Respuestas a la prueba de selección de Gridfy

Jesús García León

23 de septiembre de 2024

1. Experiencia con Python, pandas y numpy

En tu experiencia con Python, ¿has utilizado librerías como pandas o numpy para analizar datos?

He utilizado `pandas` y `numpy` en varios proyectos académicos, además de otras librerías como `Keras`, `fastai` y `scikit-learn`. Algunos de los proyectos en los que he utilizado estas herramientas son:

- **Optical Recognition of Handwritten Digits**: El objetivo de este proyecto era crear un clasificador que permitiese el reconocimiento de números manuscritos del 0 al 9. Los números eran codificados en imágenes de 8x8 pixeles, y cada pixel podía tomar un valor entre 0 y 15 en función de la intensidad del trazo.

  Para obtener dicho clasificador, apliqué `pandas` para almacenar y realizar una análisis exploratorio de los datos: descripción general del conjunto de datos, balanceo de clases, número de características, información que aporta cada característica realizando un análisis de su varianza... Con la información obtenida decidí realizar un PCA (Análisis de componentes principales), para reducir la dimensión del problema, ya que había varias características cuya varianza era cercana a 0. A continuación entrené varios modelos de clasificación con `scikit-learn`. Los modelos que decidí utilizar fueron: SVM, Regresión Logística, Random Forest y KNN. Finalmente mediante `pandas` creé una tabla comparativa de los modelos, usando las métricas habituales y una matriz de correlación para decidir que modelo era el óptimo para resolver este problema de clasificación. Para permitir que el experimento fuese replicable bajo las mismas condiciones, utilicé `numpy` para fijar la semilla del generador de números aleatorios.

- **Clasificación de imágenes en el conjunto de datos Cifar-100**: El conjunto de datos CIFAR-100 (Canadian Institute for Advanced Research, 100 clases) es un subconjunto del conjunto de datos Tiny Images y consta de 60,000 imágenes a color de 32x32 píxeles. Las 100 clases en el CIFAR-100 están agrupadas en 20 superclases. Hay 600 imágenes por clase. Cada imagen tiene una etiqueta "fina"(la clase a la

que pertenece) y una etiqueta "gruesa"(la superclase a la que pertenece). Hay 500 imágenes de entrenamiento y 100 imágenes de prueba por clase. El objetivo de este proyecto era construir una red neuronal convolucional que permitiese la clasificación de las imágenes de las 25 primeras clases. La red debía crearse desde 0 usando fastai, no estaba permitido el fine-tuning de otra red, de forma que se incentivaba la búsqueda de soluciones para evitar problemas de underfitting y overfitting.

Utilizando `fastai`, entrené un modelo de redes neuronales convolucionales para clasificar las imágenes en 25 categorías. Solo utilicé 25 categorías de las 100 disponibles para permitir que el proyecto pudiese realizarse con la limitada capacidad de memoria de la GPU que disponía en ese momento. Para este proyecto creé un modelo base, al que fui añadiendo capas para mejorar su capacidad de predicción y disminuir el error de generalización.

2. Proyecto con Python y SQL

Describe un proyecto en el que hayas utilizado Python y SQL para resolver un problema.

Un proyecto académico en el que utilicé `Python` y `SQL` fue en la elaboración de un software para gestión de hoteles. El software consistía en una interfaz programada en `Python` que permitía al usuario conectarse a una base de datos `SQL` de forma sencilla y realizar consultas sobre el número de habitaciones libres, las reservas realizadas por los clientes, los desperfectos, y las reparaciones pendientes. De esta forma, el usuario podía consultar añadir o eliminar los datos que necesitase, sin escribir comandos `SQL` y sabiendo que la base de datos se mantendría en un estado consistente en todo momento.

El mayor de los problemas que tuvimos al realizar este proyecto fue a la hora de establecer la conexión a la base de datos, ya que la mayoría de los controladores que permitían conectarse eran de pago o estaban desactualizados. Sin embargo, tras una búsqueda exhaustiva conseguimos arreglar el problema y encontrar un controlador que nos permitiese conectarnos a la base de datos.

