Descente de gradient stochastique

KOUOMEGNE T. Bertrand , SOMAVO Gloria

Master ESA

27 Janvier 2023

Sommaire

1. Présentation de la descente de gradient stochatique
   - Principe
   - Choix du taux d'apprentissage
   - Algorithme
2. Ascente de gradient et maximum de vraissemblance
   - Fonction de vraisemblance
   - Cas pratique
   - Algorithme d'optimisation numerique
3. Présentation de l'algorithme d' ascente de gradient stochastique pour déterminer les paramêtres d'un modèle de regression logistique avec python
4. Avantages et Inconvénients

MASTER ESA
Université d'Orléans

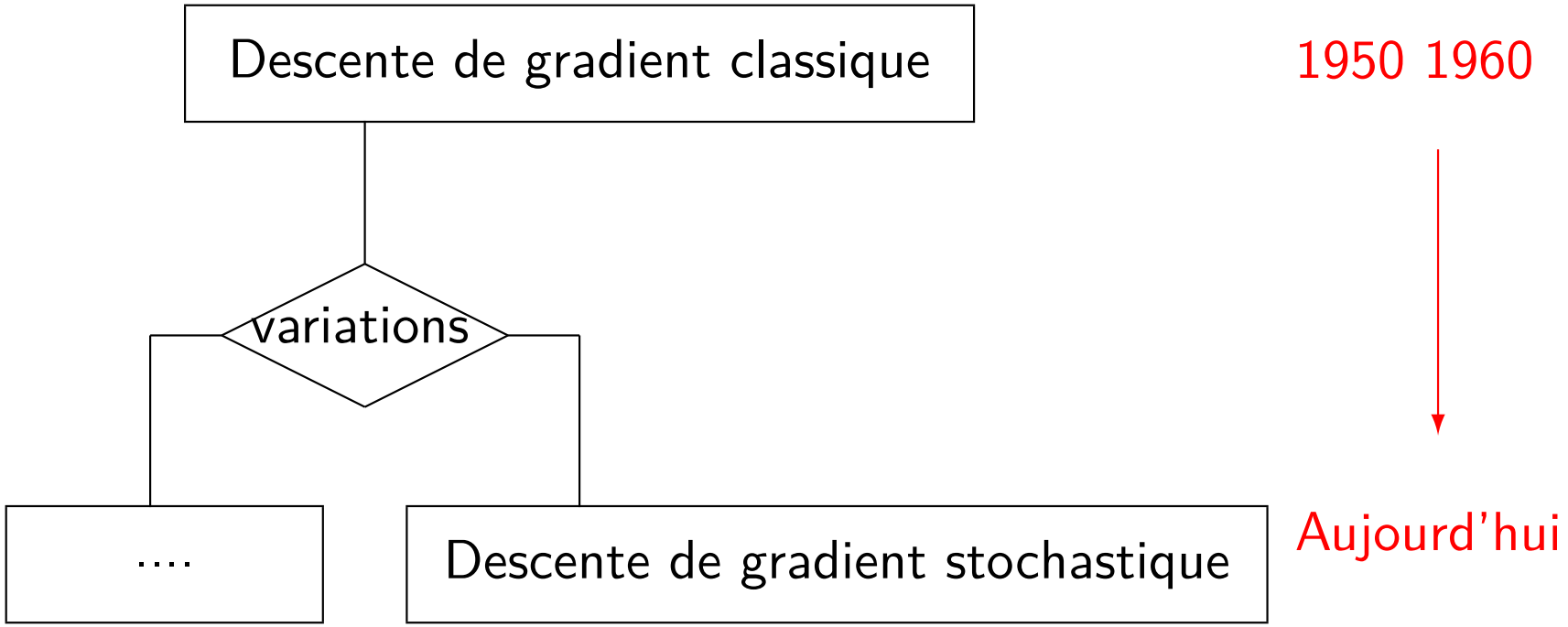
Sommaire

1. Présentation de la descente de gradient stochatique
   - Principe
   - Choix du taux d'apprentissage
   - Algorithme
2. Ascente de gradient et maximum de vraissemblance
3. Présentation de l'algorithme d' ascente de gradient stochastique pour déterminer les paramêtres d'un modèle de regression logistique avec python
4. Avantages et Inconvénients

Définition

Descente de gradient sctochastique : c'est un algorithme d'optimisation numérique utilisé pour trouver les paramètres optimaux d'un modèle d'économétrie ou de machine learning. Il s'agit d'une variante de la descente de gradiant classique.

