REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de L'Enseignement Supérieure et de la Recherche Scientifique

UNIVERSITE ABD EL HAMID IBN BADIS
−MOSTAGANEM−

Faculté des Sciences Exactes et d'Informatiques

Département de Mathématiques

PROJET DE FIN D'ETUDES EN VUE DE L'OBTENTION

DU DIPLOME DE *MASTER*

en
Mathématiques

Sujet

Machines à Vecteurs de Support et Applications

PRESENTE PAR :

Mlle. Naima DJELLOUL

ENCADRE PAR :

Mr. Abdessamad AMIR

2011/2012

Remerciements

Tout d'abord je remercie mon bon ***DIEU*** *qui m'a donné la force pour terminer ce travail. Je veux aussi remercier mes parents qui ont consacré du temps pour m'encourager pendant mes études.*

Je tiens vraiment à remercier très chaleureusement mon encadreur Mr Abdessamad AMIR pour son aide précieuse, ses conseils éclairés, et surtout pour sa patience. Je lui exprime mes profonde gratitude de m'avoir encadré et ma profonde reconnaissance pour la confiance qu'il m'a faite en me donnant la chance de travailler avec lui.

Mes remerciements vont à l'ensemble des membres du jury qui feront l'honneur de juger ce travail. Je remercie également tous ceux qui me connaissent de près ou de loin.

Dédicace

Je dédie ce modeste travail à :

-Ma famille : mes parents, mes frères et ma sœur.

-Mes collègues de la 1ère et 2ème année Master Mathématiques.

-Tout mes amis.

-Mes professeurs et enseignants de la 2ème année Master :

Mr.A. Amir, Mr.D. Bouagada

Mr.O. Belhamiti et Mr.Z. Dahmani.

- Et tout particulièrement à Mr.H. Ablaoui.

Table des matières

Chapitre 1

Introduction

Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais Support Vector Machine, SVM) sont une classe d'algorithmes basés sur la recherche de l'hyperplan de marge optimale qui, lorsque c'est possible, classe ou sépare correctement les données. Le principe est donc de trouver à partir d'un ensemble d'apprentissage un classifieur, ou une fonction classificatrice, dont la capacité de généralisation (qualité de prévision) est la plus grande possible. Les SVMs ont été développés dans les années 1990 à partir des considérations théoriques de Vladimir Vapnik sur le développement d'une théorie statistique de l'apprentissage. Les SVMs ont rapidement été adoptés pour leur capacité à travailler avec des données de grandes dimensions, le faible nombre d'hyper paramètres, leurs garanties théoriques, et leurs bons résultats en pratique. Les SVMs ont été appliqués à de très nombreux domaines (bioinformatique, recherche d'information, vision par ordinateur, finance...).

Le principe de base des SVMs consiste de ramener le problème de la classification à la recherche d'un hyperplan optimal. On a essayé dans ce mémoire de donnée une introduction aux SVMs via deux idées fondamentales. La première consiste à définir l'hyperplan comme solution d'un problème d'optimisation quadratique convexe sous contraintes linéaires 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. Nous analysons donc les principes de base de l'optimisation quadratique, le théorème classique de Karush-Khun et Tucker (K-K-T) donnera la formulation du problème dual très utile pour la résolution numérique des SVMs. Nous présenterons ensuite la deuxième idée qui nous permet de traiter le cas non linéairement séparable, le passage à la recherche de surfaces séparatrices non linéaires est obtenue 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.

Chapitre 2

La Programmation Quadratique

2.1 Conditions d'optimalités

Un programme non linéaire s'écrit d'une manière générale sous la forme :

$$\begin{cases} \min\ f\ (x), \\ x \in S. \end{cases} \tag{2.1}$$

$f\ :\ \mathbb{R}^n \longrightarrow \mathbb{R}$; une fonction differentiable, non nécessairement linéaire et S est un sous ensemble convexe de $\mathbb{R}^n$ appelé ensemble des solutions admissibles.

Le point $\overline{x}$ est une solution locale de ce problème s'il existe $\delta > 0$ tel que :

$$f\ (x) \geqslant f\ (\overline{x}),$$

pour tout x vérifiant

$$\|x - \overline{x}\| < \delta, \qquad x \in S.$$

avec $\|.\|$ est une norme de $\mathbb{R}^n$. Si $f\ (\overline{x}) \leqslant f\ (x)$, $\forall\ x \in S$, on dit que $\overline{x}$ est la solution globale du problème.

Exemple 2.1 *Soit le problème*

$$\begin{cases} \min (x_1 - 1)^2 - x_2^2 \\ x_1 \in \mathbb{R}, \\ -1 \leqslant x_2 \leqslant 2. \end{cases}$$

Ce problème admet deux solutions $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ *et* $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$. *La solution* $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$ *est la solution globale.*

Proposition 2.1 ([2]) *Soient S un ensemble convexe dans $\mathbb{R}^n$, et f une fonction continuement différentiable.*

(a) Si $\overline{x}$ est solution locale du problème (2.1). Alors,

$$\nabla f(\overline{x})^\top (x - \overline{x}) \geqslant 0,\ \forall x \in S. \tag{2.2}$$

(b) Si $\overline{x}$ satisfait (2.2) et f est convexe, alors $\overline{x}$ est la solution globale du problème.

Preuve. (a) Soit $\delta > 0$; la constante liée à l'optimalité locale. On choisi $x \in S$ arbitrairement comme suit

$$\begin{aligned} y(\lambda) &= (1-\lambda)\,\overline{x} + \lambda x, \\ &= \overline{x} + \lambda\,(x - \overline{x}), \quad \forall \lambda \in [0,1]. \end{aligned}$$

Comme S est convexe, alors $y(\lambda) \in S$, et aussi pour tout λ suffisamment petit, on a :

$$\|y(\lambda) - \overline{x}\| < \delta.$$

En utilisant l'optimalité locale

$$f(y(\lambda)) - f(\overline{x}) \geqslant 0,$$

pour tout λ suffisamment petit. En employant cette inégalité ainsi que le développement limité de la fonction f au voisinage de $\overline{x}$, on obtient :

$$\lambda \nabla f(\overline{x})^\top (x - \overline{x}) + o(\lambda) \geqslant 0,$$

tel que $\frac{o(\lambda)}{\lambda} \to 0$ quand $\lambda \to 0$. En divisant les deux membres de l'inéquation par $\lambda > 0$, on obtient (2.2).

(b) f est convexe et différentiable d'où (voir Appendice)

$$f(x) \geqslant f(\overline{x}) + \nabla f(\overline{x})^\top (x - \overline{x}), \quad \forall x \in S.$$

Comme $\overline{x}$ satisfait (2.2), on déduit que

$$f(x) \geqslant f(\overline{x}), \quad \forall x \in S.$$

Ce qui achève la démonstration. ■

Corollary 2.1 *Supposons que l'ensemble des solutions admissibles du problème est l'espace tout entier, c-à-d, $S = \mathbb{R}^n$ et f continuement différentiable, on a :*

(a) Si $f(\overline{x}) \leqslant f(x) \;\; \forall x \in \mathbb{R}^n$, alors $\nabla f(\overline{x}) = 0$.

(b) Si f est convexe et $\nabla f(\overline{x}) = 0$, alors $f(\overline{x}) \leqslant f(x)\; \forall x \in \mathbb{R}^n$.

Preuve. *(a) Comme $\overline{x}$ est une solution globale (par conséquent locale), d'après la proposition 2.1, on a*

$$\nabla f(\overline{x})^\top (x - \overline{x}) \geqslant 0, \quad \forall x \in \mathbb{R}^n.$$

En particulier pour $x = \overline{x} - \nabla f(\overline{x})$, d'où

$$\nabla f(\overline{x})^\top (x - \overline{x}) = -\|\nabla f(\overline{x})\|_2^2 \geqslant 0,$$

qui ne peut être vrai seulement si $\nabla f(\overline{x}) = 0$.

(b) Résulte immédiatement de la caractérisation d'une fonction convexe différentiable

$$f(x) \geqslant f(\overline{x}) + \nabla f(\overline{x})^\top (x - \overline{x}), \quad \forall x \in S.$$

■

Corollary 2.2 *Supposons que l'ensemble des solutions admissibles est l'orthant positif, c-à-d,* $S = \{x \in \mathbb{R}^n / x \geqslant 0\}$ *et* f *est continuement différentiable.*

(a) Si $\overline{x}$ *est la solution globale du problème (2.1), alors* $\nabla f(\overline{x}) \geqslant 0$, $\overline{x} \geqslant 0$ *et* $\nabla f(\overline{x})^\top \overline{x} = 0$.