En este caso, el proyecto no está relacionado con el tratamiento ni análisis de datos, pero combinando lo aprendido en este proyecto para manejar bases de datos `SQL` desde `Python` y los numerosos proyectos de análisis de datos que realizado en `R` y `Python`, soy capaz de realizar proyectos en los que necesite usar ambas tecnologías en conjunto.

Por último, un proyecto en el que tuve varias dificultades en el análisis de datos fue en el YearPredictionMSD. Este conjunto de datos contiene información (características sonoras) de distintas canciones, y el objetivo es predecir el año de publicación de la canción (hay 89 posibles años a predecir). Cada fila/ejemplo contiene 90 valores reales y un valor entero, correspondiente con el año de la canción, en el rango [1922,2011]). En total hay 515345 ejemplos/instancias. El problema de este conjunto es que las clases estaban altamente desbalanceadas, y la mayoría de los datos se encontraban en unas pocas clases, dificultando enormemente la clasificación de las canciones de las clases minoritarias.

Para paliar este problema, utilicé una función de pérdida con pesos, que añadía una mayor penalización a los errores cometidos en las clases minoritarias, y un sobremuestreo de las clases con menos ejemplos para aumentar artificialmente el número de ejemplos de estas clases. De esta forma, aumentó enormemente la potencia del clasificador, sin embargo, debido al enorme desbalanceo, era imposible encontrar un buen clasificador. De este proyecto aprendí la importancia de tener unos buenos datos, y la necesidad de prestar atención a la parte de recopilación de la información. Para tener buenos modelos, se necesitan buenos datos, y para ello es necesario planear con anterioridad que datos se van a recopilar siempre que sea posible. Cuando se diseña una base de datos, si se prevé que se va a utilizar modelos de inteligencia artificial sobre esa base de datos, se pueden añadir tuplas y tablas con datos adicionales que en principio no eran necesarios para resolver el problema planteado inicialmente, pero que pueden ser utilizados por modelos de IA para el análisis y previsión de errores. Por ejemplo, si se diseña una base de datos de aerogeneradores, en la que se añaden de los aerogeneradores, sus especificaciones técnicas, su localización... también se puede añadir estadísticas de uso diarias, producción de energía, condiciones climáticas... de forma que puedan ser analizadas para predecir el comportamiento del del aerogenerador a lo largo del tiempo y prevenir errores y realizar mejoras.

3. IA en Redes Eléctricas

En el contexto de redes eléctricas, ¿cómo crees que la IA puede ayudar a mejorar la operación y mantenimiento de las redes?

La inteligencia artificial tiene el potencial de hacer las redes eléctricas más seguras, baratas y eficientes. Mediante el análisis de los datos de consumo de la red se puede monitorizar y predecir el consumo de la red en un determinado momento, lo que permite aumentar o reducir la generación de energía para adaptarse con antelación a las necesidades de consumo de forma que no se desperdicie energía ni se produzca falta de potencia que puedan ocasionar cortes de suministro a los usuarios. Además mediante el uso de redes inteligentes se puede redirigir la energía de forma más eficiente en caso de que se produzca algún fallo en alguna de las vías de transmisión, se pueden optimizar las rutas ya existentes, de forma que se pueda determinar que rutas requieren de una mayor capacidad y cuales se utilizan por debajo de sus capacidades. Además, el análisis de los datos de uso, puede ayudar a prevenir errores en los dispositivos de generación y transporte de energía de forma que se pueda prevenir su rotura y se reparen antes de ser un problema. Por último la inteligencia artificial puede ayudar a mejorar los dispositivos, creando modelos más eficientes y robustos a prueba de errores. Un ejemplo de uso podría ser un parque eólico, donde las condiciones de producción de energía son cambiantes, ya que dependen directamente del viento, se podría utilizar Inteligencia Artificial para controlar cuantos aerogeneradores deben encenderse para abastecer a la red en función de las condiciones climáticas y la demanda prevista de energía. De esta forma, el sistema permitirá apagar aerogeneradores de forma automática para evitar su desgaste cuando se prevea que la demanda es baja y volver a activarlos cuando sea alta, además de planear periodos de

almacenamiento de energía para abastecerse antes de momentos de alta demanda. Una herramienta muy útil para obtener los datos necesarios para los modelos inteligentes son los gemelos digitales.

