Algorithme du Gradient Stochastique

Master MIGS

Yoann Offret

1 Principe de l'algorithme

L'objectif est de minimiser une fonctionnelle de la forme

$$F(\theta) = \mathbb{E}[f(\theta, W)],$$

où $\theta \in \mathcal{U}$ est un élément d'un ouvert de $\mathbb{R}^d$, et W est une variable aléatoire. On suppose qu'il existe un unique minimiseur θ^* et que f est différentiable. On suppose également que

$$\nabla F(\theta) = \mathbb{E}[\nabla f(\theta, W)].$$

Cependant, le calcul exact de ce gradient peut être coûteux, notamment si W prend ses valeurs dans un espace de grande dimension ou si le coût d'évaluation de ∇f est élevé. Cela rend difficile la mise en œuvre efficace de l'algorithme du gradient déterministe classique :

$$\theta_{n+1} = \theta_n - \varepsilon_{n+1} \nabla F(\theta_n),$$

où (ε_n) est une suite de pas tendant vers 0 et vérifiant certaines conditions de convergence.

Pour pallier cette difficulté, on utilise l'algorithme du *gradient stochastique*, qui remplace le gradient exact par une approximation stochastique obtenue à partir d'un échantillon de W :

$$\theta_{n+1} = \theta_n - \varepsilon_{n+1} \nabla f(\theta_n, W_{n+1}), \tag{1}$$

où (W_n) est une suite i.i.d. de même loi que W. La suite des pas (ε_n) doit satisfaire les conditions :

$$\sum_n \varepsilon_n = \infty \quad \text{et} \quad \sum_n \varepsilon_n^2 < \infty. \tag{2}$$

Un exemple typique pour (ε_n) est donné par :

$$\varepsilon_n = \frac{a}{1 + bn}, \tag{3}$$

où a et b sont des hyperparamètres à ajuster. De plus, dans de nombreux problèmes statistiques et d'apprentissage automatique, la fonctionnelle à minimiser est souvent de la forme :

$$F(\theta) = \frac{1}{N} \sum_{k=1}^{N} f(\theta, w_k), \tag{4}$$

où $(w_k)_{1 \leq k \leq N}$ est un échantillon de données expérimentales de grande taille. Cet exemple rentre parfaitement dans le cadre précédent, car on peut écrire :

$$F(\theta) = \mathbb{E}[f(\theta, W)],$$

où W suit la distribution uniforme sur l'ensemble des données $\{w_1, \ldots, w_N\}$. L'algorithme du gradient stochastique est utilisé dans des domaines variés, tels que la régression logistique, les réseaux de neurones et les systèmes de recommandation.

Exercice 1. *Implémentez l'algorithme du gradient stochastique pour minimiser la fonction*

$$F(\theta) = \frac{(\theta - 1)^2 + (\theta + 1)^2}{2}.$$

On prendra ici :

$$f(\theta, 1) = (\theta - 1)^2 \quad \textit{et} \quad f(\theta, -1) = (\theta + 1)^2.$$

Affichez l'évolution de (θ_n) au cours des itérations.

2 Un résultat de convergence

La convergence du gradient stochastique repose sur des hypothèses de convexité et de contrôle de la variance du gradient. Le théorème suivant illustre un résultat fondamental dans ce cadre.

Théorem 1. *Supposons que $\mathcal{U} = \mathbb{R}^d$ et qu'il existe $\alpha > 0$ tel que, pour tout $\theta \in \mathbb{R}^d$, on ait :*

$$\langle \nabla F(\theta), \theta - \theta^* \rangle \geq \alpha \|\theta - \theta^*\|^2. \tag{5}$$

Supposons de plus qu'il existe des constantes $A, B \geq 0$ telles que :

$$\mathbb{E}[\|\nabla f(\theta, W)\|^2] \leq A + B\|\theta - \theta^*\|^2. \tag{6}$$

Alors, les itérés (θ_n) générés par l'algorithme du gradient stochastique satisfont :

$$\lim_{n\to\infty} \|\theta_n - \theta^*\| = 0. \tag{7}$$

Remarque 2.1. *L'inégalité (5) signifie que F est fortement convexe, ce qui est essentiel pour la convergence vers θ^*. Le choix des hyperparamètres (a, b) dans (3) joue un rôle crucial dans la vitesse de convergence de l'algorithme.*

Démonstration. On a

$$\|\theta_{n+1} - \theta^*\|^2 = \|\theta_n - \theta^*\|^2 + \epsilon_{n+1}^2 \|\nabla f(\theta_n, W_{n+1})\|^2 - 2\epsilon_{n+1}\langle \nabla f(\theta_n, W_{n+1}), \theta_n - \theta^* \rangle.$$

