hugomcp@di.ubi.pt, 2025/26

Practical Project 1

Linear + Logistic Regression

1. **[Structured Data / Regression]** Consider the "Medical Cost Personal Dataset", available at "Kaggle.com", and also at the course web page. 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.

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

[Exploratory Data Analysis]

i) Convert all the data to numeric values.

ii) Check if there are any missing/NULL values

iii) Obtain the histogram of each feature, using:
   (a) Bar plots
   (b) Density estimates

iv) Analyze the correlation between features:

   (a) Observe the scatter plots between pairs of features.

   (b) Observe the scatter plots between each feature and the dependent variable.

[Implementation]

Implement a "linear_regression.py" script, that obtains the best model, according to the gradient descent algorithms.

[Empirical Validation]

Implement functions to obtain performance measures according to a "k-fold" validation scheme.

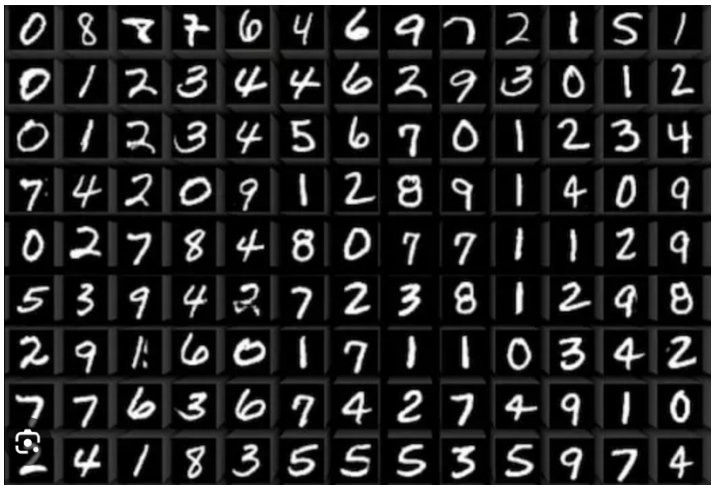
Analyze the differences in performance between the models obtained for the different folds.

[Implementation II]

Adapt the "linear_regression.py" script to fit a polynomial model (of order "p") to your data.

2. **[Unstructured Data / Classification]** Consider the "MNIST" dataset, available at the course web page. It contains 107,730 files in "png" format, each one representing one digit, from "0" to "9".

   Each image consists of 784-pixel values (a number from 0 - black to 255 - white), in the original 28 x 28 image.



[Implementation]

Implement a "logistic_regression.py" initial script, where each feature is a pixel value, top discriminate between the 10 classes in this dataset.

[Implementation II]

Implement one handcrafted features version of the classifier, where a set of features should be extracted and fed to a feed-forward neural network.

Examples:

- Compute histograms of pixel values (0–255 grayscale), as Different digits have distinct distributions of dark/light areas.
- Divide the image into zones (e.g., 4×4 or 7×7 grid), compute features like average intensity or count of nonzero pixels per zone.
- Count pixels row-wise (horizontal projection) and column-wise (vertical projection) to reveal shape outline and strokes distribution.
- Geometric Moments, HOGs, Skeleton-based, …

[Implementation III]

Implement a final version of the classifier, where the features should be extracted automatically by the neural network.