COURS APPRENTISSAGE - ENS MATH/INFO
CONVEXIFICATION DU RISQUE

FRANCIS BACH - 5 DÉCEMBRE 2014

Pour l'apprentissage supervisé, nous avons vu deux types de méthodes, certaines basées sur une une formule explicite (k-plus proche voisins, Nadaraya-Watson), d'autres sur l'optimisation du risque empirique sur une certaine classe de fonction.

Cela n'est pas toujours possible, mais il existe des cas où on sait le faire, par exemple lorsque la perte est convexe, ou quand on fait en sorte de remplacer la perte (non convexe) par une perte convexe.

1. Régression

Cout quadratique : $\ell(Y, f(X)) = \frac{1}{2}(Y - f(X))^2$ choix standard

Valeur absolue : $\ell(Y, f(X)) = |Y - f(X)|$ robustesse aux valeurs aberrantes (non lisse). Relation avec la médiane.

Huber : $\ell(Y, f(X)) = \frac{1}{2}|Y - f(X)|^2$ si $|Y - f(X)| \leqslant \varepsilon$ and $\varepsilon|Y - f(X)| - \varepsilon^2/2$ si $|Y - f(X)| \geqslant \varepsilon$ robustesse aux valeurs aberrantes (mais lisse)

ε-insensitive : $\ell(Y, f(X)) = (|Y - f(X)| - \varepsilon)_+$

2. Classification binaire (supervisée)

2.1. Convexification du risque.

Données : $(X_i, Y_i)_{i=1,\dots,n} \in \mathcal{X} \times \{-1, 1\}$ indépendantes et identiquement distribuées.
But : trouver $f : \mathcal{X} \to \{-1, 1\}$ telle que :

$$\mathcal{R}(f) = E(\mathbf{1}_{f(X) \neq Y}) = P(f(X) \neq Y)$$

soit le plus petit possible.

Remarque 1. $\mathcal{R}(f)$ désigne le risque, ou encore la probabilité d'erreur de f.

Problème : $\{-1, 1\}$ n'est pas un espace vectoriel, donc $\{f : \mathcal{X} \to \{-1, 1\}\}$ non plus, ce qui peut poser une difficulté pour résoudre un problème de minimisation.

Nouveau but : trouver $f : \mathcal{X} \to \mathbb{R}$ et considérer alors la fonction : $x \to \text{sign}(f(x))$ comme prédicteur, où :

$$\text{sign}(a) = \begin{cases} 1 & \text{si } a > 0 \\ -1 & \text{si } a < 0 \\ 0 & \text{si } a = 0 \end{cases}$$

Risque : Le risque devient alors :

$$\mathcal{R}(f) = P(\text{sign}(f(X)) \neq Y) = E(\mathbf{1}_{\text{sign}(f(X))\neq Y}) = E(\mathbf{1}_{Yf(X)\leq 0}) = E\Phi_{0-1}(Yf(X))$$

où : Φ_{0-1} est la perte $0-1$.

Risque empirique :

$$\hat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^{n} \Phi_{0-1}(Y_i f(X_i))$$

On cherche à minimiser le risque empirique, mais on ne peut pas le faire directement car Φ_{0-1} est ni continue, ni convexe.

Question : Quel lien y a-t-il entre la perte $0-1$ et les pertes convexes ?

2.2. **Exemples.**

Exemple 1.

(1) Perte quadratique : ici $\Phi(u) = (u-1)^2$

$$\Phi(Yf(X)) = (Y - f(X))^2 = (f(X) - Y)^2$$

On retrouve les moindres carrés.

$$\mathcal{R}_\Phi(f) := E\Phi(Yf(X)) = E(f(X) - Y)^2$$

$\mathcal{R}_\Phi(f)$ est appelé le Φ-risque.

(2) Perte logistique : ici $\Phi(u) = \log(1 + e^{-u})$

$$\Phi(Yf(X)) = \log(1 + e^{-Yf(X)}) = -\log(\frac{1}{1 + e^{-Yf(X)}}) = -\log(\sigma(Yf(X))$$

où : $\sigma(v) = \frac{1}{1+e^{-v}}$ fonction sigmoïde.
Lien avec les probabilités : on considère le modèle défini par :

$$q(Y = 1|X = x) = \sigma(f(x)) \text{ et } q(Y = -1|X = x) = \sigma(-f(x))$$

Alors le risque est égal à $-$la log-vraisemblance conditionnelle : $E[-\log(q(Y|X))]$

(3) Perte hinge : ici $\Phi(u) = \max(1-u, 0)$
Soit $f(x) = w^T x + b$, on appelle marge la quantité (ceci correspond à une interprétation géométrique) :

$$\frac{1}{||w||}$$

On cherchera donc à minimiser $||w||$. Un classifieur adapté à ce problème est appelé classifieur "maximum margin".