Historique


Descente de gradient classique
variations
....
Descente de gradient stochastique
1950 1960
Aujourd'hui

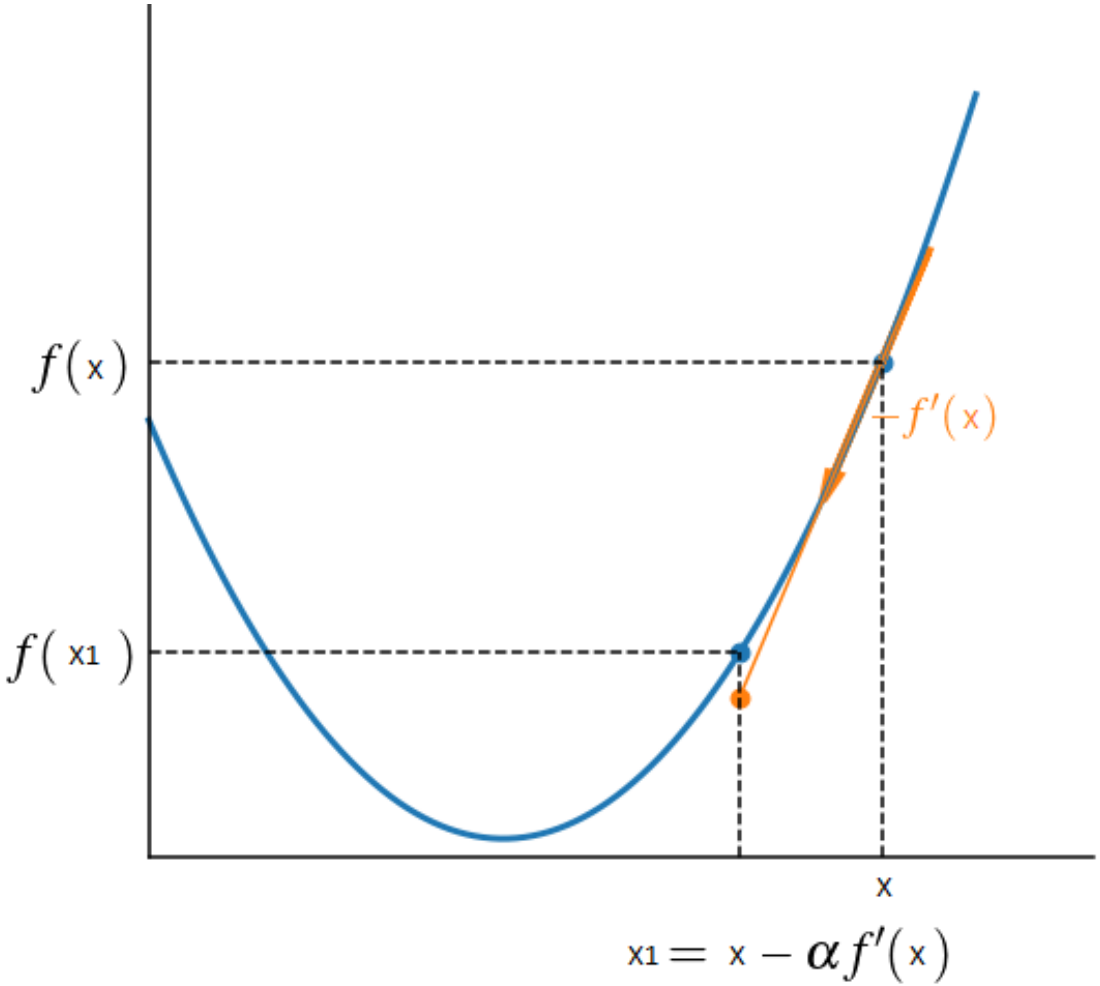
Principe

C'est un algorithme d'optimisation qui utilise la dérivée partielle d'une fonction objective pour déterminer la direction dans laquelle les paramètres du modèle doivent être mis à jour afin de réduire la valeur de la fonction objective.
En d'autres termes, il consiste à mettre à jour les paramètres en utilisant une petite quantité (appelée taux d'apprentissage) de la dérivée partielle (gradient) de la fonction objective par rapport aux paramètres.

Propriétés de la fonction objective

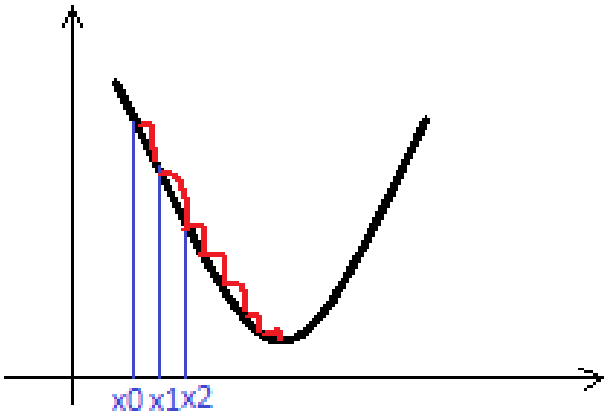
- La fonction objective doit être différentiable en tout point de son ensemble de définition.
- La fonction objective doit avoir un minimum global (convexe ou quasi-convexe).

