PRÁCTICA 1
(Parte I)
Aplicación de RNA

Inteligencia Artificial en las Organizaciones

Grado en Ingeniería Informática
Curso 2023/24

Introducción

En los últimos años, la pandemia global causada por el COVID-19 (consulte la Figura 1) ha obligado a gobiernos, organizaciones y ciudadanos a implementar medidas cruciales para combatir sus efectos. En esta batalla, el análisis de datos se ha convertido en una herramienta esencial, y en particular, el empleo de tecnologías de Aprendizaje Automático (AA), incluyendo las Redes de Neuronas Artificiales, ha desempeñado un papel fundamental.

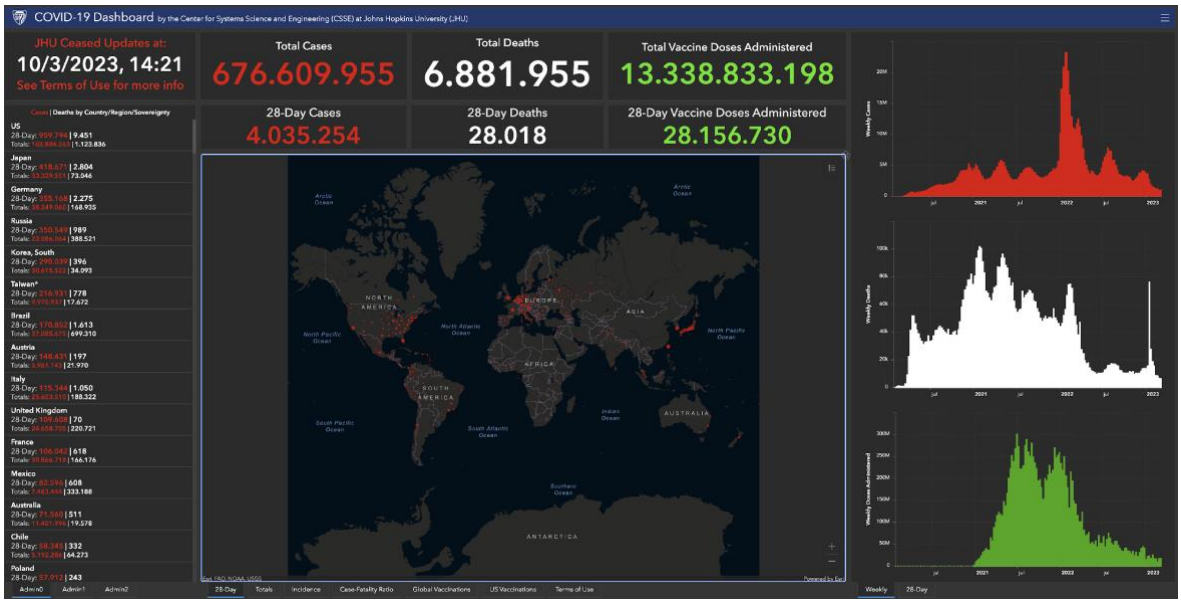

Figura 1. Situación actual de la pandemia

Existen numerosas tareas que pueden ser abordadas mediante el Aprendizaje Automático para contribuir a la lucha contra la pandemia. Por ejemplo, el AA puede ser empleado para:

- Identificar a las personas en mayor riesgo.
- Realizar diagnósticos precisos en pacientes.

- Acelerar el proceso de desarrollo de medicamentos.
- Prever la propagación de la enfermedad.
- Profundizar en la comprensión de los virus.
- Rastrear el origen de los virus.
- Anticipar posibles futuras pandemias.

El objetivo de esta práctica consiste en emplear Redes de Neuronas Artificiales (RNA) para predecir la propagación de la enfermedad. Para ello, partiendo de un conjunto de datos de entrenamiento, se construirá una RNA, definiendo aspectos clave como su arquitectura y tasa de aprendizaje. En el proceso de utilización de las RNA, se recurrirá al entorno de análisis de datos RapidMiner[1], que dispone de múltiples funcionalidades de análisis de datos.

En este enlace[2] puedes encontrar documentación sobre RapidMiner.

Descripción de los datos

A. Fuente de datos

Para crear los modelos de predicción se utilizarán los datos de *The Humanitarian Data Exchange*, en particular, el *Novel Coronavirus (COVID-19) Cases Data*[3] que recopila datos epidemiológicos del COVID-19 desde el 22 de enero de 2020 hasta el 10 de marzo del 2023. Los datos son recopilados por el *Center for Systems Science and Engineering* de la Universidad Johns Hopkins (JHU CCSE) a partir de varias fuentes.

