SY19
Séparateurs à vaste marge: cas non linéaire et régression

T. Denœux

1 Séparateurs à vaste marge non linéaires

1.1 Principe

Le principe de l'extension des SVM au cas non linéaire consiste à projeter les données dans un nouvel espace $\mathcal{H}$ grâce à une application $\Phi : \mathbb{R}^p \rightarrow \mathcal{H}$, et à rechercher l'hyperplan séparateur optimal dans ce nouvel espace. L'approche est donc similaire à celle des discriminateurs linéaires généralisés, à la différence que la fonction Φ restera implicite. Dans certains cas, l'espace $\mathcal{H}$ pourra être de dimension infinie.

Nous allons utiliser le fait que seuls apparaissent, dans l'expression du problème dual et dans la fonction de décision, des produits scalaires entre vecteurs. S'il existe une *fonction noyau* $\mathcal{K} : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}_+$ telle que

$$\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) =< \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) >,$$

$< \cdot, \cdot >$ désignant un produit scalaire dans $\mathcal{H}$, alors la fonction Φ n'a pas besoin d'être explicitée.

Par exemple, soit $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i' \mathbf{x}_j)^2$ dans le cas $p = 2$. On a

$$\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)' \Phi(\mathbf{x}_i)$$

avec

$$\Phi : \mathbf{x} \longrightarrow (x_1^2, \sqrt{2} x_1 x_2, x_2^2)'.$$

La fonction Φ est donc définie implicitement par la fonction $\mathcal{H}$.

Une fois définie une fonction noyau, il suffit de résoudre le problème d'optimisation suivant :

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j), \tag{1}$$

$$0 \leq \alpha_i \leq \gamma, \quad i = 1, \ldots, n \tag{2}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0. \tag{3}$$

Soit $\boldsymbol{\alpha}^*$ la solution de ce problème. La fonction discriminante $h^*(\mathbf{x})$ correspondante s'écrit :

$$h^*(\mathbf{x}) = \sum_{i \in S} \alpha_i^* y_i \mathcal{K}(\mathbf{x}, \mathbf{x}_i) + w_0^*,$$

ce qui induit la fonction de décision :

$$g^*(\mathbf{x}) = \text{sgn}(h^*(\mathbf{x})).$$

1.2 Choix du noyau

On peut montrer qu'une fonction noyau $\mathcal{K}$ correspond à un produit scalaire dans un espace $\mathcal{H}$ si et seulement si elle vérifie la condition suivante, appelée *condition de Mercer* :

$$\forall f : \mathbb{R}^p \to \mathbb{R} \text{ t.q. } \int f(\mathbf{x})^2 d\mathbf{x} < \infty, \quad \int \mathcal{K}(\mathbf{x},\mathbf{y}) f(\mathbf{x}) f(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0.$$

Si la condition de Mercer n'est pas vérifiée, le programme quadratique (1)-(3) peut ne pas avoir de solution. En pratique, la méthode peut quand même fonctionner même si la condition de Mercer n'est pas vérifiée.

Les fonctions noyau ci-dessous sont d'usage courant :

$$\begin{array}{lll} \mathcal{K}(\mathbf{x},\mathbf{y}) = & (\mathbf{x}'\mathbf{y}+1)^r, \quad r>0 & \text{(noyau polynomial)} \\ \mathcal{K}(\mathbf{x},\mathbf{y}) = & \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right) & \text{(noyau gaussien)} \\ \mathcal{K}(\mathbf{x},\mathbf{y}) = & \tanh(a\mathbf{x}'\mathbf{y}-b) & \text{(noyau PMC).} \end{array}$$

Les noyaux polynomial et gaussien vérifient la condition de Mercer. En revanche, le noyau PMC ne vérifie cette condition que pour certaines valeurs de a et de b. Pour ce noyau, la fonction discriminante est

$$h(\mathbf{x}) = \sum_{i \in S} \alpha_i^* y_i \tanh(a\mathbf{x}_i'\mathbf{x} - b) + w_0^*.$$

C'est donc la fonction de transfert d'un perceptron multi-couche à une couche cachée de $n_S = \text{card}(S)$ neurones. On remarque que l'architecture du réseau est ici déterminée automatiquement par apprentissage.