Explications graphique

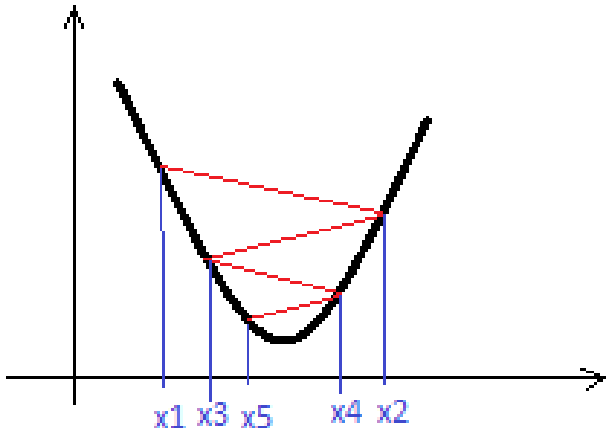


- x : Initialisation du paramètre
- $x_1 = x - \alpha f^{'}(x)$

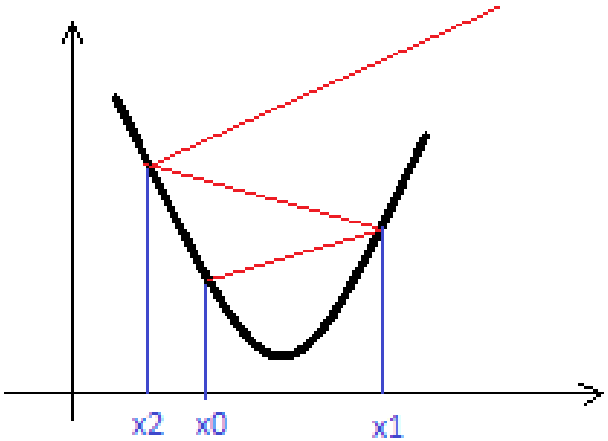
Choix du taux d'apprentissage



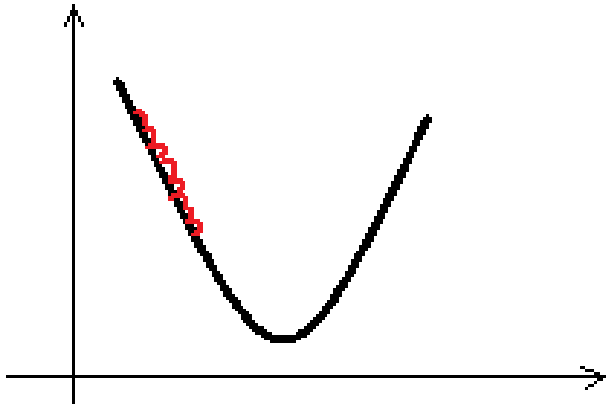
Bon choix



Trop Grand



Beaucoup trop grand



Trop petit

Descente de gradient classique

Initialisation

- Paramètre initial : θ_0
- Taux d'apprentissage : α
- Nombre d'itérations : p

Règle de passage

- Détermination du gradiant de l'échantillon au point θ_0 noté $G_n(\theta_0; x_n)$
- Actualiser la valeur de θ_0 par l'équation :
  $\theta_1 = \theta_0 - \alpha G_n(\theta_0; x_n)$
- Recommencer en prenant comme valeur initiale le point θ_1

.
.
.
.

Critère d'arrêt

L'algorithme s'arrête au bout du nombre d'itérations que vous désirez

Descente de gradient stochatique

Initialisation

- Paramètre initial : θ_0
- Taux d'apprentissage : α
- Nombre d'époque : *epoch*

