Facultad de Ciencias Exactas, Ingeniería y Agrimensura

Tecnicatura en Inteligencia Artificial

Aprendizaje Automático 1

Trabajo Práctico N°1: modelo predictivo de tarifas de Uber

Objetivos

Familiarizarse con la biblioteca scikit-learn y las herramientas que brinda para el pre-procesamiento de datos, la implementación de modelos de regresión lineal con diversos hiperparámetros y la evaluación de métricas de regresión.

Dataset

El dataset se llama uber_fares.csv y contiene información de tarifas de viajes realizados, además de distintas variables características, como se detallan a continuación:

Características de entrada:

key - un identificador único para cada viaje.
pickup_datetime - fecha y hora en que se activó el taxímetro.
passenger_count - el número de pasajeros en el vehículo (valor ingresado por el conductor).
pickup_longitude - la longitud donde se activó el taxímetro.
pickup_latitude - la latitud donde se activó el taxímetro.
dropoff_longitude - la longitud donde se desactivó el taxímetro.
dropoff_latitude - la latitud donde se desactivó el taxímetro.

Variable de salida (target):
fare_amount - el costo de cada viaje en USD

Para todos los ítems, incorporar una cantidad de texto adecuado en forma de comentarios, ya sea para la comprensión del código (usualmente una línea de comentario por cada celda) como para explicar las decisiones tomadas a lo largo del trabajo (por ejemplo, la justificación de la imputación de valores faltantes, la elección de las métricas adecuadas, entre otros). Mantener la coherencia con los comentarios.

Consignas

1. Armar grupos de tres personas para la realización del trabajo práctico. Dar aviso al cuerpo docente del equipo mediante el correspondiente formulario. En caso de no tener compañero, informar al cuerpo docente. **Se recomienda que al menos dos integrantes hayan aprobado Fundamentos de Ciencias de Datos.**

2. Crear un repositorio que se llame “AA1-TUIA-2025C1-Apellido1-Apellido2-Apellido3” en GitHub.

3. Realizar un análisis descriptivo, que ayude a la comprensión del problema, de cada una de las variables involucradas en el problema detallando características, comportamiento y rango de variación.

   Debe incluir:

- Análisis y decisión sobre datos faltantes.
- Análisis y decisión sobre datos atípicos.
- Visualización de datos (por ejemplo histogramas, scatterplots entre variables, diagramas de caja)
- Codificación de variables categóricas (si se van a utilizar para predicción).
- Matriz de correlación de variables.
- Estandarización o escalado de datos.
- Validación cruzada train - test. Realizar una división del conjunto de datos en conjuntos de entrenamiento y prueba (y si se quiere, se puede incluir validación, que luego será útil) **en el MOMENTO donde ustedes lo crean adecuado.**

4. Implementar la solución del problema de regresión con regresión lineal múltiple.
    - Probar con el método **LinearRegression**.
    - Probar con métodos de **gradiente descendiente**. ¿Algún cambio? Incorporar gráficas de Error vs Iteraciones (loss vs epochs). Agregar comentarios.
    - Probar con métodos de regularización (**Lasso, Ridge, Elastic Net**).
    - Obtener las **métricas adecuadas** (entre R2 Score, MSE, RMSE, MAE, MAPE, elegir) tanto para entrenamiento como para prueba. **¿Por qué para ambos conjuntos?**
    - ¿Creen que han conseguido un buen fitting?

5. Optimizar la selección de hiperparámetros.
    - Variar los hiperparámetros de gradiente descendiente. ¿Qué observa?
    - Variar los hiperparámetros de Lasso y Ridge. ¿Qué observa?

6. Comparación de modelos.
    - Incluyan en su análisis una comparación de modelos: de todos los modelos de regresión, ¿cuál es el mejor? **Escoger una métrica adecuada para poder compararlos**.

7. Escribir una conclusión del trabajo.

8. **Preparar una defensa del trabajo práctico**: la defensa consiste en preguntas hechas por el cuerpo docente que pueden ser: explicar una parte del código, explicar alguno de los métodos utilizados, preguntas de índole teórica, preguntas de índole práctica. Son tanto grupales como individuales.

Entrega

Las entregas parciales se realizan mediante GitHub (suben el código y nos devuelven el link del repositorio **mediante este formulario: https://forms.gle/CW85afc7NQUB4JFe9)**

El repositorio debe llamarse “AA1-TUIA-2025C1-Apellido1-Apellido2-Apellido3” **sin excepciones. No crear carpetas dentro que contengan alguno de los entregables.**

Respetar los siguientes nombres:

Notebook de trabajo: TP-regresion-AA1.ipynb

Fecha de entrega: Domingo 13/04/2024, a través de este formulario.

Cada entrega puede demorarse hasta dos días después de la fecha pactada, **con disminución de la nota final del trabajo práctico.**

No se aceptan entregas finales con fecha posterior al 15/04/2024. En caso de no tener todos los ítems entregados para esta fecha, la condición es automáticamente de desaprobado.

La defensa de los TP se hará de forma presencial , en horarios de clase, separados por turnos que la cátedra asignará según el orden en el que se fue entregando. En caso de detectar errores o una presentación en la que falten conocimientos sobre el trabajo realizado que se consideren lo suficientemente graves, se pactará una fecha para una segunda defensa en mesas de examen (donde deberán estar realizadas las correcciones y más acertada la presentación). *En caso de reprobar en esta segunda instancia de defensa, la condición es de libre*.

En caso de no aprobar ni la defensa del TP de regresión ni el de clasificación, *la condición es de libre*.