Data Mining 2 – Modélisation Statistique et Apprentissage
Master Statistique & Econométrie
UT1 Sciences Sociales, Toulouse
E-thème 5 **25 février 2018**
Enseignant: Gilles Cohen

Machine à vecteurs supports

1 Machines à Vecteurs Supports

Les Machines à vecteurs supports (MVS) représentent une nouvelle technique d'apprentissage introduite dans le cadre du principe de minimisation du risque structurel et de la théorie des bornes de la dimension de Vapnik Chervonenkis. Elles ont été ainsi développées à partir de considérations théoriques sur la complexité optimale de fonctions d'approximation à partir de données d'apprentissage. La méthode des vecteurs supports permet de résoudre des tâches de classification binaire et de régression. Dans le cas de la classification binaire, l'idée générale de la méthode est d'apprendre une séparatrice linéaire des exemples positifs et négatifs dans un espace où leurs descriptions ont été projetées.

1.1 Discriminant à marge maximale

Considérons le cas où l'on a un ensemble d'apprentissage $S_n = \{\mathrm{x}_i, y_i\}_{i=1}^n$ linéairement séparable, c'est à dire qu'il existe une fonction discriminante linéaire de la forme

$$\mathrm{x} \to \langle \mathbf{w}, \mathbf{x} \rangle + b, \quad \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R} \tag{1}$$

pour laquelle la fonction de décision correspondante $f = \text{signe}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$ à la propriété $R_n(f) = 0$. Il y a bien entendu une infinité de classifieurs linéaires qui sépare l'ensemble d'apprentissage en ne commettant aucune erreur. Un choix intuitif pour le meilleur classifieur est l'hyperplan qui se situe "exactement à mi-chemin entre les deux classes". Pour formaliser cette intuition, on va définir la notion de marge fonctionnelle et géométrique.

On définit la marge fonctionnelle de $(\mathbf{w}, b)$ par rapport à un exemple d'apprentissage $(\mathbf{x}_i, y_i)$ comme la quantité

$$\hat{\gamma}(\mathbf{x}_i, y_i) = y_i(\langle \mathbf{w}, \mathbf{x} \rangle + b)$$

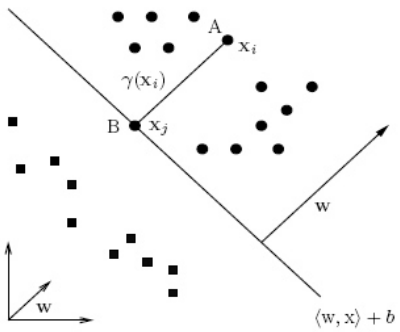

FIGURE 1 – Illustration de la marge fonctionnelle

La marge fonctionnelle est une mesure qui indique non seulement si la fonction f classifie correctement l'exemple $(\mathbf{x}_i, y_i)$ mais également avec quelle confiance. Il est donc essentiel qu'elle soit aussi grande que possible. Toutefois, si l'on utilise une fonction linéaire de la forme (1) avec les valeurs $(\alpha\mathbf{w}, \alpha\mathbf{b})$ à la place de $(\mathbf{w}, b)$ on augmente alors la marge fonctionnelle d'un facteur α sans pour cela changer f qui dépend du signe, mais pas de l'amplitude de $(\alpha\mathbf{w}, \alpha\mathbf{b})$. On va donc imposer une condition de normalisation telle que $\|\mathbf{w}\| = 1$; on peut alors remplacer $(\mathbf{w}, b)$ par $(\mathbf{w}/\|\mathbf{w}\|, b/\|b\|)$ et donc considérer la marge fonctionnelle de $(\mathbf{w}/\|\mathbf{w}\|, b/\|b\|)$.

