Machines à vecteurs supports

Résumé

Recherche d'un hyperplan, dit de marge optimale (vaste), pour la séparation de deux classes dans un espace hibertien défini par un noyau reproduisant associé au produit scalaire de cet espace. Estimation de l'hyperplan dans le cas linéaire et séparable ; les contraintes actives du problème d'optimisation déterminent les vecteurs supports. Extension au cas non séparable par pénalisation. Extension au cas non linéaire par plongement dans un espace hilbertien à noyau reproduisant.

Retour au plan du cours

1 Introduction

Les *Support Vector Machines* souvent traduit par l'appellation de Séparateur à Vaste Marge (SVM) sont une classe d'algorithmes d'apprentissage initialement définis pour la discrimination c'est-à-dire la prévision d'une variable qualitative binaire. Ils ont été ensuite généralisés à la prévision d'une variable quantitative. Dans le cas de la discrimination d'une variable dichotomique, ils sont basés sur la recherche de l'*hyperplan de marge optimale* qui, lorsque c'est possible, classe ou sépare correctement les données tout en étant le plus éloigné possible de toutes les observations. Le principe est donc de trouver un classifieur, ou une fonction de discrimination, dont la capacité de généralisation (qualité de prévision) est la plus grande possible.

Cette approche découle directement des travaux de Vapnik en théorie de l'apprentissage à partir de 1995. Elle s'est focalisée sur les propriétés de généralisation (ou prévision) d'un modèle en contrôlant sa complexité. Voir à ce sujet le chapitre **??** section **??** concernant la dimension de Vapnik Chernovenkis qui est un indicateur du pouvoir séparateur d'une famille de fonctions associé à un modèle et qui en contrôle la qualité de prévision. Le principe fondateur des SVM est justement d'intégrer à l'estimation le contrôle de la complexité c'est-à-dire le nombre de paramètres qui est associé dans ce cas au nombre de vecteurs supports. L'autre idée directrice de Vapnik dans ce développement, est d'éviter de substituer à l'objectif initial : la discrimination, un ou des problèmes qui s'avèrent finalement plus complexes à résoudre comme par exemple l'estimation non-paramétrique de la densité d'une loi multidimensionnelle en analyse discriminante.

Le principe de base des SVM consiste de ramener le problème de la discrimination à celui, linéaire, de la recherche d'un hyperplan optimal. Deux idées ou astuces permettent d'atteindre cet objectif :

- La première consiste à définir l'hyperplan comme solution d'un problème d'optimisation sous contraintes dont la fonction objectif ne s'exprime qu'à l'aide de produits scalaires entre vecteurs et dans lequel le nombre de contraintes "actives" ou vecteurs supports contrôle la complexité du modèle.
- Le passage à la recherche de surfaces séparatrices non linéaires est obtenu par l'introduction d'une fonction noyau (*kernel*) dans le produit scalaire induisant implicitement une transformation non linéaire des données vers un espace intermédiaire (*feature space*) de plus grande dimension. D'où l'appellation couramment rencontrée de machine à noyau ou *kernel machine*. Sur le plan théorique, la fonction noyau définit un espace hilbertien, dit auto-reproduisant et isométrique par la transformation non linéaire de l'espace initial et dans lequel est résolu le problème linéaire.

Cet outil devient largement utilisé dans de nombreux types d'applications et s'avère un concurrent sérieux des algorithmes les plus performants (agrégation de modèles). L'introduction de noyaux, spécifiquement adaptés à une problématique donnée, lui confère une grande flexibilité pour s'adapter à des situations très diverses (reconnaissance de formes, de séquences génomiques, de caractères, détection de spams, diagnostics...). À noter que, sur le plan algorithmique, ces algorithmes sont plus pénalisés par le nombre d'observations, c'est-à-dire le nombre de vecteurs supports potentiels, que par le nombre de variables. Néanmoins, des versions performantes des algorithmes permettent de prendre en compte des bases de données volumineuses dans des temps de calcul acceptables.

Le livre de référence sur ce sujet est celui de Schölkopf et Smola (2002). De nombreuses présentations des SVM sont accessibles sur des sites comme par exemple : `www.kernel-machines.org`. Guermeur et Paugam-Moisy (1999) et « Apprentissage Artificiel : Méthodes et Algorithmes » de Cornuéjols et Miclet (2002) en proposent en français.

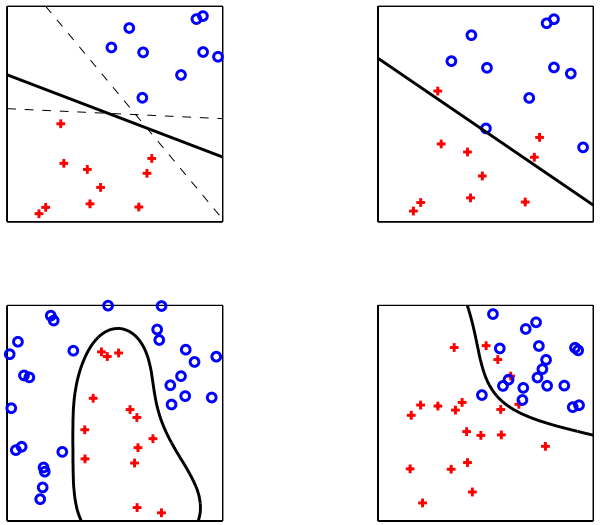


FIGURE 1 – Exemples de quatre types de problèmes de discrimination binaire où il s'agit de séparer les points bleus des croix rouges. La frontière de décision est représentée en noir.

2 Le problème de discrimination binaire

Le problème abordé est celui de la discrimination binaire. Il s'agit de trouver un moyen permettant de construire une fonction de décision associant à chaque observation sa classe. Nous allons nous traiter ce problème dans un cadre probabiliste spécifique et poser que les formes à discriminer sont des vecteurs $\boldsymbol{x} \in \mathbb{R}^p$. Le cadre probabiliste du problème consiste à supposer l'existence d'une loi inconnue $\mathbb{P}(\boldsymbol{x}, y)$ sur $(\mathbb{R}^p, \{-1, 1\})$. Le problème de discrimination vise à construire un estimateur de la fonction de décision idéale $D : \mathbb{R}^p \to \{-1, 1\}$, minimisant pour toutes les observations $\boldsymbol{x}$ la probabilité d'erreur $\mathbb{P}(D(\boldsymbol{x}) \neq y \mid \boldsymbol{x})$. Pour construire cet estimateur, on suppose l'existence d'un échantillon $\{(\boldsymbol{x}_i, y_i),\ i = 1, n\}$ (aussi appelé ensemble d'apprentissage), i.i.d. de loi parente $\mathbb{P}(\boldsymbol{x}, y)$ inconnue.

3 Les SVM linéaires

3.1 Le problème de discrimination linéaire

Un problème de discrimination est dit linéairement séparable lorsqu'il existe une fonction de décision linéaire (appelé aussi séparateur linéaire), de la forme $D(\boldsymbol{x}) = \text{signe}\big(f(\boldsymbol{x})\big)$ avec $f(\boldsymbol{x}) = \boldsymbol{v}^\top \boldsymbol{x} + a,\ \boldsymbol{v} \in \mathbb{R}^p$ et $a \in \mathbb{R}$, classant correctement toutes les observations de l'ensemble d'apprentissage $(D(\boldsymbol{x}_i) = y_i,\ i \in [1, n])$. La fonction f est appelée fonction caractéristique. C'est un problème particulier qui semble très spécifique, mais qui permet d'introduire de manière pédagogique les principaux principes des SVM : marge, programmation quadratique, vecteur support, formulation duale et matrice de gram. Nous allons ensuite généraliser au cas des observations non séparables et non linéaires par l'introduction de variables d'écart et de noyaux. Ces différent type de problèmes sont illustrés figure 1.

A toute fonction de décision et donc aux fonction de décision linéaire ont peut associer une frontière de décision :

$$\Delta(\boldsymbol{v}, a) = \{\boldsymbol{x} \in \mathbb{R}^p \mid \boldsymbol{v}^\top \boldsymbol{x} + a = 0\}$$

Tout comme la fonction de décision linéaire, cette frontière de décision est définie à un terme multiplicatif prés dans le sens ou la frontière définie par le couple $(\boldsymbol{v}, a)$ est la même que celle engendrée par $(k\boldsymbol{v}, ka)\ \forall\, k \in \mathbb{R}$. Cela est lié à la définition de l'hyperplan affine associé à la fonction caractéristique. Pour garantir l'unicité de la solution on peut soit considérer l'hyperplan standard (tel que $\|\boldsymbol{v}\| = 1$) soit l'hyperplan canonique par rapport à un point $\boldsymbol{x}$ (tel que $\boldsymbol{v}^\top \boldsymbol{x} + a = 1$).

3.2 La marge d'un classifieur

Pour un échantillon donnée, il est possible d'associer deux marges à un même classifieur linéaire : sa marge géométrique et sa marge numérique. La marge géométrique m d'un échantillon linéairement séparable est donnée par la plus petite distance d'un point de l'échantillon à la frontière de décision. La marge numérique μ est donnée elle par la plus petite valeur de la fonction de décision

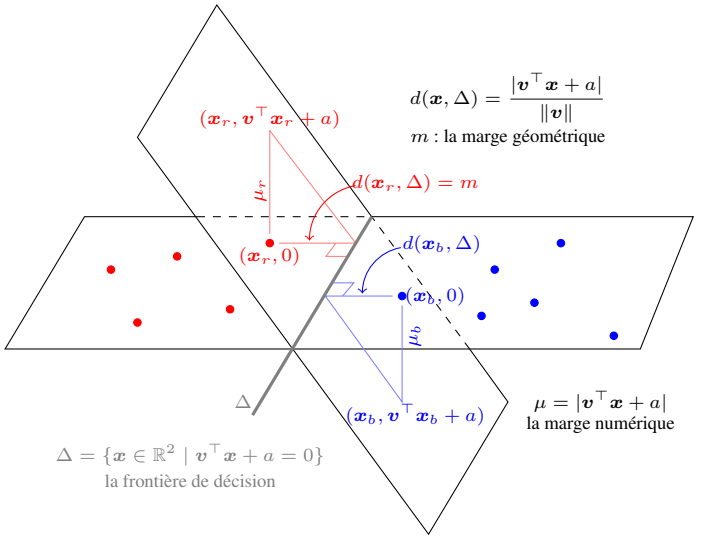


FIGURE 2 – iIllustration des deux notions de marge sur un exemple de discrimination linéaire séparable en dimension deux.

atteinte sur un point de l'échantillon. Leur définition mathématique est :

$$\underbrace{m = \min_{i \in [1,n]} \text{dist}(\boldsymbol{x}_i, \Delta(\boldsymbol{v}, a))}_{\text{marge géométrique}} \qquad \underbrace{\mu = \min_{i \in [1,n]} |\boldsymbol{v}^\top \boldsymbol{x}_i + a|}_{\text{marge numérique}}$$

La figure 2 illustre ces deux notions de marge pour un exemple en deux dimensions. On voit que pour une frontière de décision donnée, la marge géométrique m est fixée alors que la marge numérique μ dépend de la « pente » de l'hyperplan de décision (donc de $\|\boldsymbol{v}\|$). En effet, pour une observation donnée, les deux marges forment les cotés adjacents d'un triangle rectangle dont l'hypoténuse est définie par la fonction caractéristique $\boldsymbol{v}^\top \boldsymbol{x} + a$.

