Trabajo Práctico 1 — Checkpoint 4

[75.06/95.58] Organización de Datos
Segundo cuatrimestre de 2023

Grupo 15

Ayala, Tomás Gabriel - tayala@fi.uba.ar - 105336
Giacobbe, Juan Ignacio - jgiacobbe@fi.uba.ar - 109866
Olaran, Sebastian - solaran@fi.uba.ar - 109410

Docente corrector:
Pereira, Francisco

Índice

1. Introducción

Para este nuevo checkpoint, hemos decido importar el dataset proveniente del checkpoint 3. Para la construcción de las redes neuronales se necesitó estandarizar los valores que tomaban los features, de manera que nuestros modelos funcionen correctamente. Hemos estandarizado ambos datasets(train y test) para que las redes no tengan problemas a la hora de realizar predicciones.

Para la construcción de las redes neuronales, hemos buscado en un principio una arquitectura acorde para hacer las predicciones. Hemos probado 4 arquitecturas: la primera consistió en una red con una capa de entrada(de 64 neuronas), dos capas ocultas(las cuales contenían 512 neuronas cada una), y una capa de salida que tenía una neurona. Este modelo obtuvo una puntuación muy baja en la competencia de Kaggle, por lo que en las siguientes arquitecturas buscamos cambiar el número de capas ocultas, las funciones de activación para las mismas, etc.

La segunda arquitectura usada contó con 1 capa de entrada, 6 ocultas, las cuales contenían 32 neuronas cada una, y una capa de salida. Esta red obtuvo un score muy bajo en la competencia, pero logramos que duplique el score que obtuvimos con la primera arquitectura.

La tercera arquitectura usada contó con 1 capa de entrada, 8 ocultas, las cuales contenían 32 neuronas cada una, y una capa de salida. Esta red obtuvo un score de 0.70 en la competencia, pero luego encontramos una cuarta arquitectura que obtuvo la mejor performance.

La cuarta arquitectura usada(que fue la que logró obtener el score más alto en la competencia) contó con 1 capa de entrada, 5 ocultas, las cuales contenían: dos capas de 64 neuronas, otras dos con 32, y una más con 16. Obviamente, también contó con una capa de salida.

2. Construcción del modelo

La arquitectura utilizada para conseguir la mejor performance consistió en

- **Capa de entrada**: Utilizamos 128 neuronas para esta capa inicial, con una función de activación del tipo *relu*, las conexiones en esta capa son totalmente conectadas, es decir, cada neurona en esta capa está conectada a todas las neuronas en la capa siguiente.
- **Capas ocultas**: utilizamos 5 capas, las cuales contenían: dos capas de 64 neuronas, otras dos con 32, y una más con 16. Todas utilizaron una función de activación del tipo *relu*. Las conexiones en esta capa tambien son totalmente conectadas.
- **Capa de salida**: Al ser un problema de clasificación binario, hemos utilizado una sola neurona para esta capa, y una función de activación del tipo *softmax*.
- **Hiperparámetros optimizados**: Hemos optimizado un total de 2 hiperparámetros para la red, entre los cuales tenemos la cantidad de epochs y el tamaño del batch.
- **Optimizador**: Hemos utilizado un optimizador del tipo SGD, utlizando un learning rate de 0.001. Este valor para el learning rate lo vimos razonable para el problema que estábamos tratando.
- **Técnica de regularización**: Se aplicó la técnica de *Dropout* en las capas ocultas con una tasa del 50 %. El *Dropout* es una técnica de regularización que implica apagar aleatoriamente el 50 % de las neuronas durante el entrenamiento. Esto ayuda a prevenir el sobreajuste, ya que obliga a la red a aprender de manera más robusta y generalizada.
- **Ciclos de entrenamiento**: Se han realizado un total de 19 ciclos de entrenamiento para esta arquitectura.