On s'intéresse plus précisemment à la maximisation de $\hat{\gamma}$, la marge fonctionnelle de l'ensemble d'apprentissage S définie par :

$$\hat{\gamma} = \min_{i=1,\ldots,n} \hat{\gamma}(\mathbf{x}_i, y_i)$$

Considérons la figure (1), le point A de coordonnées $\mathbf{x}_i$ appartenant à la classe $y = +1$. Sa distance par rapport à la frontière de décision, $\gamma(\mathbf{x}_i)$, est donnée par le segment de droite AB. Puisque A représente $\mathbf{x}_i$, on peut alors déduire que le point B est donné par $\mathbf{x}_j : \mathbf{x}_i - \gamma(\mathbf{x}_i).(\mathbf{w}/\|\mathbf{w}\|)$. Or comme B appartient à la frontière de décision, et que tous les points éléments de cette frontière de décision satisfont l'équation $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$. On a

$$\left\langle \mathbf{w}, \underbrace{\mathbf{x}_i - \gamma(\mathbf{x}_i)\frac{\mathbf{w}}{\|\mathbf{w}\|}}_{B:\mathbf{x}_j} \right\rangle + b = 0 \quad \Rightarrow \quad \gamma(\mathbf{x}_i) = \frac{\langle \mathbf{w}, \mathbf{x}_i \rangle + b}{\|\mathbf{w}\|}$$

Plus généralement, on définit la *marge géométrique* de $(\mathbf{w}, b)$ par rapport à l'exemple d'apprentissage $(\mathbf{x}_i, y_i)$ comme étant égale à :

$$\gamma(\mathbf{x}_i, y_i) \quad = \quad \frac{y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)}{\|\mathbf{w}\|_2} \tag{2}$$

On définit finalement la marge géométrique relativement à S comme étant la plus petite des marges géométriques des exemples d'apprentissage

$$\gamma = \min_{i=1,\ldots,n} \gamma(\mathbf{x}_i, y_i)$$

On peut maintenant définir le problème du classifieur à marge maximale

$$(P) \begin{cases} \max\limits_{\gamma,\mathbf{w},b} & \gamma \\ \text{s.c.} & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq \gamma \\ & \|\mathbf{w}\| = 1 \end{cases}$$

qui correspond à maximiser la marge sous la contrainte que chacun des exemples ait une marge fonctionnelle au minimum égale à γ et sous la contrainte $\|\mathbf{w}\| = 1$ pour assurer l'égalité entre marge fonctionnelle et marge géométrique. Cette forme étant peu pratique pour l'optimisation on va utiliser une forme équivalente. Pour cela on va donc contraindre la marge fonctionnelle $\hat{\gamma}$ à être égale à 1. Plus formellement, on peut définir une paramétrisation canonique des hyperplans par rapport à un ensemble d'apprentissage $\{\mathbf{x}_i\}_{i=1}^n$ comme suit : $(\mathbf{w}, b)$ sont des paramètres canoniques si

$$\min_{\mathbf{x}_i \in X} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b| = 1 \tag{3}$$

Les hyperplans correspondant à l'équation (3) sont couramment appelés *hyperplans canoniques*. Avec la restriction imposée par Eq. (3) sur le numérateur de Eq. (2) maximiser la marge géométrique revient alors à minimiser la norme du vecteur $\mathbf{w}$. L'hyperplan qui maximise la marge géométrique minimum et satisfait la contrainte suivante :

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1, \quad i = 1, \ldots, n \tag{4}$$

est appelé *hyperplan séparateur optimal*. La marge du classifieur résultant vaut :

$$\gamma = \frac{1}{2}\left[\frac{1}{\|\mathbf{w}\|} - \frac{-1}{\|\mathbf{w}\|}\right] = \frac{1}{\|\mathbf{w}\|}$$

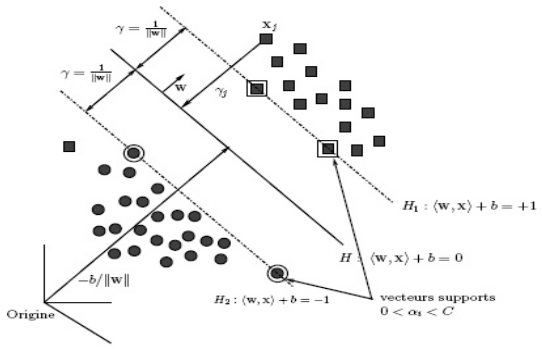

FIGURE 2 – Hyperplans séparateurs linéaires pour le cas séparable. L'hyperplan optimal est à une distance de $\frac{-b}{\|\mathbf{w}\|}$ de l'origine. Les vecteurs supports sont les points entourés qui se trouvent sur les deux hyperplans parallèles qui définissent la marge de séparation.

