	MASTER OF SCIENCE IN ENGINEERING	Teachers: *J. Hennebert, A. Perez-Uribe*
		Assistants: *J. Donzallaz, J. Rial, O. Zayene*

HES-SO   Machine Learning

Practical work 04 – 10.10.2023
Linear Regression

Summary for the organisation :

— Submit the solutions of the practical work before the date specified in Moodle.
— **Rule 1**. Submit an archive (***.zip**!) with your Python notebooks (one per exercise), including datasets and all necessary files.
— **Rule 2**. The archive file name must contain the number of the practical work, followed by the family names of the team members by alphabetical order, for example `02_dupont_muller_smith.zip`. Put also the name of the team members in the body of the notebook (in first cell). Only one submission per team.
— **Rule 3**. We give a **fail** for submissions that do not compile (missing files are a common source of errors...). So, make sure that your whole notebooks give the expected solutions by clearing all cells and running them all before submitting.

Start from the Python notebook `pw4-linear-regression-stud.ipynb` and the data provided in `lausanne-appart.csv` available on Moodle. We will play with the appartment renting price in the city of Lausanne, see Figure 1.

```
1  dataset.head()
```

	living_area	nb_rooms	rent_price
0	69	3.0	1810
1	95	3.5	2945
2	21	1.5	685
3	20	1.0	720
4	33	1.5	830

...

FIGURE 1 – Lausanne appartment data

Exercice 1 Visualize the data

a) Plot a histogram to visualize the distribution of the renting price
b) Plot a histogram to visualize the distribution of the living area
c) Plot a scatter plot of renting price as a function of living area

Exercice 2 Normal equations for linear regression

In the case of a monovariable linear regression, we want to discover the parameters θ of a simple linear model defined with

$$h_\theta(\mathbf{x}) = \theta_0 + \theta_1 x \tag{1}$$

The "best" θ's are the ones that will minimise the squared difference between the gotten outputs and the target outputs defined with

$$J(\theta) = \frac{1}{2N} \sum_{n=1}^{N} (h_\theta(\mathbf{x}_n) - y_n)^2 \tag{2}$$

A closed form solution to this problem is given by the following normal equation :

$$\theta = (X^T X)^{-1} X^T \vec{y} \tag{3}$$

With X and $\vec{y}$ defined as

$$X = \begin{pmatrix} 1 & x_{1,1} & \cdots & & x_{1,D} \\ 1 & & \ddots & & \\ 1 & \vdots & & x_{n,d} & \vdots \\ 1 & & & \ddots & \\ 1 & x_{N,1} & \cdots & & x_{N,D} \end{pmatrix} \tag{4}$$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \tag{5}$$

a) Implement Equation 3 assuming that x is the living area and y is the renting price. Use `numpy` for the vector operations. Plot the computed line on top of the scatter plot of exercise 1.
b) Compute the overall cost value according to Equation 2.

Exercice 3 Gradient descent for linear regression

Implement the *full batch* gradient descent algorithm for the previous problem. As seen in the theory, the update rules are

$$\theta_0 \leftarrow \theta_0 - \alpha \frac{1}{N} \sum_{n=1}^{N} (h_\theta(\mathbf{x}_n) - y_n) \tag{6}$$

$$\theta_1 \leftarrow \theta_1 - \alpha \frac{1}{N} \sum_{n=1}^{N} (h_\theta(\mathbf{x}_n) - y_n) x_{n,1} \tag{7}$$

Remark You need to iterate several times over the training set. If you have problems of convergence, you need to use a smaller value of α. Values such as 0.000001 are common.

a) Plot the cost value (Equation 2) as a function of the iterations. What do you observe?
b) Imagine a stopping criterion, i.e. when do we stop iterating on the training set?
c) Plot the computed line $h_\theta(\mathbf{x})$ on top of the scatter plot of exercise 1.
d) Compute the final cost value according to Equation 2 and compare it to the one of exercise 2. What can you conclude?

Exercice 4 Stochastic gradient descent for linear regression

Implement the stochastic gradient descent algorithm for the previous problem. As seen in the theory, the update rules are

$$\theta_i \leftarrow \theta_i - \alpha (h_\theta(\mathbf{x}_n) - y_n) x_{n,i} \tag{8}$$

a) Plot the computed line $h_\theta(\mathbf{x})$ on top of the scatter plot of exercise 1.
b) How many samples do you need to visit for reaching the convergence?
c) What kind of stopping criterion could we use here?
d) Compute the final cost value according to Equation 2 and compare it to the one of exercise 2 and 3. What can you conclude?

Exercice 5 Review questions

a) The linear regression has a mathematical *closed form* solution. Then, in which conditions would we prefer a gradient descent algorithm to compute the regression model?
b) Outliers in a data set can be defined as values that are out of the "usual" range in comparison with other values. They typically come from noise or anomalies in the data capturing process. What is the impact of an outlier in the stochastic gradient descent

process ? What if we have many outliers ? *Hint* : look at the equation of the MSE, and to the equation of the update rule.

c) In the case of stochastic gradient descent, what is the danger of having a too large or too small α value ? Could you think of a better (more advanced) strategy as the one stated in Slide 36 ?

d) Let's assume we expect that the target variable y has a dependency to the square and to the cube of one of the feature x_d in our multi-variable training set $(x_1, \ldots, x_d, \ldots, x_D)$. How would you proceed ? Do we need to take precautions in terms of numerical stability ?

e) *Advanced.* Could we use a descent algorithm without computing the gradient ? If yes, give a pseudo code of the algorithm to find (θ_0, θ_1) using a linear regression $h_\theta(x) = \theta_0 + \theta_1 x$.

Exercice 6   Optional – Mini-batch gradient descent for linear regression

Implement the mini-batch gradient descent algorithm for the previous problem, adding a parameter B defining the size of the mini-batch. Check that when $B = N$, you fall back on the batch gradient descent solution, and when $B = 1$, you get the behaviour of stochastic gradient descent.

Exercice 7   Optional – multi-variable linear regression

a) Implement one of the gradient descent algorithm (ex. 3-5) for the multi-variable linear regression assuming x_1 being the living area and x_2 the square of the living area. Plot the computed curve (second order) on top of the scatter plot of exercise 1.

b) Implement one of the gradient descent algorithm (ex. 3-5) for the multi-variable linear regression assuming x_1 being the living area and x_2 the number of bedrooms.

Comments:

Ex1: Well done
Ex2: Well done
Ex3: Good. The expected cost is +/- equal to 144106.4542!
Ex4: Ok. The cost value for the stochastic gradient descent should be close to 197516.621.
Your conclusions are accordingly incorrect!
Ex5: Good. b) Typically outliers perturb the stochastic gradient descent process quite a bit.
Due to the random nature in data point selection an increasing number of outliers have a bigger impact and make the process unstable.

In regards to the MSE: since the mean in unstable towards outliers it will be impacted by the outliers quite heavily. e) We could do the following: 1. Initiate with random values for theta, 2. Create small perturbations on theta, 3. Is the MSE of the new theta > MSE old theta; if yes keep it else keep the old one, 4. Repeat step 2-4 until stop criteria is met.

Grade: Pass