Proyecto de aula: Deep Learning y Series de tiempo

Marco Teran[1,2]

[1]Deep Learning y Series de tiempo
[2]Escuela de Ciencias exactas e Ingeniería

This project aims to execute an effective Deep learning project using the methodology and tools presented in the course. The project will involve planning and implementing a practical project, where the data set used will be selected by the student. The ultimate goal is to derive valuable insights through extensive experimentation with Deep learning models that can contribute to the decision-making process in a particular application domain. The project will be implemented using the Python programming language and its environment of scientific computing tools, in the form of a notebook in the iPython format.

1. Introducción

Este proyecto tiene como propósito poner en práctica los conocimientos adquiridos en el curso de Deep Learning mediante la implementación de un proyecto práctico. La realización de un proyecto práctico es fundamental para consolidar y aplicar los conceptos teóricos aprendidos en el curso, permitiendo una comprensión más profunda y significativa del tema. En este proyecto, el estudiante seleccionará un conjunto de datos para planificar e implementar el proyecto. El objetivo final del proyecto es obtener valiosas conclusiones a través de la experimentación con modelos de Deep Learning, lo que permitirá contribuir en procesos de toma de decisiones en un dominio de aplicación particular. La implementación se llevará a cabo utilizando el lenguaje de programación Python y su entorno de herramientas de computación científica en forma de un cuaderno en formato iPython.

2. Objetivos

2.1. Objetivos general

Ejecutar un proyecto de *Deep learning* de forma efectiva usando la metodología y las herramientas presentadas en el curso. Llevar a cabo un proceso de análisis de datos completo utilizando la metodología CRISP-DM y aplicar técnicas y algoritmos de Deep Learning para resolver problemas complejos de clasificación, regresión y predicción en diferentes campos, con énfasis en imágenes, series de tiempo y datos no estructurados, con el fin de obtener conclusiones valiosas a partir de un conjunto de datos seleccionado.

2.2. Objetivos específicos

- Seleccionar y comprender el conjunto de datos a utilizar, identificando las variables relevantes y posibles relaciones entre ellas.
- Realizar una exploración de los datos para detectar anomalías y posibles errores en el conjunto de datos.
- Desarrollar y ajustar un modelo de Deep learning que permita obtener conclusiones valiosas a partir de los datos y evaluar el desempeño del modelo.

3. Descripción del proyecto

Se espera que utilice la metodología de trabajo propuesta en el curso y las herramientas de modelamiento para llevar a cabo la planeación y ejecución de un proyecto aplicado. El conjunto de datos sobre el que trabajará puede ser seleccionado por ustedes (entre los conjuntos de datos propuestos) de acuerdo con sus intereses. El objetivo es que a través de un proceso de extensiva experimentación con modelos de *Deep learning* poder llegar a obtener conclusiones con información valiosa que aporte en procesos de toma de decisiones en un dominio de aplicación particular.

El proyecto se desarrollará utilizando el lenguaje de programación *Python* y su entorno de herramientas para la computación científica, en forma de *Notebook* en el formato `iPynb`. Se debe presentar el proyecto tomando como referencia las etapas previas al despliegue de la metodología *CRISP-DM* para análisis de datos (IBM, 2012).

4. Conjuntos de datos

Como se mencionó anteriormente, el planteamiento y desarrollo del proyecto se debe basar en alguno de los siguientes conjuntos de datos:

