Trimestre 21-I

Minería de datos
Práctica 1 “Regresión lineal y lineal ponderada”

La práctica se deberá realizar en equipos de 2 personas.

Datasets

1. WATER (water.csv): cuenta con 864,863 instancias, un atributo (**Salnty**) y una variable de salida (**T_degC**). **Salnty** representa los gramos de sal por kilogramo de agua marina. **T_degC** es la temperatura del agua marina en grados centígrados. Por lo tanto, el objetivo será predecir **T_degC** con base en **Salnty**.
2. MTCARS (mtcars.txt): cuenta con 32 instancias, 11 atributos (cyl, disp, hp, drat, wt, qsec, vs, am, gear, carb). De estos atributos, se ocuparán **disp** y **wt** como atributos y **hp** como la variable de salida. **disp** es el desplazamiento del motor, **wt** es el peso del autómovil y **hp** son los caballos de fuerza. Por lo tanto, el objetivo será predecir **hp** con base en **disp** y **wt**.

Procedimiento

1. Análisis exploratorio **(5 puntos)**.
    a. Explorar los datos para buscar errores (p.ej., datos faltantes).
    b. Generar el resumen con datos de estadística descriptiva del dataset.
    c. Generar los *boxplots* correspondientes para analizar el comportamiento de los datos, buscar *outliers*.
    d. Generar gráfica de dispersión.
2. Limpieza de datos **(10 puntos)**.
    a. En caso de haber datos faltantes, reparar los valores empleando la media para el atributo en cuestión.
    b. Verificar que la media antes y después del procedimiento de reparación no haya sido afectada.
3. Regresión lineal **(40 puntos)**.
    a. Investigar el uso de la función `LinearRegression` de Scikit-Learn.
    b. Aplicar la función `LinearRegression` para generar el modelo de regresión lineal.
        i. Si el dataset es muy grande, aplicar la metodología de validación simple con una proporción de 80% para el training set y 20% para el test set. Recuerda que la generación de estos conjuntos debe ser aleatoria.

ii. Si el dataset es muy pequeño, aplicar la metodología *n-fold cross validation* para generar los training sets y test sets.

4. Evaluación **(10 puntos)**.
    a. Si el dataset es muy grande, medir la precisión siguiendo la metodología de validación simple. Reportar precisión obtenida.
    b. Si el dataset es muy pequeño, medir la precisión siguiendo la metodología *n-fold cross validation*. Reportar precisiones obtenidas y la precisión final.
    c. Recuerda que la precisión se debe medir usando la función de error cuadrático medio, donde se compara el valor real de la variable de salida contra la predicción hecha con el método de regresión.
5. Gráficas de residuos **(5 puntos)**.
    a. Presentar la gráfica de residuos.
    b. Analizar si hay tendencia a una distribución uniforme.
6. Aplicar regresión lineal ponderada únicamente para el dataset Water, usando como pesos los inversos cuadráticos de los residuos **(30 puntos)**.
    a. En esta sección, el parámetro `sample_weight` del método `fit` será de suma impotancia. Esto debido a que `sample_weight` requerirá los pesos para todas las instancias. Investigar el funcionamiento de lo anterior.

Reporte de la práctica

Emplear el formato de práctica dado por el profesor y seguir las instrucciones mostradas. El archivo que se subirá a *Google Classroom* deberá estar estrictamente en formato PDF y deberá ser nombrado como `report.pdf`.

Entregar un programa en lenguaje `R` para el primer punto de la práctica. Usar el lenguaje `Python` para desarrollar los puntos 2 a 6. **Únicamente serán aceptados estos lenguajes para la generación de los programas**. Además, es forzoso el uso de `Scikit-learn` para aplicar los métodos de regresión. Entregar los programas con extensión `.R` y `.py` debidamente comentados. Entregar un archivo `README.txt` donde se exponga cómo ejecutar los programas (indicar los parámetros en caso de necesitarlos) y un ejemplo para cómo ejecutar el programa y producir así los resultados reportados.

Entrega global

Tanto el reporte y los programas deberán ser empaquetados en un archivo `.ZIP` y nombrarlo: `practice1.zip`. **Cualquier falta a las instrucciones pedidas implicará la anulación de la práctica para todos los integrantes del equipo.**

Reto por 30 puntos extra (aplicables en las calificaciones de las tareas o prácticas)

Programar el método de gradiente descendente en `Python`, empleando el gradiente del error cuadrático medio para la regresión lineal clásica. Validar el método con el dataset mtcars y comparar la precisión con aquella obtenida con `LinearRegression` de `Scikit-learn`. Entregar el programa con extensión `.py`. En el reporte, añadir todo el procedimiento empleado y los resultados.