Maximiser la marge γ est donc équivalent à minimiser $\|\mathbf{w}\|$ ou $\|\mathbf{w}\|^2$. Le problème d'optimisation est maintenant un problème avec une fonction objectif quadratique convexe et uniquement des contraintes linéaires qui s'écrit :

$$(P_{\mathcal{P}}) \begin{cases} \min\limits_{\mathbf{w},b} & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.c.} & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 \quad i = 1, \ldots, n \end{cases} \tag{5}$$

1.2 Formulation duale du discriminant linéaire à marge maximale

Pour résoudre le problème on construit d'abord le Lagrangien :

$$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1] \tag{6}$$

où les $\alpha_i \geq 0$ sont les coefficients de Lagrange. Puis on minimise $\mathcal{L}(\mathbf{w}, \alpha, b)$ par rapport aux variables primales $\mathbf{w}$ et b afin d'obtenir le dual $\theta_{\mathcal{D}}$. Les conditions à l'optimum exigent que les dérivées par rapport à $\mathbf{w}$ et b soient

nulles :

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i \mathbf{x}_i y_i = 0 \tag{7}$$

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{8}$$

En resubstituant Eq. (7) et Eq. (8) dans le Lagrangien Eq. (6) on obtient :

$$\begin{aligned} \theta_{\mathcal{D}} = \min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \alpha) &= \frac{1}{2}\|\mathbf{w}\|^2 - \mathbf{w} \underbrace{\sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i}_{(7)=\mathbf{w}} - b \underbrace{\sum_{i=1}^{n} \alpha_i y_i}_{(8)=0} + \sum_{i=1}^{n} \alpha_i \\ &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \end{aligned}$$

Ce qui conduit à la formulation duale du problème :

$$(P_{\mathcal{D}}) \begin{cases} \max\limits_{\boldsymbol{\alpha}} & \theta_{\mathcal{D}}(\boldsymbol{\alpha}) = \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{2} \sum\limits_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ \text{s.c.} & \sum\limits_{i=1}^{n} \alpha_i y_i = 0 \\ & \alpha_i \geq 0 \quad i = 1, \ldots, n \end{cases} \tag{9}$$

On remarque que le problème dual est indépendant de la dimension de $\mathbf{x}$ mais est dépendant du nombre d'observations. C'est une propriété très intéressante, spécialement lorsque $\mathcal{X}$ est de grande dimension et que le nombre d'observations reste petit.

La résolution du problème d'optimisation sous contraintes (9) donne une solution α^* optimale dont on déduit le vecteur de poids :

$$\mathbf{w}^* = \sum_{i:\alpha_i^*>0} \alpha_i^* \mathbf{x}_i y_i \tag{10}$$

qui réalise l'hyperplan à marge maximale de marge géométrique

$$\begin{aligned}\gamma^* &= \frac{1}{\|\mathbf{w}^*\|} = \left(\sum_{i,j=1}^{n} y_i y_j \alpha_i^* \alpha_j^* \langle \mathbf{x}_i, \mathbf{x}_j \rangle\right)^{-1/2} \\ &= \left(\sum_{j=1}^{n} \alpha_j^* (1 - y_i b^*)\right)^{-1/2} \\ &= \left(\sum_{i=1}^{n} \alpha_i^*\right)^{-1/2}\end{aligned}$$

et le biais optimal b^* est donné en utilisant les contraintes primales par

$$b^* = -\frac{1}{2}\left[\min_{y_i=1}(\langle \mathbf{w}^*, \mathbf{x}_i \rangle) + \max_{y_i=-1}(\langle \mathbf{w}^*, \mathbf{x}_i \rangle)\right] \tag{11}$$