3.3 Maximisation de la marge d'un classifieur

Lorsque des observations sont linéairement séparables, comme l'illustre la figure 1 (en haut à gauche) il existe dans le cas général une infinité de frontières de décision linéaires séparant cet échantillon. La notion de marge offre un critère de choix parmi toutes ces solutions en admettant que maximiser la marge c'est aussi maximiser la confiance et donc minimiser la probabilité d'erreur associée au classifieur. Nous allons résoudre le problème suivant :

$$\max_{\boldsymbol{v}, a} \underbrace{\min_{i \in [1,n]} \text{dist}(\boldsymbol{x}_i, \Delta(\boldsymbol{v}, a))}_{\text{marge : } m}$$

En introduisant explicitement la marge comme une variable, ce problème se réécrit comme un problème d'optimisation sous contraintes :

$$\begin{cases} \max\limits_{\boldsymbol{v}, a} & m \\ \text{avec} & \min\limits_{i \in [1,n]} \dfrac{|\boldsymbol{v}^\top \boldsymbol{x}_i + a|}{\|\boldsymbol{v}\|} \geq m \end{cases}$$

C'est un problème est mal posé dans le sens ou si $(\boldsymbol{v}, a)$ est une solution, $(k\boldsymbol{v}, ka), \forall\ 0 < k$ l'est aussi. Une manière de traiter cette difficulté est d'effectuer le changement de variable : $\boldsymbol{w} = \frac{\boldsymbol{v}}{m\|\boldsymbol{v}\|}$ et $b = \frac{a}{m\|\boldsymbol{v}\|}$ Le problème se réécrit alors :

$$\begin{cases} \max\limits_{\boldsymbol{w}, b} & \dfrac{1}{\|\boldsymbol{w}\|} \\ \text{avec} & y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 \quad ;\ i = 1, n \end{cases}$$

puisque $\|\boldsymbol{w}\| = \frac{1}{m}$. Cela revient à fixer à un la marge numérique du classifieur recherché (celui de norme minimale). La formulation « classique » des SVM s'obtient alors en minimisant $\|\boldsymbol{w}\|^2$ au lieu de maximiser l'inverse de la norme, ce qui donne le problème suivant qui admet la même solution que le problème précédent.

Définition 3.1 (SVM sur des données linéairement séparables)
Soit $\{(\boldsymbol{x}_i, y_i);\ i = 1, n\}$ *un ensemble de vecteurs formes étiquetées avec* $\boldsymbol{x}_i \in \mathbb{R}^p$ *et* $y_i \in \{1, -1\}$. *Un séparateur à vaste marge linéaire (SVM et support vector machine) est un discriminateur linéaire de la forme :* $D(x) = signe\big(\boldsymbol{w}^\top \boldsymbol{x} + b\big)$ *où* $\boldsymbol{w} \in \mathbb{R}^p$ *et* $b \in \mathbb{R}$ *sont donnés par la résolution du problème suivant :*

$$\begin{cases} \min\limits_{\boldsymbol{w}, b} & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ avec & y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 \qquad i = 1, n \end{cases}$$

Ce problème d'optimisation sous contraintes est un programme quadratique de la forme :

$$\left\{ \begin{array}{ll} \min\limits_{\boldsymbol{z}} & \frac{1}{2}\, \boldsymbol{z}^\top A \boldsymbol{z} - \boldsymbol{d}^\top \boldsymbol{z} \\ \text{avec} & B\boldsymbol{z} \leq \boldsymbol{e} \end{array} \right.$$

où $\boldsymbol{z} = (\boldsymbol{w}, b)^\top \in \mathbb{R}^{p+1}$, $\boldsymbol{d} = (0, \ldots, 0)^\top \in \mathbb{R}^{p+1}$, $A = \left[\begin{array}{cc} I & 0 \\ 0 & 0 \end{array} \right]$, I est la matrice identité de $\mathbb{R}^p$, $B = -[\text{diag}(\boldsymbol{y})X, \boldsymbol{y}]$, $\boldsymbol{e} = -(1, \ldots, 1)^\top \in \mathbb{R}^n$, $\boldsymbol{y} \in \mathbb{R}^n$ le vecteur des signes des observations et X la matrice $n \times d$ des observations dont la ligne i est le vecteur $\boldsymbol{x}_i^\top$. Ce problème est convexe puisque la matrice A est semidéfinie positive. Il admet donc une solution unique (qui existe puisque le problème est linéairement séparable par hypothèse) et les conditions nécessaires d'optimalité du premier ordre sont aussi suffisantes. Ce problème (dit primal) admet une formulation duale équivalente qui est aussi un programme quadratique.

La résolution du problème des SVM sur des données linéairement séparables peut se faire directement (à partir de la formulation primale) par exemple en utilisant une méthode stochastique de type Gauss-Seidel, une méthode d'ensemble actif, un algorithme de point intérieur, de Newton avec région de confiance ou type gradient conjugué. Cependant, il est intéressant de passer par la formation duale de ce problème :

- le problème dual est un programme quadratique de taille n (égal au nombre d'observations) qui peut s'avérer plus facile à résoudre que le problème primal,
- la formulation duale fait apparaître la matrice de Gram $XX^\top$ ce qui permet dans le cas général (non linéaire) d'introduire la non linéarité à travers des noyaux.

Afin de retrouver cette formulation duale nous allons maintenant expliciter le lagrangien du problème et les conditions d'optimalité de Karush, Kuhn et Tucker. Ces conditions vont nous permettre d'introduire la notion importante de vecteur support.

3.4 Conditions d'optimalité et vecteurs supports

Afin d'expliciter les conditions nécessaires d'optimalité du premier ordre il est classique lorsque l'on traite d'un problème d'optimisation sous contraintes d'expliciter son lagrangien. Dans le cas des SVM il s'écrit :

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \tfrac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i \big(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1\big)$$

où les $\alpha_i \geq 0$ sont les multiplicateurs de Lagrange associés aux contraintes. Les conditions d'optimalité de Karush, Kuhn et Tucker du programme quadratique associé aux SVM permettent de caractériser la solution du problème primal $(\boldsymbol{w}^*, b^*)$ et les multiplicateurs de lagrange α^* associés par le système d'équations suivant :

$$\begin{array}{rrll} \text{stationarité} & \boldsymbol{w}^* - \sum\limits_{i=1}^{n} \alpha_i^* y_i \boldsymbol{x}_i & = 0 & \\ & \sum\limits_{i=1}^{n} \alpha_i^*\, y_i & = 0 & \\ \text{complementarité} & \alpha_i^* \big(y_i(\boldsymbol{w}^{*\top} \boldsymbol{x}_i + b^*) - 1\big) & = 0 & i = 1, \ldots, n \\ \text{admissibilité primale} & y_i(\boldsymbol{w}^{*\top} \boldsymbol{x}_i + b^*) & \geq 1 & i = 1, \ldots, n \\ \text{admissibilité duale} & \alpha_i^* & \geq 0 & i = 1, \ldots, n \end{array}$$

Les conditions de complémentarité permettent de définir l'ensemble $\mathcal{A}$ des indices des contraintes actives (ou saturées) à l'optimum dont les multiplicateurs de Lagrange $\alpha_i^* > 0$ sont strictement positifs :

$$\mathcal{A} = \{i \in [1, n] \;\big|\; y_i(\boldsymbol{w}^{*\top} \boldsymbol{x}_i + b^*) = 1\}$$

Pour les autres contraintes, la condition de complémentarité implique que leur mutiplicateur de Lagrange est égal à zéro et que l'observation associée vérifie strictement l'inégalité $\forall j \notin \mathcal{A}, y_j(\boldsymbol{w}^{*\top} \boldsymbol{x}_j + b^*) > 1$. La solution $(\boldsymbol{w}, b, \boldsymbol{\alpha}_\mathcal{A})$ vérifie le système linéaire suivant :

$$\left\{ \begin{array}{rrrl} \boldsymbol{w} & -X_\mathcal{A}^\top D_y \boldsymbol{\alpha}_\mathcal{A} & & = \boldsymbol{0} \\ -D_y X_\mathcal{A} \boldsymbol{w} & & -b\boldsymbol{y}_\mathcal{A} & = -\boldsymbol{e}_\mathcal{A} \\ & -\boldsymbol{y}_\mathcal{A}^\top \boldsymbol{\alpha}_\mathcal{A} & & = 0 \end{array} \right.$$

avec $D_y = \text{diag}(\boldsymbol{y}_\mathcal{A})$, $\boldsymbol{\alpha}_\mathcal{A} = \boldsymbol{\alpha}(\mathcal{A})$, $\boldsymbol{y}_\mathcal{A} = \boldsymbol{y}(\mathcal{A})$ et $X_\mathcal{A} = X(\mathcal{A}; :)$. Le premier sous système (qui est la condition de stationnarité pour $\boldsymbol{w}$) est particulièrement important car il stipule que la solution du problème s'écrit :

$$\boldsymbol{w} = \sum_{i \in \mathcal{A}} \alpha_i\, y_i \boldsymbol{x}_i$$

A l'optimum, le vecteur $\boldsymbol{w}$ est une combinaison linéaire des observations $\boldsymbol{x}_i$ liées aux contraintes actives $i \in \mathcal{A}$ et pour lesquelles $|\boldsymbol{w}^{*\top}\boldsymbol{x}_i + b^*| = 1$. Ces observations sont appelées *vecteurs supports* dans le sens où elles *supportent* l'hyperplan de discrimination puisque leur marge numérique est égale à un. Les autres données n'interviennent pas dans le calcul et leur marge numérique est strictement supérieure à un. La figure 3 vient illustrer le fait que les vecteurs support définissent la frontière de décision. La fonction de décision peut alors s'écrire de deux formes différentes selon que l'on considère les variables primales ou les variables duales :

$$D(\boldsymbol{x}) = \text{sign}\big(f(\boldsymbol{x})\big) \quad \text{avec} \quad f(\boldsymbol{x}) = \sum_{j=1}^{d} w_j x_j + b = \sum_{i \in \mathcal{A}} \alpha_i \, y_i (\boldsymbol{x}^\top \boldsymbol{x}_i) + b$$