Remarque 2. Pour ce dernier exemple, il y a deux formulations possibles :
-La formulation SVM (=support vector machine) séparable :
$\min \frac{1}{2}||w||^2$ tel que $y_i(w^T x_i + b) \geq 1$
-La formulation SVM non séparable :
$\min \frac{1}{2}||w||^2 + c\sum_{i=1}^n \xi_i$ tel que

$$y_i(w^T x_i + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$

Cette dernière peut se reformuler comme suit : $\min \frac{1}{2}||w||^2 + c\sum_{i=1}^n \max(0, 1 - y_i(w^T x_i + b))$
Ou encore : $\min \frac{1}{2}||w||^2 + c\sum_{i=1}^n \Phi(y_i f(x_i)))$

Exemple 2. Si on note :

$$\mathcal{L}(w,\xi,\alpha,\beta) = \frac{1}{2}||w||^2 + c\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\beta_i\xi_i - \sum_{i=1}^{n}\alpha_i(y_i(w^T x_i + b) - 1 + \xi_i)$$

Les variables w, ξ sont liées au problème primal ; les variables α, β sont liées au problème dual.
Pour minimiser cette expression en w, en dérivant, on trouve : $0 = w - \sum_{i=1}^{n}\alpha_i y_i x_i$
En minimisant en ξ on trouve : $0 = c - \alpha_i - \beta_i$
En minimisant en b, on trouve : $0 = \sum_{i=1}^{n}\alpha_i y_i$

Et le problème dual s'écrit : $\max_\alpha \frac{-1}{2}||\sum_{i=1}^{n}\alpha_i y_i x_i||^2$ telle que :

$$\begin{gathered} 0 = \sum_{i=1}^{n}\alpha_i y_i \\ 0 \leq \alpha_i \leq c \end{gathered}$$

Remarque 3. Les conditions de KKT s'écrivent :

$$\begin{aligned} (c-\alpha_i)\xi_i &= 0, & i &= 1,\dots,n \\ \alpha_i(y_i(w^T x_i + b) - 1 + \xi_i) &= 0, & i &= 1,\dots,n \end{aligned}$$

On remarque alors que si $y_i(w^T x_i + b) - 1 > 0$, cela signifie qu'on est à droite du hinge et, puisque $\xi_i \geq 0$ implique ici $\xi_i = 0$, on a : $\alpha_i = 0$
De même, si $y_i(w^T x_i + b) - 1 = 0$, cela signifie qu'on est sur le hinge, et on a : $0 \leq \alpha_i \leq c$
Enfin, si $y_i(w^T x_i + b) - 1 < 0$, cela signifie qu'on est à gauche du hinge, et on a : $\alpha_i = c$, car $\xi_i > 0$.
Notons que ceci est un cas particulier du théorème du représentant (cours sur les noyaux).

2.3. **Liens entre risque et Φ-risque.** Plaçons nous dans le cadre initial de ce paragraphe, avec une fonction $g : \mathcal{X} \to \{-1, 1\}$. Si on note $\eta(x) = P(Y = 1|X = x)$, alors le risque de g vérifie :

$$\mathcal{R}(g) = E\Phi_{0-1}(Yg(X)) = E[E(\mathbf{1}_{(g(X))\neq Y})|X = x)] \geq E\min(\eta(X), 1 - \eta(X))$$

Et le meilleur classifieur est $g^*(x) = \text{sign}(2\eta(x) - 1)$.

On suppose pour la suite Φ convexe, $\Phi'(0) < 0$ et Φ dérivable en 0. Dans ce cas on peut montrer que Φ est "calibrée", c'est-à-dire que les prédictions optimales pour Φ sont les mêmes que les prédictions optimales pour Φ_{0-1}. Et dans ce cas, on a le théorème suivant, où Ψ est une fonction croissant convexe telle que $\Psi(0) = 0$.

Theorem 2.1.

$$\Psi(\mathcal{R}(f) - \mathcal{R}(f^*)) \leq \mathcal{R}_\Phi(f) - \mathcal{R}_\Phi^*$$

Exemple 3. (1) Perte quadratique : $\Psi(\theta) = \theta^2$

(2) SVM : $\Psi(\theta) = |\theta|$

(3) Perte logistique : $\Phi(\theta) \geq \frac{\theta^2}{2}$

3. Extensions

3.1. **Multi-class.** prediction comme $\max_{i \in \{1,\dots,k\}} f_i(x)$

one-vs-rest + vote

one-vs-all + vote

joint cost : multinomial $-\log(e^{f_y(x)} / \sum_{i=1}^k e^{f_i(x)})$, two types of SVM, i.e., $\sum_{i=1}^k (1 + f_i(x) - f_y(x))_+$ or $\max_{i=1}^k (1 + f_i(x) - f_y(x))_+$

3.2. **Ranking.** Transformation en un probleme de classification binaire

3.3. **Parcimonie par pénalisation par norme ℓ_1.**

- Exemples en bioinformatique et marketing
- Interprétabilité, temps de calcul et meilleure performance de prédiction
- Pénalité ℓ_0
- Pénalité ℓ_1 : intuition géométrique en 2D, calcul explicit en 1D
- Inférence en haute dimension : $\sigma^2 d/n$ remplacé par $(\sigma^2 k \log d)/n$

Références

[1] Convex Optimization, Boyd and Vandenberghe, Cambridge University Press, 2004

[2] Convex Analysis and Non Linear Optimization, Borwein and Lewis, Springer, 2006.