2 Extension au cas multi-classes

Dans le cas où $c > 2$, on peut construire une règle de décision en combinant plusieurs SVM binaires. On distingue deux stratégies :

- La stratégie *un-contre-tous* consiste à résoudre c problèmes binaires en opposant chaque classe aux autres. Soit h_k^* la fonction discriminante obtenue en opposant la classe ω_k aux autres. On pose

$$g(\mathbf{x}) = \arg\max_k h_k^*(\mathbf{x}).$$

- La stratégie *un-contre-un* consiste à résoudre $c(c-1)/2$ problèmes binaires en opposant la classe k à la classe ℓ, pour $k, \ell \in \{1, \ldots, c\}$, $k \neq \ell$. Chaque classifieur qui choisit une classe vote en faveur de cette classe. La classe qui recueille le plus de votes est finalement sélectionnée.

La méthode un-contre-un est souvent plus performante que la méthode un-contre-tous. Bien que le nombre de problèmes binaires à résoudre soit plus important dans cette méthode, le nombre d'exemples pris en compte dans chaque problème est plus faible. Le temps de calcul reste donc souvent acceptable, sauf si le nombre de classes est très grand.

3 SVM pour la régression

Considérons maintenant un problème de régression, pour lequel la variable à expliquer Y est numérique, à valeurs dans $\mathbb{R}$. Soit $\epsilon > 0$ une constante positive, et la fonction de coût suivante :

$$L(g(\mathbf{x}), y) = |g(\mathbf{x}) - y|_\epsilon = \begin{cases} 0 & \text{si } |g(\mathbf{x}) - y| \leq \epsilon, \\ |g(\mathbf{x}) - y| - \epsilon & \text{sinon.} \end{cases}$$

Pour cette fonction de coût, le risque empirique s'écrit :

$$\widehat{R}(g) = \frac{1}{n} \sum_{i=1}^{n} |g(\mathbf{x}_i) - y_i|_\epsilon.$$

On se propose de chercher la fonction linéaire $g(\mathbf{x}) = \mathbf{w}'\mathbf{x} + w_0$ minimisant le critère suivant :

$$\frac{1}{2}\|\mathbf{w}\|^2 + \gamma \sum_{i=1}^{n} |g(\mathbf{x}_i) - y_i|_\epsilon,$$

dans lequel le premier terme jour le rôle d'une terme de pénalisation, pénalisant les fonctions g moins « régulières », et γ est un hyperparamètre. SI $p = 1$, les examples d'apprentissage $(\mathbf{x}_i, y_i)$ tels que $|g(\mathbf{x}_i) - y_i|_\epsilon = 0$ sont situés à l'intérieur d'un « tube » de la largeur 2ϵ, centré autour de la droite $y = g(\mathbf{x})$. Par analogie, on parle de tube également lorsque $p > 1$. En introduisant des variables d'écart, ce problème peut se reformuler comme un problème d'optimisation quadratique sous contraintes :

$$\min_{\mathbf{w}, w_0, \boldsymbol{\xi}^-, \boldsymbol{\xi}^+} \frac{1}{2}\|\mathbf{w}\|^2 + \gamma \sum_{i=1}^{n} (\xi_i^- + \xi_i^+)$$

sous les contraintes :

$$\begin{aligned} \mathbf{w}'\mathbf{x}_i - y_i &\leq \epsilon + \xi_i^-, \quad i = 1, \ldots, n \\ y_i - \mathbf{w}'\mathbf{x}_i &\leq \epsilon + \xi_i^+, \quad i = 1, \ldots, n \\ \xi_i^-, \xi_i^+ &\geq 0, \quad i = 1, \ldots, n. \end{aligned}$$

En introduisant des multiplicateurs de Lagrange α_i^-, α_i^+, η_i^-, η_i^+, le lagrangien s'écrit :

$$\begin{aligned} L = \frac{1}{2}\|\mathbf{w}\|^2 + \gamma \sum_{i=1}^{n} (\xi_i^- + \xi_i^+) - \sum_{i=1}^{n} (\eta_i^- \xi_i^- + \eta_i^+ \xi_i^+) - \\ \sum_{i=1}^{n} \alpha_i^- (\epsilon + \xi_i^- + y_i - \mathbf{w}'\mathbf{x}_i - w_0) - \sum_{i=1}^{n} \alpha_i^+ (\epsilon + \xi_i^+ - y_i + \mathbf{w}'\mathbf{x}_i + w_0). \end{aligned}$$

