Practical session: Regression model

The goal is now to build a model able to linearly explain the target variable **Y** using the other available explanatory variables: X_1 , X_2 , …., X_m. The linear regression model is given by:

$$\hat{Y} = h_\theta(X) = \theta_0 + \sum_{i=1}^{m} X_i\theta_i = \boldsymbol{\theta^T X}$$

Where $h_\theta(X)$ is the hypothesis function, using the model parameters θ. In this case using:

$$h_\theta(X) = \theta_0 + \sum_{i=1}^{m} X_i\theta_i$$

Recall that training a model means setting its parameters so that the model best fits the training set. the most common performance measure of a regression model is the Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x^{(i)}) - y^{(i)})^2}$$

In practice, it is simpler to minimize the Mean Square Error (MSE) than the RMSE, and it leads to the same result (because the value that minimizes a function also minimizes its square root)

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(h(x^{(i)}) - y^{(i)})^2$$

Part 1: Simple Regression

Q1: Build a model with hypothesis $h_\theta(X) = \theta_0 + \theta_1 X$ to predict the land price from an input land area (see data_1.1.csv). Using Gradient Descent is to minimize a cost function (MSE) denoted by: $J(\theta_0, \theta_1) = J(\theta) = MSE(\theta)$

$$\frac{\partial J(\theta)}{\partial \theta_0} = \frac{2}{m}\sum_{i=1}^{m}\left(h\left(x^{(i)}\right) - y^{(i)}\right) = \frac{2}{m}\sum_{i=1}^{m}\left(\theta^T x^{(i)} - y^{(i)}\right)$$

$$\frac{\partial J(\theta)}{\partial \theta_1} = \frac{2}{m}\sum_{i=1}^{m}(h(x^{(i)}) - y^{(i)})x^{(i)}{}_1 = \frac{2}{m}\sum_{i=1}^{m}\left(\theta^T x^{(i)} - y^{(i)}\right)x^{(i)}{}_1$$

Update the parameter values: *Gradient Descent step*

$$\theta_j{}^{next\ step} = \theta_j - \eta\frac{\partial J(\theta)}{\partial \theta_j}$$

Calculate $\theta_{j=[0,1]}$, and loss function by testing learning rate η . Display graphic for loss function.

Q2: Now build a model with hypothesis: $h_\theta(X) = \theta_0 + \theta_1 X + \theta_2\sqrt{X}$ for dataset (data_1.1.csv)

Using the optimization **Adm, SGD**, to be calculate $\theta_{j=[0,1,2]}$, and loss function.

Reference: https://arxiv.org/abs/1609.04747

Part 2: Multi Regression: Boston Housing Dataset

Apply machine learning techniques using Linear Regression to be predict the values: Median value of owner-occupied homes **(MEDV).** Before build model you following many tasks below:

The Boston Housing Dataset (kaggle.com)

- **Load Data and Check Data**

- Load the dataset
- Use appropriate methods to check the structure and contents of the dataset.

- **Preprocess the Data**

- Check for duplicate entries and remove them.
- Identify and handle missing values (either by removing or imputing).
- Clean or encode categorical variables if necessary.

- **Check Correlation**

- Calculate the correlation matrix.
- Identify which features are strongly correlated with the target variable.

- **Visualization (Scatter Plot)**

- Create scatter plots to visualize the relationship between key features and the target variable.

- **Split Data (80% Train, 20% Test)**

- Define the features and target variables.
- Split the dataset into training and testing sets, using an 80/20 ratio.

Build model:

Q1: Model build pytorch.

Q2: Model Build in Scikit-learn. (Linear Regression, SVM) ********(Optional) ********