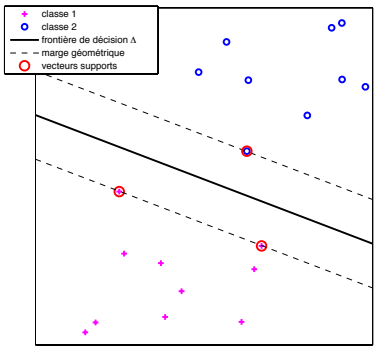


FIGURE 3 – illustration de la notion de vecteur support dans le cas d'un problème linéairement séparable. Les vecteurs supports des deux classes sont entourés par un rond rouge. Le tableau suivant récapitule les deux cas possibles.

observation inutile	observation importante
	vecteur support
contrainte non saturée	contrainte active
$i \notin \mathcal{A}$	$i \in \mathcal{A}$
$\alpha_i = 0$	$0 < \alpha_i$
$y_i(\boldsymbol{w}^{*\top}\boldsymbol{x}_i + b^*) > 1$	$y_i(\boldsymbol{w}^{*\top}\boldsymbol{x}_i + b^*) = 1$

Résoudre le problème revient à rechercher parmi les exemples disponibles ceux qui vont devenir vecteurs supports ce qui nous donne l'ensemble $\mathcal{A}$.

Une fois cet ensemble connu, la résolution du problème s'effectue en utilisant le premier sous système pour éliminer $\boldsymbol{w}$. Le problème devient :

$$\begin{cases} D_y X_{\mathcal{A}} X_{\mathcal{A}}^\top D_y \boldsymbol{\alpha}_{\mathcal{A}} + b\boldsymbol{y}_{\mathcal{A}} = \boldsymbol{e}_{\mathcal{A}} \\ \boldsymbol{y}_{\mathcal{A}}^\top \alpha_{\mathcal{A}} = 0 \end{cases}$$

le cardinal de $\mathcal{A}$ est égal au nombre d'inconnues $(p+1)$ et la solution est :

$$b = \frac{\boldsymbol{y}_{\mathcal{A}}^\top (D_y X_{\mathcal{A}} X_{\mathcal{A}}^\top D_y)^{-1} \boldsymbol{e}_{\mathcal{A}}}{\boldsymbol{y}_{\mathcal{A}}^\top (D_y X_{\mathcal{A}} X_{\mathcal{A}}^\top D_y)^{-1} \boldsymbol{y}_{\mathcal{A}}} \quad \text{et} \quad \boldsymbol{\alpha}_{\mathcal{A}} = (D_y X_{\mathcal{A}} X_{\mathcal{A}}^\top D_y)^{-1}(\boldsymbol{e} - b\boldsymbol{y}_{\mathcal{A}})$$

ce qui permet d'utiliser le fait que la matrice $D_y X_{\mathcal{A}} X_{\mathcal{A}}^\top D_y$ est définie positive. L'algorithme associé est :

ALGORITHME 1 : *calcul de la solution des SVM connaissant* $\mathcal{A}$

```
Xa = diag(y(A))*X(A, :) ; %A est l'ensemble actif
U = chol(Xa*Xa') ;        %factorisation de Choleski
a = U\ (U'\ e(A)) ;       %résolution des systèmes tri sup
c = U\ (U'\ y(A)) ;
b = (y(A)'*a)\(y(A)'*c) ;
alpha = U\ (U'\ (e(A) - b*y(A))) ;
w = Xa'*alpha ;
```

3.5 Formulation duale et biduale

Le dual de Wolfe dual du problème des SVM séparables est :

$$\begin{cases} \max\limits_{\boldsymbol{w},b,\alpha} \ \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum\limits_{i=1}^{n} \alpha_i \big(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1\big) \\ \text{avec} \quad \alpha_i \geq 0 \qquad\qquad\qquad\qquad\qquad\qquad i = 1, \ldots, n \\ \qquad\quad \boldsymbol{w} - \sum\limits_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i = 0 \quad \text{et} \quad \sum\limits_{i=1}^{n} \alpha_i \, y_i = 0 \end{cases}$$

L'élimination de la variable primale $\boldsymbol{w}$ donne la formulation duale du problème des SVM :

$$\left\{\begin{array}{ll}\min\limits_{\alpha} & \frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}\alpha_j\alpha_i y_i y_j \boldsymbol{x}_j^\top \boldsymbol{x}_i - \sum\limits_{i=1}^{n}\alpha_i \\ \text{avec} & \sum\limits_{i=1}^{n}\alpha_i\, y_i = 0 \quad \text{et} \quad 0 \leq \alpha_i \qquad i = 1,\ldots,n\end{array}\right.$$

que l'on écrit matriciellement :

$$\mathcal{D}\left\{\begin{array}{ll}\min\limits_{\boldsymbol{\alpha}\in\mathbb{R}^n} & \frac{1}{2}\boldsymbol{\alpha}^\top G\boldsymbol{\alpha} - \boldsymbol{e}^\top\boldsymbol{\alpha} \\ \text{avec} & \boldsymbol{y}^\top\boldsymbol{\alpha} = 0 \\ & 0 \leq \alpha_i \qquad i = 1,n\end{array}\right.$$

avec $G = D_y X X^\top D_y$ la matrice symétrique $n \times n$ de terme général $G_{ij} = y_i y_j \boldsymbol{x}_j^\top \boldsymbol{x}_i$. Le Lagrangien de ce problème dual est :

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\alpha},\beta,\boldsymbol{\gamma}) = \tfrac{1}{2}\boldsymbol{\alpha}^\top G\boldsymbol{\alpha} - \boldsymbol{e}^\top\boldsymbol{\alpha} + \beta\,\boldsymbol{y}^\top\boldsymbol{\alpha} - \sum_{i=1}^{n}\gamma_i\alpha_i$$

où β est le multipicateur associé à la contrainte d'égalité et $\gamma_i \geq 0$ les multipicateurs de Lagrange associés aux contraintes d'inégalité. On en déduit le problème bidual associé :

$$\left\{\begin{array}{ll}\max\limits_{\beta\in\mathbb{R},\gamma\in\mathbb{R}^n} & -\frac{1}{2}\boldsymbol{\alpha}^\top G\boldsymbol{\alpha} \\ \text{avec} & G\boldsymbol{\alpha} - \boldsymbol{e} + \beta\boldsymbol{y} - \gamma = 0 \\ & 0 \leq \gamma_i \qquad i = 1,n\end{array}\right.$$

En posant $\boldsymbol{w} = X\text{diag}(\boldsymbol{y})\boldsymbol{\alpha}$ et en identifiant b et β on retrouve la formulation initiale du problème primal. Le biais b est donné par le multiplicateur de Lagrange associé à la condition d'égalité $\boldsymbol{y}^\top\boldsymbol{\alpha} = 0$.

Ces deux formulations équivalentes primale et duale méritent d'être mis en parallèle pour mieux faire ressortir leurs avantages respectifs et permettre de décider quand il est préférable d'utiliser la formulation primale et quand utiliser le dual. Le tableau suivant récapitule les principales caractérises des SVM linéaires primales et duales.

Primal

$$\left\{\begin{array}{ll}\min\limits_{\boldsymbol{w}\in\mathbb{R}^p, b\in\mathbb{R}} & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{avec} & y_i(\boldsymbol{w}^\top\boldsymbol{x}_i + b) \geq 1 \\ & i = 1,n\end{array}\right.$$

- $p+1$ inconnues (les influences des variables)
- n contraintes linéaires
- QP de forme générale
- préférable lorsque $d < n$

Dual

$$\left\{\begin{array}{ll}\min\limits_{\alpha\in\mathbb{R}^n} & \frac{1}{2}\boldsymbol{\alpha}^\top G\boldsymbol{\alpha} - \boldsymbol{e}^\top\boldsymbol{\alpha} \\ \text{avec} & \boldsymbol{y}^\top\boldsymbol{\alpha} = 0 \\ \text{et} & 0 \leq \boldsymbol{\alpha}_i \qquad i = 1,n\end{array}\right.$$

- n inconnues (les influences des observations)
- n contraintes de boites
- G matrice des influences de chaque couple de points
- préférable lorsque $d \geq n$

Le problème est essentiellement pratique et lié à la mise en œuvre. Cette question n'est pas résolue et il existe des approches primales et des approches duales très efficaces permettant de traiter un grès grand nombre de données en grande dimension. Cependant, quelle que soit l'approche choisie, il semble important de ne pas trop pousser l'optimisation et que son arrêt prématuré peut offrir des gains de performance (technique appelée *early stoping*).