4. Gemelos Digitales en el Sector Energético

¿Qué entiendes por "gemelos digitales" y cómo crees que esta tecnología puede aplicarse en el sector energético?

Los gemelos digitales son representaciones hechas por ordenador de elementos físicos existentes o de procesos que permiten tener un modelo digital con el que realizar simulaciones, monitorizar, analizar y predecir comportamientos del objeto o proceso al que representan. Son modelos que combinan sensores para captar los datos del mundo real, que ayudaran a crear un modelo con mayor precisión y que luego serán analizados con técnicas de aprendizaje automático, para mejorar su funcionamiento. Además, es posible crear mejoras en estos modelos digitales y realizar simulaciones para predecir el impacto de estas en el objeto sin tener que construirlas físicamente, lo que ayuda a ahorrar tiempo y dinero, ya que no es necesario construir físicamente los cambios hasta que no se considere que pueden producir buenos resultados. También se puede utilizar los datos recopilados por los sensores para prevenir errores físicos o de funcionamiento usando inteligencia artificial y reducir los tiempos de inactividad.

Dentro del sector energético, pueden ser utilizados para representar una turbina eólica, una presa, e incluso una red eléctrica entera. Este último es el caso de las *SmartGrids*, con un gemelo digital de una red eléctrica inteligente, se puede optimizar el suministro eléctrico, ajustar la producción en función de la demanda, detectar anomalías, y simular el impacto de diferentes eventos, como cortes de energía, sobrecargas, picos de consumo y diferentes eventos climáticos. Esto es especialmente útil en la integración de energías renovables intermitentes (como la solar o eólica), ya que permite predecir la generación de energía en función de las condiciones climáticas y ajustarla para satisfacer la demanda en tiempo real. Las simulaciones permiten tomar decisiones informadas sobre cómo redistribuir la energía, identificar posibles cuellos de botella, o planificar el mantenimiento preventivo de los componentes críticos.

5. Propuesta de Mejora en un Proyecto

¿Puedes contarnos sobre alguna idea o mejora que hayas propuesto en un proyecto académico o personal?

En un proyecto sobre segmentación de imágenes usando transformers, queríamos realizar *transfer learning* de un modelo `mask2former` entrenado para segmentación panóptica. Para ello, decidimos hacer un *fine-tunning* del modelo, sin embargo, los recursos de GPU

disponibles no nos permitían cargar el modelo entero con conjuntos de entrenamiento grandes, y si utilizabamos lotes muy pequeños el entrenamiento era poco eficaz y no terminaba de adaptarse a los nuevos datos. Para solucionar este problema, propuse analizar las capas de la red para decidir que capas podíamos congelar y que capas debíamos permitir que se adaptasen. Al reducir el número de capas pudimos entrenar la red con lotes pequeños y obtuvimos resultados mucho más precisos.

Otra idea que he propuesto la estoy desarrollando en mi Trabajo de Fin de Grado, en el que estoy trabajando con distintos modelos de optimización para alta dimensionalidad. La mejora que se propone en el TFG es realizar un análisis de la dependencia de de las variables y dividir estas en grupos más pequeños de variables independientes, de forma que se puedan aplicar los algoritmos a estos grupos reducidos y disminuir la complejidad del problema. El objetivo del trabajo es decidir si es preferible utilizar algoritmos diseñados para alta dimensionalidad, o es mejor observar previamente las dependencias que se establecen entre las variables que intervienen en el proceso de optimización, lo que añade un coste adicional al problema, y aplicar algoritmos de alta o baja dimensionalidad a cada grupo por separado.