INSA - ModIA, 4^e année. **Années 2020-2021**

Régression linéaire et gradient stochastique

Ce TD-TP a pour objectif de plonger de manière un peu plus individualisée dans le cours d'optimisation stochastique. Il n'est pas noté, mais j'encourage très vivement les étudiants à rédiger consciencieusement leurs réponses et leurs idées. La rédaction force à mieux présenter les choses et surtout à mieux les cerner. Ca me permettra aussi plus facilement de corriger d'éventuelles incompréhensions.

1 Introduction

L'objectif de ce TP est d'illustrer la première partie du cours d'optimisation stochastique sur les erreurs en apprentissage et d'implémenter les premières versions du gradient stochastique. Il a aussi pour objectif de vous entraîner à effectuer des calculs avec des variables aléatoires pour les rendre plus accessibles et mieux comprendre les cours à venir.

Nous allons nous placer dans le cadre de travail le plus simple : la régression linéaire. Ce cadre présente plusieurs avantages :

— C'est probablement le plus simple d'un point de vue théorique et il permet d'appréhender de nombreux phénomènes avec des mathématiques relativement élémentaires.

— C'est probablement encore le plus utilisé dans les applications, et il me semble nécessaire de le comprendre profondément.

2 Le cadre

Soit X un vecteur aléatoire de $\mathbb{R}^d$, pour $d \in \mathbb{N}$ suivant une certaine distribution de probabilité inconnue P_X. Pour un certain vecteur $\theta \in \mathbb{R}^d$, on construit une variable aléatoire $Y \in \mathbb{R}$ définie par $Y = \langle \theta, X \rangle + B$ où $B \sim \mathcal{N}(0, \sigma^2)$ est une variable aléatoire gaussienne indépendante de X.

L'objectif de ce TP est d'apprendre le vecteur θ inconnu à partir de $n \in \mathbb{N}$ observations $(x_i, y_i)_{1 \leq i \leq n}$ tirées indépendamment suivant la loi P.

Pour ce faire, on peut simplement résoudre le problème de minimisation du risque empirique suivant :

$$\inf_{w \in \mathbb{R}^d} E_n(w) = \frac{1}{2n} \sum_{i=1}^{n} (\langle w, x_i \rangle - y_i)^2 \tag{1}$$

On notera w_n^* n'importe quel minimiseur (sous réserve d'existence) du problème ci-dessus.

3 Questions préliminaires

Ces questions introductives sont posées pour se forcer à revoir un peu le cours et les notions introduites.

1. Déterminer les espaces $\mathcal{X}$ et $\mathcal{Y}$ du cours. Quelle est la fonction perte l utilisée ici ? Quelle est la famille $\mathcal{H}$ de prédicteurs utilisés ?
2. Déterminer la loi conditionnelle $P(Y|X)$.
3. Quelle est la définition du risque moyen ici ?
4. Quel est le prédicteur optimal w^* ? Est-il unique ? Que vaut $E(w^*)$?
5. Si $X \sim \mathcal{N}(0, I_d)$ que vaut $E(w)$?
6. Déterminer l'erreur d'approximation $\mathcal{E}_{app}$ pour ce problème.
7. Est-ce que la fonction E_n est convexe ou non convexe ?
8. Calculer $\nabla E_n(w)$.
9. Est-ce que le problème (1) possède une solution ? Une solution unique ?

4 Erreur d'estimation

Dans cette partie, on se propose de borner l'erreur d'estimation

$$\mathcal{E}_{est} = E(w_n^*) - E(w^*),$$

pour se faire une idée de la vitesse de convergence du risque empirique.

Soient $(Z_i)_{1 \leq i \leq n}$ un ensemble de variables aléatoires i.i.d. de variance σ^2.

1. Que vaut $\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n Z_i\right)$?
2. Que peut-on en déduire sur la différence $E_n(w) - E(w)$?

Malheureusement, ce résultat est valable pour un vecteur $w \in \mathbb{R}^d$ quelconque, mais ne permet pas de contrôler $E(w_n^*) - E(w^*)$ directement. On va donc affiner ce résultat. On suppose que les paramètres optimaux w^* et w_n^* vivent dans une boule de rayon $R > 0$. On utilise la décomposition suivante.

$$\mathcal{E}_{est} = [E(w_n^*) - E_n(w_n^*)] + [E_n(w_n^*) - E_n(w^*)] + [E_n(w^*) - E(w^*)].$$

1. Que pouvez-vous dire de $E_n(w_n^*) - E_n(w^*)$?
2. En déduire que
$$\mathcal{E}_{est} \leq 2 \sup_{\|w\|_2 \leq R} |E(w) - E_n(w)|.$$
3. Etablir l'identité suivante :
$$E_n(w) - E(w) = \frac{1}{2}\left\langle w, \left(\frac{1}{n}\sum_{i=1}^n x_i x_i^T - \mathbb{E}(XX^T)\right) w \right\rangle - \left\langle w^T, \frac{1}{n}\sum_{i=1}^n y_i x_i - \mathbb{E}(YX) \right\rangle + \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^n y_i^2 - \mathbb{E}(Y^2)\right).$$
4. Borner la valeur absolue des erreurs ci-dessus en espérance. Vous pourrez par exemple utiliser des inégalités de Bernstein (scalaires, vectorielles et matricielles). Que conclure sur le taux de convergence de $\mathcal{E}_{est}$ vers 0 ?