Règle de passage

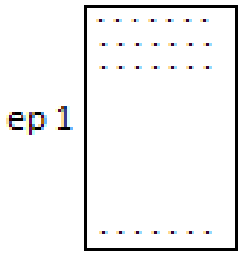
- Détermination du gradient individuel au point θ_0 noté $G_i(\theta_0; x_i)$, l'individu k est tiré au hazard
- Actualiser la valeur de θ_0 par l'équation :
  $\theta_1 = \theta_0 - \alpha G_i(\theta_0; x_i)$
- Recommencer en prenant comme valeur initiale le point θ_1
- Après avoir parcouru l'ensemble de l'échantillon, correspondant à une époque, recommencer

Critère d'arrêt

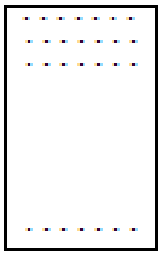
L'algorithme s'arrête au bout du nombre d'époques que vous désirez

MASTER ESA Université d'Orléans

Descente de gradiant classique

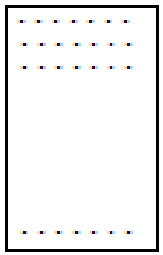


=> itération 1



=> itération 2

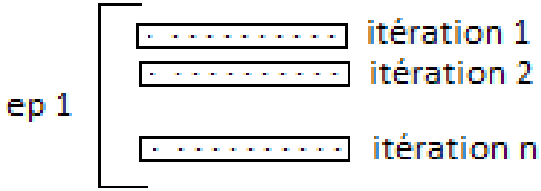
. . . . . .

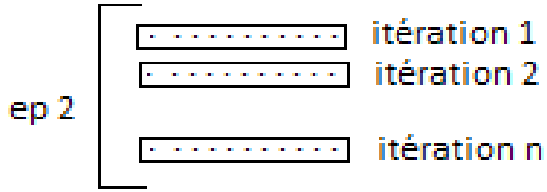


=> itération p

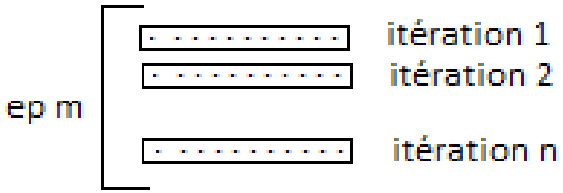
1 époque = 1 itération

Descente de gradiant stochastique





. . . . . .