Los datos para la realización de la práctica están disponibles en:

https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases

B. Formato de los datos

El fichero que se debe utilizar es el fichero diario de casos confirmados[4]. El conjunto de datos que tiene un total de 289 registros que corresponden a Provincias/Estados de distintos países/regiones. El conjunto de datos recopila el número de infectados totales desde el día 22 de enero de 2020 hasta el día 10 de marzo de 2023. Hay un total de 1147 atributos (columnas) por cada registro. Los primeros cuatro atributos corresponden al nombre de la provincia/estado, la región/país y la longitud y latitud de esta. El resto de los

[1] Contenidos relacionados con RapidMiner en https://academy.rapidminer.com/

[2] Se puede encontrar más información en https://docs.rapidminer.com/latest/studio/

[3] https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases

[4] time_series_covid19_confirmed_global.csv (Nota: aunque se indique que la fecha de actualización es el 24/3/2022, contiene todos los datos hasta el 9/3/2023).

atributos reflejan el número de contagiados acumulados por día. El conjunto de datos está en formato separado por comas (.csv).

C. Procesamiento de los datos

Una vez descargado el fichero de datos, se debe llevar a cabo el procesamiento de los mismos. Para ello, se puede:

- Importar el fichero csv en Excel para renombrar los atributos correspondientes a las fechas. Esto se realiza para que el modelo sea compatible con nuevos datos a la hora de realizar predicciones.
- Realizar ingeniería de características (e.g. número de días, medias, varianza, etc.)
- Exportar el fichero Excel a csv.
- Verificar que el fichero se haya generado correctamente[5].

I Parte: Regresión (RNA)

Para resolver un problema de regresión, como el que se plantea en este caso, se pueden utilizar los *k* días anteriores de las regiones/países seleccionados[6] para obtener el valor de los próximos días de cualquier región/país. Para ello se utilizarán los casos diarios, no los casos acumulados.

Cada uno de estos *k* valores podrán ser una entrada de la red de neuronas artificiales, así como los cuatro primeros atributos del conjunto de datos[7]. El valor que estimar será el valor de casos positivos del próximo día.

Una vez generado el fichero de datos, se utilizará el operador basado en una red neuronal artificial multicapa *feed-forward* entrenada con descenso de gradiente estocástico utilizando *back-propagation* implementada en RapidMiner (Deep Learning) para construir el modelo[8]. Hay que tener en cuenta que RapidMiner puede generar la arquitectura de la red. Sin embargo, esto no garantiza que se obtengan los mejores resultados. Por esta razón, es necesario experimentar con distintas arquitecturas y valores de los meta-parámetros. Por ejemplo, distintos números de entradas (e.g. variando el número de días), distintos números de nodos en la capa oculta, número de capas ocultas, etc[9].

El entrenamiento y testeo de la red se deberá realizar utilizando *Cross-Validation.* Los resultados de este proceso se deben analizar en detalle.

[5] El preprocesamiento de datos se puede llevar a cabo directamente en RapidMiner

[6] Países con datos muy dispares, darán un modelo de baja precisión para la realización de las predicciones.

[7] Se debe tener en cuenta que el proceso de recopilación de datos de los distintos países ha sufrido cambios a lo largo del tiempo y que existen variables exógenas que podrían afectar el comportamiento de los casos reportados (e.g. vacunación). La selección de datos de entrada tiene que estar justificada.

[8] Se puede utilizar el perceptrón multicapa (Neural Network) si no se utilizan atributos polinomiales.

[9] Se deben utilizar, al menos, 10 configuraciones de redes distintas.

Según la configuración experimental utilizada, al finalizar el experimento se puede obtener la siguiente información:

- Modelo que genera el algoritmo
- Coeficiente de correlación.
- Error cuadrático medio, error absoluto y error cuadrático relativo.
- Etcetera.

Por último, y dado que se quiere obtener pronósticos sobre la evolución de los contagios a nivel de país/región, se debe utilizar el modelo generado (Apply Model) para hacer la **predicción para TRES días consecutivos en España y en un país adicional a elección del grupo de trabajo.**

II Parte: Series temporales

El ingrediente fundamental de la minería de datos son los datos. Una de las formas en que se pueden presentar los datos de un dominio o problema en particular es mediante una serie temporal. Una serie temporal es una sucesión de datos medidos en determinados momentos y ordenados cronológicamente.

Para resolver un problema de regresión en una serie temporal, se utilizan los *k* valores anteriores para obtener el próximo valor (x_{t+1}). Cada uno de estos *k* valores serán una entrada del algoritmo de regresión (e.g. una red de neuronas artificiales) como se aprecia en la Fig 2.

