Proyecto Final Statistical Learning II

César Luis Polanco Cortez *
[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 referente a las primas que se ofrecen en una aseguradora. El objetivo es predecir si en base a atributos como IMC, Edad, Genero, gastos se puede predecir si una persona es fumadora o no.

La segunda aplicación que se presenta es una red neuronal CNN (Convolutional Neural Network) para la cual se utiliza un dataset de fotos de plantas saludables y plantas con plagas o enferemedades, la cantidad de imagenes de cada categoría supera las cien imágenes lo cual se considera 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 sentieminto de un script de la pelicula Harry Potter y la piedra filosofal. El entrenamiento se apoyó de un set de datos externo donde se pudo obtener los sentimientos en base a twits publicados.

***Keywords*—** Deep Learning, Plant Desease, Sentiment Analysis, Text Mining

1 Definición del problema

Para el proyecto elaborado, se presentó un problema para cada red.

1.1 Red Neuronal MLP

Para las aseguradoras es de vital importancia el precio de la prima de cada asegurado. En muchas ocasiones, se tiende a estimar hacia arriba el monto asegurado a personas que por naturaleza debería realizar ajuste a su monto de prima por casos particulares como ser propensos a enfermedades crónicas o bien ser fumadores activos y recurrentes. La mala clasificación de dichos asegurados genera pérdidas monetarias a la organización.

1.2 Red Neuronal CNN

En el mundo de la agricultura existen enfermedades y plagas comunes las cuales se puede mitigar siempre que el agricultor aplique los distintos tratamientos recomendados por un experto. Estas plagas dañan múltiples cultivos y generalmente un agricultor no cuenta con el dinero necesario para estar invirtiendo en la llegada de un experto que analice dicha planta lo cual se traduce a pérdidas de cultivos y baja producción de recursos.

1.3 Red Neuronal RNN

En Guatemala el tema de la inclusión en niños es un reto, ya que existen niños con retos especiales como es el caso de captar y procesar emociones. Como un pequeño aporte a la inclusión para personas con esta discapacidad se busca realizar un análisis de sentimiento de la obra literaria Harry Potter y la piedra filosofal.

*pccl87@galileo.edu

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

2 Metodología

Sin importar el tipo de red neuronal se realizó el siguiente trabajo:

2.1 Análisis exploratorio de datos

Como en todo proyecto de datos, se aplicó los conocimientos adquiridos en el trascurso de los últimos trimestres para poder identificar las variables, el tipo de datos con el que se trabaja, correlaciones en caso aplicara, revisión de medidas de tendencia central como media, mediana, moda. etc.

2.2 Tratamiento de datos

Se aplica el tratamiento particular 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 aplicó One-Hot-Encoding para la variable a predecir.

2.2.2 Red CNN

Para el caso de este proyecto se realizó transformaciones espaciales cambiando el tamaño de la imagen, aplicando rotaciones, trasposiciones.

2.2.3 Red RNN

Para el caso de este proyecto se realizó el proceso de Stemming, lemmatization, Tokenization, Embedding del texto a procesar.

2.3 Desarrollo de modelo

2.3.1 Red MLP

Para el caso de este proyecto se realizó una implementación de 6 neuronas como input, adicional, 3 capas ocultas de 256, 128, y 32 nuronas respectivamente para finalizar 1 neurona de salida para clasificar si la persona era o no fumadora.

2.3.2 Red CNN

Para el caso de este proyecto se realizó data augmentation y se reescaló la imagen a 224x224x3. se implementó Droup out como técnica de regulización para no realizar un over-fitting de la información y se creó capas ocultas de 32 y 64 neuronas con un kernel de 3x3.

2.3.3 Red RNN

Para el caso de este proyecto se realizó una red con una capa LSTM (Long Short-Term Memory) se aplicó embedding del vocabulario y adicional se creó una capa de salida de 13 neuronas para la clasificación del sentimiento.

2.4 Entrenamiento del modelo

2.4.1 Red MLP

Para el caso de este proyecto se realizó una categorización Binary Crossentropy como una función de costo, además se aplicó un optimizador Nadam. El modelo utilizó 200 iteraciones realizando callbacks para conservar el mejor modelo utilizado.

2.4.2 Red CNN

Para el caso de este proyecto se realizó una categorización Sparse Categorical Crossentropy como una función de costo, además se aplicó un optimizador Adam. El modelo utilizó 10 iteraciones por limitantes computaciones. Se realizó callbacks para conservar el mejor modelo utilizado.

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 Nadam. El modelo utilizó 100 iteraciones segmentadas realizando callbacks para conservar el mejor modelo utilizado.

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 de avanza las iteraciones. Adicional, la precisión del set de datos de validación se ajusta de buena manera al set de datos de entrenamiento.