(b) Si f *est convexe,* $\nabla f(\overline{x}) \geqslant 0$, $\overline{x} \geqslant 0$ *et* $\nabla f(\overline{x})^\top \overline{x} = 0$, *alors* $\overline{x}$ *est la solution globale du problème (2.1).*

Preuve. D'après la proposition 2.1, on a :

$$\nabla f(\overline{x})^\top (x - \overline{x}) \geqslant 0, \quad \forall x \geqslant 0. \tag{2.3}$$

Pour $x = 0$, on a $\nabla f(\overline{x})^\top \overline{x} \leqslant 0$ alors que pour $x = 2\overline{x}$, on a $\nabla f(\overline{x})^\top \overline{x} \geqslant 0$, par conséquent $\nabla f(\overline{x})^\top \overline{x} = 0$. Pour montrer que $\nabla f(\overline{x}) \geqslant 0$, raisonnons par absurde ; supposons qu'il existe un indice i telle que $[\ \nabla f(\overline{x})]_i < 0$. On définit $x \geqslant 0$ comme suit :

$$x_j = \overline{x}_j \qquad (j \neq i) \qquad\qquad x_i = \overline{x}_i - [\ \nabla f(\overline{x})]_i > 0.$$

En remplaçant dans l'inéquation (2.3), on obtient :

$$0 \leqslant \nabla f(\overline{x})^\top (x - \overline{x}) = \sum_{j=1}^{n} [\ \nabla f(\overline{x})]_j (x_j - \overline{x}_j) = -[\ \nabla f(\overline{x})]_i^2 < 0.$$

Contradiction. Par conséquent, il n'y a aucun indice i telle que $[\ \nabla f(\overline{x})]_i < 0$, et ainsi $\nabla f(\overline{x}) \geqslant 0$.

(b) La convexité de f implique que :

$$f(x) \geqslant f(\overline{x}) + \nabla f(\overline{x})^\top (x - \overline{x}), \quad \forall x \in S,$$

le second terme est non négatif par définition et donc $f(x) \geqslant f(\overline{x}),\ \forall\ x \geqslant 0.$ ■

2.2 La programmation quadratique

Un programme quadratique est un exemple de la programmation non linéaire, il s'écrit sous la forme suivante :

$$\begin{cases} \min\ f\ (x) = \frac{1}{2} x^\top Q x + p^\top x, \\ x \in S. \end{cases} \tag{2.4}$$

où $Q \in \mathbb{R}^{n \times n}$ est une matrice symétrique, et S est un sous ensemble polyédrique de $\mathbb{R}^n$, c-à-d un ensemble définit par un nombre fini de contraintes linéaires d'égalité et / ou d'inégalité. Le programme (2.4) est convexe si Q est semi définie positive (voir Appendice).

On donne le *Gradient* de la fonction f par :

$$\nabla f(x) = Qx + p,$$

et la matrice *Hessienne* par :

$$\nabla^2 f(x) = Q.$$

Exemple 2.2 *Soit*

$$f(x_1, x_2, x_3) = x_1 - x_3 + 3x_1^2 - 2x_1x_2 + x_2^2 - 2x_2x_3 + 4x_1x_3 + 4x_3^2.$$

Alors

$$\nabla f(x_1, x_2, x_3) = \begin{pmatrix} 1 + 4x_1 - 2x_2 + 4x_3 \\ -2x_1 + 2x_2 - 2x_3 \\ -1 - 2x_2 + 4x_1 + 8x_3 \end{pmatrix},$$

d'où

$$p = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix},$$

et

$$\nabla^2 f(x_1, x_2, x_3) = \begin{pmatrix} 6 & -2 & 4 \\ -2 & 2 & -2 \\ 4 & -2 & 8 \end{pmatrix}.$$

Par conséquent

$$f(x_1, x_2, x_3) = \frac{1}{2} \begin{pmatrix} x_1, & x_2, & x_3 \end{pmatrix} \begin{pmatrix} 6 & -2 & 4 \\ -2 & 2 & -2 \\ 4 & -2 & 8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 1, & 0, & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

2.3 Existence de la solution

Avant d'étudier la solution (ou les solutions) du problème (2.4), il faut s'assurer de son existence.

Définition 2.1 *Une fonction $f : S \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ est dite coercive si*

$$f(x) \longrightarrow +\infty \text{ quand } \|x\| \longrightarrow +\infty. \tag{2.5}$$

Exemple 2.3 *La fonction $x \longmapsto x^2$ est coercive alors que $x \longmapsto x^3$ ne l'est pas.*

Définition 2.2 *On dit que $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ est propre si*

$$f > -\infty \quad et \quad f \neq +\infty.$$

Théorème 2.1 ([1]) *Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ une fonction propre, continue et coercive. Alors il existe un point $\overline{x} \in \mathbb{R}^n$ tel que :*

$$f(\overline{x}) \leqslant f(x), \quad \forall x \in \mathbb{R}^n.$$

Preuve. Soit $d = \inf_{x\in\mathbb{R}^n} f(x) < +\infty$. Soit $(x_n)_{n\in\mathbb{N}}$ une suite minimisante c-à-d telle que :

$$\lim_{n\to+\infty} f(x_n) = d < +\infty. \tag{2.6}$$

Montrons que la suite $(x_n)_{n\in\mathbb{N}}$ est bornée. Supposons par absurde qu'elle ne l'est pas c-à- d qu'il existe une sous-suite notée $(x_{\varphi(n)})_n$ de $(x_n)_n$ telle que

$$lim_{n\to+\infty} \|x_{\varphi(n)}\| = +\infty.$$

Par coercivité de f, on a alors : $lim_{n\to+\infty} f(x_{\varphi(n)}) = +\infty$, ce qui contredit (2.6).La suite $(x_n)_{n\in\mathbb{N}}$ est donc bornée : il existe alors une suite extraite notée $(x_{\zeta(n)})_n$ de$(x_n)_n$, qui converge vers $\overline{x} \in \mathbb{R}^n$. En utilisant maintenant la continuité de f, on a alors :

$$f(\overline{x}) = \lim_{n\to+\infty} f(x_{\zeta(n)}) = d,$$

d'où le résultat. ■

2.3.1 Les conditions de K-K-T (Karush, Kuhn et Tucker)

On va considérer une représentation particulière de l'ensemble polyédrique S,

$$\begin{array}{rcl} S & = & \{x \ / \ Ax \geqslant b, \ x \geqslant 0\}, \\ A & \in & \mathbb{R}^{m\times n}, \ b \in \mathbb{R}^m. \end{array}$$

Le problème (2.4) devient :

$$\left\{ \begin{array}{c} \min \ f \ (x) = \frac{1}{2} x^\top Q x + p^\top x, \\ Ax \geqslant b, \\ x \geqslant 0. \end{array} \right. \tag{2.7}$$

La proposition 2.1 est essentielle pour établir les conditions de K-K-T pour la programmation quadratique. Ces conditions sont nécessaires pour le programme quadratique (2.7) ; c-à-d, n'importe quelle solution locale les satisfera. Elles sont suffisantes pour les programmes quadratiques convexes (Q semi définie positive) et x et les variables duales $u \in \mathbb{R}^m$ (appelés "multiplicateurs de Lagrange") satisfassent les conditions de K-K-T, alors x est une solution globale de (2.7).

Théorème 2.2 (conditions de K-K-T pour la programmation quadratique [2])
Si $\overline{x}$ est une solution locale du problème (2.7), alors il existe $\overline{u} \in \mathbb{R}^m$ tel que :

$$0 \leqslant \overline{x} \perp Q\overline{x} - A^\top \overline{u} + p \geqslant 0, \tag{2.8}$$

$$0 \leqslant \overline{u} \perp A\overline{x} - b \geqslant 0. \tag{2.9}$$

Réciproquement, si une paire $(\overline{x}, \overline{u}) \in \mathbb{R}^n \times \mathbb{R}^m$ satisfait (2.8) et (2.9) et Q est semi définie positive, alors $\overline{x}$ est la solution globale du problème (2.7).