3.6 Le cas des données non séparables

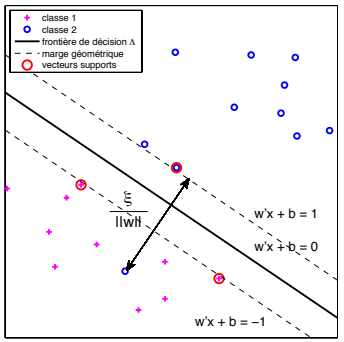
Dans le cas où les données ne sont pas linéairement séparables, il est possible de suivre la même démarche adoptée dans le cas séparable au prix de l'introduction de variables d'écart. L'idée est de modéliser les erreurs potentielles par des variables d'écart positives ξ_i associées à chacune des observations $(x_i, y_i), i \in [1,n]$. Dans le cas où le point vérifie la contrainte de marge $y_i(\boldsymbol{w}^\top\boldsymbol{x}_i + b) \geq 1$, la variable d'écart est nulle. On a donc les deux cas suivants :

$$\begin{array}{lll}\text{pas d'erreur :} & y_i(\boldsymbol{w}^\top\boldsymbol{x}_i + b) \geq 1 \Rightarrow & \xi_i = 0 \\ \text{erreur :} & y_i(\boldsymbol{w}^\top\boldsymbol{x}_i + b) < 1 \Rightarrow & \xi_i = 1 - y_i(\boldsymbol{w}^\top\boldsymbol{x}_i + b) > 0\end{array}$$

On associe à cette définition une fonction cout appelée « cout charnière » du fait de la forme de son graphe représentée figure 4 :

$$\xi_i = \max\big(0, 1 - y_i(\boldsymbol{w}^\top\boldsymbol{x}_i + b)\big)$$

Le problème consiste alors à simultanément maximiser la marge et minimiser



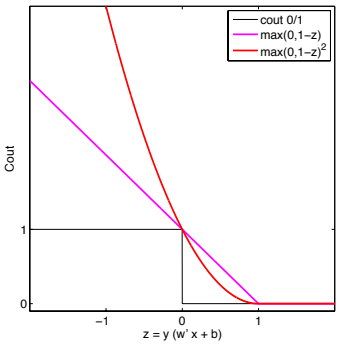


FIGURE 4 – illustration de la notion d'écart (à gauche) et de la fonction de cout charnière (droite). Dans cet exemple tous les écarts sont nuls sauf un, celui du point bleu mal classé. Cet écart mesure la distance du point à la marge numérique de l'hyperplan séparateur.

la somme des terme d'erreurs à la puissance d (avec typiquement $d = 1$ ou 2). Il s'écrit alors :

$$\left\{ \begin{array}{ll} \min\limits_{\boldsymbol{w},b,\xi} & \left\{ \begin{array}{l} \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \frac{1}{d}\sum\limits_{i=1}^{n} \xi_i^d \end{array} \right. \\ \text{avec} & y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \quad i = 1, n \end{array} \right.$$

On constate que si toutes les variables d'écart $\xi_i = 0$, on retrouve le problème des SVM linéairement séparables. La résolution du problème précédent peut s'effecteur grâce à l'introduction d'un terme d'équilibrage $C > 0$ fixé qui permet de poser le problème sous la forme suivante :

$$(\textit{Primal} \text{ linéaire SVM}) \quad \left\{ \begin{array}{ll} \min\limits_{\boldsymbol{w},b,\xi} & \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{d}\sum\limits_{i=1}^{n} \xi_i^d \\ \text{avec} & y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i \quad i = 1, n \\ & \xi_i \geq 0 \quad i = 1, n \end{array} \right.$$

Comme nous l'avons fait précédemment, on recherche un point selle du lagrangien :

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{d}\sum_{i=1}^{n} \xi_i^d - \sum_{i=1}^{n} \alpha_i \big(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1 + \xi_i\big) - \sum_{i=1}^{n} \beta_i \xi_i$$

avec des multiplicateurs de Lagrange $\alpha_i \geq 0$ et $\beta_i \geq 0$. Le calcul des gradients est identique par rapport à $\boldsymbol{w}$ et b. Dans le cas où $d = 1$, la dérivée partielle du lagrangien par par rapport aux variables d'écart s'écrit : $\partial_{\xi_i}\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha}) = C - \alpha_i - \beta_i$. Cette condition supplémentaire de stationnarité permet d'éliminer les β car :

$$\beta_i \geq 0 \text{ et } C - \alpha_i - \beta_i = 0 \quad \Rightarrow \quad \alpha_i \leq C$$

L'ensemble de ces conditions nous permet d'obtenir la formulation duale qui est le programme quadratique à résoudre pour obtenir la solution des SVM.

Définition 3.2 (formulation duale des SVM L1) *Le problème dual des SVM avec une variable d'écart et le cout charnière (appelé aussi L1 SVM) s'écrit :*

$$(\textit{Dual SVM L1}) \quad \left\{ \begin{array}{ll} \min\limits_{\boldsymbol{\alpha} \in \mathbb{R}^n} & \frac{1}{2}\boldsymbol{\alpha}^\top G \boldsymbol{\alpha} - \boldsymbol{e}^\top \boldsymbol{\alpha} \\ \textit{avec} & \boldsymbol{y}^\top \boldsymbol{\alpha} = 0 \\ & 0 \leq \alpha_i \leq C \qquad i = 1, n \end{array} \right.$$

avec G la matrice symétrique $n \times n$ de terme général $G_{ij} = y_i y_j \boldsymbol{x}_j^\top \boldsymbol{x}_i$.

Ce problème reste un programme quadratique avec des contraintes de boites. Les multiplicateurs $\boldsymbol{\alpha}$ représentant l'influence de chacun des exemples sont maintenant bornés supérieurement par C qui représente l'influence maximale permise pour un exemple.

Dans le cas où $d = 2$, la contrainte de positivité sur les variables d'écart est inutile et le problème dual s'écrit (L2 SVM) :

$$(Dual \text{ L2 SVM}) \quad \left\{ \begin{array}{ll} \min\limits_{\boldsymbol{\alpha}\in\mathbb{R}^n} & \frac{1}{2}\boldsymbol{\alpha}^\top(G+\frac{1}{C}I)\boldsymbol{\alpha} - \boldsymbol{e}^\top\boldsymbol{\alpha} \\ \text{avec} & \boldsymbol{y}^\top\boldsymbol{\alpha} = 0 \\ & 0 \leq \alpha_i \qquad\qquad i = 1,n \end{array} \right.$$

avec I la matrice identité. Dans ce cas, la matrice G est « préconditionnée » d'un facteur $\frac{1}{C}$.

3.7 Autres formulations et généralisation

Il existe d'autres manières de poser le problème permettant d'arriver au même résultat et qui permettent d'éclairer différents aspects des SVM.

Si l'on considère la minimisation de la distance entre les enveloppes convexes de chacune des classes on retrouve le problème dual des SVM :

$$\left\{ \begin{array}{ll} \min\limits_{\boldsymbol{\alpha}\in\mathbb{R}^n} & \|u-v\|^2 \\ \text{avec} & u = \sum\limits_{\{i|y_i=1\}} \alpha_i \boldsymbol{x}_i, \qquad v = \sum\limits_{\{i|y_i=-1\}} \alpha_i \boldsymbol{x}_i \\ & \sum\limits_{\{i|y_i=1\}} \alpha_i = 1, \sum\limits_{\{i|y_i=-1\}} \alpha_i = 1, \quad 0 \leq \alpha_i \quad i = 1,n \end{array} \right.$$

et

$$f(\boldsymbol{x}) = \frac{2}{\|u-v\|^2}\left(u^\top\boldsymbol{x} - v^\top\boldsymbol{x}\right) \text{ et } b = \frac{\|u\|^2 - \|v\|^2}{\|u-v\|^2}$$

Une autre manière de poser le problème consiste à utiliser une formulation de type « régularisation du risque empirique ». Dans ce formalisme on considère que m'on cherche à minimiser une forme pénalisée du cout charnière. Le cout charnière est vu comme un terme d'attache aux données et la pénalité un terme de régularisation :

$$\min_{\boldsymbol{w},b} \quad \frac{1}{q}\|\boldsymbol{w}\|_q^q + \frac{1}{d}\sum_{i=1}^{n} C_i \ \max\left(1 - y_i(\boldsymbol{w}^\top\boldsymbol{x}_i + b) \geq 1, 0\right)^d$$

avec $q = 2$, $d = 1$ et les C_i identiques pour les SVM que nous avons introduits, $q = 1$ ($\|\boldsymbol{w}\|_1 = \sum_{j=1}^{d} |w_j|$) et $d = 1$ LP SVM (les SVM en programmation linéaire avec sélection de variable). Il existe de nombreuses autres variantes. Citons le cas $q = 0$ L0 SVM (le problème est alors non convexe), et *elastic net* SVM avec à la fois un terme avec $q = 1$ et un autre avec $q = 2$ qui débouche aussi sur un programme quadratique. Les C_i permettent de moduler l'influence des observations par exemple en fonction de leur classe (utile dans le cas de classes déséquilibrées).

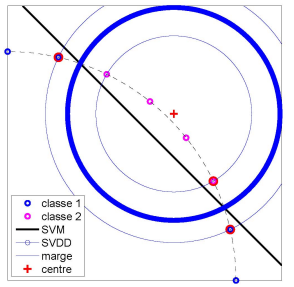


FIGURE 5 – Exemple de discrimation de données normalisées par SVM (la frontière est en noir) et par SVDD avec une marge de 0.05 (la frontière de décision est en bleu).

Une autre approche (appelée en anglais *support vector data description ou SVDD*) consiste à rechercher l'enveloppe sphérique de rayon minimal englobant les points d'une classe et excluant ceux de l'autre classe. Le rayon de cette enveloppe est noté R et m désigne une marge de confiance sur cette classification. Le problème posé s'écrit pour C et m fixés :

$$\left\{ \begin{array}{lll} \min\limits_{\boldsymbol{c},R,\xi\in\mathbb{R}^n} & R^2 + C\Big(\sum\limits_{y_i=1} \xi_i^+ + \sum\limits_{y_i=-1} \xi_i^-\Big) & \\ \text{avec} & \|\boldsymbol{x}_i - \boldsymbol{c}\|^2 \leq R^2 - m + \xi_i^+, \quad \xi_i^+ \geq 0 & \text{quand } y_i = 1 \\ & \|\boldsymbol{x}_i - \boldsymbol{c}\|^2 \geq R^2 + m - \xi_i^-, \quad \xi_i^- \geq 0 & \text{quand } y_i = -1 \end{array} \right.$$

C'est un programme linéaire avec des contraintes quadratique dont le dual est un programme quadratique analogue à celui des SVM. Pour $m = 0$ on a la forme originale des SVDD. Dans le cas particulier où les observations sont normées ($\|\boldsymbol{x}_i\| = 1$) et pour $m = 1$, le programme quadratique dual des SVDD est (pratiquement) identique à celui des SVM (voir l'illustration figure 5).

4 Le cas non linéaire : les noyaux

Dans le cas général, la frontière optimale est non linéaire. Dans le cadre des SVM, la prise en compte de non linéarités dans le modèle s'effectue par l'introduction de noyaux non linéaires. De manière inattendue, l'utilisation de noyaux ne modifie pas fondamentalement la nature des SVM (pourvu que l'on travaille dans le dual).