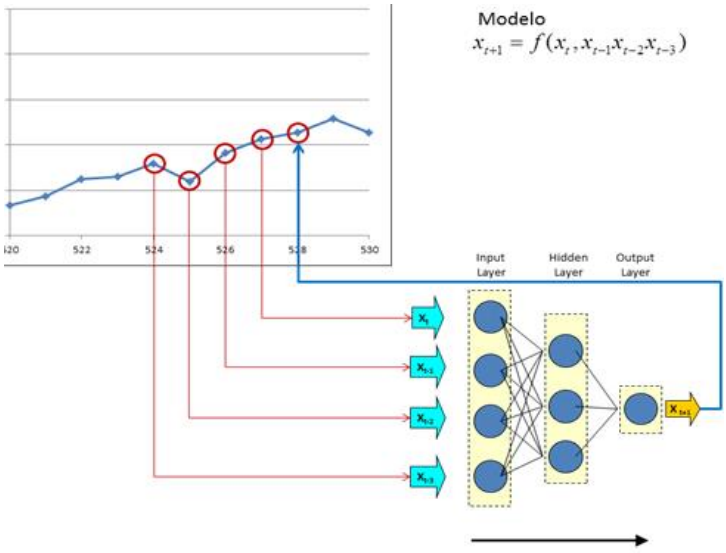

Figura 2. Predicción en series temporales con RRNN

Lo primero que hay que hacer es generar el conjunto de entrenamiento para el algoritmo de regresión (e.g. RNA). Este conjunto se crea a partir de los datos de la serie temporal teniendo en cuenta los k valores que formarán parte de la entrada (ver Figura 3).

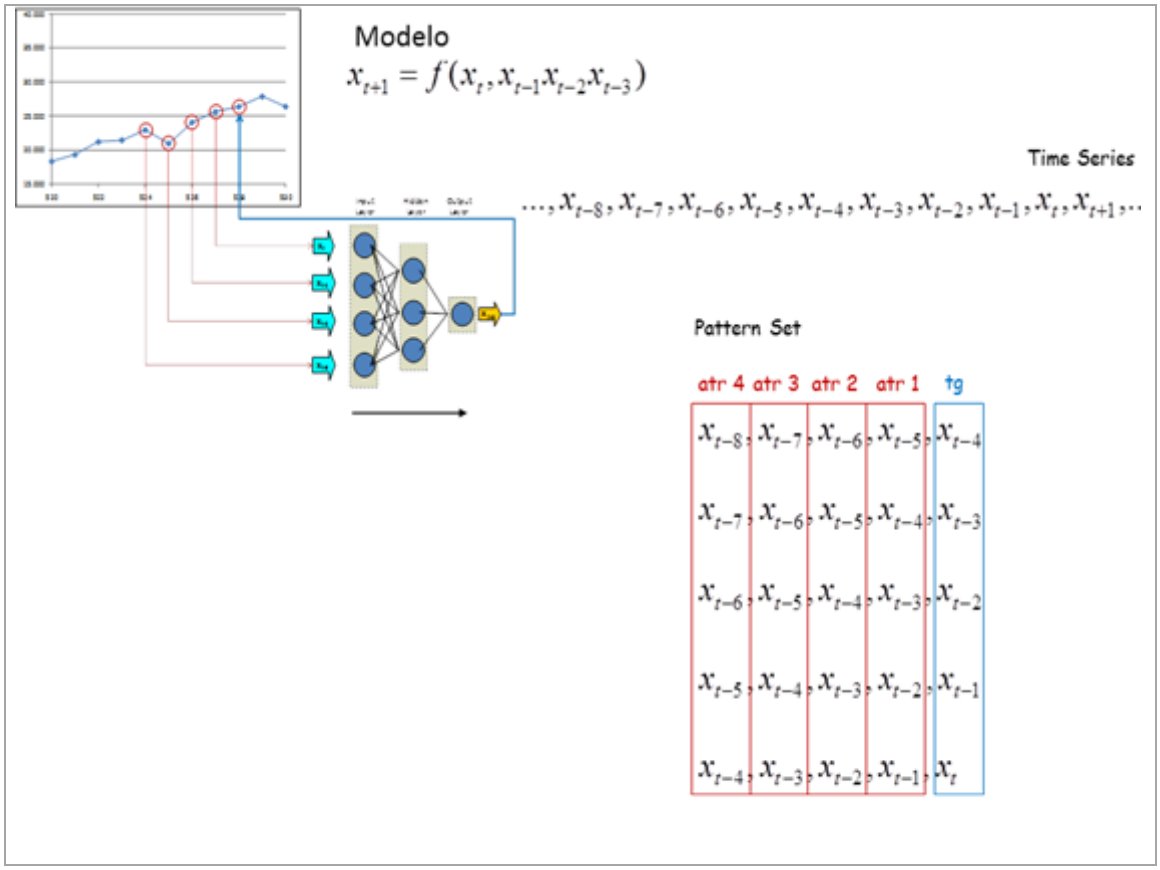

Figura 3. Generación patrones entrada para una RNA

A pesar de que la tipología de las RNA es muy variada, en este caso, al igual que en el apartado anterior, se utiliza el modelo de RNA denominado Perceptrón y con una topología Multicapa.

