THÉORIE DE L'APPRENTISSAGE ET BORNES PAC.

COURS D'APPRENTISSAGE, ECOLE NORMALE SUPÉRIEURE, AUTOMNE 2014

Rémi Lajugie
remi.lajugie@ens.fr

RÉSUMÉ. Dans ce TD on va chercher à illustrer certains points de théorie vus en cours. On commencera par donner une preuve de la célèbre ingalité de Hoeffding. Ensuite on s'intéressera à l'équivalence entre optimisation d'un objectif régularisé et d'un objectif contraint. Enfin on visualisera sur de vraies données deux des notions vues en cours (risque fréquentiste et borne PAC sur l'erreur de généralisation).

1. Exercice 1 : Inégalité de Hoeffding

Dans ce TP, on appelle transformée de Laplace d'une variable aléatoire comme étant la fonction suivante, définie sur $\mathbb{R}$, $f(s) = \mathbb{E}[\exp(sX)]$.

On va chercher à démontrer l'inégalité de Hoeffding.
On rappelle que l'inégalité d'Hoeffding s'énonce comme suit :
Inégalité : *Soit $X_1, \dots, X_n$ des variables alatoires réelles et indépendantes presque surement bornées par a_i et b_i. Si on appelle $S_n = \sum_i X_i$, alors la probabilité que la somme dévie de sa moyenne est majorée de la manière suivante :*

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq \exp(-2\frac{t^2}{\sum_i (b_i - a_i)^2})$$

La preuve se déroule en deux étapes, on va commencer par démontrer le lemme suivant (appelé le lemme d'Hoeffding) :
Lemme : *Soit X est une variable aléatoire centrée presque surement bornée, i.e, $X \in [a, b] p.s$ (en particulier $a \leq 0$ et $b \geq 0$), sa transformée de Laplace va être bornée par $\exp(\frac{s^2(b-a)^2}{8})$ quel que soit $s > 0$.*

1) Établir que, presque sûrement, on a $\exp(sX) \leq \frac{b-X}{b-a} \exp sa + \frac{X-a}{b-a} \exp sb$. En déduire une inégalité sur la transformée de Laplace.

2) Introduire $\theta = \frac{-a}{b-a}$. Considérer la fonction $\phi(u) = -\theta u + \log(1 - \theta + \theta \exp(u))$. Trouver un majorant quadratique à ϕ

3) En déduire le lemme d'Hoeffding.

On considère désormais des variables aléatoires $X_1, \dots, X_n$ presque surement bornées entre a_i et b_i. On appelle $S_n = \sum_i X_i$
4) En appliquant la méthode de Chernoff (inégalité de Markov combinée avec l'exponentielle), borner $\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t)$ par une quantité dépendant de la transformée de Laplace de $S_n - \mathbb{E}[S_n]$.

5) En optimisant sur la famille de bornes obtenues, retrouver le résultat de Hoeffding.

6) Comment retrouve-t-on l'inégalité de Chernoff à partir de la question précédente ?

2. Exercice 2 : Correspondance entre optimisation sous contrainte et optimisation régularisée.

En apprentissage statistique, un des problèmes majeurs est le surapprentissage comme on a pu le voir au cours des précédents TDs. Pour éviter ce problème, il faut contrôler la complexité de la fonction que l'on apprend. Dans le cas des plus proches voisins, un tel contrôle d'effectue naturellement en choisissant le nombre de plus proches voisins pris en compte.
Une autre grande classe d'algorithme cherche à minimiser une forme de risque empirique régularisé (c'est le cas de la régression ridge). Dans ce cas, le contrôle de la complexité s'effectue par le contrôle de la norme de la fonction (ou du paramètre w encodant la fonction). Intuitivement, une fonction de petite norme est une fonction moins complexe. Cette notion sera mieux formalisée lors du cours sur les méthodes à noyaux.
Pour fixer les idées, nous allons nous placer dans le cas de la régression de $\mathbb{R}^p$ dans $\mathbb{R}$ lorsque la fonction de décision est de la forme $f(x) = w^T x$ où $x \in \mathbb{R}^p$ et $w \in \mathbb{R}^p$ est le vecteur de paramètre. Étant donné un n-échantillon d'apprentissage $(x_1, y_1), \ldots, (x_n, y_n)$, une manière naturelle de contrôler la norme consiste à apprendre w comme solution d'un problème d'optimisation du type :

$$w_R \in \text{argmin}_{w \in \mathbb{R}^p} \sum_{i=1}^{n} \ell(w^T x, y_i) \quad (\mathcal{C}_R)$$
$$t.q,\ \|w\|_2^2 \leq R^2.$$

