UNIVERSIDAD AUTONOMA DE OCCIDENTE
FACULTAD DE INGENIERÍA
APRENDIZAJE AUTOMATICO
PROFESOR: JOSE LUIS PANIAGUA JARAMILLO
TAREA 1: PROBLEMAS DE REGRESIÓN

Anotaciones

El objetivo de este proyecto es analizar los datos de un data set para un problema de regresión y desarrollar un algoritmo de regresión lineal.

Tener en cuenta lo siguiente:

- Escoger un data set de los posibles
- Primero realizar un análisis de las diferentes variables estimadoras para verificar cuales son la más idóneas para el proceso de regresión
- Dividir el data set en dos partes, una para la obtención del modelo y otra para la validación
- Escoger al menos tres variables estimadoras o todas las del dataset para generar una
- Realizar la estimación de los parámetros de la regresión por tres métodos:
    - Usando el cálculo mediante la Ecuación Normal.
    - Usando la librería scikit-learn.
        - Incluir regularización Lasso
        - Incluir regularización Ridge
    - Usando el cálculo del gradiente descendente
- Calcular el error de las predicciones con el conjunto de entrenamiento y validación
- Calcular el coeficiente de regresión, el R2_score, el MSE y RMSE del modelo obtenido con los datos de entrenamiento y validación.
- Entregar un notebook de Google Colab debidamente documentado.

Responder las siguientes preguntas:

- ¿Es posible que se presente el fenómeno de sobreajuste (overfitting) en la regresión lineal? ¿Se presento el fenómeno de sobreajuste en el modelo entrenado?
- ¿El data set seleccionado presenta valores atípicos (outliers)? ¿Cómo se identificaron los valores atípicos?
- ¿Es necesario eliminar los valores atípicos (outliers)? ¿Por qué sí o por qué no?
- ¿Qué es el escalado (Normalization)? ¿Cuándo es necesario?
- ¿En qué escenario es preferible utilizar el Descenso del Gradiente en lugar de la Regresión Ordinaria por Mínimos Cuadrados (Ecuación Normal) y por qué?

Data sets posibles

Se proponen como posibles data set el siguiente listado.

Airfoil Self-Noise Data Set
https://archive.ics.uci.edu/ml/datasets/Airfoil+Self-Noise

Auto MPG Data Set
https://archive.ics.uci.edu/ml/datasets/Auto+MPG

Energy efficiency Data Set (seleccionar una de las dos salidas disponibles y1= Heating Load, y2 = Cooling Load)
https://archive.ics.uci.edu/ml/datasets/Energy+efficiency

Concrete Compressive Strength Data Set
https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength

Yacht Hydrodynamics Data Set
https://archive.ics.uci.edu/ml/datasets/Yacht+Hydrodynamics

Stock portfolio performance Data Set (seleccionar una salida de las seis posibles para realizar la regresión) Y1=Annual Return,Y2=Excess Return Y3=Systematic Risk,Y4=Total Risk ,Y5=Abs. Win Rate ,Y6=Rel. Win Rate
https://archive.ics.uci.edu/ml/datasets/Stock+portfolio+performance

Daily Demand Forecasting Orders Data Set
https://archive.ics.uci.edu/ml/datasets/Daily+Demand+Forecasting+Orders

Real estate valuation data set Data Set
https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set

Electrical Grid Stability Simulated Data Data Set
https://archive.ics.uci.edu/ml/datasets/Electrical+Grid+Stability+Simulated+Data+

Behavior of the urban traffic of the city of Sao Paulo in Brazil Data Set
https://archive.ics.uci.edu/ml/datasets/Behavior+of+the+urban+traffic+of+the+city+of+Sao+Paulo+in+Brazil

QSAR fish toxicity Data Set
https://archive.ics.uci.edu/ml/datasets/QSAR+fish+toxicity