Ce noyau est une fonction k qui associe à tout couple d'observations $(\boldsymbol{x}, \boldsymbol{x}')$ une mesure de leur « influence réciproque » calculée à travers leur corrélation ou leur distance. Des exemples typiques de noyaux sont le noyaux polynômial $k(\boldsymbol{x}, \boldsymbol{x'}) = (c + \langle \boldsymbol{x}, \boldsymbol{x'} \rangle)^p$ et le noyau gaussien $k(\boldsymbol{x}, \boldsymbol{x'}) = e^{-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\sigma^2}}$. Etant donné le noyau k, la fonction de décision s'écrit $D(\boldsymbol{x}) = \text{signe}(f(\boldsymbol{x}) + b)$ et :

$$\text{cas linéaire : } f(\boldsymbol{x}) = \sum_{i \in \mathcal{A}} \alpha_i y_i \boldsymbol{x}_i^\top \boldsymbol{x} \qquad \text{non linéaire : } f(\boldsymbol{x}) = \sum_{i \in \mathcal{A}} \alpha_i y_i k(\boldsymbol{x}_i, \boldsymbol{x})$$

La fonction de discrimination est une combinaison linéaire des noyaux dont le signe de l'influence dépend de la classe. L'ensemble actif $\mathcal{A}$, les coefficients α_i associés et le biais b sont donné par la résolution du même problème dual que dans le cas des SVM linéaires (définition 3.2). Seule la définition de la matrice G change (mais pas sa taille) :

$$\text{cas linéaire : } G_{ij} = y_i y_j \, \boldsymbol{x}_i^\top \boldsymbol{x}_j \qquad \text{non linéaire : } G_{ij} = y_i y_j \, k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

L'idée ici est de considérer dans le dual l'influence de chacune des observations dans la construction de la solution. Le calcul de cette influence passe par la résolution d'un programme quadratique de taille n (le nombre d'observations). L'utilisation de noyaux permet d'introduire de la non linéarité sans modifier la complexité algorithmique du problème à résoudre.

Nous allons maintenant détailler le cadre théorique et le mécanisme d'action des noyaux.

4.1 Les noyaux

Définition des noyaux

Outre le passage au non linéaire, l'utilisation de noyaux permet une autre généralisation très utile dans la pratique : la définition des SVM sur des objets complexes comme des images, des graphes, des protéines où des automates pour ne citer qu'eux. Nous n'avons donc pas besoin de faire d'hypothèse sur la nature du domaine des observations $\mathcal{X}$ et un noyau k est défini de manière générale comme une fonction de deux variables sur $\mathbb{R}$:

$$k : \mathcal{X}, \mathcal{X} \longrightarrow \mathbb{R} \qquad (\boldsymbol{x}, \boldsymbol{x}') \longmapsto k(\boldsymbol{x}, \boldsymbol{x}')$$

Définition 4.1 (Matrice de Gram) *La matrice de Gram du noyau $k(.,.)$ pour les observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_n\}$ (pour tout entier n fini) est la matrice carrée K de taille n et de terme général $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.*

Définition 4.2 (Noyau positif) *Un noyau k est dit positif si, pour tout entier n fini et pour toutes les suite de n observations possibles $\{\boldsymbol{x}_i, i = 1, n\}$, la matrice de Gram associée est une matrice symétrique définie positive.*

L'intéret des noyaux positifs c'est qu'il est possible de leur associer un produit scalaire. Les noyaux dit de Mercer sont des noyaux positifs définis sur un ensemble compact.

Construction des noyaux

Il existe deux façons de construire un noyau positif :

1. soit on s'appuie sur une transformation $\Phi(\boldsymbol{x})$ de $\mathcal{X}$ sur un espace $\mathcal{H}$ muni d'un produit scalaire et l'on défini le noyau à travers ce produit scalaire : $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x'}) \rangle_{\mathcal{H}}$. Par exemple si l'on considère $\boldsymbol{x} = (x_1, x_2)$ dans $\mathbb{R}^2$ et $\Phi(\boldsymbol{x}) = (x_1^2, \sqrt{2} x_1 x_2, x_2^2)$ est explicite. Dans ce cas, $\mathcal{H}$ est de dimension 3 et le produit scalaire dans $\mathbb{R}^3$ s'écrit :

$$\begin{aligned} \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}') \rangle &= x_1^2 x_1'^2 + 2 x_1 x_2 x_1' x_2' + x_2^2 x_2'^2 \\ &= (x_1 x_1' + x_2 x_2')^2 \\ &= \left(\boldsymbol{x}^\top \boldsymbol{x}'\right)^2 \\ &= k(\boldsymbol{x}, \boldsymbol{x'}). \end{aligned}$$

2. soit on utilise les propriétés algèbriques des noyaux positifs :
   - un noyau séparable est un noyau positif,
   - la somme de deux noyaux positifs est un noyau positif,
   - le produit de deux noyaux positifs est un noyau positif,

- le produit tensoriel ainsi la somme directe de deux noyaux positifs est un noyau positif,
- le passage à la limite conserve la positivité : si la limite d'une suite de noyaux positif existe, c'est aussi un noyau positif.

On peux ainsi par exemple démontrer la positivité du noyau gaussien.

$$\begin{aligned}\exp(-\|\boldsymbol{x}-\boldsymbol{x'}\|^2) &= \exp(-\|\boldsymbol{x}\|^2 - \|\boldsymbol{x'}\|^2 + 2\boldsymbol{x}^\top\boldsymbol{x'}) \\ &= \exp(-\|\boldsymbol{x}\|^2)\exp(-\|\boldsymbol{x'}\|^2)\exp(2\boldsymbol{x}^\top\boldsymbol{x'})\end{aligned}$$

Puisque $\boldsymbol{x}^\top\boldsymbol{x'}$ est un noyau positif et la fonction exp est la limite d'un développement en série à coefficients positifs, $\exp(2\boldsymbol{x}^\top\boldsymbol{x'})$ est aussi un noyau positif. La fonction $\exp(-\|\boldsymbol{x}\|^2)\exp(-\|\boldsymbol{x'}\|^2)$ étant séparable c'est un noyau positif, ce qui permet de conclure sur la positivité du noyau gaussien, le produit de deux noyaux positifs étant un noyau positif.

Exemples de noyaux

Les noyaux positifs se divisent en deux grandes familles principales : les noyaux radiaux qui dépendent d'une distance et les noyaux projectifs qui sont définis à partir d'un produit scalaire.

type	nom	$k(s,t)$
radial	gaussien	$\exp\left(-\frac{r^2}{\sigma}\right),\ r=\|s-t\|$
radial	laplacien	$\exp(-r/\sigma)$
radial	rationnel	$1-\frac{r^2}{r^2+\sigma}$
radial	loc. gauss.	$\max\left(0,1-\frac{r}{3\sigma}\right)^p \exp(-\frac{r^2}{\sigma})$
non stationnaire	χ^2	$\exp(-r/\sigma),\ r=\sum_k \frac{(s_k-t_k)^2}{s_k+t_k}$
projectif	polynomial	$(s^\top t)^p$
projectif	affine	$(s^\top t+\sigma)^p$
projectif	cosinus	$s^\top t/\|s\|\|t\|$
projectif	correlation	$\exp\left(\frac{s^\top t}{\|s\|\|t\|}-\sigma\right)$

TABLEAU 1 : Exemples de noyaux pour $\mathcal{X}=\mathbb{R}^p$.

La plus part de ces noyaux dépendent d'un paramètre σ appelé « largeur de bande » dont le réglage est souvent critique pour le bon fonctionnement de la méthode.

La construction de noyaux sur des structures complexes donne lieu à de nombreux développement « d'ingénierie de noyaux » spécifique aux domaines d'applications concernés qui utilisent les techniques de construction présentées ci-dessus.

4.2 Cadre fonctionnel des noyaux

Etant donné un noyau k, il est possible de représenter des observations $\boldsymbol{x}$ par une fonction définie par la transformation $\Phi(\boldsymbol{x}) = k(x,\bullet)$ (appelée *kernel map*). Il se trouve alors que l'espace $\mathcal{H}$ dans lequel on plonge les observations peut être complètement défini ainsi que sa topologie par le noyau considéré ce qui présente de nombreux avantages. Plus précisément, $\mathcal{H}$ admet une structure d'espace de Hilbert à noyau reproduisant (EHNR).

Définition 4.3 (Espace de Hilbert à noyau reproduisant) *Soit $\mathcal{H}$ un espace de Hilbert muni de son produit scalaire $\langle\bullet,\bullet\rangle_{\mathcal{H}}$. $\mathcal{H}$ est un EHNR si , $\forall\boldsymbol{x}\in\mathcal{X}$ fixé, il existe une fonction $k(\boldsymbol{x},\bullet)\in\mathcal{H}$ symétrique telle que :*

$$\forall f\in\mathcal{H},\quad f(\boldsymbol{x}) = \langle f, k(\boldsymbol{x},\bullet)\rangle_{\mathcal{H}}$$

Le noyau joue ici le rôle de fonctionnelle d'évaluation et permet d'accéder à la valeur de la fonction en tout point $\boldsymbol{x}$.

Théorème 4.4 (EHNR et noyaux positifs) *A tout noyau positif on peut associer un Espace de Hilbert à noyau reproduisant et réciproquement.*

C'est ce théorème qui permet d'affirmer que choisir de travailler avec un noyau k consiste à rechercher f la partie fonctionnelle de la solution dans le EHNR associé à ce noyaux. Ce cadre fonctionnel présente certains avantages :

- l'universalité puisque, pour certaines familles de noyaux, toute fonction continue peut être approchée arbitrairement par un élément de l'EHNR,
- une certaine « confiance » puisque la convergence des fonctions implique la convergence ponctuelle car $\forall\boldsymbol{x}\in\mathcal{X}$:

$$\forall f, f_n\in\mathcal{H},\quad |f(\boldsymbol{x})-f_n(\boldsymbol{x})| \le \|k(\boldsymbol{x},.)\|_{\mathcal{H}}\ ||f-f_n||_{\mathcal{H}}$$

- la régularité des fonctions est contrôlée par leur norme

$$|f(\boldsymbol{x})-f(\boldsymbol{x'})| \le \|f\|_{\mathcal{H}}\ ||k(\boldsymbol{x},.)-k(\boldsymbol{x'},.)||_{\mathcal{H}}$$

- la possibilité de représenter les fonction comme une combinaison linéaire d'exemples (équation 1),
- la possibilité de construire des noyaux à travers la définition d'un EHNR.