De la condition complémentaire de Karush-Kuhn-Tucker, il découle que les points $\mathbf{x}_i$ pour lesquels les α_i sont positifs doivent satisfaire

$$y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*) = 1$$

ce qui veut dire que seulement les points pour lesquels la marge fonctionnelle vaut 1 et donc se trouvent les plus proches de l'hyperplan séparateur optimal ont des multiplicateurs de Lagrange α_i^* non nuls. Donc dans l'expression du vecteur de poids $\mathbf{w}^*$ seulement ces points sont pris en compte. C'est la raison pour laquelle on les appelle les *vecteurs supports*. On notera l'ensemble des indices des vecteurs supports par sv. L'hyperplan optimal s'écrit alors dans la représentation duale en terme de ce sous-ensemble :

$$f(\mathbf{x}, \boldsymbol{\alpha}^*, b^*) = \sum_{i \in \mathrm{SV}} \underbrace{y_i \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle}_{\langle \mathbf{w}^*, \mathbf{x} \rangle} + b^*$$

1.3 Le cas non linéairement séparable (marge souple)

Pour une tâche de classification séparable, un tel hyperplan optimal existe mais très souvent, les points seront presque linéairement séparables dans le sens que seulement une petite fraction de l'ensemble de ces points rendra le problème non linéairement séparable. Pour traiter ces cas, on introduit des variables ressorts au niveau des contraintes permettant ainsi à ces vecteurs particuliers d'être mal classés. La marge de l'hyperplan est ainsi relâchée en

pénalisant les points d'apprentissage mal classés par le système. ξ est une valeur non négative définie par :

$$\xi_i := (1 - y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*))_+ \tag{12}$$

où $x_+ := \max\{0, x\}$. On a une mauvaise classification lorsque $\xi_i > 1$. Formellement l'hyperplan optimal est défini comme étant l'hyperplan qui maximise la marge et minimise le nombre des erreurs $\mathcal{E} = \mathbb{1}_{\{\xi>1\}}$. Comme c'est un problème non linéaire, on cherchera donc à minimiser une certaine fonctionnelle $\theta(\xi)$ qui représente une expression analytique d'une borne supérieure du nombre des erreurs. On a donc l'inégalité suivante : $\mathcal{E} = \mathbb{1}_{\{\xi>1\}} \leq \theta(\xi)$. La fonctionnelle est habituellement de la forme

$$\theta(\boldsymbol{\xi}) = \sum_{i=1}^{n} \xi_i^\sigma \tag{13}$$

où σ est une constante positive. Le problème d'optimisation devient

$$(P_\mathcal{P}) \begin{cases} \min\limits_{\mathbf{w},\xi,b} & \dfrac{1}{2}\sum\limits_{i=1}^{n} \|\mathbf{w}\|^2 + C\sum\limits_{i=1}^{n} \xi_i^\sigma \\ \text{s.c.} & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i,\ i = 1, \ldots, n \end{cases} \tag{14}$$

Pour $\sigma > 1$ le coût attribué à une erreur croît plus que linéairement en fonction de son écart par rapport à sa vraie valeur, mais pour des exemples qui ne sont pas des erreurs ($0 < \xi < 1$) la contribution à la somme des erreurs est minime. Pour $\sigma \to 0$ toutes les $\xi > 0$ y compris les exemples qui ne sont pas des erreurs contribuent avec le même coût qui est de 1. Cette configuration est une meilleure approximation du nombre des erreurs que $\sigma > 1$. Habituellement les valeurs $\sigma = 1$ et $\sigma = 2$ sont utilisées parce que le problème d'optimisation résultant est quadratique et pour $\sigma = 1$ le dual correspondant ne prend pas en compte $\boldsymbol{\xi}$ et simplifie ainsi le problème d'optimisation.

On va donc considérer le cas de la marge souple pour les valeurs $\sigma = 1$ et $\sigma = 2$.

Marge souple norme L_1. Si l'on prend $\sigma = 1$ le Lagrangien du problème d'optimisation (14) s'écrit :

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i[y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b) - 1 + \xi] - \sum_{i=1}^{n} \beta_i \xi_i \tag{15}$$

Les conditions pour obtenir un point selle exigent que les dérivées par rapport aux variables primales s'annulent :