Preuve. Comme $\overline{x}$ est une solution locale du problème (2.7), et d'après la proposition 2.1, $\overline{x}$ résout le problème linéare :

$$\begin{cases} \min\limits_{x} \nabla f(\overline{x})^\top x, \\ x \in S. \end{cases}$$

Ce dernier est un programme linéaire parce que S est un ensemble polyédrique et la fonction objectif est linéaire et est égal à

$$\begin{cases} \min\limits_{x} (Q\overline{x} + p)^\top x, \\ Ax \geqslant b, \\ x \geqslant 0. \end{cases} \tag{2.10}$$

Comme $\overline{x} \in S$, on a

$$\begin{aligned} A\overline{x} &\geqslant b, \\ \overline{x} &\geqslant 0. \end{aligned}$$

En appliquant le théorème de dualité forte d'un programme linéaire (voir Appendice) sur le dernier système, il existe $\overline{u} \in \mathbb{R}^m$ qui est dual admissible

$$\begin{aligned} A^\top \overline{u} &\leqslant Q\overline{x} + p, \\ \overline{u} &\geqslant 0. \end{aligned}$$

En outre, la condition de complémentarité donne

$$\overline{x}^\top (Q\overline{x} + p - A^\top \overline{u}) = 0 \quad et \quad \overline{u}^\top (A\overline{x} - b) = 0.$$

En combinant toutes les relations ci-dessus, on obtient (2.8) et (2.9). Par conséquent, on a établi la condition nécessaire de KKT.Réciproquement, supposons que $\overline{x}$ et $\overline{u}$ satisfassent (2.8) et (2.9), alors, d'après le théorème de complémentarité linéaire (voir Appendice) $\overline{x}$ résout (2.10), et par conséquent

$$\nabla f(\overline{x})^\top (x - \overline{x}) \geqslant 0, \ \forall x \in S.$$

Donc, d'après la proposition 2.1, $\overline{x}$ est la solution globale du problème (2.7). ■

2.3.2 La dualité

La programmation quadratique, en particulier dans le cas convexe admet une théorie riche qui etend un grand nombre de résultats comme en programmation linéaire. En première étape définissons la fonction *Lagrangienne*, qui est une combinaison de la fonction objectif et les contraintes du problème. Pour le problème (2.7), le Lagrangien :

$$\mathcal{L}(x, u, v) = \frac{1}{2} x^\top Q x + p^\top x + u^\top (Ax - b) - v^\top x,$$

tel que u et v sont les variables duales, désigné plus fréquemment sous le nom de multiplicateurs de Lagrange des contraintes $Ax \geqslant b$ et $x \geqslant 0$ respectivement.

Le dual du problème (2.7) est défini comme suit :

$$\begin{cases} \max\limits_{x,u,v} \mathcal{L}(x, u, v) \\ \nabla_x \mathcal{L}(x, u, v) = 0, \\ u \geqslant 0, \\ v \geqslant 0. \end{cases} \tag{2.11}$$

Théorème 2.3 (La dualité faible : programmation quadratique [2]) . *Soit Q une matrice semi définie positive et supposons que $\overline{x}$ est une solution admissible du problème (2.7) et $(\widehat{x}, \widehat{u}, \widehat{v})$ est admissible pour le problème (2.11). Alors*

$$\frac{1}{2}\overline{x}^\top Q\overline{x} + p^\top \overline{x} \geqslant \mathcal{L}(\widehat{x}, \widehat{u}, \widehat{v}).$$

Preuve. Comme $\overline{x}$ est primal admissible et $(\widehat{x}, \widehat{u}, \widehat{v})$ est dual admissible. Alors :

$$\begin{gathered} A\overline{x} \geqslant b, \\ \overline{x} \geqslant 0, \\ Q\widehat{x} - A^\top \widehat{u} + p - \widehat{v} = 0, \\ \widehat{u} \geqslant 0, \\ \widehat{v} \geqslant 0. \end{gathered}$$

En conséquence, on a :

$$\begin{aligned} &\tfrac{1}{2}\overline{x}^\top Q\overline{x} + p^\top \overline{x} - \mathcal{L}(\widehat{x}, \widehat{u}, \widehat{v}) \\ &\quad = \tfrac{1}{2}(\overline{x} - \widehat{x})^\top Q(\overline{x} - \widehat{x}) - \widehat{x}^\top Q\widehat{x} + \widehat{x}^\top Q\overline{x} + p^\top \overline{x} - p^\top \widehat{x} + \widehat{u}^\top (A\widehat{x} - b) + \widehat{v}^\top \widehat{x} \\ &\quad = \tfrac{1}{2}(\overline{x} - \widehat{x})^\top Q(\overline{x} - \widehat{x}) + \widehat{x}^\top Q\overline{x} + p^\top \overline{x} - \widehat{u}^\top b \\ &\quad \geqslant \widehat{x}^\top Q\overline{x} + p^\top \overline{x} - b^\top \widehat{u} \\ &\quad \geqslant \overline{x}^\top A^\top \widehat{u} - b^\top \widehat{u} \\ &\quad \geqslant 0. \end{aligned}$$

La première égalité est simplement une identité, la deuxième égalité découle de $\widehat{v} = Q\widehat{x} - A^\top \widehat{u} + p$, la première inégalité découle de fait que Q est semi définie positive, et la deuxième inégalité découle de fait que $\overline{x} \geqslant 0$ et $Q\widehat{x} - A^\top \widehat{u} + p \geqslant 0$. L'inégalité finale découle de fait que $\widehat{u} \geqslant 0$ et $A\overline{x} - b \geqslant 0$. ■

Théorème 2.4 (La dualité forte : programmation quadratique [2]) . *Soit Q une matrice semi définie positive. Si $\overline{x}$ est une solution de (2.7), alors ils existent $\overline{u} \in \mathbb{R}^m$ et $\overline{v} \in \mathbb{R}^n$ tels que $(\overline{x}, \overline{u}, \overline{v})$ résout le problème (2.11) et les deux problèmes sont égaux en valeur.*

Preuve. Comme $\overline{x}$ résout le problème (2.7), alors le théorème 2.2 assure l'existence de $\overline{u}$ tel que $(\overline{x}, \overline{u}, \overline{v})$ est dual admissible (où $\overline{v} = Q\overline{x} - A^\top \overline{u} + p$), en raison des conditions (2.8 et 2.9). Par conséquent :

$$\begin{aligned} \frac{1}{2}\overline{x}^\top Q\overline{x} + p^\top \overline{x} &\geqslant \mathcal{L}(\overline{x}, \overline{u}, \overline{v}) \\ &= \frac{1}{2}\overline{x}^\top Q\overline{x} + p^\top \overline{x} - \overline{u}^\top (A\overline{x} - b) - \overline{x}^\top (Q\overline{x} - A^\top \overline{u} + p) \\ &= \frac{1}{2}\overline{x}^\top Q\overline{x} + p^\top \overline{x}. \end{aligned}$$

L'inégalité découle du théorème de la dualité faible (théorème 2.3), la première égalité est une identité simple, et la deuxième égalité découle des conditions de complémentarité de (2.8 et 2.9). On conclu que la valeur de la fonction objectif duale $\mathcal{L}(\overline{x}, \overline{u}, \overline{v})$ au point admissible dual $(\overline{x}, \overline{u}, \overline{v})$ est égale à sa borne supérieure $\frac{1}{2}\overline{x}^\top Q\overline{x} + p^\top \overline{x}$, et donc $(\overline{x}, \overline{u}, \overline{v})$ doit résoudre le problème (2.11). ■

Dans ce paragraphe, on s'intéresse a calculer la distance entre un point et sa projection sur un hyperplan donné, le résultat est une conséquence du théorème de K-K-T. Il est important de signaler que l'étude sera faite avec une norme quelconque.

2.4 Projection sur un hyperplan avec une norme arbitraire

Définition 2.3 *Soit $\|.\|$ une norme quelconque de $\mathbb{R}^n$, alors sa norme duale est définit par :*

$$\|x\|' = \max_{\|y\|=1} x^\top y.$$

L'inégalité généralisée de Cauchy Schwarz est :

$$\pm x^\top y \leqslant \mid x^\top y \mid \leqslant \|x\|'\|y\|.$$

Comme exemple, $\forall\ \alpha\ ,\beta \in [1,\infty]$ et $\frac{1}{\alpha}+\frac{1}{\beta}=1$, alors la norme α et la norme β sont des normes duales d'aprés l'inégalité classique de Hölder.

