SY19
Séparateurs à vaste marge: cas linéaire

T. Denœux

1 Notations

Nous nous intéresserons dans ce chapitre au problème de la discrimination en deux classes. Nous supposerons disposer d'un ensemble d'apprentissage $\mathcal{L} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ avec $y_i \in \{-1, 1\}$.

Un hyperplan H a pour équation $h(\mathbf{x}) = 0$ avec $h(\mathbf{x}) = w_0 + \mathbf{w}'\mathbf{x}$. La distance entre un point $\mathbf{x}$ de $\mathbb{R}^p$ et H est :

$$d(\mathbf{x}, H) = \frac{|h(\mathbf{x})|}{\|\mathbf{w}\|},$$

et $h(\mathbf{x})$ est positif d'un côté de l'hyperplan, et négatif de l'autre côté. A un hyperplan H peut être associée une fonction de décision :

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

2 Hyperplan séparateur optimal (cas séparable)

2.1 Formalisation du problème

On dit que l'ensemble d'apprentissage $\mathcal{L}$ est *linéairement séparable* s'il existe un hyperplan H d'équation $h(\mathbf{x}) = 0$ vérifiant $h(\mathbf{x}_i) > 0$ pour tout i tel que $y_i = +1$ et $h(\mathbf{x}_i) < 0$ pour tout i tel que $y_i = -1$, c'est-à-dire vérifiant

$$h(\mathbf{x}_i) y_i > 0, \quad \forall i \in \{1, \ldots, n\}.$$

Soit H un hyperplan séparant les deux classes. La distance entre un vecteur d'apprentissage $\mathbf{x}_i$ est H peut s'écrire :

$$d(\mathbf{x}_i, H) = \frac{h(\mathbf{x}_i) y_i}{\|\mathbf{w}\|}.$$

On appelle *marge* de H (relativement à $\mathcal{L}$) la plus petite distance entre un vecteur d'apprentissage et H :

$$\rho = \min_i d(\mathbf{x}_i, H).$$

Le problème considéré dans cette section consiste, étant donné un ensemble d'apprentissage linéairement séparable, à trouver un hyperplan H maximisant la marge. On peut montrer qu'à un tel hyperplan correspond généralement une

règle de décision ayant une probabilité d'erreur faible, comparée aux autres règles linéaires ayant un taux d'erreur apparent nul. En effet, supposons que chaque exemple de test soit à une distance au plus égale à r d'un exemple d'apprentissage de la même classe. Si $r \leq \rho$, le taux d'erreur de test est nul.

Ce problème peut être formalisé comme un problème de maximisation sous contraintes :

$$\max_{\mathbf{w},w_0} \rho$$

sous les contraintes

$$\frac{y_i(\mathbf{w}'\mathbf{x}_i + w_0)}{\|\mathbf{w}\|} \geq \rho, \quad i = 1, \ldots, n.$$

Ce problème peut être simplifié en remarquant que, si $(\mathbf{w}, w_0)$ est solution, il en est de même pour $(\beta\mathbf{w}, \beta w_0)$ avec $\beta \geq 0$ quelconque : on peut donc fixer arbitrairement la norme de $\mathbf{w}$, par exemple en posant $\|\mathbf{w}\| = 1/\rho$. Le problème d'optimisation peut alors s'exprimer de la manière suivante :

$$\min_{\mathbf{w},w_0} \frac{1}{2}\|\mathbf{w}\|^2,$$

sous les contraintes

$$y_i(\mathbf{w}'\mathbf{x}_i + w_0) \geq 1, \quad i = 1, \ldots, n.$$

Les contraintes définissent une marge $\rho = 1/\|\mathbf{w}\|$ de part et d'autre de l'hyperplan séparateur. Les vecteurs $\mathbf{x}_i$ tels que $y_i(\mathbf{w}'\mathbf{x}_i + w_0)$ sont situés au bord de cette marge et sont appelés *vecteurs de support*. Nous verrons que la règle de décision correspondant à l'hyperplan solution du problème peut s'exprimer à l'aide des seuls vecteurs de support.

Le problème précédent est un problème d'optimisation quadratique. La fonction à minimiser étant strictement concave et les contraintes linéaires définissant une région convexe, ce problème admet un seul minimum global.

2.2 Rappels d'optimisation

Soit le problème de minimisation suivant :

$$\min_{\mathbf{w}} f(\mathbf{w}) \tag{1}$$

sous les contraintes $c_i(\mathbf{w}) \geq 0$, $i = 1, \ldots, n$. Dans la méthode des multiplicateurs de Lagrange, on associe à chaque contrainte c_i un coefficient (*multiplicateur de Lagrange*) $\alpha_i \in \mathbb{R}$, et on définit le *lagrangien* par :

$$L(\mathbf{w}, \boldsymbol{\alpha}) = f(\mathbf{w}) - \sum_{i=1}^{n} \alpha_i c_i(\mathbf{w}),$$

avec $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ le vecteur des multiplicateurs de Lagrange.