1 époque = n itérations

Sommaire

1. Présentation de la descente de gradient stochatique
2. Ascente de gradient et maximum de vraissemblance
   - Fonction de vraisemblance
   - Cas pratique
   - Algorithme d'optimisation numerique
3. Présentation de l'algorithme d' ascente de gradient stochastique pour déterminer les paramêtres d'un modèle de regression logistique avec python
4. Avantages et Inconvénients

MASTER ESA Université d'Orléans

Fonction de vraisemblance

- C'est la probabilité jointe d'apparition des données en fonction des paramètres de la loi statistique choisie.
  $L_n(\theta; x_1, ..., x_n) = P_\theta(X_1 = x_1, ..., X_n = x_n)$ avec $X_i \sim \mathcal{D}(\theta)$
- Vraisemblance conditionnelle
  $L_n(\theta; y|x) = P_\theta(Y = y|x)$ avec $Y|X \sim \mathcal{D}(\theta)$
- La méthode du maximum de vraisemblance permet d'estimer les paramètres d'une loi statistique ;
- Elle permet aussi d'estimer les paramètres d'un modèle de regression.

Cas continue

Vraissemblance

- $L_i(\theta; x_i) = f_\theta(x_i)$
- $L_n(\theta; x) = \prod_{i=1}^{n} f_\theta(x_i)$

Estimateurs de MLE

- $\hat{\theta} = argmax L_n(\theta; x_1, x_2, ..., x_n)$

ou

- $\hat{\theta} = argmax ln(\theta; x_1, x_2, ..., x_n)$

/ Gradient

- $g_n(\theta; x) = \frac{\partial ln(L(\theta; x))}{\partial \theta}$

Hessienne

- $H_n(\theta; x) = \frac{\partial^2 ln(L(\theta; x))}{\partial \theta \, \partial \theta^T}$

Conditions

- FOC : $g_n(\hat{\theta}; x) = 0$
- SOC : La hessienne $H_n(\hat{\theta}; x)$ est définit négative

Cas discret

Vraissemblance

- $L_i(\theta; x_i) = P_\theta(X = x_i)$
- $L_n(\theta; x) = \prod_{i=1}^{n} P_\theta(x_i)$

Estimateurs de MLE

- $\hat{\theta} = argmax L_n(\theta; x_1, x_2, ..., x_n)$

ou

- $\hat{\theta} = argmax ln(\theta; x_1, x_2, ..., x_n)$

Gradient

- $g_n(\theta; x) = \frac{\partial ln(L(\theta; x))}{\partial \theta}$

Hessienne

- $H_n(\theta; x) = \frac{\partial^2 ln(L(\theta; x))}{\partial \theta \, \partial \theta^T}$

Conditions

- FOC : $g_n(\hat{\theta}; x) = 0$
- SOC : La hessienne $H_n(\hat{\theta}; x)$ est définit négative

MASTER ESA
Université d'Orléans

Cas pratique

Exemple1 : Considérons un échantillon de $X = (X_1, X_2, ..., X_n)$ de variables aléatoires indépendantes et identiquement distribué suivant une loi normale standard $N(\mu, \sigma^2)$ de densité : $f_x(x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ avec $\theta = (\mu, \sigma^2)$

- Vraisemblance
  $L(\theta; x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$
- Log- Vraisemblance
  $l_n(\theta; x) = \sum_{i=1}^{n} ln(L(\theta; x_i)) = -\frac{n}{2} ln(\sigma^2) - \frac{n}{2} ln(2\pi) - \frac{n}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$
- Gradient
  $\frac{\partial ln(L(\theta;x))}{\partial \theta} = \begin{pmatrix} \frac{\partial ln(L(\theta;x))}{\partial \mu} \\ \frac{\partial ln(L(\theta;x))}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} \frac{n}{\sigma^2} \sum_{i=1}^{n} (x_i - \hat{\mu}) \\ -\frac{n}{2\hat{\sigma}^2} + \frac{n}{2\hat{\sigma}^4} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \end{pmatrix}$
- Estimateur du Maximum de vraisemblance
  $\frac{\partial ln(L(\theta;x))}{\partial \theta} = 0 \rightarrow \hat{\theta} = \begin{pmatrix} \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \end{pmatrix}$

