Universidad Intercontinental de la Empresa

PLE02

Aprendizaje Automático

Borja González Seoane

S08, 23 de octubre de 2024

1. Descripción de la práctica

En la PLE02, la segunda de las prácticas de laboratorio evaluadas de la asignatura de Aprendizaje Automático, se trabajará en la selección y evaluación de modelos, intentando además abarcar las otras cuestiones propias de la Unidad III al respecto de la persistencia y la gestión de proyectos.

Para la realización de la práctica, se empleará el mismo entorno de laboratorio habitual de la asignatura, basado en Python sobre Jupyter Notebooks, y las librerías de análisis de datos y aprendizaje automático ya introducidas en las prácticas anteriores: Scikit-Learn, Pandas, NumPy, Matplotlib, Seaborn, etc.

En este caso, el estudiante deberá plantear un proyecto de AA muy sencillo, pero compuesto por varios pasos distintos que se integren en una canalización de datos. Véanse las transparencias de la sesión S07 relativas a gestión de proyectos de AA. El estudiante tiene libertad para diseñar la estructura del proyecto, pero se espera que incluya al menos los siguientes pasos en sendos *notebooks* separados:

1. Breve análisis exploratorio de datos (EDA) del conjunto de datos, en el que necesariamente se deberá analizar la correlación entre las variables, además de cualquier otro análisis que se considere relevante. Puede tomarse la PLE01 como referencia para este paso, si bien no es necesario un análisis tan exhaustivo.

2. Preprocesamiento de datos: limpieza de datos, codificación de variables categóricas en numéricas, normalización de variables, eliminación de variables irrelevantes o correlacionadas, sustitución de variables correlacionadas por una sola que las represente, etc. No se trata de realizar transformaciones por el mero hecho de hacerlas, sino de incluir aquellas que se consideren necesarias para mejorar la calidad de los datos. Es posible que se necesite volver atrás desde los pasos posteriores para modificar el preprocesamiento.

3. Partición de los datos en conjuntos de entrenamiento y test. Probablemente este paso sea interesante incluirlo en el *notebook* anterior, de preprocesamiento, pero se deja a criterio del estudiante.

4. Selección y evaluación de modelos: se deberán probar al menos tres modelos distintos. Se recomienda emplear modelos de Scikit-Learn, pero se permite el uso de implementaciones propias. En el entrenamiento de los modelos se deberá incluir la validación cruzada y la exploración de hiperparámetros. Con respecto a estos últimos, se espera que se documente brevemente y en el propio *notebook* los hiperparámetros a estudiar. Se espera que se incluya algún tipo de control al respecto del sobreajuste de los modelos.

5. Evaluación de los modelos: se deberán comparar los modelos en base a las métricas seleccionadas, justificando la elección de las mismas. Se espera que el *notebook* permita seleccionar el mejor modelo de entre los probados.

1.1. Persistencia de artefactos

Además de los *notebooks* con los pasos anteriores, se espera que el estudiante persista los artefactos generados en la ejecución de los mismos. En concreto, se espera que se persistan los datos transformados tras el preprocesamiento y la partición, y los modelos entrenados. Aunque se reitera la libertad del estudiante para diseñar la estructura del proyecto, es necesario al menos llegar a persistir tanto un CSV —o soporte equivalente para datos tabulares— como algún archivo de pesos de modelo.

Con respecto al protocolo de serialización de los modelos, el estudiante puede referirse a las transparencias de la sesión S07, en las que a su vez se cita la documentación de Scikit-Learn al respecto. Dentro de lo expuesto en esa documentación, el estudiante tiene libertad para elegir el protocolo de persistencia que considere más adecuado.

1.2. Sugerencia de nombrado de archivos

Los nombres en concreto de los *notebooks* y de los archivos persistidos se dejan a criterio del estudiante, puesto que dependen en gran medida de la estructura de la canalización que haya diseñado. Se ha indicado ya que en las transparencias de la sesión S07 se puede encontrar una sugerencia típica de la que se puede partir. En cualquier caso, se recomienda que los nombres sean descriptivos y se empleen índices para ordenar los pasos de la canalización. Una propuesta opcional sería la siguiente: `ple02_<indice>_<descripcion>.ipynb` para los *notebooks* y `ple02_<descripcion>.<formato>` para los archivos persistidos, donde `<indice>` sería un número que indicase el orden de ejecución de los *notebooks*, `<descripcion>` sería una descripción breve del contenido del archivo y `<formato>` sería la extensión del archivo. Ejemplo: `ple02_00_eda.ipynb`, `ple02_data_prep.csv`.

1.3. Conjunto de datos y línea de trabajo de AA

En el Campus Virtual se disponibilizará un conjunto de datos en formato CSV: `ple02_gimnasio.csv`. Este conjunto de datos contiene información recopilada en un gimnasio sobre ciertos parámetros de las sesiones de entrenamiento de sus clientes. El objetivo de la práctica será **predecir la cantidad de calorías quemadas** en una sesión de entrenamiento en función de otras variables.

2. Entrega y evaluación de la práctica

La entrega de la práctica se realizará a través del Campus Virtual, donde se habilitará una tarea específica para ello. Los entregables de la tarea serían todos los *notebooks* empleados en la canalización de datos y los archivos persistidos generados. Además, se deberá añadir un archivo `README.md` en el que se describa someramente la estructura de la canalización de datos y se indique cómo ejecutarla. Se deberán comprimir todos los archivos en un único archivo ZIP para su entrega, cuyo nombre deberá seguir el siguiente patrón: `ple02_<apellidos>_<nombre>.zip`.

Además de la propia sesión S08, en la que el estudiante podrá trabajar en el ejercicio y consultar las dudas que le surjan, se dejará una semana adicional para completar la práctica y enviarla a través del Campus Virtual. Así pues, la tarea se cerrará una semana después de la sesión S08.

Una vez depositada la tarea, el profesor procederá a su revisión, para lo que podría convocar a los estudiantes a una sesión de reunión de defensa de haber aclaraciones que matizar. En caso de que aplicase se notificaría debidamente a los estudiantes. A este respecto, cabe destacar que se le podría solicitar al estudiante **explicar cualquier parte de la práctica pormenorizadamente**, incluyendo **cualquier fragmento del código entregado**.

La evaluación de la práctica se realizará basándose en los siguientes criterios:

- Buen diseño de la canalización del proyecto: estructura clara, pasos bien definidos, persistencia de artefactos, etc.

- Profundidad en el entrenamiento y evaluación de los modelos: selección de modelos, validación cruzada, exploración de hiperparámetros, control de sobreajuste, etc.
- Calidad del modelo final obtenido.
- Calidad del código y de la documentación distribuida en los *notebooks*.

Se recuerda que la calificación de esta práctica supondrá un 5 % de la calificación global de la asignatura, tal y como figura en el Contrato de Enseñanza-Aprendizaje.

Versión del documento: 20 de octubre de 2024