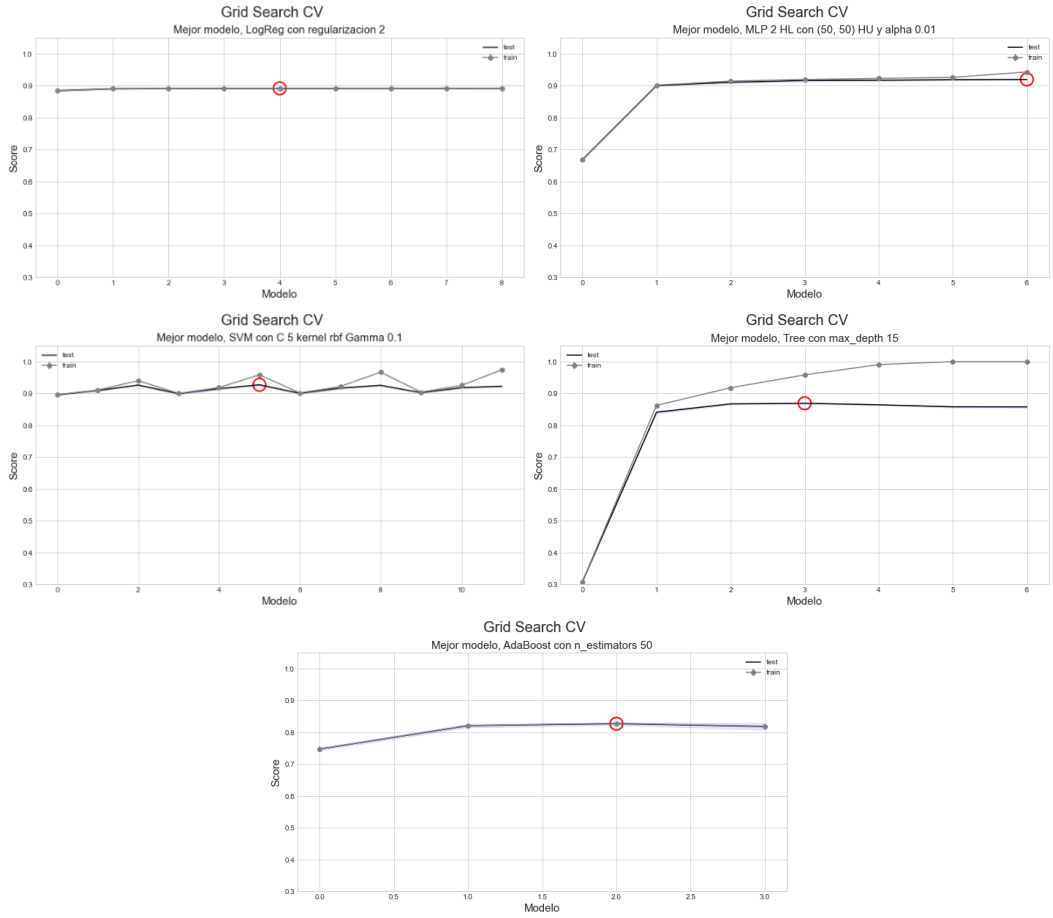

Figura 1.3: Gráficos del desempeño de los modelos Regresión logística (LogReg), Redes neuronales (MLP), Maquinas de soporte vectorial (SVM), Arboles de clasificación (Tree), AdaBoost. El desempeño del mejor modelo viene remarcado en el circulo rojo, y sus correspondientes parámetros se mencionan en el título.

Por último, estos resultados, se probaron utilizando la aplicación interactiva ajustando el modelo de máquinas de soporte vectorial con el conjunto de entrenamiento y en efecto se notó una gran mejoría en el desempeño (Figura 1.4).

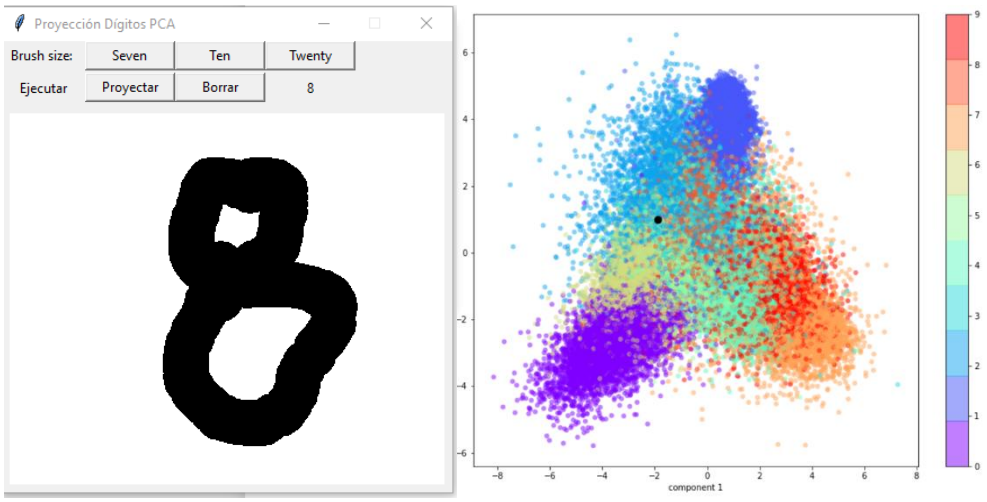

Figura 1.4: a) Aplicación interactiva para dibujar un digito y te dice cuál es, b) la proyección del digito (punto negro) en los primeros 2 componentes de la representación LDA.