4.3 Les SVM : le cas général

Définition 4.5 (SVM (L1)) *Soit $\{(\boldsymbol{x}_i, y_i);\ i = 1, n\}$ un ensemble de vecteurs formes étiquetées avec $\boldsymbol{x}_i \in \mathcal{X}$ et $y_i \in \{1, -1\}$. Soit $\mathcal{H}$ un EHNR de noyau k. Un séparateur à vaste marge (SVM) est un discriminateur de la forme : $D(x) = signe\big(f(\boldsymbol{x}) + b\big)$ où $f \in \mathcal{H}$ et $b \in \mathbb{R}$ sont donnés par la résolution du problème suivant pour un $C \geq 0$ donné :*

$$(SVML1) \quad \left\{ \begin{array}{lll} \min\limits_{f\in\mathcal{H}, b\in\mathbb{R}} & \frac{1}{2}\|f\|^2_{\mathcal{H}} + C\sum_{i=1}^n \xi_i & \\ avec & y_i(f(\boldsymbol{x}_i) + b) \geq 1 - \xi_i & i = 1, n \\ & 0 \leq \xi_i & i = 1, n \end{array} \right.$$

Théorème 4.6 (Solution des SVM) *La partie fonctionnelle de la solution du problème des SVM s'écrit comme une combinaison linéaire des fonctions noyau pris aux points supports :*

$$f(\boldsymbol{x}) = \sum_{i\in\mathcal{A}} \alpha_i y_i k(\boldsymbol{x}, \boldsymbol{x}_i) \tag{1}$$

où $\mathcal{A}$ désigne l'ensemble des contraintes actives et les α_i les solutions du programme quadratique suivant :

$$(SVMdual) \quad \left\{ \begin{array}{lll} \min\limits_{\boldsymbol{\alpha}\in\mathbb{R}^n} & \frac{1}{2}\boldsymbol{\alpha}^\top G\boldsymbol{\alpha} - \boldsymbol{e}^\top\boldsymbol{\alpha} & \\ avec & \boldsymbol{y}^\top\boldsymbol{\alpha} = 0 & \\ & 0 \leq \alpha_i \leq C & i = 1, n \end{array} \right. \tag{2}$$

où G est la matrice $n \times n$ de terme général $G_{ij} = y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Le biais b à la valeur du multiplicateur de Lagrange de la contrainte d'égalité à l'optimum.

La démonstration de ce théorème reprend la démarche adoptée dans le cas linéaire puisque le lagrangien s'écrit :

$$\mathcal{L} = \tfrac{1}{2}\|f\|^2_{\mathcal{H}} + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i\big(y_i(f(\boldsymbol{x}_i) + b) - 1 + \xi_i\big) - \sum_{i=1}^n \beta_i\xi_i$$

où les $\alpha_i \geq 0$ et les $\beta_i \geq 0$ sont les multiplicateurs de Lagrange associés aux contraintes. Les conditions d'optimalité de Karush, Kuhn et Tucker du problème qui permettent de caractériser la solution du problème primal (f^*, b^*, ξ^*) et les multiplicateurs de lagrange $\boldsymbol{\alpha}^*$ et $\boldsymbol{\beta}^*$ associés s'écrivent :

$$\begin{array}{rrll} \text{stationarité} & f^*(\bullet) - \sum_{i=1}^n \alpha_i^* y_i k(\boldsymbol{x}_i, \bullet) & = 0 & \\ & \sum_{i=1}^n \alpha_i^*\ y_i & = 0 & \\ & C - \alpha_i^* - \beta_i^* & = 0 & i = 1, n \\ \text{complementarité} & \alpha_i^*\big(y_i(f^*(\boldsymbol{x}_i) + b^*) - 1 + \xi_i^*\big) & = 0 & i = 1, n \\ & \beta_i^*\xi_i^* & = 0 & i = 1, n \\ \text{admissibilité primale} & y_i(f^*(\boldsymbol{x}_i) + b^*) + \xi_i^* & \geq 1 & i = 1, n \\ & \xi_i^* & \geq 0 & i = 1, n \\ \text{admissibilité duale} & \alpha_i^* & \geq 0 & i = 1, n \\ & \beta_i^* & \geq 0 & i = 1, n \end{array}$$

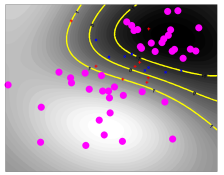
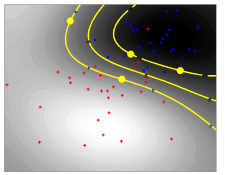
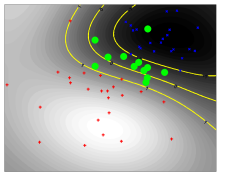
Les conditions de complémentarité $\beta_i^*\xi_i^* = 0$ permettent de définir l'ensemble I_C des indices pour lesquels $\beta_i^* = 0$. Pour cet ensemble, la troisième condition de stationnarité implique $\alpha_i = C$. Pour les indices n'appartenant pas à I_C, β_i^* peut être éliminé et cette même condition de stationnarité couplée aux conditions d'admissibilité duale stipule que $0 \leq \alpha_i \leq C$. Une fois les β_i éliminés, le système d'équation se traite comme dans le cas linéaire et conduit au problème dual (2).

Les contraintes initiales (donc les exemples) sont maintenant réparties dans trois ensembles illustrés par la figure 6 : I_C quand $\alpha_i = C$, I_0 pour les $\alpha_i = 0$ et I_α lorsque $0 \leq \alpha_i \leq C$, et l'ensemble des contraintes actives $\mathcal{A} = I_\alpha \cup I_C$.

La plus part des variantes présentées dans le cas linéaire s'étend au cas non linéaire en remplaçant le vecteur $\boldsymbol{w}$ par la fonction f :

linéaire	non linéaire
$\boldsymbol{x}$	$k(\bullet, \boldsymbol{x})$
$\|\boldsymbol{w}\|$	$\|f\|_{\mathcal{H}}$
$\boldsymbol{w}^\top\boldsymbol{x}_i$	$f(\boldsymbol{x}_i) = \langle f, k(\bullet, \boldsymbol{x}_i)\rangle_{\mathcal{H}}$

Ce mécanisme très général et s'adapte à de nombreuses autres méthodes linéaires comme la régression linéaire, l'analyse en composantes principale ou

données inutiles bien classées	données importantes support	données suspectes
$I_0\ ,\ \ \alpha_i = 0$	$I_\alpha\ ,\ \ 0 < \alpha_i < C$	$I_C\ ,\ \ \alpha_i = C$
$y_i\big(f(\boldsymbol{x}_i) + b\big) > 1$	$y_i\big(f(\boldsymbol{x}_i) + b\big) = 1$	$y_i\big(f(\boldsymbol{x}_i) + b\big) < 1$

FIGURE 6 – Exemple de répartition des observations en trois ensembles. Les observations « inutiles » sont en violet sur la figure de gauche, les points supports sont en jaune sur la figure du milieu et les points mal classés sont en vert sur la figure de droite.

le filtrage linéaire. L'extension est plus délicate dans le cas de norme 1 ou 0. Le recours au noyau passe alors par l'utilisation d'un dictionnaire de fonctions.

5 Autour des SVM

5.1 SVM, probabilités et convergence

Il est parfois utile d'associer des probabilités aux SVM. Pour ce faire, il est possible de s'appuyer sur le fait que la fonction de discrimination idéale $\log \frac{\mathbb{P}(Y=1|\boldsymbol{x})}{\mathbb{P}(Y=-1|\boldsymbol{x})}$ devrait avoir le même signe que la fonction de discrimination $f(\boldsymbol{x}) + b$. En posant :

$$\log \frac{\widehat{\mathbb{P}}(Y=1|\boldsymbol{x})}{\widehat{\mathbb{P}}(Y=-1|\boldsymbol{x})} = a_1 f(\boldsymbol{x}) + a_2$$

il est possible d'obtenir un estimateur de la probabilité à postériori de la forme :

$$\widehat{\mathbb{P}}(Y=1|\boldsymbol{x}) = 1 - \frac{1}{1+\exp^{a_1 f(\boldsymbol{x})+a_2}}$$

où a_1 et a_2 sont des paramètres estimés par maximum de vraisemblance.

L'estimation des probabilités a posteriori permet d'énoncer quelques faits à propos des SVM :

- dans leur forme générale, avec le noyau adéquat, les SVM sont universellement consistent (l'erreur de classification des SVM converge vers l'erreur de Bayes)
- avec les mêmes hypothèses, les SVM convergent vers la règle de bayes
- mais il n'est pas possible de construire un estimateur consistant des probabilités a posteriori du fait de la parcimonie),
- on peut en revanche donner un intervalle de confiance sur cette probabilité qui nécessite l'estimation d'un autre paramètre fonctionnel.

5.2 Les SVM multiclasses

L'adaptation des SVM bi classes au cas multiclasse peut se faire de trois façons différentes. Le choix va dépendre de la taille du problème.

1. L'approche *un contre tous* consiste à entrainer un SVM biclasse en utilisant les élément d'une classe contre tous les autres. Il s'agit de résoudre de l'ordre de c problèmes SVM chacun de taille n.
2. L'approche *un contre un* : consiste à entrainer $\frac{c(c-1)}{2}$ SVM sur chacun des couples de classes, puis à décider la classe gagnante soit par un vote majoritaire soit en post traitent les résultats grâce à l'estimation de probabilités a posteriori. Le nombre de classifieurs SVM à entraine peut être réduit en utilisant un codage astucieux pour les classes à travers un code correcteur d'erreur ou un graphe directe acyclique (DAGSVM).
3. L'approche *globale* consiste à traiter le problème en une seule fois. Cela peut se faire en posant formellement le problème, par exemple si l'on note $f_\ell(\boldsymbol{x}) - b_\ell$ la fonction de discrimination associée à la classe ℓ :

$$\left\{\begin{array}{ll} \min\limits_{\boldsymbol{f}\in\mathcal{H},\boldsymbol{b},\xi\in\mathbb{R}^n} & \frac{1}{2}\sum\limits_{\ell=1}^{c}\|f_\ell\|^2_{\mathcal{H}} + \frac{C}{d}\sum\limits_{i=1}^{n}\sum\limits_{\ell=1,\ell\neq y_i}^{c}\xi_{i\ell}^d \\ \text{avec} & y_i\big(f_{y_i}(\boldsymbol{x}_i) + b_{y_i} - f_\ell(\boldsymbol{x}_i) - b_\ell\big) \geq 1 - \xi_{i\ell} \\ & \xi_{i\ell} \geq 0 \text{ pour } i = 1,...,n;\ \ \ell = 1,...,c;\ \ \ell \neq y_i \end{array}\right.$$

Le dual de ce problème est un programme quadratique de même nature que celui des SVM et de taille $n \times c$. L'estimateur associé est non consistant mais donne de bons résultats en pratique. Il existe là aussi de nombreuses autres formulation (dont certaines consistantes).