Le problème dual consiste à maximiser le langrangien sous les contraintes

$$\alpha_i^- \geq 0, \alpha_i^+ \geq 0, \eta_i^- \geq 0, \eta_i^+ \geq 0, \quad i = 1, \ldots, n,$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0, \frac{\partial L}{\partial w_0} = 0, \frac{\partial L}{\partial \xi_i^-} = \frac{\partial L}{\partial \xi_i^+} = 0, i = 1, \ldots, n.$$

Ces contraintes permettent d'obtenir les relations suivantes :

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{n}(\alpha_i^+ - \alpha_i^-) = 0, \\
\frac{\partial L}{\partial w_0} &= \sum_{i=1}^{n}(\alpha_i^+ - \alpha_i^-) = 0, \\
\frac{\partial L}{\partial \xi_i^-} &= \gamma - \alpha_i^- - \eta_i^- = 0, \quad i = 1, \ldots, n \\
\frac{\partial L}{\partial \xi_i^+} &= \gamma - \alpha_i^+ - \eta_i^+ = 0, \quad i = 1, \ldots, n.
\end{aligned}$$

En utilisant ces relation, l'expression du lagragien se simplifie et l'on obtient :

$$L = \frac{1}{2}\sum_{i,j}(\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-)\mathbf{x}_i'\mathbf{x}_j - \epsilon\sum_{i=1}^{n}(\alpha_i^+ + \alpha_i^-) + \sum_{i=1}^{n} y_i(\alpha_i^+ - \alpha_i^-),$$

à maximiser sous les contraintes :

$$\sum_{i=1}^{n}(\alpha_i^+ - \alpha_i^-) = 0$$

$$\begin{aligned}
0 \leq \alpha_i^- \leq \gamma, \quad i = 1, \ldots, n, \\
0 \leq \alpha_i^+ \leq \gamma, \quad i = 1, \ldots, n.
\end{aligned}$$

Il s'agit à nouveau d'un problème d'optimisation quadratique qui peut être résolut à l'aide d'un solveur standard. Soient $\alpha_i^{-*}, \alpha_i^{+*}$, $i = 1, \ldots, n$ les solutions de ce problème. Les vecteurs d'appentissage $\mathbf{x}_i$ tels que $\alpha_i^{-*} > 0$ ou $\alpha_i^{+*} > 0$ sont appelés *vecteurs de support*. (Remarque : on n'a jamais simultanément $\alpha_i^{-*} > 0$ et $\alpha_i^{+*} > 0$). Soit S l'ensemble des vecteurs de support. On a

$$\mathbf{w}^* = \sum_{i=1}^{n}(\alpha_i^{+*} - \alpha_i^{-*})\mathbf{x}_i.$$

D'après les conditions de KT, les vecteurs de supports vérifient

$$\mathbf{w}'\mathbf{x}_i - y_i = \epsilon + \xi_i^{-*}$$

pour $\xi_i^{-*} \geq 0$ ou

$$y_i - \mathbf{w}'\mathbf{x}_i = \epsilon + \xi_i^{+*},$$

pour $\xi_i^{+*} \geq 0$. Ils sont donc situés en dehors du tube, ou à la frontière. Les vecteurs de support situés à la frontière sont ceux pour lesquels $\eta_i^{-*} > 0$, ce qui implique $\xi_i^{-*} = 0$, ou $\eta_i^{+*} > 0$, ce qui implique $\xi_i^{+*} = 0$. Dans le premier cas on a

$$0 < \alpha_i^{-*} < \gamma$$

et dans le second

$$0 < \alpha_i^{+*} < \gamma.$$

Pour calculer w_0^*, il suffit de considérer un tel vecteur de support et de poser, selon le cas

$$w_0^* = \epsilon + y_i - \mathbf{w}^{*\prime}\mathbf{x}_i$$

ou

$$w_0^* = y_i - \mathbf{w}^{*\prime}\mathbf{x}_i - \epsilon.$$

Enfin, on peut remarquer que, comme dans le cas de la discrimination, la solution obtenue ne dépend que des produits scalaire, ce qui rend possible une généralisation par l'introduction d'une fonction noyau.