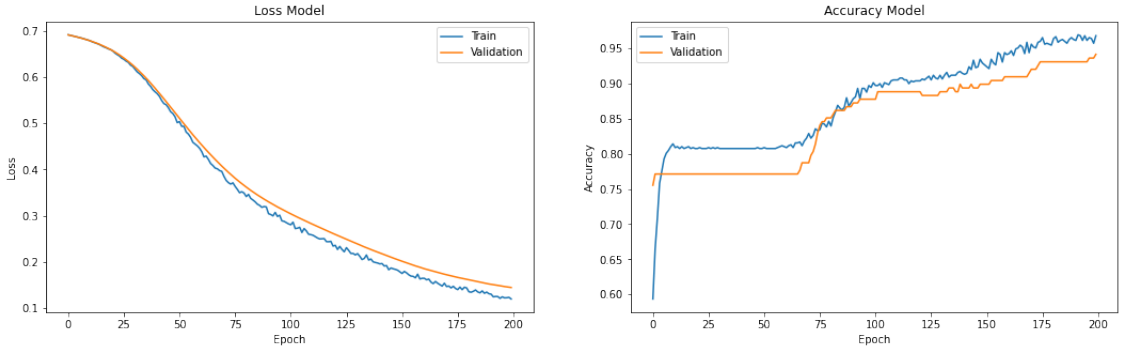

Figure 1: Gráficas de modelo MLP

2.5.2 Red CNN

En el caso de este proyecto se puede observar como la curva de Loss decrece a medida de avanza las iteraciones. En este proyecto la precisión no expresa excelencia por la cantidad de iteraciones utilizadas en el proyecto.

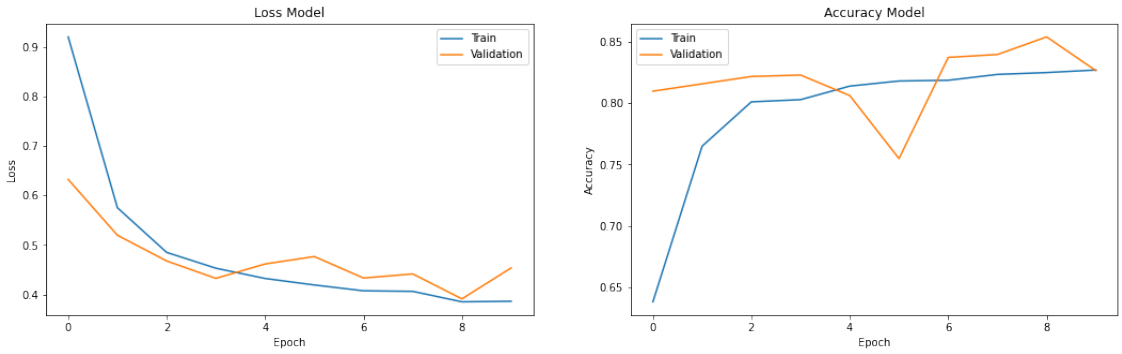

Figure 2: Gráficas de modelo CNN

2.5.3 Red RNN

Para este modelo, tanto la precisión y el erorr se puede mejorar, pero las predicciones realizadas con el modelo se considera exitosas

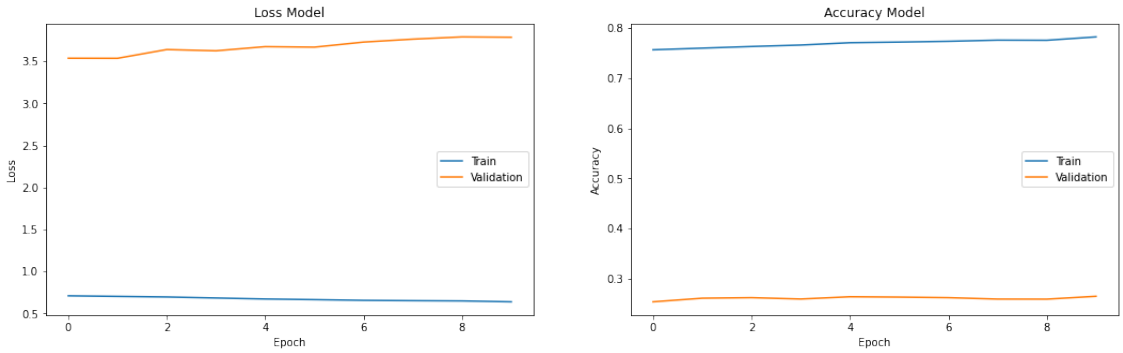

Figure 3: Gráficas de modelo RNN

3 Conclusiones

3.1 Red MLP

Como en todo modelo de ciencia de datos siempre es importante realizar un análisis exploratorio de los datos. No olvidar que los optimizadores y las funciones de activación se debe elegir de acuerdo al caso que se desee estudiar.

3.2 Red CNN

Importante la selección del kernel a utilizar, es uno de los modelos más poderosos a nivel de caso de uso pero también de los más costosos computacionalmente hablando.

3.3 Red RNN

El modelo que más tiempo consumió en su desarrollo, importante establecer el modelo de entrenamiento para poder realizar una análisis de sentimiento y considerar siempre el preprocesamiento del texto.

References

Referencias utilizadas.

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

[2] Kathuria, Ramandeep Singh, et al. "Real time sentiment analysis on twitter data using deep learning (Keras)." 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). IEEE, 2019.

[3] Jose, R. (2019). A convolutional neural network (cnn) approach to detect face using tensorflow and keras. International Journal of Emerging Technologies and Innovative Research, ISSN, 2349-5162.