Proyecto Final Statistical Learning II

Luis Adolfo Martnes Ortiz *
[1]Instituto de Investigación de operaciones, Universidad Galileo

Abstract

El presente artículo muestra la aplicación de 3 redes neuronales más utilizadas en Deep Learning, de forma separada e independiente, la primera aplicación que se presenta es una red neuronal MLP (Multi-Layer Perceptron), para la cual se utiliza un dataset que contiene información de los precios de habitaciones en airbnb de New York. La construcción de este modelo tiene como objetivo predecir los precios de las habitaciones.

La segunda aplicación que se presenta es una red neuronal CNN (Convolutional Neural Network) para la cual se utiliza un dataset que contiene fotos de flores agrupadas por la categoria a que pertenece, estas categorias son: rosas, margaritas, tulipanes y dandileon. Se tiene una coleccin de fotos que superan las 600 cada categoría, por lo que se tiene una muestra representativa para la red convolucional.

La tercera aplicación que se presenta es una red neuronal recurrente RNN (Recurrent Neural Network) donde se aplica el análisis de un dataset que contiene noticias, el objetivo desarrollar un modelo para predecir si la noticia es falsa o verdadera.

1 Definición del problema

Para este proyecto se desarrollo un problema para cada red.

1.1 Red Neuronal MLP

Para los empresarios hoteleros y páginas que ofrecen el servicio de hospedaje en muchas ocasiones se les complica predecir el cambio de precios en las diferentes ubicaciones de la ciudad de New York. La predicción de precios de hospedaje es vital para brindar este servicio para los clientes, ya que supondría para ellos una ventaja que les ayudaría ajustar su presupuesto revisando la proyección de los mismos.

1.2 Red Neuronal CNN

Hoy en día en empresas y personas individuales se les complica identificar los tipos de flores que se encuentran en su territorio, muchas de ellas han cometido errores cuando desean aplicar algún químico para evitar plagas, esto por confusión del tipo de flor. Por lo que se busco generar el modelo dada una imagen determine que tipo de planta es.

1.3 Red Neuronal RNN

Hoy en día en las redes sociales ya sea en Guatemala y en cualquier otro país hay una cantidad de personas que se dedica a difundir información falsa, lo cual provoca alteración en la población, por lo que se busco generar un modelo que valida si una noticia es falsa o verdadera.

*adolfo.martinez@galileo.edu

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

2 Metodología

Para los tres tipo de red neuronal se realizo la misma fases:

2.1 Análisis exploratorio de datos

Una fase muy importante en todo proyecto, realizamos análisis estadístico de la información a utilizar en cada problema planteado, como identificar variables, el tipo de dato de nuestro dataset, correlaciones, y medidas de tendencia central.

2.2 Tratamiento de datos

Se aplica el tratamiento de los datos

2.2.1 Red MLP

Para el caso de este proyecto se realizó una escala en los datos y se segmentó la variable dependientes. Se realizo etiquetado LabelEncoder para las variable categoricas para poder utilizarlas en nuestro modelo.

2.2.2 Red CNN

Para el caso de este proyecto se realizó transformaciones en la imagenes como el tamaño, se realizo LabelEncoder, se realizo una separación para tener dataset de entrenamiento y test.

2.2.3 Red RNN

Para el caso de este proyecto se realizó el proceso de Tokenization, se removio los NA, Embedding del texto a procesar, se separo el texto con etiqueta a predecir, se determino longitud del texto y se realizo un word index.

2.3 Desarrollo de modelo

2.3.1 Red MLP

Para el caso de este proyecto se realizó una escala en los datos y se segmentó la variable dependientes. Se aplicó LabelEncoder para las variables con mayor correlación

2.3.2 Red CNN

Para el caso de este proyecto se realizó transformaciones espaciales cambiando el tamaño de la imagen, se aplico LabelEncoder, se realizo una separación de los datos y luego se realizo la red nueronal.

2.3.3 Red RNN

Para el caso de este proyecto se realizó el proceso de Stemming, Tokenization, Embedding del texto a procesar y word index, para luego crear nuestra red nueronal ltsm ya que esta tiene una memoria a largo plazo.

2.4 Entrenamiento del modelo

2.4.1 Red MLP

Para el caso de este proyecto se realizó una categorización mean squared error como una función de costo, además se aplicó un optimizador Nadam. El modelo utilizó 98 iteraciones realizando callbacks para conservar el mejor modelo utilizado, obteniendo un r2 y RSME optimo.

