Ejercicios clase: Regresión Lineal

Regresión lineal y polinomial con Boston Housing dataset

1. Carga el dataset de Boston Housing dataset. En cada ejercicio puedes dividir en train y test. **(Tarde)**

2. Dentro de la pagina de Kaggle de Boston Housing, en la vista general se encuentra la descripción de las distintas variables contenidas en el dataset. La variable objetivo a predecir es *medv.*
    a. Haz un histograma de la variable objetivo con la función *hist*.
    b. Ahora haz un histograma de la variable objetivo mediante la función *distplot* de la librería *seaborn*.
    c. Verifica que no hay ningún missing en ninguna de las variables del dataset.
    d. Haz un scatter con cada una de las variables *age*, *dis*, *lstat* y *rm* con la variable a predecir *medv*. Entiende las variables e interpreta los gráficos. Si estimaramos un modelo de regresión lineal con estas 4 variables:
        i. ¿qué signo esperas para cada uno de los coeficientes de estas variables?
        ii. ¿tiene sentido el signo esperado teniendo en cuenta la relación económica entre la variable y la variable objetivo?

3. Estima 4 regresiones lineales simples distintas, una para cada una de las 4 variables anteriores.
    a. Obtén el coeficiente de la pendiente de cada recta y grafica la recta de ajuste junto a la nube de puntos. ¿Coincide el signo con lo que esperabas en el apartado d anterior?
    b. ¿Cuál de los 4 modelos obtiene mejor MSE? ¿Cuál obtiene mejor R2?

4. Ahora estima una regresión incorporando solamente las variables *lstat* y *rm*.
    a. Obtén los coeficientes de cada variable. ¿Siguen presentando el mismo signo que por separado?
    b. ¿El nuevo modelo presenta mejor MSE y mejor R2?
    c. Gráfica la variable objetivo (la y real) contra la predicción y analiza el gráfico.
    d. ¿Cual es el efecto marginal de cada variable?

5. Ahora estima una regresión incorporando solamente las variables *lstat*, *rm* y *age*.
    a. Obtén los coeficientes de cada variable. ¿Siguen presentando el mismo signo que por separado?

b. ¿El nuevo modelo presenta mejor MSE y mejor R2 que el anterior?
c. ¿Qué conclusión sacas de incorporar la variable *age*?

6. Ahora estima una regresión polinomial de orden 2, solamente incorporando las variables *lstat* y *rm*.
    a. Obtén los 6 coeficientes del polinomio. Identifica cada coeficiente con el término del polinomio al que pertenece: *lstat*, *lstat*^2, *lstat*rm*…
    b. ¿El nuevo modelo presenta mejor MSE y mejor R2 que el anterior?
    c. Gráfica la variable objetivo contra la predicción y analiza el gráfico. Compara este gráfico con el del apartado 4c.

7. Haz todas las pruebas que se te ocurran para obtener la mejor regresión lineal posible.