5.3 Le choix du noyau : SVM à noyaux multiples

Le choix du noyau et de ses paramètres comme la largueur de bande est souvent un problème pratique et critique lors de la mise en œuvre des SVM. Une manière d'aborder cette question consiste à utiliser des noyaux multiples. L'idée est d'optimiser le choix du noyau parmi une collection possible. Ainsi, si l'on note $k_m, m \in [1, M]$ une famille de M noyaux positifs, cette optimisation s'effectue à travers le réglage de M coefficients de sorte que le noyau choisi s'exprime comme une combinaison positive des noyaux du dictionnaire $k = \sum_m d_m k_m$, $0 \leq d_m, m \in [1, M]$. En imposant en plus une contrainte de type $\sum_m d_m = 1$ favorisant la parcimonie, la solution optimale va chercher à sélectionner les noyaux les plus utiles. En notant $\mathcal{H}_m$ l'EHNR associé au noyau k_m et $f_m \in \mathcal{H}_m$ une fonction de cet EHNR, le problème s'écrit :

$$(MKL) \quad \left\{ \begin{array}{lll} \min\limits_{\{f_m\},b,\xi,d} & \frac{1}{2}\sum_m \frac{1}{d_m}\|f_m\|^2_{\mathcal{H}_m} + C\sum_i \xi_i & \\ \text{avec} & y_i\Big(\sum_m f_m(\boldsymbol{x}_i) + b\Big) \geq 1 - \xi_i & i = 1, n \\ & 0 \leq \xi_i, & i = 1, n \\ & \sum_m d_m = 1, \quad 0 \leq d_m, & m = 1, M \end{array} \right.$$

Ce problème peut se résoudre itérativement en utilisant une méthode d'optimisation alterné. Pour des $d_m, m = 1, M$ fixées, les $f_m, m = 1, M$ et b sont la solution d'un problème de SVM L1 dans le cas général. Lorsque les $f_m, m = 1, M$ et b sont fixés, les d_m peuvent être optimisée via une itération de gradient réduit. Lorsqu'un d_m est suffisamment petit, le noyau associé est éliminé.

5.4 Règlage du C : chemin de régularisation

Outre le choix du noyau, l'autre hyperparamètre à régler dans les SVM est le terme d'équilibrage C. Une méthode classique mais couteuse consiste à faire de la validation croisée sur une grille prédéfinie de valeurs de C candidates. Il existe une alternative particulièrement élégante : la technique du chemin de régularisation qui permet de calculer, pour un cout algorithmique analogue à celui du calcul d'une seule solution, un ensemble de solutions possibles pour un ensemble de valeurs de C remarquables.

En effet, le formulation duale des SVM est un programme quadratique paramétrique en fonction de C. Connaissant une solution pour un C_0 donnée, on peut montrer que cette solution varie linéairement lorsque C varie dans un voisinage de C_0. Afin d'illustrer ce fait et pour simplifier les calculs on considère le problème sans biais en posant $\lambda_0 = \frac{1}{C_0}$. La fonction cout à minimiser s'écrit :

$$J(f, \lambda_0) \quad \sum_{i=1}^{n} h\big(y_i f(\boldsymbol{x}_i)\big) + \frac{\lambda_o}{2}\|f\|^2_{\mathcal{H}}$$

où $h(z) = \max(1 - z, 0)$ désigne la fonction charnière représentée figure 4. La solution f_o de ce problème est caractérisée par $0 \in \partial_f J(f_o)$ où $\partial_f J(f_o)$ désigne la sous différentielle du cout au point f_o. La sous différentielle de J par rapport à f dépend des trois ensembles d'indices I_0, I_α et I_1 et s'écrit :

$$\partial_f J(f, \lambda_0) = \{\sum_{i \in I_\alpha} \alpha_i y_i K(\boldsymbol{x}_i, \bullet) - \sum_{i \in I_1} y_i K(\boldsymbol{x}_i, \bullet) + \lambda_0\ f(\bullet)\}$$

avec $\alpha_i \in]-1, 0[$, la sous différentielle de la fonction $h(z)$ au point $z = 1$.

On considère maintenant λ_n un élément du voisinage de λ_0 pour lequel les ensembles I_0, I_α and I_C demeurent inchangés à l'optimum. En particulier aux points $\boldsymbol{x}_j, j \in I_\alpha$ $(f_o(\boldsymbol{x}_j) = f_n(\boldsymbol{x}_j) = y_j)$: $\partial_f J(f)(\boldsymbol{x}_j) = 0$. En écrivant ces équations puis en en faisant la différence on obtient :

$$\begin{array}{rcll} \sum_{i \in I_\alpha} \alpha_{io} y_i K(\boldsymbol{x}_i, \boldsymbol{x}_j) & = & \sum_{i \in I_1} y_i K(\boldsymbol{x}_i, \boldsymbol{x}_j) - \lambda_o\ y_j & \\ \sum_{i \in I_\alpha} \alpha_{in} y_i K(\boldsymbol{x}_i, \boldsymbol{x}_j) & = & \sum_{i \in I_1} y_i K(\boldsymbol{x}_i, \boldsymbol{x}_j) - \lambda_n\ y_j & \\ \hline G(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_o) & = & (\lambda_o - \lambda_n)\boldsymbol{y} & \text{avec } G_{ij} = y_i K(\boldsymbol{x}_i, \boldsymbol{x}_j) \end{array}$$

Cette dernière équation nous permet de faire apparaître l'évolution linéaire de la solution lorsque λ varie puisque :

$$\boldsymbol{\alpha}_n = \boldsymbol{\alpha}_o + (\lambda_o - \lambda_n)\boldsymbol{d} \qquad \text{avec } \boldsymbol{d} = G^{-1}\boldsymbol{y}$$

Pour une valeur critique de λ les ensembles I_0, I_α et I_1 changent, la matrice G change aussi ainsi que le vecteur $\boldsymbol{d}$.

L'ensemble de toutes les valeurs critiques de λ et les solutions α associées forment un chemin de régularisation que l'on peut initialiser pour des valeurs très grandes de λ avec tous les points lorsque les deux classes sont équilibrées et qui évolue lorsque λ décroît en éliminant ou échangeant des exemples.

6 D'autres machines à noyau

Les principes des SVM ont été appliqué à de nombreux autres problèmes et ont permis la création d'autres machines à noyau. Ces principes requiérent :

- un problème initialement linéaire,
- l'utilisation de noyaux et la formulation initiale du problème dans un EHNR comme un problème de minimisation d'une norme,
- un théorème de représentation qui permet de représenter la fonction solution comme une combinaison linéaire des observations vues à travers le noyau,
- la résolution du problème dans le dual qui donne un programme quadratique de taille n (le nombre d'exemples),
- la parcimonie de la solution qui sélectionne les exemples importants.

Nous avons choisi les SVDD pour faire le lien avec la géométrie algorithmique, la régression pour le lien avec les splines et l'ordonnancement.

6.1 La plus petite boule englobante

Dans certains cas les étiquettes y_i ne sont pas disponibles. Etant donné un ensemble d'observations $\boldsymbol{x}_i, i \in [1, n]$ i.i.d. selon une loi $\mathbb{P}$. Il est intéressant de décrire ces données. Le problème de *détection de nouveauté* consiste à décider si une nouvelle observation $\boldsymbol{x}$ est proche ou non de cet ensemble. Une manière d'aborder ce problème consiste à rechercher une frontière de décision sous la forme de la plus petite boule englobante (*minimum enclosing ball*) définie par son centre f et son rayon R. C'est un problème ancien déjà traité dans le plan par Sylvester en 1857. Dans le cadre des EHNR le problème est aussi connu sous le nom de *suport vector data description* (SVDD) :

$$(SVDD) \quad \left\{ \begin{array}{ll} \min\limits_{f\in\mathcal{H}, b\in\mathbb{R}, \xi} & R^2 + C\sum_{i=1}^n \xi_i \\ \text{avec} & \frac{1}{2}\|k(\bullet, \boldsymbol{x}_i) - f\|^2_{\mathcal{H}} \leq R^2 + \xi_i \quad 0 \leq \xi_i \quad i = 1, n \end{array} \right.$$

Le paramètre $C = \frac{1}{\nu n}$ permet de régler la proportion ν de points que l'on désire maintenir en dehors de la boule (*outliers*). Le dual de ce problème est le programme quadratique suivant analogue à celui des SVM :

$$(SVDD \text{ dual}) \quad \left\{ \begin{array}{ll} \min\limits_{\boldsymbol{\alpha}\in\mathbb{R}^n} & \frac{1}{2}\boldsymbol{\alpha}^\top K\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^\top diag(K) \\ \text{avec} & \boldsymbol{e}^\top\boldsymbol{\alpha} = 1 \\ & 0 \leq \alpha_i \leq C \qquad\qquad i = 1, n \end{array} \right.$$

où K est la matrice de gram $n \times n$ de terme général $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. La figure 7 illustre le fonctionnement de cette méthode.

Remarques :

- la solution de ce problème est liée à la loi parente de l'échantillon puisqu'elle converge dans certains cas vers la frontière de l'ensemble de mesure minimal $\min_E$ vol$(E)\{E \subset \mathcal{X} \mid \mathbb{P}(E) = \nu\}$.
- une autre manière presque équivalent de poser le problème est le *one class SVM* dans le quel on cherche à maximiser ρ la valeur minimale de f :

$$(OCSVM) \quad \left\{ \begin{array}{ll} \min\limits_{f\in\mathcal{H}, b\in\mathbb{R}, \xi} & \frac{1}{2}\|f\|^2_{\mathcal{H}} - \rho + C\sum_{i=1}^n \xi_i \\ \text{avec} & f(\boldsymbol{x}_i) \geq \rho - \xi_i \qquad\qquad 0 \leq \xi_i \quad i = 1, n \end{array} \right.$$

  dont le dual est le même que celui des SVDD avec le terme linéaire de la fonction cout en moins.
- comme dans le cas des SVM, le programme quadratique est paramétrique et il existe une chemin de régularisation linéaire par morceau permettant d'estimer un ensemble de courbes de niveau de la mesure de probabilité à un moindre cout.