Posons alors $D_{n+1} = \langle \nabla f(\theta_n, W_{n+1}) - \nabla F(\theta_n), \theta_n - \theta^* \rangle$. On peut alors réécrire l'égalité précédente de la manière suivante :

$$\|\theta_{n+1} - \theta^*\|^2 = \|\theta_n - \theta^*\|^2 + \epsilon_{n+1}^2 \|\nabla f(\theta_n, W_{n+1})\|^2 - 2\epsilon_{n+1}\langle \nabla F(\theta_n), \theta_n - \theta^* \rangle - 2\epsilon_{n+1} D_{n+1}.$$

Puis, par (5), on obtient

$$\|\theta_{n+1} - \theta^*\|^2 \leq (1 - 2\varepsilon_{n+1}\alpha)\|\theta_n - \theta^*\|^2 + \epsilon_{n+1}^2 \|\nabla f(\theta_n, W_{n+1})\|^2 - 2\epsilon_{n+1} D_{n+1}.$$

Par ailleurs, par construction, W_{n+1} est indépendante de $\theta_0, \cdots, \theta_n$, on peut alors montrer que D_{n+1} est centrée car

$$\mathbb{E}[D_{n+1}] = \mathbb{E}[\mathbb{E}[D_{n+1}|\theta_n]] = 0.$$

On en déduit que

$$\mathbb{E}[\|\theta_{n+1} - \theta^*\|^2] \leq (1 - 2\varepsilon_{n+1}\alpha)\mathbb{E}[\|\theta_n - \theta^*\|^2] + \epsilon_{n+1}^2 \mathbb{E}[\|\nabla f(\theta_n, W_{n+1})\|^2].$$

Enfin, en utilisant (6), on a

$$\mathbb{E}[\|\theta_{n+1} - \theta^*\|^2] \leq (1 - 2\varepsilon_{n+1}\alpha + B\varepsilon_{n+1}^2)\mathbb{E}[\|\theta_n - \theta^*\|^2] + A\epsilon_{n+1}^2.$$

Pour conclure, posons $a_{n+1} = (1 - 2\varepsilon_{n+1}\alpha + B\varepsilon_{n+1}^2)$ et $b_{n+1} = A\varepsilon_{n+1}^2$. Soit $N \geq 1$ tel pour tout $n \geq N$ on ait $0 \leq a_{n+1} \leq 1$. On montre alors par récurrence que pour $n \geq k \geq N$,

$$\mathbb{E}[\|\theta_{n+1} - \theta^*\|^2] \leq \left(\prod_{i=k+1}^{n+1} a_i \right) \mathbb{E}[\|\theta_k - \theta^*\|^2] + \sum_{i=k+1}^{n+1} b_i.$$

Or, en utilisant (2), on a

$$\lim_{n\to\infty} \prod_{i=k+1}^{n+1} a_i = 0 \quad \text{et} \quad \sum_{i=1}^{\infty} b_i < +\infty.$$

Quel que soit $\varepsilon > 0$, il existe $k \geq N$ tel que

$$\forall n \geq k, \quad \sum_{i=k+1}^{n+1} b_i \leq \sum_{i=k+1}^{+\infty} b_i \leq \varepsilon.$$

Il existe également $m \geq k$ tel que pour

$$\forall n \geq m, \quad \left(\prod_{i=k+1}^{n+1} a_i \right) \mathbb{E}[\|\theta_k - \theta^*\|^2] \leq \varepsilon.$$

Ce qui implique

$$\forall n \geq m, \quad \mathbb{E}[\|\theta_{n+1} - \theta^*\|^2] \leq 2\varepsilon,$$

d'où le résultat. □

3 Application à la régression logistique

La régression logistique est une méthode statistique utilisée pour modéliser la probabilité qu'une variable aléatoire Y prenne l'une des deux valeurs possibles, par exemple $+1$ et -1, en fonction de p variables explicatives réelles $X = (X_1, \ldots, X_p)$.

3.1 Modèle logistique

On suppose que, quel que soit le p-uplet $x = (x_1, \ldots, x_p)$, la probabilité conditionnelle de X donnée Y vérifie la relation log-linéaire suivante :

$$\ln \left(\frac{\mathbb{P}(X = x \mid Y = +1)}{\mathbb{P}(X = x \mid Y = -1)} \right) = a_0 + a_1 x_1 + \cdots + a_p x_p, \tag{8}$$

où $a_0, \ldots, a_p$ sont des constantes spécifiques au modèle. Cette hypothèse sous-tend que le logarithme du rapport des vraisemblances s'exprime comme une combinaison linéaire des variables explicatives $x_1, \ldots, x_p$. Ce modèle est particulièrement utile car il transforme un problème de classification binaire en un problème de régression linéaire sur une échelle logarithmique.

3.2 Transformation des probabilités

On peut montrer que cette relation log-linéaire est équivalente à une expression des probabilités conditionnelles de Y donnée X, sous la forme :

$$\ln \left(\frac{\mathbb{P}(Y = +1 \mid X = x)}{1 - \mathbb{P}(Y = +1 \mid X = x)} \right) = b_0 + b_1 x_1 + \cdots + b_p x_p, \tag{9}$$

où $b_0, \ldots, b_n$ sont d'autres constantes qui dépendent linéairement de $a_0, \ldots, a_n$. En pratique, ces constantes b_i représentent les coefficients du modèle estimés à partir des données.

