COURS APPRENTISSAGE - ENS MATH/INFO
CONVEXIFICATION DU RISQUE - KMEANS/ACP

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 fonctions.

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

Coût quadratique : $\ell(Y, f(X)) = \frac{1}{2}(Y - f(X))^2$. C'est le choix standard et donne lieu aux "moindres carrés".

Valeur absolue : $\ell(Y, f(X)) = |Y - f(X)|$ robustesse aux valeurs aberrantes (non lisse). Relation avec la médiane qui minimise $\frac{1}{n}\sum_{i=1}^n |x_i - \mu|$ pour $x_i \in \mathbb{R}$.

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)

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 :

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

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.**

(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 :

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

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

(3) Perte hinge : ici $\Phi(u) = \max(1-u, 0)$
Si $f(x) = w^T x + b$, on retrouve la formulation de la 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 : $\min \frac{1}{2}||w||^2 + c\sum_{i=1}^n \Phi(y_i f(x_i)))$. On a bien un Φ-risque.

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 (Bartlett, Jordan, McAuliffe, 2005)**.**

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

Exemple 1. (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 : permet une reformulation en k problèmes de classification binaire.

coût joint : multinomial $-\log(e^{f_y(x)}/\sum_{i=1}^k e^{f_i(x)})$

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

4. K-means

Pour parler d'estimation de paramêtres "cachés", les Français et les Anglais utilisent des appelations qui peuvent porter à confusion. Dans un cadre supervisé, les Anglais parleront de *classification*, alors que les Français utiliseront *discrimination*. Dans un contexte non-supervisé, les Anglais parleront cette fois de *clustering*, alors que les Français utiliseront *classification*.

K-means est un algorithme de quantification vectorielle (clustering en anglais). K-means est un algorithme de minimisation alternée qui, étant donné un entier K, va chercher à séparer un ensemble de points en K clusters (Figure 1).

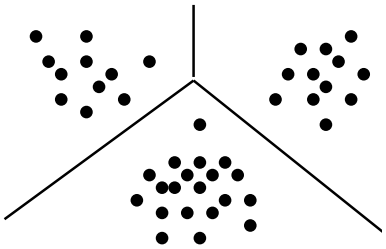

Figure 1. Clustering sur un ensemble de points 2D, 3 clusters.

4.1. **Notations, Mesure de distorsion.** On utilise les notations suivantes :
- Les $x_i \in \mathbb{R}^p$, $i \in \{1, ..., n\}$ sont les points à séparer.
- Les z_i^k sont des variables indicatrices associées aux x_i telles que $z_i^k = 1$ si x_i appartient au cluster k, $z_i^k = 0$ sinon. z est la matrice des z_i^k.
- μ est le vecteur des $\mu_k \in \mathbb{R}^p$, où μ_k est le centre du cluster k.

On définit de plus *la mesure de distorsion* $J(\mu, z)$ par :

$$J(\mu, z) = \sum_{i=1}^{n}\sum_{k=1}^{n} z_i^k \|x_i - \mu_k\|^2$$

4.2. **Algorithme.** Le but de l'algorithme est de minimiser $J(\mu, z)$, il se présente sous la forme d'un algorithme de minimisation alternée :

- Etape 0 : "choisir le vecteur μ"
- Etape 1 : on minimise J par rapport à z : $z_i^k = 1$ pour $k \in \arg\min \|x_i - \mu_k\|$, *i.e.* on associe à x_i le centre μ_k le plus proche.
- Etape 2 : on minimise J par rapport à μ : $\mu_k = \frac{\sum_i z_i^k x_i}{\sum_i z_i^k}$.
- Etape 3 : retour à l'étape 1 jusqu'à convergence.

Remarque 2. L'étape de minimisation par rapport à z revient à répartir les x_i selon les cellules de Voronoï dont les centres sont les μ_k.

Remarque 3. Dans l'étape de minimisation selon μ, μ_k est obtenu en annulant la k-ième coordonnée du gradient de J selon μ.

4.3. **Convergence et initialisation.** On peut montrer que cet algorithme converge en un nombre fini d'opérations. Cependant la convergence et locale, ce qui pose le problème de l'initialisation.

Une méthode classique consiste à lancer plusieurs fois l'algorithme en prenant les moyennes μ_k aléatoirement à chaque fois, puis on compare leur mesure de distorsion. On choisit la répartition qui possède la distorsion minimale.

Dans le pire des cas, cet algorithme peut se révéler arbitrairement mauvais, mais dans la pratique, il réalise de très bons résultats.

4.4. **Choix de K.** Le choix de K n'est pas universel, on remarque que si on augmente K, la distorsion diminue, et s'annule lorsque chaque point est centre de son cluster. Pour pallier à ce phénomène il possible de rajouter un terme en fonction de K dans l'expression de J, mais là encore son choix est arbitraire.

5. Analyse en composantes principales

5.1. **Formulation analytique.** Données $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathbb{R}^{n \times p}$ où chaque ligne représente un échantillon iid, et les colonnes les descripteurs. On va chercher la direction de l'espace telle qu'une fois qu'on a projeté les données dans cette direction, la variance est maximale.

On veut donc $\max_u \mathrm{Var}_n\left(u^T x_i\right)$ (variance empirique des données) avec $\|u\|_2 = 1$.

Si on suppose les données centrées :

$$\frac{1}{n}\sum_{i=1}^n x_i = 0 \quad \text{d'où} \quad \mathrm{Var}\left(u^T x_i\right) = \sum_{i=1}^n u^T x_i x_i^T u = u^T X^T X u$$

Le maximum est donc atteint pour u vecteur propre associé à la plus grande valeur propre de $X^T X$. On peut itérer ce procédé pour trouver d'autres directions principales :

5.1.1. *Déflation.* On projette x_i sur l'orthogonal de u :

$$x_i' \leftarrow x_i - \left(x_i^T u\right) u, \quad \text{d'où} \quad X' \leftarrow X - X v v^T = X\left(Id - v v^T\right)$$

.

On obtient la séquence des *composantes principales* qui sont en fait les vecteurs propres de X^TX par ordre décroissant des valeurs propres associées.

Si on fait la décomposition en valeurs singulières (SVD) de X, en écrivant $X = USV^T$, avec U et V orthogonales et $S = \text{Diag}(\sigma_1, \ldots, \sigma_n)$ avec $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_n$, on sait que $X^TX = VS^2V^T$, et les vecteurs de l'ACP cherchés sont donc les vecteurs trouvés par la SVD.

5.1.2. *Composantes principales.* Les composantes principales sont donc les k premières colonnes de V. Si on veut projeter les données sur cette base de composantes principales, on calcule $XV = USV^TV = US$.

5.1.3. *Variables principales.* Les variables principales sont les k premières colonnes de U, et une image de la projection des variables initiales sur les variables principales est donnée par $X^TU = VS$. (Notons que si les données sont centrées réduites, les vecteurs représentant les variables sont sur la sphère unité).

5.2. **Formulation de synthèse.** Un problème a priori différent conduit à une solution liée à l'ACP : étant donnée la matrice X, on voudrait trouver une matrice de rang faible $\tilde{X}$ qui approche bien les données. Soit en utilisant la norme de Froebenius :

$$\min_{\tilde{X}} \left\| X - \tilde{X} \right\|_F^2 \qquad \text{avec} \quad \text{rang} \qquad \left(\tilde{X} \right) \leqslant k$$

La solution $\tilde{X}$ est obtenue en projetant X sur ses k premières composantes principales.

Références

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

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