Théorème 2.5 ([6]) *Soit $q \in \mathbb{R}^n$, qui ne vérifie pas l'équation de l'hyperplan*

$$P = \{x \mid w^\top x = \gamma\},\ 0 \neq w \in \mathbb{R}^n,\ \gamma \in \mathbb{R}.$$

Alors, la projection $p(q) \in P$ employant une norme quelconque $\|.\|$ de $\mathbb{R}^n$ est donnée par :

$$p(q) = q - \frac{w^\top q - \gamma}{\|w\|'} y(w), \tag{2.12}$$

où $\|.\|'$ est la norme duale de $\|.\|$ et :

$$y(w) \in \arg\max_{\|y\|=1} w^\top y.$$

Par conséquent, la distance entre q et sa projection $p(q)$ est :

$$\|q - p(q)\| = \frac{\mid w^\top q - \gamma \mid}{\|w\|'}. \tag{2.13}$$

Preuve. On doit montrer que $p(q) \in P$ et qu'il réalise le minimum du problème

$$\begin{cases} \min\limits_x \|q - x\| \\ x \in P. \end{cases}$$

ce qui est équivalent à satisfaire la condition suffisante d'optimalité de K-K-T pour un certain multiplicateur de Lagrange $\lambda \in \mathbb{R}$:

$$\|q - p(q)\| \leqslant \|x - q\| - \lambda(w^\top x - \lambda),\ \forall x \in \mathbb{R}^n. \tag{2.14}$$

$p(q) \in P$, en effet

$$w^\top p(q) - \gamma = w^\top q + \frac{\gamma - w^\top q}{\|w\|'} w^\top y(w) - \gamma = w^\top q + \frac{\gamma - w^\top q}{\|w\|'}\|w\|' - \gamma = 0. \tag{2.15}$$

Par conséquent $p(q) \in P$. Pour prouver (2.14), définissons

$$\lambda = \frac{1}{\|w\|'} \frac{\mid \gamma - w^\top q \mid}{(\gamma - w^\top q)}. \tag{2.16}$$

Ainsi on doit montrer que :

$$\|\frac{\gamma - w^\top q}{\|w\|'} y(w)\| \leqslant \|x - q\| - \frac{1}{\|w\|'} \frac{\mid \gamma - w^\top q \mid}{\gamma - w^\top q}.(w^\top x - \gamma),\ \forall x \in \mathbb{R}^n, \tag{2.17}$$

ou d'une manière équivalente :

$$\mid \gamma - w^\top q \mid .\|y(w)\| + (w^\top x - \gamma)\frac{\mid \gamma - w^\top q \mid}{(\gamma - w^\top q)} \leqslant \|w\|'.\|x - q\|,\ \forall x \in \mathbb{R}^n. \tag{2.18}$$

Comme $\|y(w)\| = 1$, l'inégalité (2.18) est équivalente à :

$$\pm w^\top (x - q) \leqslant \|w\|'.\|x - q\|,\ \forall x \in \mathbb{R}^n, \tag{2.19}$$

ce qui découle immédiatement de la généralisation de l'inégalité de Cauchy-Schwarz ou d'une manière équivalente de la définition de la norme duale. Par conséquent, (2.14) est réalisable et d'aprés la condition suffisante de K-K-T, $p(q)$ est la projection de q sur P. En remplaçant $\|y(w)\| = 1$, on aura la distance entre q et sa projection $p(q)$. ■

Corollary 2.3 ([6]) *Pour la norme 1 ;* $\| . \|_1,\ w^\top y(w) =\| w \|_\infty$ *et par conséquent*

$$\left.\begin{array}{c} y_i(w) = 0 \quad si \mid w_i \mid \neq \| w \|_\infty \\ y_i(w) = v_i \quad si \quad w_i = \| w \|_\infty \\ y_i(w) = -v_i \quad si\ w_i = -\| w \|_\infty \end{array}\right\}, v_i \geqslant 0\ ;\ \sum_{i=1}^{n} v_i = 1.$$

En outre, comme $\| y(w) \|_1 = 1$, *et d'aprés (2.12) et (2.13) on a :*

$$p(q) = q - \frac{w^\top q - \gamma}{\|w\|_\infty} y(w),\ \ et \quad \| p(q) - q \|_1 = \frac{\mid w^\top q - \gamma \mid}{\|w\|_\infty}.$$

Corollary 2.4 ([6]) *Pour la norme 2 ;* $\| . \|_2 \quad w^\top y(w) =\| w \|_2$ *et par conséquent*

$$y(w) = \frac{w}{\| w \|_2}.$$

En outre, comme $\| y(w) \|_2 = 1$, *et d'aprés (2.12) et (2.13) on a :*

$$p(q) = q - \frac{w^\top q - \gamma}{\|w\|_2} y(w),\ \ et \quad \| p(q) - q \|_2 = \frac{\mid w^\top q - \gamma \mid}{\|w\|_2}.$$

Corollary 2.5 ([6]) *Pour la norme* ∞ *;* $\| . \|_\infty,\ \ w^\top y(w) =\| w \|_1$ *et par conséquent*

$$\left.\begin{array}{c} y_i(w) = 1\ si\ w_i > 0 \\ y_i(w) = -1\ si\ w_i \leqslant 0 \end{array}\right\}.$$

En outre, comme $\| y(w) \|_\infty = 1$, *et d'aprés (2.12) et (2.13) on a :*

$$p(q) = q - \frac{w^\top q - \gamma}{\|w\|_1} y(w),\ \ et \quad \| p(q) - q \|_\infty = \frac{\mid w^\top q - \gamma \mid}{\|w\|_1}.$$

Chapitre 3

Machines à Vecteurs de Support

3.1 Introduction

Les machines à vecteurs de support (SVMs) sont des algorithmes ayant comme but de résoudre les problèmes de classification à deux classes. On appelle problème de classification à deux classes un problème dans lequel on tente de déterminer la classe à laquelle appartient un individu (individu est ici employé au sens de constituant d'un ensemble) parmi deux choix possibles. Plusieurs méthodes ont été suggérées en littérature pour étendre l'application des SVMs aux problèmes de discrimination à plus de deux classes.

Pour ce faire, on utilise les caractéristiques connues de cet individu. Ces n caractéristiques sont représentées par un vecteur $x \in \mathbb{R}^n$. La classe à laquelle appartient l'individu est représentée par $y \in \{-1, 1\}$, où une des classes possible est représentée par -1 et l'autre par 1. Par conséquent, avec cette notation, le problème est de déterminer la valeur de y en se servant de x.

Pour y parvenir, les machines à vecteurs de support utilisent un ensemble de données pour les quelles le classement est déjà connu et s'en servent pour construire une règle qui permet d'efectuer une bonne classification. Cet ensemble de données est appelé l'ensemble d'apprentissage. La règle trouvée avec l'ensemble d'apprentissage doit être la plus générale possible, puisqu'il faut aussi qu'elle soit bonne pour de nouvelles données qui n'étaient pas dans l'ensemble d'apprentissage. Nous présentons ici comment les SVMs font pour trouver cette règle d'abord dans le cas le plus simple possible, c'est-à-dire le cas où les données sont linéairement séparables.

3.2 Hyperplan séparateur

Supposons que nous disposons d'un ensemble d'apprentissage de m données de la forme $(x_i, y_i) \in \mathbb{R}^n \times \{-1, 1\}$ $(i = 1, ..., m)$, dont nous voulons nous servir pour déterminer une règle permettant de classer les données. Supposons aussi que ces données sont linéairement séparables, c-à-d qu'il existe un hyperplan dans $\mathbb{R}^n$ tel que toutes les données appartenant à la classe 1 (P_+) se retrouvent d'un côté de l'hyperplan alors que celles de la classe -1 (P_-) se situent de l'autre côté.

figure 1. Des donnees lineairement separables

Plus formellement, les données sont dites linéairement séparables s'il existe un hyperplan

$$w^\top x - \gamma = 0$$

tel que $w^\top x - \gamma > 0$ pour tout x appartenant à P_+, et $w^\top x - \gamma < 0$ pour tout x appartenant à P_-, avec $w = (w_1, ..., w_n) \in \mathbb{R}^n$ le vecteur des coefficients de l'hyperplan et $\gamma \in \mathbb{R}$ un scalaire appelé le biais (remarquons que tout hyperplan peut s'écrire sous cette forme). Nous dirons d'un tel hyperplan qu'il sépare les données.