Exercice 2. *Montrer l'équivalence précéndente.*

Cette équation montre que le logarithme du rapport des probabilités (*logit*) de la classe $+1$ par rapport à -1 est également une combinaison linéaire des variables explicatives. Enfin, on peut réécrire la probabilité $\mathbb{P}(Y = \varepsilon \mid X = x)$ pour $\varepsilon \in \{-1, +1\}$ sous une forme plus pratique pour les calculs :

$$\mathbb{P}(Y = \varepsilon \mid X = x) = \frac{1}{1 + e^{-\varepsilon(b_0 + b_1 x_1 + \cdots + b_p x_p)}}. \tag{10}$$

Cette expression est connue sous le nom de *fonction sigmoïde*, qui projette la sortie linéaire $b_0 + b_1 x_1 + \cdots + b_p x_p$ dans l'intervalle $(0, 1)$, correspondant à une probabilité.

Exercice 3. *Montrer l'expression précédente.*

Remarque 3.1. *Les paramètres $b_1, \ldots, b_n$ indiquent l'influence de chaque variable explicative sur la probabilité de la classe $+1$. Un coefficient positif b_i signifie que X_i augmente la probabilité de $Y = +1$, tandis qu'un coefficient négatif la réduit. Le modèle suppose que les effets des variables explicatives sur le logit sont additifs. La régression logistique est utilisée dans de nombreux domaines, notamment la médecine (prédiction de la présence d'une maladie), la finance (évaluation du risque de défaut), et l'apprentissage automatique (classification supervisée).*

3.3 Maximum de vraisemblance

Soient $y_1, \cdots, y_N$ un échantillon de la variable à expliquer et $x_1^{(k)}, \cdots, x_p^{(k)}$ les variables explicatives correspiondants à y_k pour $1 \leq k \leq N$. Notons $\Theta = (b_0, \cdots, b_p)$ les paramètres du modèle. La vraisemblance des donnée sécrit alors

$$L_\Theta = \prod_{k=1}^{N} \frac{1}{1 + e^{-y_k(b_0 + b_1 x_1^{(k)} + \cdots + b_p x_p^{(k)})}},$$

et la log-vraisemblance

$$\log L_\Theta = -\sum_{k=1}^{N} \log\left(1 + e^{y_k(b_0 + b_1 x_1^{(k)} + \cdots + x_p^{(k)})}\right).$$

Quitte à diviser par N, la fonction à miniser est de la même forme que (4), à savoir

$$F(\Theta) = \frac{1}{N} \sum_{k=1}^{N} \log\left(1 + e^{-y_k(b_0 + b_1 x_1^{(k)} + \cdots + b_p x_p^{(k)})}\right).$$

3.4 Une application au diagnostic précoce du cancer

Un hôpital souhaite développer un modèle pour prédire si un patient est atteint d'un cancer donné (par exemple, le cancer du poumon) en s'appuyant sur des données cliniques, génétiques, biologiques et d'imagerie. La variable cible Y est binaire : $Y = 1$ si le patient est atteint du cancer, $Y = -1$ sinon. Les données des patients proviennent de plusieurs sources, générant un ensemble de variables explicatives $X = (X_1, X_2, \ldots, X_p)$ (voir Table 1). Une étude réalisée sur une population représentative de 100 000 patients, composée d'environ 50% de patients atteints et 50% de patients non atteints, a produit les résultats contenus dans le fichier `design_matrix.csv`.

Exercice 4. *Déterminer les paramètres du modèle logistique, effectuer une classification en utilisant les valeurs obtenues, et comparer les résultats aux véritables données d'entraînement.*

Catégorie	**Variable**	**Description**
Cliniques	X_1	Âge du patient
	X_2	Sexe du patient
	X_3	Indice de masse corporelle (IMC)
	X_4	Antécédents familiaux de cancer
	X_5	Nombre d'années de tabagisme
	X_6	Niveau d'activité physique
	X_7	Antécédents médicaux de maladies chroniques
Biologiques	X_8	Niveau de glucose
	X_9	Concentration de protéines
	X_{10}	Taux de cholestérol total
	X_{11}	Niveaux hormonaux
	X_{12}	Inflammation systémique
	X_{13}	Analyse sanguine
Génétiques	X_{14} à X_{113}	Marqueur SNPs
Imagerie	X_{114}	Volume pulmonaire
	X_{115}	Taille de la plus grande lésion identifiée
	X_{116}	Densité moyenne des tissus pulmonaires
	X_{117}	Présence de calcifications
	X_{118}	Score d'emphysème
Mode de vie	X_{119}	Quantité de fruits et légumes consommés
	X_{120}	Niveau de stress autodéclaré
	X_{121}	Consommation d'alcool (en unités par semaine)

TABLE 1 – Liste des variables explicatives pour le diagnostic du cancer.