Semestral AD2021

TC3020 Aprendizaje Automático
Práctica 1 “Regresión lineal”

Entrega: Lunes 23 de agosto a más tardar a las 23:59 hrs

La práctica se deberá realizar en equipos de 2 personas.

Datasets

1. GENERO (genero.txt): cuenta con 10,000 instancias, donde se ocupará a la columna **Height** como atributo y a la columna **Weight** como variable de salida. El objetivo será predecir el peso de una persona dada su altura.
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 **(10 puntos)**.
    a. Generar el resumen con datos de estadística descriptiva del dataset.
    b. Generar los *boxplots* correspondientes para analizar el comportamiento de los datos.
    c. Generar gráfica de dispersión.
2. Regresión lineal **(30 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.
3. Evaluación **(10 puntos)**.

   a. Si el dataset es muy grande, medir la precisión siguiendo la metodología de validación simple. Reportar el MSE obtenido.
   b. Si el dataset es muy pequeño, medir el MSE siguiendo la metodología *n-fold cross validation*. Reportar los MSEs obtenidos de forma parcial y el MSE final.
   c. Recuerda que el MSE es 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.
4. Batch Gradient Descent **(45 puntos)**
   a. Programar el gradiente descendente para la regresión lineal y probarlo en los dos datasets dados.
   b. Además, comparar su desempeño con respecto a lo que entrega Scikit-Learn.
5. Reflexión **(5 puntos).** Agregar una reflexión por cada integrante del equipo sobre los conceptos aprendidos y los retos enfrentados durante el desarrollo de la práctica.

Reporte de la práctica

Emplear el formato de práctica dado por el profesor y seguir las instrucciones mostradas. El archivo que se subirá a *Canvas* 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 4. **Únicamente serán aceptados estos lenguajes para la generación de los programas**. Además, es forzoso el uso de `Scikit-learn` para aplicar el método 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.**