Sous l'hypothèse que les données sont linéairement séparables, trouver une règle pour les classer est très simple. En effet, il suffit de prendre un hyperplan qui sépare les classes, puis de classer les données selon le côté de l'hyperplan où elles se trouvent. Plus formellement, soit

$$w^\top x - \gamma = 0 \tag{3.1}$$

un hyperplan qui sépare les données. Alors, il suffit d'utiliser la fonction suivante (appelée la fonction $classificatrice$) pour effectuer la classification :

$$f(x) = w^\top x - \gamma \tag{3.2}$$

$$\begin{aligned} x \in P_+ &\Longrightarrow f(x) \geqslant 0, \\ x \in P_- &\Longrightarrow f(x) \leqslant 0. \end{aligned} \tag{3.3}$$

Cette fonction classe les données par rapport au côté de l'hyperplan où elles se trouvent. On remarque que si un ensemble de données est séparé par un hyperplan, il sera parfaitement classé par cette fonction. Notons que si une donnée est directement sur l'hyperplan (ce qui peut arriver en considérant des données qui ne sont pas dans l'ensemble d'apprentissage), elle sera assignée à la classe 0, ce qui signifie qu'elle ne peut être classée par le modèle actuel.

Dans ce cas, il est possible de la laisser inclassée, d'utiliser une autre règle ou de l'assigner aléatoirement à l'une des deux classes.

Grâce à la fonction classificatrice, on constate qu'il suffit de trouver un hyperplan qui sépare les données pour déterminer une règle permettant de les classer.

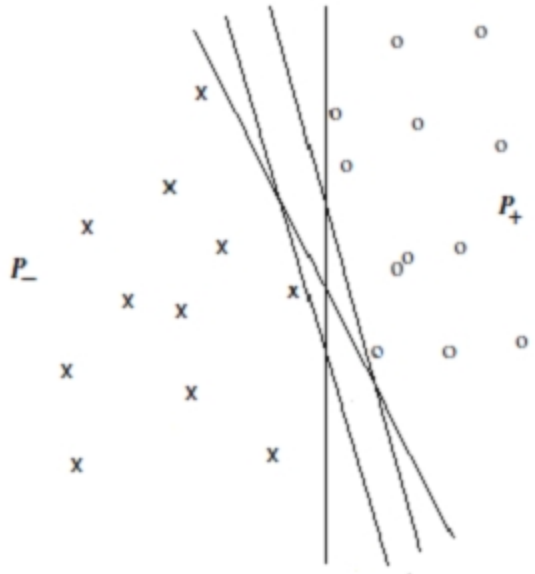

figure 2. Il existe une infinite d'hyperplans pouvant
separer les donnees

Cependant, si les données sont linéairement séparables, il existe une infinité d'hyperplans qui peuvent servir de séparateurs. L'idée des machines à vecteurs de support est de choisir le meilleur hyperplan, c-à-d celui qui donnera la règle qui se généralisera le mieux à d'autres données que celles de l'ensemble d'apprentissage. Afin de déterminer ce qui caractérise le meilleur hyperplan, introduisons le concept de marge.

3.3 Marge et hyperplan canonique

Définissons la marge d'un hyperplan comme étant la distance entre l'hyperplan et la donnée la plus proche. Plus formellement (si on note par s la donnée de l'échantillon le plus proche et on prend la norme 2) du corollaire 2.4, on a

$$\|s - f(x)\|_2 = \frac{|w^\top s - \gamma|}{\|w\|_2}. \tag{3.4}$$

D'après un résultat de la théorie de l'apprentissage statistique [Vapnik], l'hyperplan qui aura la meilleure généralisation est celui qui possède la plus grande marge. Ce concept est à la base des machines à vecteurs de support. Dans le cas le plus simple, c-à-d celui où les données sont linéairement séparables, les SVMs trouvent l'hyperplan qui sépare les données avec la plus vaste marge possible, puis utilisent cet hyperplan pour classer de nouvelles

données à l'aide de la fonction classificatrice donnée plus haut. Toutefois, le problème de trouver l'hyperplan avec la marge maximale est mal posé, puisqu'il existe en réalité une infinité de manières différentes d'écrire le même hyperplan.

En effet, supposons que l'hyperplan

$$w^\top x - \gamma = 0$$

soit un hyperplan dont la marge est maximale, et soit $\lambda \in \mathbb{R}^+ \backslash \{0\}$. Alors, l'hyperplan

$$\lambda w^\top x - \lambda\gamma = 0$$

est en réalité le même hyperplan et sépare les données, puisque λ est positif. Par conséquent, $\lambda w^\top x - \lambda\gamma = 0$ correspond aussi à l'hyperplan dont la marge est maximale, mais possède un vecteur des coefficients et un biais différents (si $\lambda \neq 1$). Le nombre infini de manières d'écrire la solution du problème de l'hyperplan avec la plus vaste marge complique sa résolution. Afin de rendre le problème bien posé et pour lutter contre une solution insignifiante (car $w = 0_{\mathbb{R}^m}$ et $\gamma = 0$ vérifient l'équation de l'hyperplan), introduisons le concept d'hyperplan canonique.

Et pour cela, on redéfini (w, γ) comme suit :

$$\frac{(w, \gamma)}{\min\limits_{x \in P_+ \cup P_-} \mid w^\top x + \gamma \mid}, \tag{3.5}$$

on aura :

$$\begin{aligned} x \in P_+ &\Longrightarrow f(x) = w^\top x - \gamma \geqslant 1, \\ x \in P_- &\Longrightarrow f(x) = w^\top x - \gamma \leqslant -1. \end{aligned} \tag{3.6}$$

et pour les vecteurs s (la donnée de l'échantillon le plus proche)

$$\begin{aligned} s \in P_+ &\Longrightarrow f(s) = w^\top s - \gamma = 1, \\ s \in P_- &\Longrightarrow f(s) = w^\top s - \gamma = -1. \end{aligned}$$

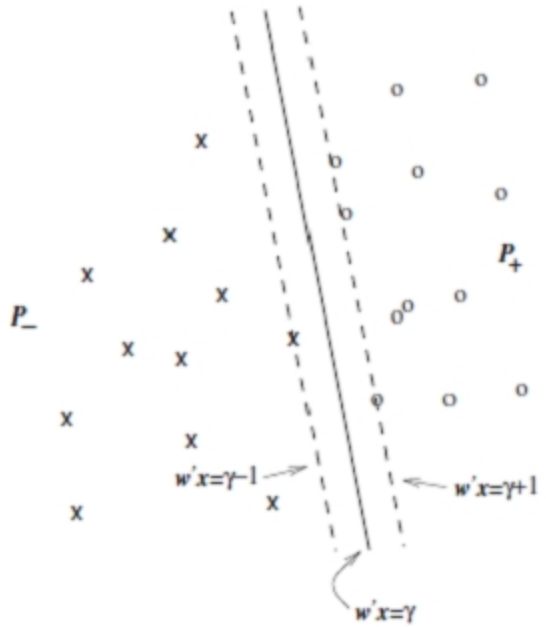

figure 3. Les hyperplans de bondissement et la marge

On peut aussi montrer que tout hyperplan qui sépare les données peut s'écrire sous forme canonique et qu'il n'existe qu'une seule façon d'écrire un hyperplan pour qu'il soit canonique. Ainsi, en ne considérant que les hyperplans canoniques, chaque hyperplan s'écrit de manière unique. De plus, il n'existe qu'un seul hyperplan pour lequel la marge est maximale. Ceci deviendra évident un peu plus loin, puisque le vecteur des coefficients de l'hyperplan sera exprimé comme étant le point qui minimise une fonction strictement convexe (rappelons que les fonctions strictement convexes n'ont qu'un unique minimum global). Par conséquent, en ne considérant que les hyperplans canoniques, le problème de trouver l'hyperplan avec la plus grande marge est bien posé.

3.4 Trouver l'hyperplan

On peut montrer que pour un hyperplan canonique, la marge est donnée par l'expression

$$\frac{1}{\|w\|_2}, \tag{3.7}$$

on voit donc que plus $\|w\|_2$ est petite, plus la marge de l'hyperplan canonique correspondant est grande. Ainsi, afin de trouver l'hyperplan qui sépare le mieux les données, il faut trouver celui qui respecte les conditions d'un hyperplan canonique et pour lequel $\|w\|_2$ est minimale.

La recherche du meilleur hyperplan peut donc s'écrire sous la forme du problème d'optimisation suivant :

$$\begin{cases} \min\limits_{w,\gamma} & \|w\|_2 \\ D(Aw - e\gamma) \geqslant e, \end{cases} \tag{3.8}$$

où A est une matrice de taille $m \times n$, qui assemble tous les points x_i $(i = 1, ..., m)$ de l'échantillon

$$A = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} x_1^1 & x_1^2 & \cdots & x_1^n \\ x_2^1 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \cdots & \vdots \\ x_m^1 & x_m^2 & \cdots & x_m^n \end{pmatrix},$$

D est une matrice diagonale

$$D_{ii} = y_i,$$

et

$$e = (1, 1, ..., 1)^\top.$$

Nous avons ainsi formulé un problème d'optimisation dont la solution optimale est l'hyperplan canonique séparant les données avec la plus vaste marge possible. Cependant, il est possible de formuler un problème équivalent, mais avec une fonction objectif plus simple. En effet, comme

$$\|w\|_2 = \sqrt{w^\top w},$$

minimiser $\|w\|_2$ est équivalent à minimiser $\|w\|_2^2$. Évidemment, minimiser $w^\top w$ est équivalent à minimiser $\frac{1}{2} w^\top w$ (cette petite modification permet d'éviter d'avoir une constante dans la représentation duale du problème, comme nous le verrons un peu plus loin).