2.4.2 Red CNN

Para el caso de este proyecto se realizó una categorizaciÓCategorical binary Crossentropy como una función de costo, además se aplicó un optimizador Adam con un lr de 0.0001. El modelo utilizó 50 epoch y un batch size de 128, con un activador softmax. Además se utilizo parametros para evitar el sobre ajuste.

2.4.3 Red RNN

Para el caso de este proyecto se realizó una categorización Categorical Crossentropy como una función de costo, además se aplicó un optimizador Adam. Se realizaron dos modelos Lstm con diferentes capas, el primer modelo utilizó 4 iteraciones y el segundo 3 iteraciones con activador relu ambos.

2.5 Prueba de rendimiento

Para las disintas redes neuronales se aplicó la técnica de separar los datos por entrenamiento y validación.

2.5.1 Red MLP

En el caso de este proyecto se puede observar como la curva de Loss decrece a medida que avanza las iteraciones. Adicional, podemos observar la gráfica de la proyección y las observaciones.

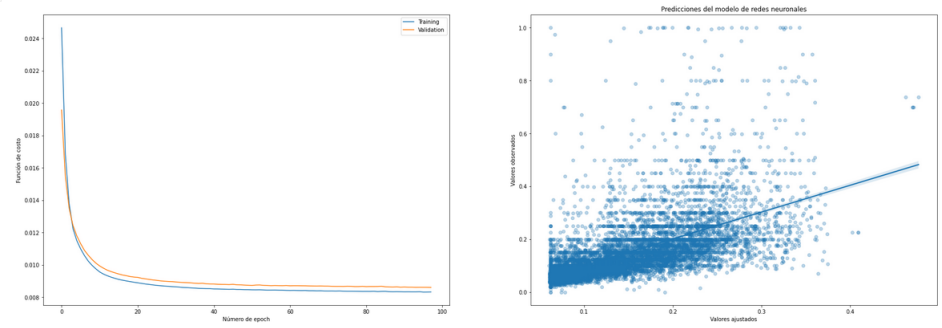

Figure 1: Gráficas de modelo MLP

2.5.2 Red CNN

En el caso de este proyecto se puede observar como la curva de Loss decrece a medida que avanza las iteraciones. Y observamos como la gráfica de precición aumenta.

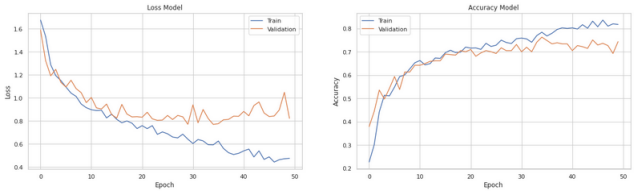

Figure 2: Gráficas de modelo CNN

2.5.3 Red RNN

Las predicciones realizadas con el modelo se considera exitosas, Pero requiere de una mejora en la precisión.

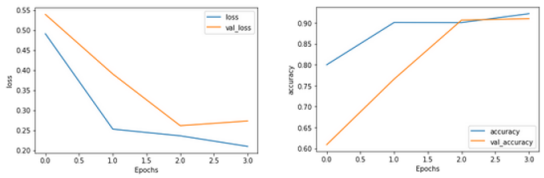

Figure 3: Gráficas de modelo RNN

3 Conclusiones

3.1 Red MLP

Como en todo modelo de ciencia de datos siempre es importante realizar un análisis exploratorio de los datos. Algo muy importante es saber las funciones de activación a utilizar, este modelo obtuvimos un r2 y rmse optimo, pero que se puede mejorar.

3.2 Red CNN

Importante la selección del kernel a utilizar, observamos que un modelo muy poderoso y requiere un costo computacional para implementarlo y entrenarlo, en nuestro problema se logro entrenarlo reduciendo nuestro dataset, pero tomo un tiempo fuerte para lograrlo.

3.3 Red RNN

Un modelo complejo que tomo mucho tiempo en realizarlo y entrenarlo, pero que al final logramos realizarlo, pero requiere de más información para aumentar la eficacia del mismo y buscar un dataset mas grande.

References

Referencias utilizadas.

[1] RAMASUBRAMANIAN, Karthik; SINGH, Abhishek. Deep learning using keras and tensorflow.En Machine Learning Using R. Apress, Berkeley, CA, 2019. p. 667-688.

[2] MOOLAYIL, Jojo. An introduction to deep learning and keras. En Learn Keras for Deep Neural Networks. Apress, Berkeley, CA, 2019.

[3] Torres, Jordi. DEEP LEARNING Introducción práctica con Keras. Lulu. com, 2018.