

# Practical Project 1

## Linear Regression

1. Consider the “Medical Cost Personal Dataset”, available at “Kaggle.com”. The data consists of age, sex, BMI(body mass index), children, smoker, region and charges features.

|   | age | sex    | bmi    | children | smoker | region    | charges     |
|---|-----|--------|--------|----------|--------|-----------|-------------|
| 0 | 19  | female | 27.900 | 0        | yes    | southwest | 16884.92400 |
| 1 | 18  | male   | 33.770 | 1        | no     | southeast | 1725.55230  |
| 2 | 28  | male   | 33.000 | 3        | no     | southeast | 4449.46200  |
| 3 | 33  | male   | 22.705 | 0        | no     | northwest | 21984.47061 |
| 4 | 32  | male   | 28.880 | 0        | no     | northwest | 3866.85520  |

The goal is to check whether the charges billed by health insurance can be predicted from the remaining features.

$$h_{\theta}(x_i) = \theta_0 + \theta_1 \text{age} + \theta_2 \text{sex} + \theta_3 \text{bmi} + \theta_4 \text{children} + \theta_5 \text{smoker} + \theta_6 \text{region}$$

2. Exploratory Data Analysis.
  - a) Convert all the data to numeric values.
  - b) Check if there are any missing/NULL values
  - c) Obtain the histogram of each feature, using:
    - a. Bar plots
    - b. Density estimates
  - d) Analyze the correlation between features:
    - a. Observing the scatter plots between pairs of features.
    - b. Observing the scatter plots between each feature and the dependent variable.
3. Obtain Implement the “linear\_regression.py” script, that obtains the best model, according to gradient descent.
4. Implement functions to obtain performance measures according to a “k-fold” validation scheme.



hugomcp@di.ubi.pt, 2023/24

5. Analyze the differences in performance between the models obtained for the different folds.
6. Adapt the “linear\_regression.py” script to fit a polynomial model (of order “p”) to your data. Repeat the analysis of step 5.