- **Google Play Store Apps:** Datos de 10 mil aplicaciones de la *App Store* obtenidas a través de *web scraping* con el objetivo de analizar el mercado de *Android*. [acceder]
- **Trip Advisor Hotel Reviews:** 20 mil reseñas de hoteles extraídas de *Tripadvisor*. Se puede usar este conjunto de datos para descubrir cómo son los mejores hoteles o usarla en sus propios viajes. [acceder]
- **Netflix Movies and TV Shows:** Este conjunto de datos consiste en programas de televisión y películas disponibles en Netflix a partir de 2019. El conjunto de datos se recoge de Flixable, que es un motor de búsqueda de Netflix de terceros. En 2018, publicaron un interesante informe que muestra que el número de programas de televisión en Netflix casi se ha triplicado desde 2010. Utilizando este conjunto de datos, se puede averiguar: qué tipo de contenido se produce en qué país, identificar contenido similar a partir de la descripción y muchas más tareas interesantes. [acceder]
- **Avocado Prices:** Datos históricos de los precios del aguacate y volumen de ventas en múltiples mercados de estados unidos. Se puede modelar como una serie de tiempo. [acceder]
- **Fashion MNIST:** Un conjunto de datos similar a *MNIST* con 70 mil imágenes con tamaño $28x28$ de prendas de ropa. Presenta una tarea de clasificación. [acceder]
- **Students Performance in Exams:** Notas obtenidas por estudiantes en varias asignaturas.Estos datos se basan en la demografía de la población. Los datos contienen varias características como el tipo de comida que se le da al estudiante, el nivel de preparación para el examen, el nivel de educación de los padres y el rendimiento de los estudiantes en Matemáticas, Lectura y Escritura. Con los datos se pueden resolver varios tipos de problemas de regresión y clasificación. También se puede utilizar para encontrar qué factores pueden conducir a mejores resultados en los exámenes. [acceder]
- **Credit Card Fraud Detection:** Este conjunto de datos ayuda a las empresas y equipos a reconocer las transacciones fraudulentas con tarjetas de crédito. El conjunto de datos contiene transacciones realizadas por titulares de tarjetas de crédito europeas en septiembre de 2013. El conjunto de datos presenta detalles de 284807 transacciones, incluidos 492 fraudes, ocurridos durante dos días. [acceder]
- **Melbourne Housing Market:** El conjunto de datos del mercado de la vivienda de Melbourne es un recurso de aprendizaje favorito para los principiantes en la ciencia de los datos. Tiene muchas características: datos numéricos, categóricos e incluso geográficos (latitud y longitud). Por tanto, también puede utilizarse para el análisis geoespacial y otros problemas de agrupación. Del mismo modo, también se pueden realizar tareas de regresión y clasificación con este conjunto de datos. También hay numerosos ejemplos de código y guías disponibles para este conjunto de datos, lo que lo convierte en el conjunto de datos ideal para los estudiantes. [acceder]
- **IBM HR Analytics Employee Attrition & Performance:** Prediga el desgaste de sus empleados más valiosos. Descubra los factores que conducen al desgaste de los empleados y explora cuestiones importantes como *La relación entre la distancia de la casa al trabajo por puesto de trabajo y el desgaste* o *La relación entre el ingreso mensual promedio por educación y desgaste*. Este es un conjunto de datos ficticio creado por científicos de datos de *IBM*. [acceder]
- **UJIIndoorLoc:** Muchas aplicaciones del mundo real necesitan conocer la localización de un usuario para ofrecer sus servicios. La localización de usuarios ha sido un tema de investigación de interés en los últimos años. La localización de usuarios consiste en estimar la posición del usuario (latitud, longitud y altitud) mediante un dispositivo electrónico, normalmente un teléfono móvil. El problema de la localización en exteriores puede resolverse con gran precisión gracias a la tecnología GPS en los dis-

positivos móviles. Sin embargo, la localización en interiores sigue siendo un problema abierto, principalmente debido a la pérdida de la señal GPS en entornos interiores. Aunque existen algunas tecnologías y metodologías de posicionamiento en interiores, esta base de datos se centra en las basadas en huellas digitales WLAN (también conocidas como *WiFi Fingerprinting*). [acceder]

- **COVID19 Global Forecasting (Week 5):** En este desafío, tendrás que predecir el número diario de casos confirmados de COVID19 en varios lugares del mundo, así como el número de víctimas mortales resultantes, para fechas futuras. Este último reto incluye datos de condados del estado de EE.UU. El dataset se refiere a la quinta semana del desafío de pronóstico de COVID-19 en Kaggle, que es una competencia en la que se deben desarrollar pronósticos precisos para casos confirmados y fallecimientos relacionados con COVID-19 en diferentes regiones. El desafío tiene como objetivo identificar factores que parecen influir en la tasa de transmisión del COVID-19, no solo producir pronósticos precisos. La competencia se lanzó en colaboración con el grupo de investigación de la Oficina de Política Científica y Tecnológica de la Casa Blanca y otras organizaciones, y se basa en un conjunto de datos llamado COVID-19 Open Research Dataset (CORD-19), que fue diseñado para abordar preguntas científicas abiertas relacionadas con COVID-19. [acceder]

5. Entregables

5.1. Contenido de la primera entrega del proyecto

Para la primera entrega se requiere la extracción, preprocesamiento, visualización y análisis de los datos. Se deberá encontrar las principales características estadísticas de estos utilizando las herramientas vistas en clases. Estos se deberán representar y visualizar.
El archivo `ZIP` debe incluir los siguientes archivos:

Jupyter Notebook: con todo el código del proyecto. El *Notebook* debe estar debidamente explicado usando celdas de texto. Todos los pasos de carga, preprocesamiento y visualización de los datos, así como los respectivos archivos adicionales del modelo (si existen). Asegúrese de que el *Notebook* se visualiza correctamente y está libre de errores antes de enviarlo.