Cas pratique

Exemple2 : Considérons une variable dichotomique Y_i prenant deux valeurs 0 ou 1, telle que :

$y_i = 1$ avec P(Yi=1|xi) et $y_i = 0$ avec $1 - P(Yi = 1|xi)$

$P(Y_i = 1|x_i) = \wedge(x_i\prime\theta)) = \frac{e^{x_i\prime\theta}}{1+e^{x_i\prime\theta}}$

- Vraisemblance
  $L_i(\theta; y_i) = \wedge(x_i{}^T\theta)^{y_i}(1 - \wedge(x_i{}^T\theta))^{1-y_i}$
  $L_n(\theta; y) = \prod_{i=1}^{n} \wedge(x_i{}^T\theta)^{y_i}(1 - \wedge(x_i{}^T\theta))^{1-y_i}$

- Log- Vraisemblance
  $\ell_n(\theta; y|x) = ln(L\ (\theta; y|x)) = \sum_{i=1}^{n} y_i ln(\wedge(x_i{}^T\theta)) + \sum_{i=1}^{n} (1 - y_i) ln(\wedge(1 - x_i{}^T\theta))$

- Gradient
  $\frac{\partial ln(L\ (\theta;y|x))}{\partial\theta} = \sum_{i=1}^{n} (y_i - \wedge(\theta; x_i))x_i$

- Estimateur du Maximum de vraisemblance
  $\frac{\partial ln(L\ (\theta;y|x))}{\partial\theta} = 0 \rightarrow \sum_{i=1}^{n} (y_i - \frac{e^{x_i\prime\theta}}{1+e^{x_i\prime\theta}})x_i = 0$ (impossible ! ! !)

Lien avec l'ascente de gradient stochastique

- Algorithme d'ascent de gradient stochastique
  $\theta_j := \theta_j + \alpha g_i(\theta_j; y|x_{ij}) = \theta_j - \alpha(y_i - \wedge(x_i\prime\theta))x_{ij}$
- Algorithme de Gauss Newton
  $\theta_j = \theta_j - H_n^{-1}(\theta_k; x)\ g_n(\theta_k; x)$

Sommaire

1. Présentation de la descente de gradient stochatique
2. Ascente de gradient et maximum de vraissemblance
3. Présentation de l'algorithme d' ascente de gradient stochastique pour déterminer les paramêtres d'un modèle de regression logistique avec python
4. Avantages et Inconvénients

Sommaire

1. Présentation de la descente de gradient stochatique
2. Ascente de gradient et maximum de vraissemblance
3. Présentation de l'algorithme d' ascente de gradient stochastique pour déterminer les paramêtres d'un modèle de regression logistique avec python
4. Avantages et Inconvénients

MASTER ESA
Université d'Orléans

Avantages et Inconvénients

Avantages

- Convergence en moins d'une époque pour un echantillon large ;
- l'algorithme converge presque surement vers un maximum global.
- Facilement implémentable sur les logiciels .

Inconvénients

- Faible vitesse de convergence ;
- Un taux d'apprentissage trop élevé risque de conduire à la non convergence de l'algorithme ;
- Un taux d'apprentissage trop faible augmente la durée de convergence.

MASTER ESA Université d'Orléans

Bibliographie

- Maximum Likelihood, Logistic Regression, and Stochastic Gradient Training **Charles Elkan**
- Introduction au Machine Learning **Chloé-Agathe Azencott**
- Chapter 2 : Maximum Likelihood Estimation Advanced Econometrics - Master ESA **M. Christophe Hurlin**
- @deepmaths