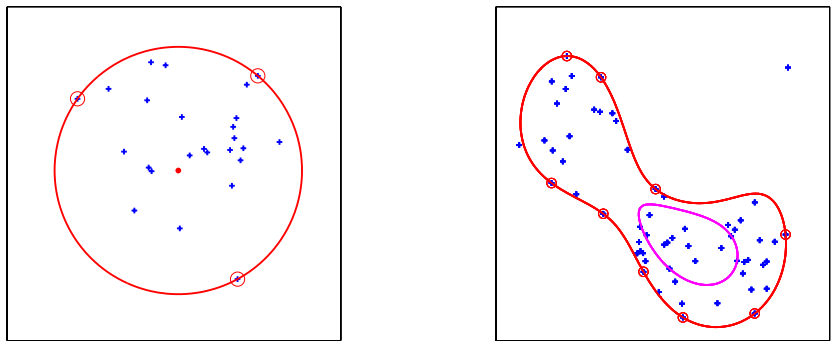


FIGURE 7 – Exemple de SVDD linéaire (à gauche) et pour un noyau gaussien (à droite). Les points entourés de rouge sont des vecteurs support. Sur la figure de droite, la solution à été calculée pour deux valeurs de C (en rouge pour 10% d'*outliers* et en magenta pour la courbe de niveau à $\nu = 0,8$). On constate que le point en haut à droite est placé en dehors de l'enveloppe calculée.

6.2 SVM et regression

Les SVM peuvent également être mis en oeuvre en situation de régression, c'est-à-dire pour l'approximation de fonctions quand Y est quantitative. Dans le cas non linéaire, le principe consiste à rechercher une estimation de la fonction par sa décomposition sur une base fonctionnelle. Si l'on se considère un terme d'attache aux données de type « moindres carrés » pénalisés, la formulation avec contraintes dans l'EHNR $\mathcal{H}$ s'écrit :

$$\left\{ \begin{array}{ll} \min\limits_{f\in\mathcal{H}} & \frac{1}{2}\|f\|^2_{\mathcal{H}}+\frac{1}{2\lambda}\sum\limits_{i=1}^{n}\xi_i^2 \\ \text{avec} & f(x_i)=y_i+\xi_i, \qquad i=1,n \end{array} \right.$$

La solution prend la forme d'une fonction *spline*. C'est une combinaison linéaires des noyaux (comme dans le cas des SVM équation 1). Les coefficients $\boldsymbol{\alpha}$ sont les variables du problème dual données par la résolution du système linéaire suivant :

$$(K+\lambda I)\boldsymbol{\alpha}=\boldsymbol{y}$$

Dans cette formulation (qui peut être reprise pour la classification et l'estimation de densité) *toutes* les observations sont vecteur support ($\mathcal{A}=\{1,\ldots,n\}$).

Afin d'obtenir de la parcimonie (et espérer un petit nombre de vecteurs support), il faut modifier le terme d'attache aux données. Parmi toutes les solutions envisageables, celle retenue dans le cadre de la *support vector regression*, est une fonction de déviation en valeur absolue tronquée (appelé aussi le cout t-insensible). Le problème avec contraintes s'écrit :

$$(SVR) \quad \left\{ \begin{array}{ll} \min\limits_{f\in\mathcal{H},b\in\mathbb{R}} & \frac{1}{2}\|f\|^2_{\mathcal{H}}+C\sum\xi_i \\ \text{avec} & |f(\boldsymbol{x}_i)+b-y_i|\leq t+\xi_i, \qquad 0\leq\xi_i, \qquad i=1,n \end{array} \right.$$

et la formulation équivalente en terme de cout pénalisé est :

$$\min_{f\in\mathcal{H},b\in\mathbb{R}} \frac{1}{2}\|f\|^2_{\mathcal{H}}+C\sum_{i=1}^{n}\max\big(0,|f(\boldsymbol{x}_i)+b-y_i|-t\big)$$

Le dual est un programme quadratique bi-paramétrique et il existe deux chemins de régularisation linéaires, en fonction des paramètres C et t la largeur de la zone insensible considérée.

Dans cette situation, les noyaux k utilisés sont ceux naturellement associés à la définition de bases de fonctions. Noyaux de splines, noyaux d'ondelettes ou encore noyau de Dirichlet associé à un développement en série de Fourier sont des grands classiques. Ils expriment les produits scalaires des fonctions de la base.

6.3 SVM et ordonnancement

Il arrive que les étiquettes disponibles y_{ij} expriment une préférence de l'observation $\boldsymbol{x}_i$ sur l'observation $\boldsymbol{x}_j$. On définit alors un ensemble $\mathcal{P}$ de couples d'indices i,j pour tous les $\boldsymbol{x}_i$ préférables à $\boldsymbol{x}_j$. La tâche associée consiste à construire une fonction score (de préférence) f permettant d'ordonner des observations candidates. Ce problème est important dans le cadre de la recherche d'information et de la recommandation. Dans le cadre des machines à noyaux, l'apprentissage de cette fonction f peut se formuler de la manière suivante :

$$(rankSVM) \quad \left\{ \begin{array}{ll} \min\limits_{f\in\mathcal{H}} & \frac{1}{2}\|f\|^2_{\mathcal{H}}+\frac{C}{d}\sum_{i,j\in\mathcal{P}}\xi_{ij}^d \\ \text{avec} & f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j)\geq 1-\xi_{ij}, \qquad 0\leq\xi_{ij}, \quad i,j\in\mathcal{P} \end{array} \right.$$

Pour le dual de ce problème non linéaire, les $\text{card}(\mathcal{P})\leq n^2$ contraintes vont correspondre à autant de variables ce qui peut s'avérer trop. Il est cependant toujours possible de résoudre le problème linéaire dans le primal, c'est-à-dire rechercher le vecteur $\boldsymbol{w}$ tel que $f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j)=\boldsymbol{w}^\top(\boldsymbol{x}_i-\boldsymbol{x}_j)$. Dans ce cas, les contraintes s'écrivent matriciellement $\boldsymbol{w}^\top\Delta_f\geq\boldsymbol{e}-\boldsymbol{\xi}$ où Δ_f, la matrice des différences est de taille $n^2\times p$ se décompose $\Delta_f=AX$ avec X la matrice des observations $n\times p$ et A une matrice creuse avec au plus un couple de valeurs (1,-1) par ligne. Ce problème peut être lors résolu pour $d=2$ par un algorithme de Newton à partir du calcul du gradient $\boldsymbol{g}$ et de la matrice hessiène H du problème :

$$\begin{array}{ll} \boldsymbol{g} & =\boldsymbol{w}+C\,X^\top A_{\mathcal{A}}^\top A_{\mathcal{A}}X\boldsymbol{w} \\ H & =I+CX^\top A_{\mathcal{A}}^\top A_{\mathcal{A}}X \end{array}$$

où $\mathcal{A}$ désigne l'ensemble des paires qui violent les contraintes. Ainsi le calcul d'une direction de descente $\boldsymbol{d}=H^{-1}\boldsymbol{g}$ peut être approchée par la méthode de gradient conjugué qui requière le calcul du produit de la hessiène par un vecteur $\boldsymbol{v}$ qui ne nécessite pas la construction explicite de H puisque :

$$H\boldsymbol{v}=\boldsymbol{v}+CX^\top A_{\mathcal{A}}^\top A_{\mathcal{A}}X\boldsymbol{v}$$

Il existe de nombreuse extensions et adaptations. Citons l'utilisation de fonction cout mieux adaptés comme le gain cumulé normalisé (*normalized discounted cumulative gain NDCG@k*) pour $\boldsymbol{s} = X\boldsymbol{w}$ ordonnés et π une permutation de ces scores tronquées à k éléments :

$$DCG@k(\boldsymbol{s},\pi) = \sum_{i=1}^{k} \frac{2^{s_{\pi_i}} - 1}{log(i+2)} \quad \text{et} \quad NDCG@k(\boldsymbol{s},\pi) = \frac{DCG@k(\boldsymbol{s},\pi)}{DCG@k(\boldsymbol{s},\pi_s)}$$

où π_s désigne la permutation maximisant $DCG@k(\boldsymbol{s},\pi)$.

Pour plus de détails, le lecteur intéressé par se type de problème pourra consulter les résultats de la compétition *Yahoo Learning to Rank Challenge* [1].

7 Exemples de mise en œuvre

Même si les SVM s'appliquent à un problème de régression, nous n'illustrons que le cas plus classique de la discrimination.

7.1 Cancer du sein

La prévision de l'échantillon test par un Séparateur à Vaste marge conduit à la matrice de confusion :

```
ign malignant
  benign         83          1
  malignant       3         50
```

et donc une erreur estimée de 3%.

7.2 Concentration d'ozone

Un modèle élémentaire avec noyau par défaut (gaussien) et une pénalisation de 2 conduit à une erreur de prévision estimée à 12,0% sur l'échantillon test. La meilleure prévision de dépassement de seuil sur l'échantillon test initial est fournie par des SVM d'ε-régression. Le taux d'erreur est de 9,6% avec la matrice de confusion suivante :

```
           0    1
```

1. learningtorankchallenge.yahoo.com

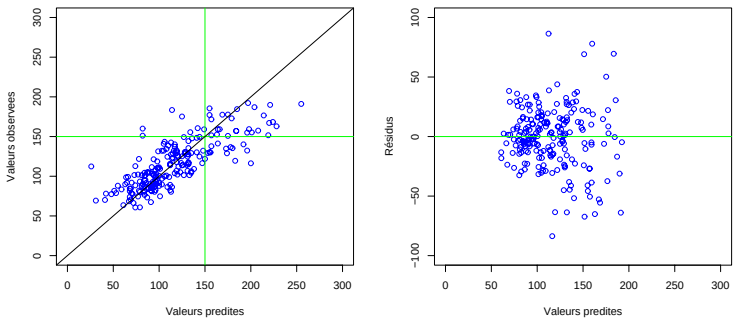


FIGURE 8 – Ozone : Valeurs observées et résidus en fonction des valeurs prédites pour l'échantillon test.

```
FALSE 161   13
TRUE    7   27
```

Ce résultat serait à confirmer avec des estimations sytématiques de l'erreur. Les graphiques de la figure 8 montre le bon comportement de ce prédicteur. Il souligne notamment l'effet "tunnel" de l'estimation qui accepte des erreurs autour de la diagonale pour se concentrer sur les observations plus éloignées donc plus difficiles à ajuster.

7.3 Carte Visa

Les données bancaires posent un problème car elles mixent variables quantitatives et qualitatives. Celles-ci nécessiteraient la construction de noyaux très spécifiques. Leur traitement par SVM n'est pas détaillé ici.