Par conséquent, afin de trouver l'hyperplan canonique qui sépare les données avec la plus grande marge possible, il suffit de résoudre le problème d'optimisation suivant :

$$\begin{cases} \min\limits_{w,\gamma} & \frac{1}{2} w^\top w \\ D(Aw - e\gamma) \geqslant e. \end{cases} \tag{3.9}$$

qui est un problème d'optimisation avec une fonction objectif strictement convexe. Ceci assure qu'il n'y a pas de minimum relatif et qu'il n'existe qu'une unique solution optimale.

3.5 Les vecteurs de support

Comme la fonction $w \longmapsto w^\top w$ est une fonction convexe continue et dérivable, que les contraintes $D(Aw - e\gamma) \geqslant e$ sont des fonctions affines et que le domaine du problème est $\mathbb{R}^n$, la solution optimale trouvée respecte nécessairement les conditions de K-K-T. En particulier, elle respecte la condition de complémentarité de K-K-T, c-à-d que

$$\widetilde{u}_i(D_{ii}(A_{i.}\overline{w} - \overline{\gamma}) - 1) = 0, \quad i = 1, ..., m$$

où $\widetilde{u}$ représente la solution optimale du problème dual et $(\overline{w}, \overline{\gamma})$ représente celle du problème primal. Cette condition implique que si $D_{ii}(A_{i.}\overline{w} - \overline{\gamma}) - 1 \neq 0$, alors $u_i = 0$. Par conséquent, les seuls cas où u_i peut ne pas être nul sont ceux où $D_{ii}(A_{i.}\overline{w} - \overline{\gamma}) - 1 = 0$, c-à-d ceux où

$$D_{ii}(A_{i.}\overline{w} - \overline{\gamma}) = 1.$$

Or, les seuls points où $D_{ii}(A_{i.}\overline{w} - \overline{\gamma}) = 1$ sont ceux qui sont sur la marge. Par conséquent, seuls les points sur la marge peuvent avoir des u_i non nuls. Ces points sont appelés les vecteurs

de support. La raison de ce nom est que ce sont les seuls points utiles pour déterminer l'hyperplan. En effet, rappelons que le vecteur des coefficients de l'hyperplan est donné par

$$w = A^\top Du.$$

Ainsi, tout point qui n'est pas sur la marge n'apporte aucune contribution, puisque u_i est alors nul. Si tous les points sauf les vecteurs de support étaient retirés de l'ensemble d'apprentissage, on retrouverait le même hyperplan. Les vecteurs de support peuvent donc être vus comme les points contenant toute l'information essentielle du problème.

3.6 Marges souples

3.6.1 Machines à vecteurs de support et bruit

En pratique, les données sont rarement parfaites. Il y a souvent du bruit, c-à-d des données qui sont mal classées par un modèle qui est toutefois excellent en général. Il s'agit donc d'erreurs qui sont inévitables, même pour les meilleurs modèles. Toutefois, les machines à vecteurs de support ne permettent pas de tenir compte de ce phénomène, puisque dans les contraintes, toutes les données doivent être correctement classées. Supposons par exemple qu'un ensemble de données serait très bien séparé par un hyperplan, mais qu'il n'est pas linéairement séparable dû à la présence d'un certain bruit dans les données. Dans un tel cas, il serait impossible de construire une SVM linéaire, car il est impossible que toutes les contraintes soient respectées.

Afin de contourner ce problème, nous introduisons des variables artificielles (slack variables), mais la règle trouvée risque de très mal se généraliser, puisqu'elle va tenir compte de toutes les petites variations et ainsi généraliser des phénomènes qui sont en réalité bien spécifiques à l'ensemble de données actuel.

3.6.2 Marge souple

Un meilleur moyen serait de permettre à quelques données d'être à l'intérieur de la marge ou du mauvais côté de l'hyperplan. Il s'agit du concept de marge souple (soft margin). Une première idée serait de tenter de maximiser la marge tout en minimisant le nombre de données mal classées. Toutefois, le nombre de données mal classées peut être trompeur, puisqu'il ne permet pas de déterminer si une donnée était presque correctement classée ou si elle était en réalité très loin de l'hyperplan. Une meilleure idée est d'attribuer à chaque donnée x_i une valeur z_i qui représente à quel point la donnée est éloignée d'un bon classement, puis de tenter de minimiser la somme des z_i. Plus formellement, au lieu d'imposer

$$D(Aw - e\gamma) \geqslant e,$$

ce qui oblige les données à être bien classées, les contraintes seront plutôt

$$D(Aw - e\gamma) + z \geqslant e, \quad z \geqslant 0.$$

où $z = (z_1, z_2, ..., z_m)^\top$.

Par conséquent, il est possible pour une donnée d'être du mauvais côté de la marge, si z_i est non nul. On dira d'une donnée qu'elle est du mauvais côté de la marge si elle est mal

classée ou si sa distance par rapport à l'hyperplan séparateur est plus petite que la marge (remarquons que les points pour lesquels $z_i \neq 0$ ne sont pas considérés dans le calcul de la marge). L'objectif est ainsi de maximiser la marge tout en minimisant la somme des z_i. Le problème d'optimisation devient alors

$$\begin{cases} \min\limits_{w,\gamma,z} \quad \frac{1}{2} w^\top w + v e^\top z \\ D(Aw - e\gamma) + z \geqslant e, \qquad\qquad z \geqslant 0. \end{cases} \tag{3.10}$$

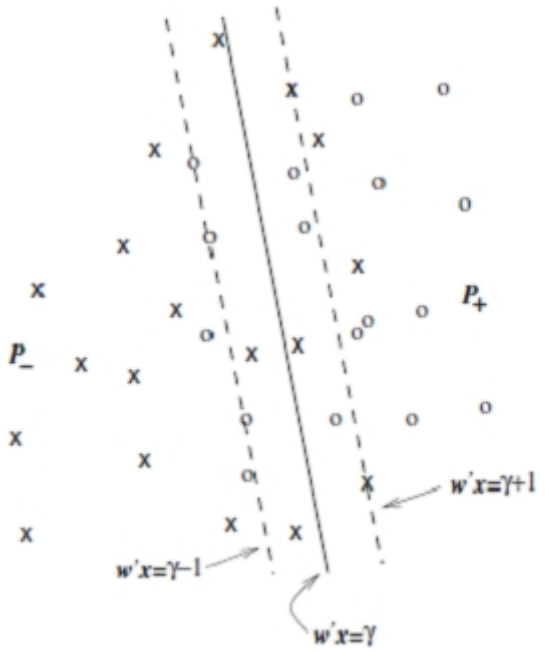

figure 4. Marge souple

où $v > 0$ est une constante qui représente la pénalité d'avoir des données mal classées. Lorsque v est très élevée, il y aura très peu de données mal classées, alors qu'il y en aura plus pour une valeur plus faible de cette constante. Le choix de v a une grande influence sur le modèle. En pratique, plusieurs modèles sont souvent construits, avec différentes valeurs de v, puis le meilleur est choisi.

3.7 Représentation duale

En employant u pour dénoter les multiplicateurs de Lagrange pour les contraintes $D(Aw - e\gamma) + z \geqslant e$, les conditions de K-K-T pour ce problème peuvent être écrites comme suit :

$$\begin{aligned} 0 &= w - A^\top D u, \\ 0 &= e^\top D u, \\ 0 &\leqslant v e - u \perp z \geqslant 0, \\ 0 &\leqslant D(Aw - e\gamma) + z - e \perp u \geqslant 0. \end{aligned}$$

Il n'est pas difficile de voir que le problème dual associé au problème est le suivant :

$$\begin{cases} \min\limits_{u} \frac{1}{2} u^\top DAA^\top Du - e^\top u \\ \qquad 0 \leqslant u \leqslant ve, \\ \qquad e^\top Du = 0. \end{cases} \tag{3.11}$$

En effet, il est souvent plus commode de résoudre cette forme du problème pour la variable duale u et puis de récupérer w en remplaçant $w = A^\top Du$ (la 1 ére condition de K-K-T, récupérant γ en tant que multiplicateur de Lagrange pour la contrainte $e^\top Du$).