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 \tag{16}$$

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{17}$$

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \tag{18}$$

Les conditions complémentaires de Karush-Kuhn-Tucker sont :

$$\alpha_i [y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \xi] = 0 \tag{19}$$

$$\beta_i \xi_i = 0 \tag{20}$$

En resubstituant les conditions à l'optimum Eq. (16)–(18) dans Eq. (15) on obtient la fonction objectif duale :

$$\begin{aligned}
\theta_{\mathcal{D}} = \min_{\mathbf{w}, b, \xi} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \xi_i - \mathbf{w} \underbrace{\sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i}_{(16=\mathbf{w})} - b \underbrace{\sum_{i=1}^{n} \alpha_i y_i}_{(17=0)} \\
&\quad + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i \xi_i - \sum_{i=1}^{n} \beta_i \xi_i \\
&= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} C\xi_i - \alpha_i \xi_i - \beta_i \xi_i \\
&= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \sum_{i=1}^{n} \xi_i \underbrace{(C - \alpha_i - \beta_i)}_{(18)=0} \\
&= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle
\end{aligned}$$

L'égalité $C - \alpha_i - \beta_i = 0$ (Eq. (18)) avec l'inégalité $\beta_i \geq 0$ (Eq. (20)) impose l'inégalité $\alpha_i \leq C$. D'où la formulation duale du problème d'optimisation :

$$(P_{\mathcal{D}}) \begin{cases} \max\limits_{\alpha} & = \theta_{\mathcal{D}}(\alpha) \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{2} \sum\limits_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\langle \mathbf{x}_i, \mathbf{x}_j \rangle), \\ \text{s.c.} & \sum\limits_{i=1}^{n} \alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n \end{cases} \tag{21}$$

La seule différence par rapport au problème (9) est la borne supérieure C pour les multiplicateurs de Lagrange. Les conditions de KKT impliquent que lorsque les variables ressorts sont non nulles, on ait l'égalité $\alpha_i = C$. Pour les points correspondants, la distance à l'hyperplan est plus petite que $1/\|\mathbf{w}\|$ comme l'indique la première contrainte dans Eq. (14). Le paramètre C contrôle directement la taille maximale des α_i et limite l'influence des exemples atypiques (outliers), qui auraient autrement des valeurs élevés pour les multiplicateurs de Lagrange. Pour α_i compris entre 0 et C les points correspondants se trouvent sur un des deux hyperplans.

Marge souple Norme L_2. Si l'on prend $\sigma = 2$ le Lagrangien du problème d'optimisation (14) s'écrit :

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\alpha_i[y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) - 1 + \xi_i] \quad (22)$$

En prenant le Lagrangien et en posant les conditions pour obtenir un point selle on obtient la formulation duale du problème d'optimisation :

$$(P_{\mathcal{D}}) \begin{cases} \max\limits_{\boldsymbol{\alpha}} & \theta_{\mathcal{D}}(\boldsymbol{\alpha}) = \sum\limits_{i=1}^{n}\alpha_i - \frac{1}{2}\sum\limits_{i,j=1}^{n}\alpha_i\alpha_j((\mathrm{x}_i.\mathrm{x}_j) + \frac{1}{C}\delta_{i,j})y_i y_j. \\ \text{s.c.} & \sum\limits_{i=1}^{n}\alpha_i y_i = 0 \\ & \alpha_i \geq 0 \qquad i = 1, \ldots, n \end{cases} \quad (23)$$

L'unique différence par rapport à la norme L_1 est l'addition de $1/C$ à la diagonale de la matrice $G = (\mathbf{x}_i.\mathbf{x}_j)$. Ceci a pour effet de rajouter $1/C$ aux valeurs propres de la matrice rendant ainsi le problème mieux conditionné.

1.4 MVS non linéaire

En pratique, la plupart des applications de classification réelles ne sont pas linéairement séparables et le classifieur optimal est non linéaire. Les MVS peuvent être étendues à la classification non linéaire en utilisant ce que l'on appelle couramment *l'astuce du noyau* (*kernel trick*). On a vu précédemment que dans l'algorithme des MVS les données impliquées dans les processus d'apprentissage et de classification apparaissaient uniquement sous forme de produits scalaires. Cette particularité est exploitée afin de construire des versions non linéaires des classifieurs à marge maximale.