Dado que las RNA tienen múltiples meta-parámetros, para determinar cuál es el modelo que mejor se ajusta a los datos, deben analizarse varias configuraciones del RNA.

II Parte: Desarrollo

En esta parte de la práctica, se realizará la tarea de predicción utilizando los operadores para series temporales que proporciona la herramienta *RapidMiner*.

En esta parte de la práctica se deberán, utilizando Redes de Neuronas Artificiales, **predecir los 3 valores siguientes** de dos series relacionadas con la evolución de contagios del COVID-19 en dos países (España y el país seleccionado en la Parte I). Además, se analizarán en conjunto **dos series temporales de cada país** con el propósito de modelar conjuntamente las series temporales y capturar posibles dependencias entre ellas. Debido a esto, modelar varias series simultáneamente puede dar resultados diferentes para cada serie que modelarlas de forma individual. Para este apartado de la práctica, se utilizarán los datos de la evolución (diarios) de los contagios junto a los datos de vacunación (datos utilizados para la creación de las dos series temporales) para ambos países.

A. Obtención de los Datos:

- Los datos de evolución de los contagios se pueden obtener del sitio web de *The Humanitary Data Exchange*[10] (ficheros *confirmed global).*
- Los datos sobre la evolución de la campaña de vacunación mundial se pueden encontrar en el sitio web *Our World in Data*[11]*.*
- Seleccionar dos países (España y el país seleccionado en la Parte I) sobre los que se realizará el análisis.
- Generar un fichero csv por cada país que se analizará. Cada fichero debe contener las dos series temporales que se desean analizar (casos diarios y vacunados diarios).

B. Proceso de entrenamiento:

- Una vez seleccionadas las dos series temporales (de cada país) que se van a analizar se creará un conjunto de atributos (Windowing) que serán la entrada al algoritmo de regresión seleccionado (en este caso, deberá ser una red de neuronas).
- La mayoría de los atributos corresponden a valores anteriores de la variable dependiente. Sin embargo, se pueden incorporar otro tipo de atributos como medias, desviaciones, etc. En el caso de utilizar las dos series, los valores corresponderán a valores anteriores de las variables analizadas.
- En el proceso de entrenamiento es necesario probar con distintas arquitecturas; es decir, distintos números de entradas (para ello se seleccionarán diferentes valores en la opción *window size*), distintos números de nodos en la capa oculta, etc.

C. Predicción:

- Con cada uno de los dos modelos de RNA elegidos, se realizará la predicción de los siguientes 3 valores.
- Estos resultados deberán ser analizados y comparados.

[10] https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases
[11] Datos disponibles en este enlace: https://ourworldindata.org/covid-vaccinations

II Parte: Requisitos

Es importante incluir en la documentación de esta parte de la práctica lo siguiente:

- La descripción de las series temporales utilizadas.
- El proceso de entrenamiento: Las arquitecturas probadas, los problemas encontrados, etc. Además, se debe justificar la solución tomada para realizar la predicción.
- Mostrar los resultados: Mostrar gráficamente los resultados obtenidos en la fase de entrenamiento con las RNA.
- Comparación de los resultados: Pueden compararse los resultados obtenidos con una línea de tendencia adecuada.

Evaluación de la práctica

Aspectos para evaluar en la corrección de la práctica:

- Planteamiento y desarrollo del problema: 25%
- Resultados del problema: 25%
- Análisis de resultados y conclusiones: 25%
- Presentación: 15%
- Contexto de la práctica (información complementaria sobre el desarrollo de la práctica): 10%

Debe darse importancia a la presentación para de los resultados, el análisis de éstos, conclusiones, etc. Es importante tener en cuenta que el contexto de la práctica en la que se incluirá cualquier información relevante conocida por los estudiantes, casos similares, noticias al respecto o cualquier otra información de interés y relacionada con la práctica.

Entrega de la práctica

- La práctica deberá realizarse en grupos de **4 personas** (y entregarse únicamente por uno de los integrantes del grupo)**.**
- Esta práctica está dividida en dos partes. En este documento se detalla la primera parte. Sin embargo, la entrega de la práctica será un único documento en el que se detallen y analicen y relacionen las dos partes de la Práctica 1. La entrega será un documento. Además, ***se debe incluir en el documento un enlace a un fichero comprimido con todos los ficheros utilizados para la realización de la práctica.***
- La entrega de esta Práctica 1 (Parte I y II) se realizará con fecha máxima de entrega (por Aula Global):
  - Grupo 85: **3 de octubre – 23:50h**

- Grupos 843: **4 de octubre – 23:50h**

- No hay un formato de documento establecido. Sin embargo, es importante que toda la información se muestre de forma clara y su presentación también se considera como parte de la nota de la práctica.