Remarque 3.1 *Les conditions de K-K-T tiennent toujours dans le cas de la marge souple. Par conséquent, d'après la condition complémentaire, pour la solution optimale, les égalités suivantes sont vérifiées :*

$$u^\top (D(Aw - e\gamma) + z - e) = 0 \Leftrightarrow u_i (D_{ii}(A_{i.}w - \gamma) + z_i - 1) = 0, \quad i = 1, ..., m,$$

$$(ve - u)^\top z = 0 \Leftrightarrow (v - u_i) z_i = 0, \ i = 1, ..., m.$$

Ceci implique que si $z_i \neq 0$, alors $v - u_i = 0$ (puisque $(v - u_i)z_i = 0$), et donc $v = u_i$. De plus, si un point est tel que $z_i \neq 0$, alors il est du mauvais côté de la marge, ce qui découle directement du rôle de z_i dans le problème d'optimisation. À l'opposé, tous les points pour lesquels $z_i = 0$ sont du bon côté de la marge, et ainsi nécessairement bien classés.

D'autre part, si, pour une certaine donnée, on a $0 < u_i < v$, alors celle-ci est exactement sur la marge. En effet, on a alors $0 < u_i < v$, $u_i \neq v$, et donc il faut que $z_i = 0$ pour que $(v - u_i)z_i = 0$. De plus, $u_i \neq 0$, ce qui implique que $D_{ii}(A_{i.}w - \gamma) + z_i - 1 = 0$ afin de respecter l'égalité $u_i(D_{ii}(A_{i.}w - \gamma) + z_i - 1) = 0$. Comme $z_i = 0$, il s'ensuit que

$$D_{ii}(A_{i.}w - \gamma) = 1$$

et donc que $A_{i.}$ (qui représente le point x_i) est directement sur la marge.

Enfin, les points pour lesquels $z_i = 0$ et $D_{ii}(A_{i.}w - \gamma) - 1 \neq 0$ ont un u_i nul, afin de respecter l'égalité $u_i(D_{ii}(A_{i.}w - \gamma) + z_i - 1) = 0$. Les points directement sur la marge sont appelés vecteurs de support libres (free support vectors), ou encore vecteurs de support non-bornés (unbounded support vectors). Les points pour lesquels $u_i = v$ sont quant à eux appelés vecteurs de support bornés (bounded support vectors). Ici encore, les vecteurs de support sont les seuls points qui sont vraiment important pour déterminer l'hyperplan optimal, puisque ce sont les seuls points pour lesquels $u_i = 0$.

3.8 Machines à vecteurs de support pour données non linéairement séparables

3.8.1 Transformations

Jusqu'à présent, les machines à vecteurs de support permettent de trouver une règle pour classer les données lorsque celles-ci sont linéairement séparables. Cependant, il existe bien des cas pour lesquels il est impossible de séparer entièrement les données avec un hyperplan. Afin de régler ce problème, il est possible d'appliquer une transformation aux données de sorte qu'une fois transformées, elles soient linéairement séparables. L'espace où

se trouvent les données avant d'être transformées est appelé l'espace d'entrée (input space), alors qu'après avoir appliqué la transformation, les données se trouvent dans ce qu'on appelle l'espace de redescription (feature space). Il suffit alors de trouver l'hyperplan dans l'espace de redescription qui sépare le mieux ces données transformées. De retour dans l'espace d'entrée, le séparateur n'est pas linéaire.

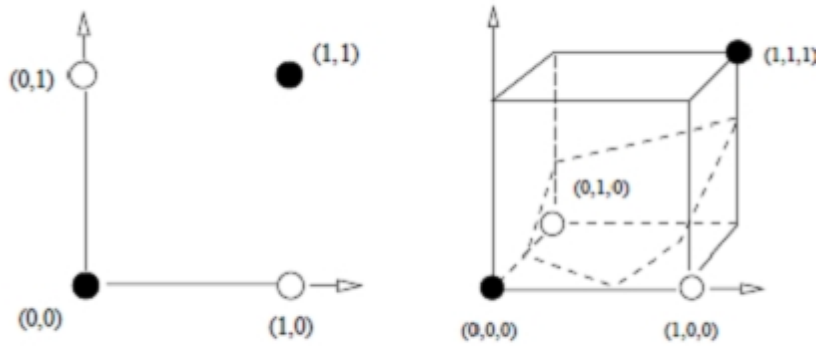

figure 5. Plonger les donnes dans un espace de dimension superieure permet de les rendre lineairement separables.

Soit

$$\begin{array}{rcl} \phi & : & \mathbb{R}^n \longrightarrow \mathbb{R}^{n'} \\ & & x \longrightarrow \phi(x) \end{array} \tag{3.12}$$

la transformation appliquée aux données pour les rendre linéairement séparables, avec n' la dimension de l'espace de redescription. Très souvent, $n' > n$, ce qui signifie que la transformation amène les données dans un espace de dimension supérieure afin de mieux pouvoir les séparer (voir figure 5).

Pour trouver le séparateur, on procède de la même manière que précédemment, mais en substituant $\phi(A_{i.}^\top)$ à $A_{i.}$ $(i = 1, ..., m)$. Il s'agit donc de résoudre le problème suivant

$$\left\{ \begin{array}{c} \min\limits_{w,\gamma,z} \quad \frac{1}{2} w^\top w + v e^\top z \\ D_{ii}(w^\top \phi(A_{i.}^\top) - \gamma) + z_i \geqslant 1, \\ z_i \geqslant 0, \qquad i = 1, ..., m. \end{array} \right. \tag{3.13}$$

Le dual de ce problème est

$$\left\{ \begin{array}{c} \min_u \frac{1}{2} \sum\limits_{i=1}^{m} \sum\limits_{j=1}^{n} u_i D_{ii} \phi(A_{i.}^\top)^\top \phi(A_{i.}^\top) D_{jj} u_j - \sum\limits_{i=1}^{m} u_i \\ 0 \leqslant u \leqslant ve, \\ e^\top D u = 0. \end{array} \right. \tag{3.14}$$

La fonction classificatrice associée à ce problème est par conséquent

$$f(x) = w^\top \phi(x) - \gamma = \sum_{i=1}^{m} D_{ii} u_i \phi(A_{i.}^\top)^\top \phi(x) - \gamma. \tag{3.15}$$

Si la transformation utilisée est appropriée, la résolution d'un de ces problèmes (le primal ou le dual) permet de trouver un séparateur non linéaire avec la marge la plus grande possible, permettant ainsi d'utiliser les machines à vecteurs de support dans le cas où les données ne peuvent pas être séparées linéairement.

3.8.2 Les noyaux

Toutefois, l'utilisation des transformations pose certains problèmes. En effet, outre le fait qu'il faille choisir une bonne transformation, il faut l'appliquer à toutes les données, puis effectuer les calculs avec ces données transformées, c-à-d dans l'espace de redescription. Or, comme la dimension de cet espace est bien souvent beaucoup plus grande que celle de l'espace d'entrée, les calculs requis peuvent devenir extrêmement longs à effectuer. C'est ici que la formulation duale du problème d'optimisation prend toute son importance. En effet, on remarque que lorsque le problème est sous sa forme duale, les données de l'ensemble d'apprentissage n'apparaissent que dans un produit scalaire avec d'autres données du même ensemble. Il en est de même dans la fonction classificatrice duale. Ceci amène à définir comme suit une fonction appelée noyau (kernel) :

$$\begin{array}{rcl} K & : & \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R} \\ & & (x_i, x_j) \longrightarrow \phi(x_i)^\top \phi(x_j) \end{array} \tag{3.16}$$

alors $K_{ij} = \phi(A_{i.}^\top)^\top \phi(A_{j.}^\top) = K(A_{i.}^\top, A_{j.}^\top)$.

Cette fonction prend en entrée deux points dans l'espace d'entrée et calcule leur produit scalaire dans l'espace de redescription. L'avantage d'une telle fonction est qu'il n'est pas nécessaire d'appliquer une transformation aux données afin de calculer leur produit scalaire dans l'espace de redescription. Ce calcul peut se faire directement à partir des données de l'espace d'entrée.

