Práctica 1
Aplicación de RNA

Inteligencia Artificial en las Organizaciones

Grupo 83

Miguel Gutierrez Pérez
100383537@alumnos.uc3m.es

Mario Lozano Cortés
100383511@alumnos.uc3m.es

Alba Reinders Sánchez
100383444@alumnos.uc3m.es

Alejandro Valverde Mahou
100383383@alumnos.uc3m.es

10 de octubre de 2020

Índice

1. Introducción

2. Parte 1: Aproximación usando Weka

2.1. Planteamiento y desarrollo del problema a tratar

El objetivo de la asignatura de Inteligencia Artificial en las Organizaciones es el poner en relieve el encaje de las diversas técnicas de IA en contextos reales. Con esta motivación en mente, parece evidente que es imprescindible lograr obtener soluciones a problemas apremiantes con el conjunto de técnicas disponibles. Por su parte, la **epidemia** del **SARS-CoV-2** supone un **reto** para la humanidad y constituye una oportunidad para demostrar el potencial de las nuevas herramientas IA de las que disponemos para hacer frente a este nuevo reto.

Uno de los **modelos computacionales** cuya **aplicación** resulta **atractiva** es el de las Redes de Neuronas Artificiales. La capacidad de aprender de grandes conjuntos de datos hacen de esta una opción perfecta para comprender y predecir los contagios causados por el virus.
Para comprender cómo cumplir con este objetivo debemos tener en cuenta cómo es posible que una red de neuronas aprenda. La respuesta se haya en su estructura. A nivel básico la estructura de una RNA supone un conjunto de entradas multiplicadas por unos pesos a modo de impulso nervioso al que se le puede aplicar determinadas funciones de activación. Dicha estructura de una neurona puede ser vista a continuación.

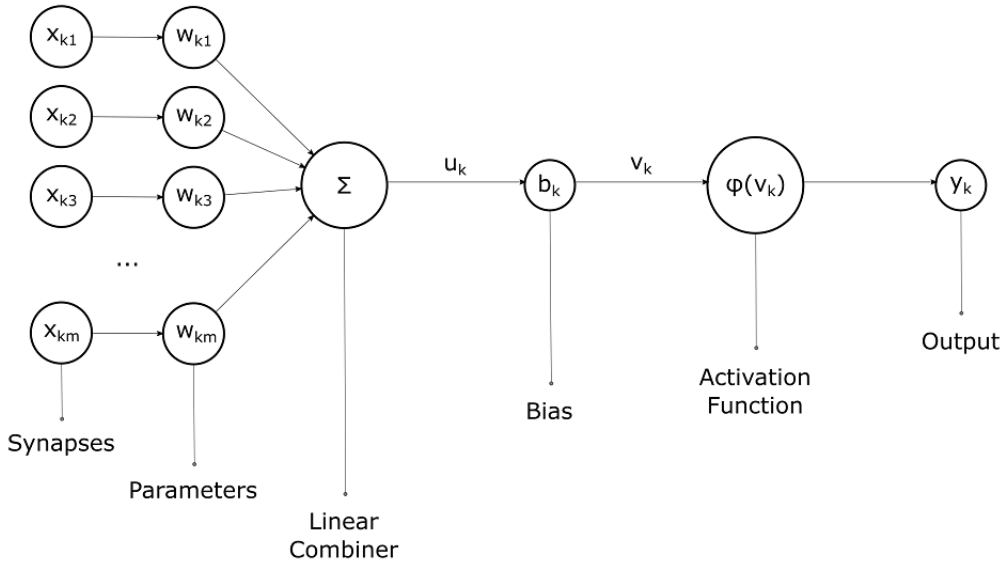

Figura 1: Estructura de una neurona

Por lo tanto, hemos trasladado el problema a la **selección de los pesos** precisos para realizar predicciones lo más exactas posibles. El modelo a emplear se tratará del **Perceptrón Multicapa**. Consecuentemente el trabajo realizado en esta práctica consiste en *tratar los datos para la aplicación del modelo elegido*, *realizar los experimentos oportunos para dar con una configuración de la red apropiada* y *analizar los resultados* obtenidos haciendo una reflexión crítica del proceso seguido.