On montre que, si la fonction f admet un minimum pour une valeur $\mathbf{w}^*$ dans la région de faisabilité, alors les conditions suivantes, appelées *conditions*

de Kuhn et Tucker (KT) sont vérifiées pour des nombres α_i^*, $i = 1, \ldots, n$:

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}}(\mathbf{w}^*, \alpha^*) &= 0 && (2) \\ c_i(\mathbf{w}^*) &\geq 0, \quad i = 1, \ldots, n && (3) \\ \alpha_i^* c_i(\mathbf{w}^*) &= 0 \quad i = 1, \ldots, n && (4) \\ \alpha_i^* &\geq 0 \quad i = 1, \ldots, n. && (5) \end{aligned}$$

De plus, on montre que le problème (1) est équivalent au problème suivant (dual de Wolfe) :

$$\max_{\boldsymbol{\alpha}} L(\mathbf{w}, \boldsymbol{\alpha}) \tag{6}$$

sous les contraintes

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}} &= 0 && (7) \\ \alpha_i &\geq 0 \quad i = 1, \ldots, n. && (8) \end{aligned}$$

Les contraintes actives $c_i(\mathbf{w}^*) = 0$ correspondent aux indices i tels que $\alpha_i^* > 0$.

2.3 Application

Ici le lagrangien s'écrit

$$L(\mathbf{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\mathbf{w}'\mathbf{x}_i + w_0) - 1], \tag{9}$$

et la condition (7) devient :

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 && (10) \\ \frac{\partial L}{\partial w_0} &= -\sum_{i=1}^{n} \alpha_i y_i = 0 && (11) \end{aligned}$$

En développant (9) et en utilisant (10)-(11), on obtient l'expression :

$$L(\mathbf{w}, w_0, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i' \mathbf{x}_j, \tag{12}$$

qui ne dépend plus de $\mathbf{w}$ ni de w_0. On cherche donc maintenant à maximiser (12) par rapport à $\boldsymbol{\alpha}$, sous les contraintes

$$\begin{aligned} \alpha_i &\geq 0, \quad i = 1, \ldots, n \\ \sum_{i=1}^{n} y_i \alpha_i &= 0. \end{aligned}$$

Ce problème peut être résolu par n'importe quel algorithme de programmation quadratique. Soit $\boldsymbol{\alpha}^*$ la solution. La condition de KT (4) devient ici

$$\alpha_i^* [y_i(\mathbf{w}^{*\prime}\mathbf{x}_i + w_0^*) - 1] = 0, \quad i = 1, \ldots, n.$$

Si $\alpha_i^* > 0$, la contrainte $y_i({\mathbf{w}^*}'\mathbf{x}_i + w_0^*) = 1$ est active, et $\mathbf{x}_i$ est donc un vecteur de support. Inversement, si $\mathbf{x}_i$ n'est pas un vecteur de support, $y_i({\mathbf{w}^*}'\mathbf{x}_i + w_0^*) > 1$ et donc $\alpha_i^* = 0$. Nous noterons

$$S = \{i \in \{1, \ldots, n\} | \alpha_i^* > 0\}$$

l'ensemble des indices des vecteurs de support.

Une fois résolu le problème dual, on en déduit facilement la solution $(\mathbf{w}^*, w_0^*)$ du problème initial. On a en effet

$$\mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i = \sum_{i \in S} \alpha_i^* y_i \mathbf{x}_i,$$

d'après (10). Par ailleurs, pour tout $i \in S$,

$$y_i({\mathbf{w}^*}'\mathbf{x}_i + w_0^*) = 1 \Rightarrow w_0^* = \frac{1}{y_i} - {\mathbf{w}^*}'\mathbf{x}_i = y_i - {\mathbf{w}^*}'\mathbf{x}_i.$$

La solution du problème initial s'exprime donc uniquement en fonction des vecteurs de support.

L'hyperplan séparateur optimal a donc pour équation $h^*(\mathbf{x}) = 0$ avec

$$\begin{aligned} h^*(\mathbf{x}) &= {\mathbf{w}^*}'\mathbf{x} + w_0^* && (13) \\ &= \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i'\mathbf{x} + w_0^* && (14) \\ &= \sum_{i \in S} \alpha_i^* y_i \mathbf{x}_i'\mathbf{x} + w_0^*. && (15) \end{aligned}$$

Par conséquent, la fonction de décision correspondante $g^*(\mathbf{x}) = \operatorname{sgn}(h^*(\mathbf{x}))$ s'exprime uniquement à l'aide des produits scalaires entre $\mathbf{x}$ et les vecteurs de support. Une telle règle de décision est appelée *séparateur à vaste marge* (SVM).