3. Cuadro de Resultados

A continuacion podemos observar un cuadro de las metricas de cada mejor predictor de cada modelo:

Modelo	F1-Test	Precision Test	Recall Test	Accuracy	Kaggle
1° Arquitectura	0.8670	0.8557	0.8786	0.8640	0.3314
2° Arquitectura	0.8389	0.8216	0.8569	0.8339	0.60446
3° Arquitectura	0.7982	0.6684	0.9906	0.7472	0.70638
4° Arquitectura**(Mejor predictor)**	0.8164	0.9309	0.7271	0.8350	0.80193

La primera arquitectura muestra un buen equilibrio en términos de F1-Test, Precision y Recall. La métrica de Accuracy es alta, indicando un buen rendimiento general en la clasificación. Sin embargo, la métrica Kaggle sugiere que podría haber margen para la mejora en una competencia específica,y que nuestra red neuronal no seria consisitente frente a nuevos datos.

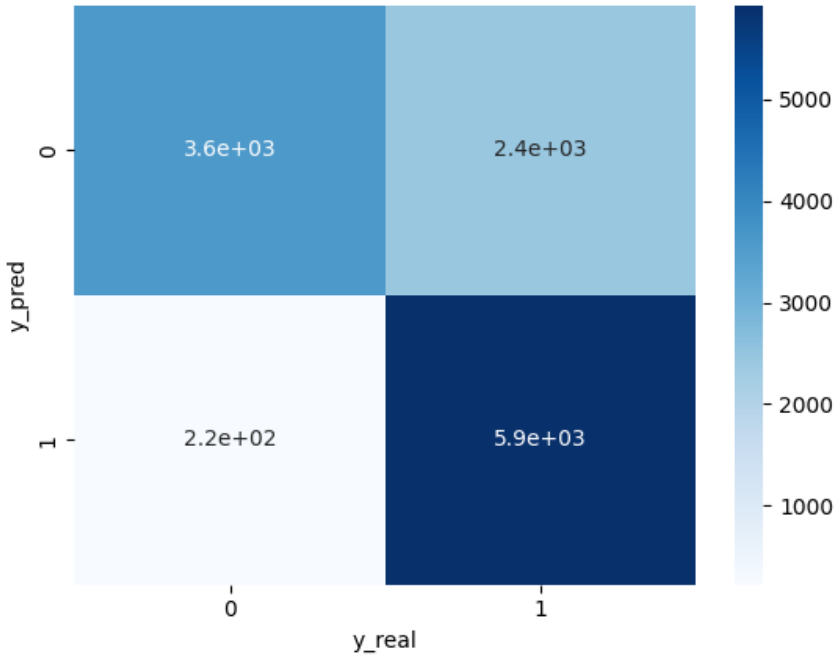
La segunda arquitectura muestra valores de F1-Test(los mas altos en conjuntos de TEST), Precision y Recall ligeramente más bajos que la primera, pero sigue manteniendo un buen equilibrio.Obtuvimos valores de Kaggle mas competitivos pero aun asi el rendimiento de la nueva arquitectura demosotro no ser optimo

La tercera arquitectura muestra un alto valor de Recall, lo que indica una buena capacidad para identificar verdaderos positivos. Sin embargo, la Precision es bastante baja, lo que podría llevar a un alto número de falsos positivos.

La cuarta arquitectura, identificada como el **Mejor Predictor**, se caracteriza por alcanzar un equilibrio sólido entre Precision y Recall. Esto significa que es capaz de hacer predicciones precisas sin dejar de capturar la mayoría de los casos positivos. Además, su alto puntaje en Kaggle la coloca como la mejor opción en una competencia específica, demostrando su rendimiento sólido en un conjunto de datos de prueba desconocido.

4. Matriz de Confusión

A continuación podemos ver la matriz de confusión de nuestro mejor predictor



Como podemos apreciar, y tal como indica nuestro puntaje, aproximadamente el 83 por ciento de las predicciones son exactas. No se observa una clara inclinación hacia uno u otro parámetro, ya sea en los falsos positivos, falsos negativos, verdaderos positivos o verdaderos negativos. El modelo muestra un equilibrio en su capacidad para predecir tanto casos positivos como negativos, lo que sugiere que no está sesgado hacia ninguna de las dos clases y es efectivo en su capacidad de clasificación.

5. Tareas realizadas

Integrante	Tarea
Tomas Gabriel Ayala	Armado de Reporte , Armado y entrenamiento de Redes Neuronales
Sebastian Olaran	Armado de Reporte y creacion de arquitecturas
Juan Ignacio Giacobbe	Creación de arquitecturas y búsqueda de hiperparámetros