Considérons que les données se trouvent dans un certain espace $\mathcal{X}$, et soit

$$\phi : \mathcal{X} \to \mathcal{F} \tag{24}$$

une fonction qui transforme l'espace initial $\mathcal{X}$ en un espace de Hilbert $\mathcal{F}$ de dimension pouvant être infinie. On peut maintenant essayer de classifier les nouvelles données dans le nouvel espace $\mathcal{F}$ (l'espace de caractéristiques) en utilisant la même approche de marge maximale que précédemment. La seule modification de l'algorithme implique le remplacement des produits scalaires $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ par $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$. Cependant, cette approche nécessite de définir la fonction ϕ, de vérifier qu'elle engendre un espace de Hilbert, et de projeter toutes les données dans le nouvel espace puis de construire le classifieur. C'est une approche difficile, voire même impossible dans le cas où $\mathcal{F}$ est de dimension très élevée ou infinie.

Au lieu d'explicitement projeter les vecteurs dans l'espace de caractéristiques et de calculer le produit scalaire dans cet espace, il est possible sous certaines conditions d'utiliser une fonction $k(\mathbf{x}, \mathbf{z})$ dont la valeur donne directement le produit scalaire entre deux vecteurs $\phi(\mathbf{x})$ et $\phi(\mathbf{z})$. En d'autres termes, il doit exister une transformation ϕ et un espace $\mathcal{F}$ muni d'un produit scalaire, avec $\langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle = k(\mathbf{x}, \mathbf{z})$, correspondant à la *fonction noyau* choisie $k(.,.)$. On donne les détails d'une telle fonction dans la section suivante, on note ici qu'avec une fonction noyau, le nouveau problème d'optimisation, dans sa formulation duale, devient

$$(P_{\mathcal{D}}) \begin{cases} \max\limits_{\alpha} \theta_{\mathcal{D}}(\alpha) & = \sum\limits_{i=1}^{n} \alpha_i - \dfrac{1}{2} \sum\limits_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j), \\ \text{s.c.} & \sum\limits_{i=1}^{n} \alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n \end{cases} \tag{25}$$

C'est, encore, un problème d'optimisation convexe, si la matrice $Q = [y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)]_{i,j}$ est définie positive. La résolution de ce problème conduit à la fonction de décision :

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i^* y_i k(\mathbf{x}, \mathbf{x}_i) + b^* \right) \tag{26}$$

1.4.1 Les noyaux

On a mentionné précédemment que pour qu'une fonction $k(.,.)$ soit une fonction noyau il doit exister une fonction $\phi : \mathcal{X} \to \mathcal{F}$ telle que $\mathcal{F}$ soit un espace de Hilbert avec un produit scalaire défini par $\langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle = k(\mathbf{x}, \mathbf{z})$, où $\mathbf{x}, \mathbf{z} \in \mathcal{X}$. On commence par définir la terminologie pour définir ensuite les conditions assurant qu'une fonction est une fonction noyau.

Définition 1.1 *(Matrice de Gram). Soit une fonction $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ et un ensemble de points $S_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathcal{X}$, la matrice $K \in \mathcal{M}_{n \times n}[\mathbb{R}]$ défini par*

$$K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j) \tag{27}$$

est appelée matrice de Gram ou matrice noyau de k relativement à S_n

Définition 1.2 *(Matrice définie positive).Une matrice réelle symétrique $K \in \mathcal{M}_{n \times n}[\mathbb{R}]$ satisfaisant*

$$\langle \boldsymbol{\alpha}, K\boldsymbol{\alpha} \rangle = \boldsymbol{\alpha}^T K \boldsymbol{\alpha} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K_{ij} \geq 0 \tag{28}$$

$\forall \boldsymbol{\alpha} \in \mathbb{R}^n$, est dite définie positive

Le lemme suivant donne un moyen pratique de tester si une matrice est une matrice noyau

Lemme 1.3 *Une matrice est définie positive si et seulement si ses valeurs propres sont non négatives.*

Définition 1.4 *(Fonction noyau). Une fonction $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est une fonction noyau si pour tout ensemble $S_n \subset \mathcal{X}$ la matrice noyau correspondante est une matrice définie positive.*

1. noyau polynomial :

$$k(\mathbf{x}, \mathbf{z}) = (a\langle \mathbf{x}, \mathbf{z} \rangle + b)^d \tag{29}$$

   où a, b, c sont des scalaires réels.

2. noyau RBF :

$$k(\mathbf{x}, \mathbf{z}) = \exp(-\gamma^2 \|\mathbf{x} - \mathbf{z}\|^2). \tag{30}$$