3 Extension au cas non linéairement séparable

Dans le cas où l'ensemble d'apprentissage n'est pas linéairement séparable, le problème d'optimisation précédent n'a pas de solution. En revanche, on peut introduire des *variables d'écart* ξ_i, $i = 1, \ldots, n$ et considérer le problème suivant, qui par construction admet toujours une solution :

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

sous les contraintes

$$\begin{aligned} y_i(\mathbf{w}'\mathbf{x}_i + w_0) &\geq 1 - \xi_i, \quad i = 1, \ldots, n \\ \xi_i &\geq 0, \quad i = 1, \ldots, n, \end{aligned}$$

γ étant un hyperparamètre fixé. Le Lagrangien s'écrit :

$$\begin{aligned} L(\mathbf{w}, w_0, \xi, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{w}\|^2 + \gamma \sum_{i=1}^{n} \xi_i \\ - \sum_{i=1}^{n} \alpha_i [y_i(\mathbf{w}'\mathbf{x}_i + w_0) - (1 - \xi_i)] - \sum_{i=1}^{n} \mu_i \xi_i, \quad (16) \end{aligned}$$

et la condition (7) devient :

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 \\
\frac{\partial L}{\partial w_0} &= -\sum_{i=1}^{n} \alpha_i y_i = 0 \\
\frac{\partial L}{\partial \xi_i} &= \gamma - \alpha_i - \mu_i = 0, \quad i = 1, \ldots, n,
\end{aligned}$$

ce qui est équivalent à

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \tag{17}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{18}$$

$$\mu_i = \gamma - \alpha_i, \quad i = 1, \ldots, n. \tag{19}$$

En développant le lagrangien (16) et en simplifiant grâce aux contraintes, on obtient :

$$L(\mathbf{w}, w_0, \xi, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i' \mathbf{x}_j, \tag{20}$$

c'est-à-dire exactement la même expression que (12), indépendante de $\boldsymbol{\mu}$. Le problème dual consiste donc à maximiser (20) par rapport à $\boldsymbol{\alpha}$, sous les contraintes

$$\begin{aligned}
\alpha_i &\geq 0, \quad i = 1, \ldots, n \\
\mu_i &\geq 0, \quad i = 1, \ldots, n \\
\mu_i &= \gamma - \alpha_i, \quad i = 1, \ldots, n, \\
\sum_{i=1}^{n} \alpha_i y_i &= 0,
\end{aligned}$$

qui se ramènent à

$$0 \leq \alpha_i \leq \gamma, \quad i = 1, \ldots, n$$

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

Soit $\boldsymbol{\alpha}^*$ la solution de ce problème. Les conditions de KT (4) s'écrivent :

$$\alpha_i^* [y_i(\mathbf{w}^{*\prime} \mathbf{x}_i + w_0^*) - (1 - \xi_i^*)] = 0, \quad i = 1, \ldots, n \tag{21}$$

$$\mu_i^* \xi_i^* = 0, \quad i = 1, \ldots, n \tag{22}$$

avec $\mu_i^* = \gamma - \alpha_i^*$. Les vecteurs de supports $\mathbf{x}_i$ correspondent aux α_i^* strictement positifs. Ils vérifient

$$y_i(\mathbf{w}^{*\prime} \mathbf{x}_i + w_0^*) = 1 - \xi_i^*.$$

Ce sont donc soit des exemples mal classés (si $1 - \xi_i^* < 0$), soit des vecteurs situés à l'intérieur de la marge, c'est-à-dire à une distance de l'hyperplan inférieure ou égale à $\rho = 1/\|\mathbf{w}\|$. Si de plus $\mu_i^* > 0$, c'est-à-dire $0 < \alpha_i < \gamma$, alors $\xi_i^* = 0$

(d'après (22)) et le vecteur de support $\mathbf{x}_i$ est situé sur le bord de la marge (sa distance à H est exactement égale à ρ).

Comme précédemment, notons $S = \{i \in \{1, \ldots, n\} | \alpha_i^* > 0\}$ l'ensemble des indices des vecteurs de support. Grâce à (17) on obtient

$$\mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i = \sum_{i \in S} \alpha_i^* y_i \mathbf{x}_i.$$

Pour obtenir w_0^*, il suffit de considérer un vecteur de support tel que $0 < \alpha_i^* < \gamma$ et d'utiliser (21). On obtient alors

$$w_0^* = \frac{1}{y_i} - \mathbf{w}^{*\prime} \mathbf{x}_i = y_i - \mathbf{w}^{*\prime} \mathbf{x}_i.$$

En pratique, il est recommandé pour des raisons numériques de prendre la moyenne de ces quantités pour tous les i tels que $0 < \alpha_i^* < \gamma$.

La fonction de décision a exactement la même expression que dans le cas linéairement séparable :

$$g^*(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i \in S} \alpha_i^* y_i \mathbf{x}_i' \mathbf{x} + w_0^* \right).$$

Elle s'exprime donc uniquement en fonction des produits scalaires de $\mathbf{x}$ avec les vecteurs de support.

La méthode dépend du seul hyperparamètre γ, qui peut être déterminé par validation croisée.