2.2. Tratamiento de datos

Los datos que se han utilizado pertenecen al *Novel Coronavirus (COVID-19) Cases Data* de la página *The Humanitarian Data Exchange*, se trata de una recopilación de **datos epidemiológicos del COVID-19** desde el día 22 de enero de 2020.

En concreto, los datos de interés son los que se encuentran en el fichero diario de casos confirmados, este está compuesto de **266** provincias/estados de distintos países/regiones, de los cuales se recopila el **número de infectados totales** desde el 22 de enero hasta la fecha.

Los primeros atributos son: nombre de la provincia/estado, país/región, longitud y latitud (del país). Seguidos del número de contagiados acumulados por día.

El tratamiento de los datos que se ha llevado a cabo es el siguiente:

- En primer lugar, se ha eliminado la comilla simple del nombre de un país del fichero, para evitar errores en *Weka*.
- Se han renombrado los atributos correspondientes a las fechas para que el fichero sea compatible con nuevos datos, para ello se renombra el día actual a *Día 0* y el resto a *Día -1*, *Día -2*,...
- Por último, se convierte el fichero *.csv* en *.arff*.

Además, se ha generado un nuevo fichero para abordar la generación de resultados sobre los territorios concretos en los que se focaliza la atención de la práctica, **Brasil** y **España**. Dicho fichero se obtiene borrando del fichero *.arff* anterior las filas que no corresponden a estos países.

2.3. Configuración de experimentos con MLP

2.4. Análisis de resultados

3. Parte 2: Series temporales

3.1. Descripción de las series temporales utilizadas

Podemos definir las series temporales como un una **sucesión de datos o muestras medidos en intervalos de tiempo regulares** de los cuales resulta especialmente interesante el factor de predicción de los valores en instantes de tiempo futuros. Dichas predicciones son posibles al detectar patrones o tendencias en los datos a lo largo del tiempo.

La pregunta que nos debemos hacer es si se corresponden los datos de contagios acumulados y muertes por territorios del *SARS-CoV-2* con una serie temporal de la que podamos realizar predicciones fiables. Para contestar a esta pregunta vamos a fijarnos en las variables de las que disponemos y la relación entre ambas. Como variable independiente nos encontramos con tiempo, mientras que el número de contagios (o muertes, según el caso) se identifica con la variable dependiente puesto que esta cifra depende en gran medida del la cifra medida en el instante de tiempo anterior. Por lo tanto **se trata de una serie temporal**.

Sin embargo, existen más características en las que nos podemos fijar. [2]

- **Estacionalidad**: Se refiere a fluctuaciones periódicas. Un análisis en profundidad de los datos de la epidemia debería considerar este factor, no obstante, dado que la pandemia no lleva entre nosotros un periodo de tiempo lo suficientemente grande como para detectar un comportamiento estacional a lo largo de las distintas épocas del año no nos fijaremos en este parámetro.
- **Estacionaridad**: No debe ser confundido con la estacional dado que hace referencia al mantenimiento de los valores estáticos tales como la media y varianza. El proceso de crecimiento descontrolado de una pandemia no puede ser considerado estacionario debido a la tendencia alcista y volátil de los datos.

A continuación se muestra un ejemplo de una serie temporal estacional y estacionaria para comprobar cómo difiere de la serie temporal de los datos del *COVID* mundiales.

