Dans le cadre d'un sujet d'exercice à l'Ecole 42 Paris, je dois faire deux programmes :

Le premier est chargé de lire des données provenant d'un fichier csv, Ce fichier contenant des lignes avec deux informations (Kilométrage, et prix de voiture). Il doit ensuite utiliser une régression linéaire avec descente de gradient. (Ce projet assez simple nous donne déjà l'hypothèse à entraîner ainsi que le type de fonction que l'on doit trouver)

$$estimatePrice = \theta_0 + (\theta_1 * mileage)$$

il sauvegarde $\theta_0 \ et \ \theta_1$ pour être utiliser par le deuxième programme.
Celui-ci ne fait que récupérer ces infos pour que l'utilisateur rentre un kilométrage et que le programme lui donne l'estimation.
Mes programmes fonctionnent très bien et voici ma solution

**<u>Algo de Régression Linéaire</u>**

Hypothèse imposé :
les données sont organisées par : $\hat{y} = f(x) = \theta_0 + \theta_1 x$

pseudo code :

repeat until convergence* : {

$\nabla(J)$ compute
$\theta_0 := \theta_0 - \alpha\nabla(J)_0$
$\theta_1 := \theta_1 - \alpha\nabla(J)_1$
}

avec :
$\alpha$ learning rate (usualy between 0 and 1)

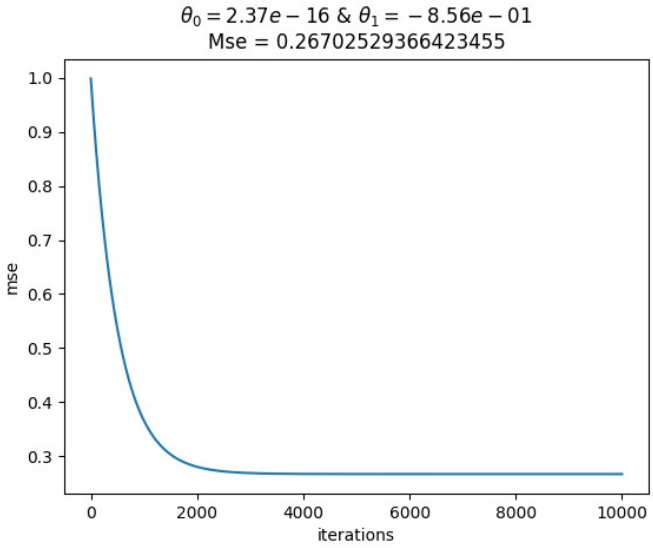
*Pour l'instant on boucle un certain nombre de fois.

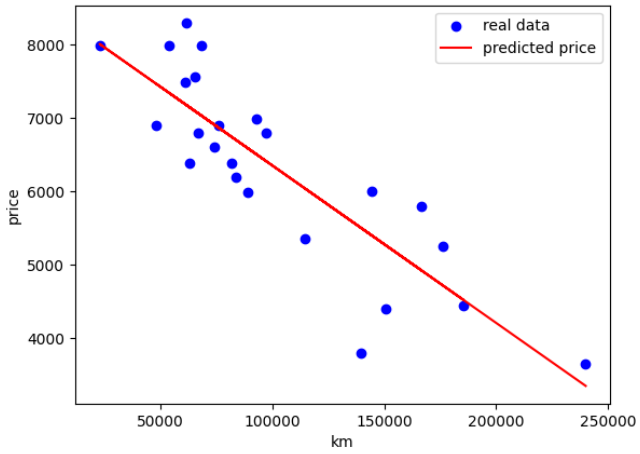
le gradient je le calcul comme ça :

$$\nabla(J) = \frac{1}{m} X'^T (X'\theta - y)$$

- $\nabla(J)$ vecteur de dimension 2 x 1.
- $X'$ matrice (m x 2)
- $X'^T$ matrice transposée de $X'$ (2 x m)
- $y$ vecteur de dimension m
- $\theta$ vecteur

en principe j'ai un *objet* **LinearRegression** qui a entre autre:
un *attribut* vecteur $\theta$
une *méthode* **Predict(x)** qui calcul les prédictions avec les $\theta$ mémorisés.
Une *méthode* **fit()** qui : récupère la prédiction avec les $\theta$, calcul le gradient $\nabla$ (un tableau de deux float) et diminue $\theta_0, \theta_1$ avec le facteur $\alpha$. il me renvoi un tableau des loss (MSE) pour afficher sa progression (On est sur une régression très simple et classique)




Ma question vient du fait que pour que je dérive correctement, j'ai dû Normaliser les donner avec la formule

$$x'^i = \frac{x^i - \mu}{\sigma} \; avec \; i \in [1, m]$$

x un vecteur de dimension m
x' la version normalisé du vecteur x
$\mu$ la moyenne des x
$\sigma$ la variation standard
(Je normalise x et y)

donc mon programme 1 sauvegarde $\theta_0 \; et \; \theta_1$ 'Normalisé' PLUS $\mu \; et \; \sigma$ des données x et y, pour que le deuxième programme fonctionne correctement.
Le programme 2 récupère le kilométrage( (en km), le normalise et calcul un prix 'normalisé' par

$$\hat{y} = f(x) = \theta_0 + \theta_1 x$$

puis 'dé-normalse' pour donner à l'utilisateur un prix en euros…

et voici enfin ma question :
Peut on, dans le premier programme, calculer $\theta_0 \; et \; \theta_1$ qui seraient déjà 'dénormalisé' et ainsi ne pas enregistrer les moyennes et variations standards des vecteurs x, y ?