Reporte: un informe del trabajo en forma de artículo científico en formato `PDF` generado en LaTeX que documente los pasos de la metodología **CRISP** relacionados. El trabajo debe tener al menos estas secciones: una introducción que describa el problema y trabajo relevante asociado (no menos de 3 fuentes de literatura indexada donde se haya utilizado el repositorio); la descripción del método; visualización y análisis de los datos; y una sección de conclusiones.

Repositorio Repositorio `GIT` (el enlace debe estar al final del documento `PDF` antes de la Bibliografía): el repositorio debe contener carpetas: códigos, LaTeX, etc.

5.2. Contenido de la segunda entrega del proyecto

La aplicación de dos técnicas de *Deep learning*, sus respectivas métricas de evaluación y comparativa.
El archivo `ZIP` debe incluir los siguientes archivos:

Jupyter Notebook: con todo el código del proyecto. El *Notebook* debe estar debidamente explicado usando celdas de texto. Todos los pasos de carga, preprocesamiento, entrenamiento y prueba deben incluirse con el código del modelo, así como los respectivos archivos adicionales del modelo (si existen). Asegúrese de que el *Notebook* se visualiza correctamente y está libre de errores antes de enviarlo.

Reporte: un informe del trabajo en forma de artículo científico en formato `PDF` generado en LaTeX que documente todos los pasos de la metodología **CRISP**. El trabajo debe tener al menos estas secciones: una introducción que describa el problema y trabajo relevante asociado (no menos de 3 fuentes de literatura indexada donde se haya utilizado el repositorio); la descripción del método; la evaluación experimental, incluyendo la descripción de los conjuntos de datos, la configuración experimental, los resultados y la discusión; y una sección de conclusiones.

Repositorio Repositorio `GIT` (el enlace debe estar al final del documento `PDF` antes de la Bibliografía): el repositorio debe contener carpetas: códigos, LaTeX, etc.

5.3. Contenido de la tercera entrega del proyecto

Para la tercera entrega se requiere la aplicación de una técnica de *Deep Learning*, sus respectivas métricas de evaluación y comparativa.
El archivo `ZIP` debe incluir los siguientes archivos:

Jupyter Notebook: con todo el código del proyecto. El *Notebook* debe estar debidamente explicado usando celdas de texto. Todos los pasos de carga, preprocesamiento, entrenamiento y prueba deben in-

cluirse con el código del modelo, así como los respectivos archivos adicionales del modelo (si existen). Asegúrese de que el *Notebook* se visualiza correctamente y está libre de errores antes de enviarlo.

Reporte: un informe del trabajo en forma de artículo científico en formato `PDF` generado en LaTeX que documente todos los pasos de la metodología **CRISP**. El trabajo debe tener al menos estas secciones: una introducción que describa el problema y trabajo relevante asociado (no menos de 3 fuentes de literatura indexada donde se haya utilizado el repositorio); la descripción del método; la evaluación experimental, incluyendo la descripción de los conjuntos de datos, la configuración experimental, los resultados y la discusión; y una sección de conclusiones.

Repositorio Repositorio `GIT` (el enlace debe estar al final del documento `PDF` antes de la Bibliografía): el repositorio debe contener carpetas: códigos, LaTeX, vídeo

Póster: un archivo PDF con un póster que presente sus resultados. El póster debe mostrar de forma visual el problema, el método y los resultados experimentales.

- **Por favor, no incluya imágenes o archivos binarios diferentes a los solicitados.** Todos los archivos de la presentación deben estar comprimidos en un único archivo `ZIP`.
- El archivo debe ser nombrado como `dl-project-username1-username2-username3.zip`, donde nombre de usuario es el nombre de usuario asignado por la universidad en su correo (incluir los nombres de usuario de todos los miembros del grupo).
- El archivo debe ser enviado a mi correo antes de la medianoche de la **fecha límite**.

5.4. Recomendaciones para tener éxito en el proyecto:

- Planificar el proyecto con suficiente tiempo para cada una de las etapas del proceso de análisis de datos.
- Familiarizarse con las herramientas de machine learning y Python antes de comenzar el proyecto.
- Asegurarse de que el conjunto de datos seleccionado es adecuado para el problema que se busca resolver.
- Documentar todo el proceso de análisis de datos para facilitar la evaluación y comunicación de los resultados.
- Trabajar en equipo y aprovechar las habilidades y conocimientos individuales de cada miembro para lograr los objetivos del proyecto de manera efectiva.

6. Bibliografía

IBM. *"Manual CRISP-DM de IBM SPSS Modeler."* CRISP-DM, 2012. [Descargar]