Homework 1 — **Introduction to Deep-Learning**
Linear classification and regression
`https://allauzen.github.io/cours/IDL_ESPCI/`
E.S.P.C.I | P.S.L
06/01/2020

Part 1 | Submission

To know the deadline and how to submit, please refer to the website of the course. For this first homework, please submit a single pdf report with the detailed answers (it can be handwritten and scanned if you are lazy, but you can also use latex or a note book to write math or whatever). This assignment can be done in teams of 2 or 3 (if really you insist, you can do it alone) The file to submit should be like (put the correct number of lastname of course):

hw1-idl-lastname1-lastname2-lastname3.pdf.

Part 2 | Logistic regression: online learning

Assume $\mathcal{D}$, a set of training examples made of $\boldsymbol{x}_{(i)} \in \mathbb{R}^D$ and the associated answer $c_{(i)}$. We consider a binary classification task and $c_{(i)} \in \{0, 1\}$. The goal is to train a logistic regression model on this dataset $\mathcal{D} = (\boldsymbol{x}_{(i)}, c_{(i)})$. We define the following quantities of interest during the training process:

the prediction:	$y_{(i)}$	$= \frac{1}{1+e^{-(w_0+\mathbf{w}^t\boldsymbol{x}_{(i)})}} = \sigma(w_0 + \mathbf{w}^t\boldsymbol{x}_{(i)})$;
the pointwise loss:	$l(c_{(i)}, \boldsymbol{x}_{(i)}, \boldsymbol{\theta})$	$= -\big(c_{(i)}log(y_{(i)}) + (1 - c_{(i)})log(1 - y_{(i)})\big)$;
the full loss:	$\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$	$= \sum_{i=1}^{n} l(c_{(i)}, \boldsymbol{x}_{(i)}, \boldsymbol{\theta})$.

For a training example i, based on the values of the parameters $\boldsymbol{\theta}$, we infer the prediction $y_{(i)}$ to then compute the pointwise loss l which contributes to $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$. The goal is to minimize $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$. Before the training process, the parameters are randomly initialized at the values $\mathbf{w}_{(0)}$ for $\mathbf{w}$ and $w_{0(0)}$ for w_0. In an online scenario, at the i-th iteration, we pick a training example $(\boldsymbol{x}_{(i)}, c_{(i)})$ and carry out the following steps:

a. Make the prediction with the current set of parameters $\boldsymbol{\theta}_{(i)} = (w_{0(i)}, \mathbf{w}_{(i)})$

b. Compute the gradient of the pointwise loss $l(c_{(i)}, \boldsymbol{x}_{(i)}, \boldsymbol{\theta}_{(i)})$

c. Update the parameters to get $\boldsymbol{\theta}_{(i+1)}$ with a learning rate of η

See the course for more details. The following exercises provide insights on the training algorithm.

Exercise 1. Online gradient descent: impact of an update

1. Compute the gradient of the pointwise loss $w.r.t$ $\mathbf{w}$

$$\frac{\partial l(c_{(i)}, \boldsymbol{x}_{(i)}, \boldsymbol{\theta}_{(i)})}{\partial \mathbf{w}} =?$$

   It can be easier to distinguish the cases $c = 1$ and $c = 0$.

2. Write the update in both cases ($c = 1$ and $c = 0$). Provide an interpretation of these updates.

3. Show that the update indeed improves the pointwise loss:

$$l(c_{(i)}, \boldsymbol{x}_{(i)}, \boldsymbol{\theta}_{(i+1)}) \leq l(c_{(i)}, \boldsymbol{x}_{(i)}, \boldsymbol{\theta}_{(i)})$$

Exercise 2. Unroll the updates

The training process iteratively updates the parameters $\boldsymbol{\theta}_{(i)}$ (the value of the parameters $\boldsymbol{\theta}$ at "time" i of the training process) based on the value at time $i-1$ ($\boldsymbol{\theta}_{(i-1)}$). We can consider the training algorithm as a sequence by recursion. It starts with randomly initialized values $\mathbf{w}_{(0)}$ for $\mathbf{w}$ and $w_{0(0)}$ for w_0. Except where explicitly stated, the bias term w_0 is omitted in the remaining questions, without loss of generality.

1. We know pick the first training example $(\boldsymbol{x}_{(1)}, c_{(1)})$, write the updated parameters, $\mathbf{w}_{(1)}$ as a function of $\mathbf{w}_{(0)}$, $\boldsymbol{x}_{(1)}$, $c_{(1)}$ and the value predicted by the model $y_{(1)}$

2. Move on to the next step by computing $\mathbf{w}_{(2)}$ as a function of $\mathbf{w}_{(1)}$, $\boldsymbol{x}_{(2)}$, $c_{(2)}$ and $y_{(2)}$. Then replace $\mathbf{w}_{(1)}$ by the expression found in the previous question.

3. At the i-th iteration of the online training process, express the updated value $\mathbf{w}_{(i)}$ as a function of $\mathbf{w}_{(0)}$ and $(\boldsymbol{x}_{(k)}, c_{(k)}, y_{(k)})$ for $k \leq i$.

4. Provide an interpretation of this last formula.