Facultad de Ciencias Exactas, Ingeniería y Agrimensura

Tecnicatura en Inteligencia Artificial

Aprendizaje Automático 1

Trabajo Práctico: Predicción de precios de casas

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.

Dataset

El dataset se llama house-prices.csv y contiene información de precios de casas de Boston, además de otras variables características, como se detallan a continuación:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222.0	18.7	396.90	5.33	36.2

Características de entrada en orden:

1) CRIM: tasa de criminalidad per cápita por ciudad
2) ZN: proporción de terrenos residenciales zonificados para lotes de más de 25,000 pies cuadrados
3) INDUS: proporción de acres de negocios no minoristas por ciudad
4) CHAS: variable dummy del río Charles (1 si el tramo limita con el río; 0 de lo contrario)
5) NOX: concentración de óxidos de nitrógeno (partes por 10 millones) [parts/10M]
6) RM: número promedio de habitaciones por vivienda
7) AGE: proporción de unidades ocupadas por sus propietarios construidas antes de 1940
8) DIS: distancias ponderadas a cinco centros de empleo de Boston
9) RAD: índice de accesibilidad a las autopistas radiales
10) TAX: tasa de impuesto sobre la propiedad a valor completo por $10,000 [$/10k]
11) PTRATIO: proporción alumno-maestro por ciudad
12) B: El resultado de la ecuación B=1000(Bk - 0.63)^2 donde Bk es la proporción de negros por ciudad
13) LSTAT: % de población de menor estatus socioeconómico

Variable de salida (target):

14) MEDV: Valor mediano de las viviendas ocupadas por sus propietarios en miles de dólares [k$]

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 hasta tres personas para la realización del trabajo práctico. Dar aviso al cuerpo docente del equipo. 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-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.
   - 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-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: Viernes 20/09/2024, 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 22/09/2024. 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*.