ℓ est une fonction de perte, admettant des valeurs finies, pas forcément convexe à valeurs dans $\mathbb{R}_+$.
Une autre manière de procéder consiste à régulariser l'objectif plutot qu'à le contraindre, i.e, d'optimiser :

$$w_\lambda \in \text{argmin}_{w \in \mathbb{R}^p} \sum_{i=1}^{n} \ell(w^T x, y_i) + \lambda \|w\|_2^2 \quad (\mathcal{R}_\lambda).$$

$\lambda \geq 0$ est appelé le paramètre de régularisation. On remarque que si λ tend vers l'infini, la solution du problème tend vers 0.
Le but de l'exercice est de montrer l'équivalence de ces deux formulations dans le cas convexe, c'est à dire que si w_λ est solution d'un problème régularisé $\mathcal{R}_\lambda$, on pourra trouver un problème contraint $\mathcal{C}_R$ dont w_λ soit solution et rciproquement. Dans le cas général, on se proposera de déterminer quels types de liens unissent ces deux classes de problèmes .
Il faut bien remarquer que l'on ne cherche pas à trouver le R ou le λ optimal.

7) Dans le cas général, commencer par montrer que, quelque soit le paramètre de régularisation $\lambda \geq 0$, on peut faire correspondre un problème $(\mathcal{C}_R)$ au problème $(\mathcal{R}_\lambda)$.

On considère la fonction g définie définie sur $\mathbb{R}_+$ définie par $g(R) = \min_{\|w\|^2 = R^2} \sum_{i=1}^n \ell(w^T x_i, y_i)$.
8) Montrer que $f(R) = \min_{\|w\|^2 \leq R^2} \sum_{i=1}^n \ell(w^T x_i, y_i)$ décroit. Comment se déduit le graphe de f à partir de celui de g ? Le graphe de cette fonction f est appelé la frontière de Pareto entre $\|w\|^2$ et ℓ. Sur ce graphe on voit apparaitre les solutions des problèmes $\mathcal{C}_R$.

9) Dans le cas où ℓ est convexe, montrer que l'ensemble des points au delà de la frontière de Pareto, i.e, $\mathcal{P} = \{(x, y), \exists w \in \mathbb{R}^p, x \leq \sum_i \ell(w^T x_i, y_i), y \leq \|w\|^2\}$ est convexe.

10) *Méthode de scalarisation :* Soit $\Lambda = (1, \lambda)$, on considère le problème d'optimisation scalaire à λ fixé :

$$\min_{w \in \mathbb{R}^p} \Lambda^T (\sum_i \ell(w^T x_i, y_i), \|w\|_2^2).$$

Vérifier qu'un w solution de ce problème est sur la frontière de Pareto. Comment peut on résoudre graphiquement ce problème en supposant la frontière de Pareto connue ?

11) Graphiquement, à quelle condition un point de la frontière de Pareto peut-il être obtenu par optimisation d'un problème scalaire comme celui de la question précédente ?

12) Déduire que dans le cas convexe toute solution à un problème régularisé est solution d'un certain problème $\mathcal{C}_R$ et vice-versa.

3. Exercice 3 : Visualisation des erreurs de généralisation

Le jeu de données recommandé pour cet exercice est le jeu de données BREAST, disponible en ligne sur le site de l'UCI (University of California at Irvine, prend). Néanmoins l'exercice est transposable sur n'importe quel jeu de données associé à une tâche de classification. Vous pouvez donc tester l'exercice sur votre jeu de données préféré.

13) Préparez des fonctions permettant de séparer aléatoirement le jeu de données entre données d'entrainement et de test tout en permettant de faire varier la taille de l'ensemble d'entrainement.

14) On considère les classes d'hypothèses $\mathcal{H}_k$ données par les règles du plus proche voisin pour différents k. Dans tout l'exercice on suppose que k est fixé (ce qui veut dire que l'on s'épargne le problème de la sélection de modèle). Quel est la classe d'hypothèses dont le code est le plus court ?

15) En reprenant le code du TP 2 (on en utilisant un code Matlab/Octave disponible en ligne) et le code écrit pour les questions précédentes, représentez graphiquement la distribution empirique de l'erreur de généralisation pour des tailles d'échantillon d'entrainement de 200, 60 et 15.
Faire également varier k le nombre de plus proches voisins. Que remarque-t-on ?

16) Comment peut se lire sur ce graphique le risque fréquentiste ? La borne PAC sur l'erreur de généralisation ?