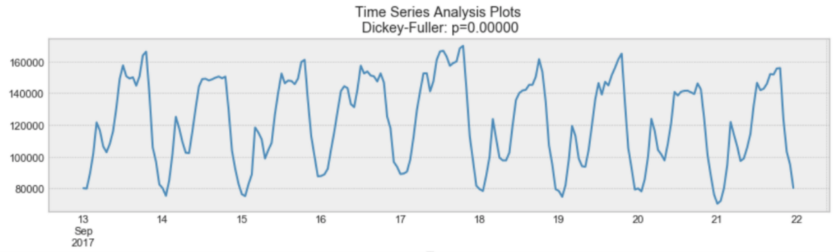

Figura 2: Serie temporal estacional y estacionaria

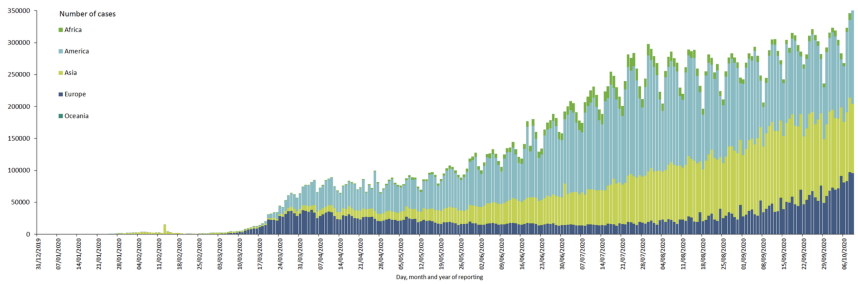

Figura 3: Casos mundiales de *COVID*. [3]

4. Proceso de entrenamiento

La red de neuronas generada seguirá el algoritmo de *Multilayer Perceptron* para la predicción de series temporales en el apartado *timeseriesForecasting* de *Weka*. Por lo tanto, el siguiente paso consiste en la selección de los parámetros de la red que permitan encontrar la mejor predicción posible.

4.1. Arquitecturas probadas

En el caso de *Multilayer Perceptron* existen multitud de parámetros de los cuales debemos considerar su variación y combinación para obtener una buena arquitectura de red. Algunos de ellos (*hidden layers, learning rate y training time*) han sido descritos en la la sección Parte 1 de esta memoria y por tanto se obvia la explicación. Sin embargo, la aproximación con series temporales introduce un nuevo parámetro a explicar. La definición del mismo se da a continuación.

- **Lag Lenght**: Las variables *lag* son el mecanismo por el cual se determina la **relación entre las variables pasadas y las tratadas actualmente**. Una forma de pensar en este parámetro consiste en equipararlo a la longitud de la ventana de tiempo en la que se tiene en cuenta el valor de las variables pasadas. Por consiguiente, **un valor apropiado de este parámetro en el caso considerado es 7 días** dado que los estados suelen tener problemas en las notificaciones del fin de semana, siendo apropiado estudiar el efecto de la pandemia en porciones de una semana. Por lo tanto, se usarán 7 y 14 días para las pruebas.

Se concluye que es necesario realizar la siguiente aproximación a la arquitectura de la red mostrada en la tabla contigua, combinando adecuadamente valores lógicos de los parámetros propuestos.

ID experimento	Lag Lenght	Hidden Layers	Learning Rate	Training Time
1	7	a	0,1	500
2	7	a	0,1	5000
3	7	a	0,3	500
4	7	a	0,3	5000
5	7	t	0,3	500
6	7	t	0,3	5000
7	7	t	0,1	500
8	7	t	0,1	5000
9	14	a	0,1	500
10	14	a	0,1	5000
11	14	a	0,3	500
12	14	a	0,3	5000
13	14	t	0,3	500
14	14	t	0,3	5000
15	14	t	0,1	500
16	14	t	0,1	5000

Tabla 1: Configuración de experimentos

Finalmente, es interesante destacar que cada uno de los experimentos debe ser dividido a su vez en *4 sub-experimentos* debido a que la herramienta de *forecasting* permite tomas los datos

de contagios y muertes de manera aislada o de manera cruzada. Es decir, para cada una de las configuraciones de los experimentos debemos realizar uno con los datos de confirmados acumulados aislados, uno con los datos de los muertos aislados y otros dos con las series cruzadas. Los resultados de los experimentos se tratan en el siguiente apartado.

4.2. Arquitectura seleccionada

5. Contexto de la práctica

6. Conclusiones

7. Referencias

1. Introduction to Neurons in Neural Networks. Medium. Consultado en Octubre 2020. Url: https://medium.com/artificial-neural-networks
2. The complete guide to time series. Consultado en Octobre 2020. Url: https://towardsdatascience.com/the-complete-guide-to-time-series
3. COVID Data European Union. Consultado en Octubre 2020. Url: https://www.ecdc.europa.eu