5 Questions d'optimisation stochastique

Dans cette partie, on suppose que i_k est un indice aléatoire tiré uniformément dans l'ensemble $\{1, \dots, n\}$.

1. Est-ce que E_n est fortement convexe ?
2. Soit $f_i(w) = \frac{1}{2}\|\langle w, x_i\rangle - y_i\|_2^2$. Calculer $\nabla f_i(w)$.
3. Calculer $\mathbb{E}_{i_k}(\nabla f_{i_k}(w))$.
4. Calculer la constante de Lipschitz L de $\nabla E_n(w)$
5. Calculer $\mathbb{E}_{i_k}(\|\nabla f_{i_k}(w)\|_2^2)$ en fonction de $\|\nabla E_n(w)\|_2^2$.

6 Travail pratique (théorie)

Dans ce travail nous supposerons simplement que $X \sim \mathcal{N}(0, I_d)$. Nous supposons aussi que $\theta_i = 1$ pour tout i.

1. Définir une fonction

   `X, Y = generate_data(d,n,theta,sigma)`
2. Définir une fonction qui renvoit le risque moyen

   `E(w,theta,sigma).`
3. Définir une fonction qui renvoit le risque empirique

   `En(w,X,Y)`
4. Définir une fonction qui renvoit le gradient du risque empirique

   `grad_En(w,X,Y)`
5. Définir une fonction qui renvoit le gradient stochastique suivant la loi uniforme

   `grad_sto_En(w,X,Y,n_batch)`
6. Calculer la constante de Lipschitz de ∇E_n numériquement.
7. Si vous aviez un choix, quelle méthode d'optimisation vous semblerait la plus efficace pour minimiser E_n ?
8. Effectuer une descente de gradient à pas constant sur E_n et stocker les suites $E_n(w_k)$ et $E(w_k)$.
9. Etudier l'erreur d'approximation $\|w_n^* - \theta\|_2$ en fonction de n. Expliquez vos observations à partir du cours et des questions théoriques.
10. Effectuer un algorithme de gradient stochastique à pas constant sur E_n et stocker les suites $E_n(w_k)$ et $E(w_k)$.
11. Effectuer un algorithme de gradient stochastique à pas décroissant sur E_n et stocker les suites $E_n(w_k)$ et $E(w_k)$.
12. Comparer les taux de convergence pour le risque empirique et le risque moyen en fonction du nombre d'epoch pour chaque méthode.
13. Implémenter un algorithme de gradient stochastique online et comparer aux précédents.
14. Implémenter la méthode SAGA et comparer.
15. Relier les observations au cours. Quelle méthode devrait être privilégiée ?
16. Etudier l'influence de n, de σ.

7 Travail pratique (pratique)

Pour finir ce TP, nous proposons de tester les algorithmes de gradient stochastique dans un cadre moderne, avec la librairie PyTorch et des réseaux de Neurones. De très nombreux tutoriels sur ces logiciels existent et sont bien réalisés.

Je vous propose ici de suivre le tutoriel suivant : pytorch-mnist.

Une fois que vous avez réussi à reproduire les expériences suggérées, comparez différents algorithmes d'optimisation disponibles sous PyTorch (SGD simple, SGD Momentum, RMSProp, SAGA, ADAM).

8 L'inégalité de Bernstein (inégalité de concentration)

Théorème 1 (Inégalité de Bernstein)**.** *Soient $Z_1, \dots, Z_n$ un ensemble de n vecteurs aléatoires indépendants et identiquement distribués tels que $|Z_i| \leq c$ presque sûrement, $\mathbb{E}(Z_i) = \mu$ et $\mathrm{Var}(Z_i) = \sigma^2$. Alors*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right| \geq t\right) \leq 2\exp\left(\frac{nt^2}{2\sigma^2 + 2ct/3}\right). \tag{2}$$

Ce type d'inégalité est appelé *inégalité de concentration*. Il indique que la probabilité que la moyenne empirique dévie de la moyenne est très faible si on a un nombre d'observations suffisant.

1. Etablir la proposition suivante : l'inégalité suivante est valide avec une probabilité supérieure à $1 - \delta$.
2. On aimerait montrer que E_n est proche de E. Comment utiliser l'inégalité de Bernstein ou le corollaire précédent ?