Grâce au concept de noyau, il est possible de réécrire le problème dual de cette manière :

$$\begin{cases} \min_u \frac{1}{2} u^\top DKDu - e^\top u \\ 0 \leqslant u \leqslant ve, \\ e^\top Du = 0. \end{cases} \tag{3.17}$$

La fonction classificatrice s'écrit ainsi :

$$f(x) = \sum_{i=1}^{m} D_{ii} u_i K(A_{i.}^\top, x) - \gamma.$$

On remarque que de cette manière, lorsque la fonction noyau est connue, la transformation $\phi(x)$ n'apparaît nulle part, ni dans le problème, ni dans l'application de la solution. Par conséquent, grâce à la fonction noyau, il n'est pas nécessaire d'effectuer la transformation sur les données. Cette fonction permet donc de faire tous les calculs nécessaires sans avoir à se préoccuper de la dimension de l'espace de redescription.

3.9 Exemples de noyaux

Il est bien de savoir qu'un noyau est tout ce qui est nécessaire pour utiliser les SVMs dans le cas non linéaire, mais cette information est inutile sans la connaissance des noyaux qu'il

est possible d'utiliser. Nous présentons maintenant les manières de construire des noyaux, ainsi que les noyaux les plus fréquemment utilisés pour les machines à vecteurs de support sont :

- Noyau linéaire
$$K(x_i, x_j) = (x_i^\top x_j).$$
- Noyau polynomial de degré d
$$K(x_i, x_j) = (x_i^\top x_j + 1)^d.$$
- Noyau Gaussein
$$K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}).$$
- Noyau multi quadratique inverse
$$K(x_i, x_j) = \frac{1}{\sqrt{\|x_i - x_j\|_2^2 + \beta}}.$$

3.10 Avantages des SVMs

SVM est une méthode de classification intéressante car le champ de ses applications est large, parmis ses avantages nous avons :

- Un grand taux de classification et de généralisation par rapport aux méthodes classiques.
- Elle nécessite moins d'effort pour designer l'architecture adéquate (petit nombre de paramètre à régler ou à estimer).
- La résolution du problème est convertie en résolution d'un problème quadratique strictement convexe dont la solution est unique et donnée par des méthodes de programmation quadratique.

Chapitre 4

Appendice

Définition 4.1 (Ensemble convexe [1]) *Soit $S \subset \mathbb{R}^n$. L'ensemble S est convexe si*

$$\forall (x,y) \in S^2, \quad \forall \lambda \in]0,1[, \quad \lambda x + (1-\lambda)y \in S,$$

c-à-d, si x et y sont deux éléments de S alors le segment qui relie x à y est inclus dans S.

Exemple 4.1 *$\mathbb{R}^n$ est convexe.*

Définition 4.2 (Fonction convexe/strictement convexe [1]) *Soient $S \subset \mathbb{R}^n$ convexe et $f : S \to \mathbb{R}$.*

-f est convexe si

$$\forall (x,y) \in S^2, \quad \forall \lambda \in]0,1[, \quad f(\lambda(x) + (1-\lambda)y) \leqslant \lambda\ f(x) + (1-\lambda) f(y).$$

-f est strictement convexe si

$$\forall (x,y) \in S^2,\ x \neq y, \quad \forall \lambda \in]0,1[, \quad f(\lambda(x) + (1-\lambda)y) < \lambda\ f(x) + (1-\lambda) f(y).$$

Théorème 4.1 ([1]) *Soit $S \subset \mathbb{R}^n$ convexe et $f : S \longrightarrow \mathbb{R}$ différentiable. La fonction f est convexe si et seulement si :*

$$\forall (x,y) \in S^2, \qquad f(y) \geqslant f(x) + < \nabla f(x), y - x > . \tag{4.1}$$

Théorème 4.2 ([1]) *Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ une fonction de classe C^2.*

Si la Hessienne $\nabla^2 f(x)$ est une matrice symétrique définie positive pour tout $x \in \mathbb{R}^n$, alors f est strictement convexe.

Si $\nabla^2 f(x)$ est une matrice symétrique semi définie positive pour tout $x \in \mathbb{R}^n$, alors f est convexe.

Définition 4.3 (Programme Linéaire [9]) *est un problème dans lequel les variables sont des réels qui doivent satisfaire un ensemble d'équations et/ou d'inéquations linéaire (dites "contraintes") et la valeur d'une fonction linéaire de ces variables (appelée "fonction objectif") doit être rendue minimum (ou maximum). La forme générale d'un programme linéaire est :*

$$(P) \begin{cases} \max f(x) = cx \\ Ax \leqslant b, \\ x \geqslant 0. \end{cases} \Leftrightarrow \begin{cases} \max f(x) = \sum_{j=1}^{n} c_j x_j \\ \sum_{j=1}^{n} a_{ij} x_j \leqslant b_i, \quad i = 1, ..., m, \\ x_j \geqslant 0, \qquad\qquad j = 1, ..., n. \end{cases} \tag{4.2}$$

Définition 4.4 (Programme Linéaire dual [9]) *On appelle "dual" du problème linéaire (4.2), le programme linéaire (D) qui s'écrit sous la forme :*

$$(D)\begin{cases} \min b^\top y \\ A^\top y \geqslant c^\top, \\ y^\top \geqslant 0. \end{cases}$$

Remarque 4.1 *Le concept de dualité est un concept fondamental en programmation linéaire. En effet, il faut considérer que deux programmes linéaires duaux ne constituent pas deux problèmes distincts mais deux aspects du même problème ; car quand on résout un programme linéaire, on résout aussi son dual.*

Théorème 4.3 (Théorème fondamental de la dualité [9]) *Soit $(\overline{x}, \overline{y})$ un couple de solution admissibles de, respectivement, (P) et (D) alors :*

$$c\overline{x} \leqslant \overline{y}b.$$

Corollary 4.1 *Soit $(\overline{x}, \overline{y})$ un couple de solutions admissibles de, respectivement, (P) et (D) vérifiant*

$$c\overline{x} \geqslant \overline{y}b,$$

alors :

$$c\overline{x} = \overline{y}b,$$

et $(\overline{x}, \overline{y})$ est un couple de solutions optimales.

Théorème 4.4 (Théorème des écarts complémentaire [9]) *Soit $(\overline{x}, \overline{y})$ un couple de solutions optimales de, respectivement (P) et (D) alors :*

$$[(\sum_{j=1}^{n} a_{ij}\overline{x}_j) - b_i]\overline{y}_i = 0, \qquad i = 1, ..., m.$$

$$[(\sum_{i=1}^{m} a_{ij}\overline{y}_i) - c_j]\overline{x}_j = 0, \qquad j = 1, ..., n.$$

Corollary 4.2 ([9]) *Soit $(\overline{x}, \overline{y})$ un couple de solutions optimales de, respectivement (P) et (D) alors :*

$$(\overline{x}_j > 0) \implies (\sum_{i=1}^{m} a_{ij}\overline{y}_i = c_j),$$

$$(\sum_{j=1}^{n} a_{ij}\overline{x}_j < b_i) \implies (\overline{y}_i = 0),$$

$$(\overline{y}_i > 0) \implies (\sum_{j=1}^{n} a_{ij}\overline{x}_j = b_i),$$

$$(\sum_{i=1}^{m} a_{ij}\overline{y}_i > c_j) \implies (\overline{x}_j = 0).$$

Conséquence directe du théorème 5.4.

Bibliographie

[1] M.S. Bazaraa, H.D. Sherali et C.M. Shetty "Nonlinear Programming Theory and Algorithms", AJohn Wiley & Sons, INC.,Publication, 2006.

[2] M.C. Ferris, O.L. Mangasarian et S.J. Wright "Linear Programming with MATLAB", MPS-SIAM Series on Optimization, 2007.

[3] G. Fung and O.L. Mangasarian " Breast Tumor Susceptibility to Chemotherapy via Support Vector Machines", 2006.

[4] L. Juntao, J. Yingmin and L. Wenlin "Adaptive Huberized Support Vector Machine and its Application to Microarray Classification", 2011.

[5] Y.J. Lee and O.L. Mangasarian " Reduced Support Vector Machines", 2001.

[6] O.L. Mangasarian " Arbitary-Norm Separating Plane, 1999.

[7] E. Marchiori and M. Sebag "Bayesian Learning with Local Support Vector Machines for Cancer Classification with Gene Expression Data", 2005.

[8] J. Thongkam, G. Xu, Y. Zhang, and F. Huang "Support Vector Machine for Outlier Detection in Breast Cancer Survivability Prediction", 2008.

[9] R.J. Vanderbei "Linear Programming, Foundations and Extensions", Springer, 2008.