Rapport sur le projet d'Optimisation
Support-Vector Machines

Kawisorn Kamtue & Clémence Réda

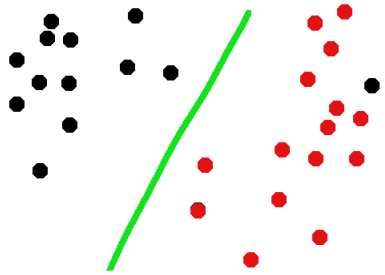
December 11, 2016

1 Support Vector Machine

Les *Support Vector Machine solvers* (SVM) sont une catégorie d'algorithmes d'apprentissage statistique supervisé. Ils permettent de résoudre le problème de classification binaire suivant :

Etant donnés $(x_i)_{i\leq m}$ des points dans $\mathbb{R}^n$, et $(y_i)_{i\leq m}$ les étiquettes des points tels que l'étiquette de x_i soit $y_i \in \{-1, 1\}$, on cherche la droite qui sépare "le mieux possible" les points dans différentes classes, autrement dit, la frontière de Voronoi entre les deux classes.

Figure 1: Exemple de frontière pour deux classes : celles des points noirs et celle des points rouges



La frontière que l'on recherche est une fonction linéaire, donc de la forme (avec deux paramètres de dimension 1 ω et b) :

$$f : X -> \omega^T X + b$$

telle que :

$$\forall i, y_i = -1 \Rightarrow f(x_i) \leq -1$$
$$\forall i, y_i = 1 \Rightarrow f(x_i) \geq 1$$
$$\Leftrightarrow \forall i, y_i \times f(x_i) \geq 1 \quad (1)$$

Pour simplifier le problème, on peut prendre $\omega' = \begin{bmatrix} \omega \\ 1 \end{bmatrix}$ (que l'on note par la suite ω) pour supposer $b = 0$ sans perte de généralité.

Pour obtenir un résultat robuste, on souhaite que les deux droites $f(X) = 1$ et $f(X) = -1$ soient les plus distantes possibles. En effet, si ces deux droites sont trop proches, cela signifie que la probabilité d'erreur quant à la prédiction de la classe d'un point proche de ces droites sera importante.

La distance γ entre ces deux droites se calcule de la façon suivante : soient u, v deux points tels que $f(v) = 1$ et $f(u) = -1$. Alors :

$$\|f(v) - f(u)\| = \|\omega \times (v - u)\| = \|\omega\| \times \|(v - u)\| = \|\omega\| \times \|\gamma\| = \|1 - (-1)\| = 2$$

Finalement, le problème d'optimisation à résoudre pourrait être :

$max_w\ \gamma = \frac{2}{\|w\|}$ avec (1)
$\Leftrightarrow min_w\ \|w\|$ avec (1)
$\Leftrightarrow min_w\ \frac{1}{2} \times \|w\|^2$ avec (1) pour faciliter les calculs

Un autre problème se pose si on s'arrête ici : par exemple, dans l'exemple de la frontière de Voronoi que l'on a vu ci-dessus, il n'existe pas de droite telle qu'il n'y ait que de points noirs d'un côté et que des points rouges de l'autre côté, ce qui contredit la condition (1). Le problème est alors infaisable. Pourtant, la droite dessinée en vert peut sembler acceptable comme frontière pour cet ensemble de points.

On tient compte de cette erreur en introduisant les variables $(z_i)_{i \leq m}$. Pour que la condition (1) soit toujours vérifiée, il faut que quand $y_i \times f(x_i) \geq 1$, $z_i = 0$ et lorsque $y_i \times f(x_i) < 1$, $z_i = 1 - y_i \times f(x_i)$. Le but étant de minimiser le nombre de ces erreurs, ie. points mal classés, on utilise un paramètre C constant qui permet d'insister plus ou mins sur la minimisation de ces erreurs :

(P) $min_w\ \frac{1}{2} \times \|w\| + C \times \sum_{i \leq m} z_i$
avec $\forall i, z_i \geq 0$
$\forall i, y_i \times (\omega^T x_i) \geq 1 - z_i$

Les fonctions que l'on a introduites sont toutes convexes. Si la dimension des points $(x_i)_i$ est petite, nous allons pouvoir utiliser la méthode de Newton pour résoudre ce problème. On verra par la suite le *kernel trick* qui permettra de ne pas tenir compte de la dimension, mais seulement du nombre d'échantillons $(x_i)_i$.

2 Calcul du dual

Calculons le lagrangien du problème (P). Soit λ le multiplicateur de Lagrange de dimension $1 \times m$:

$$\begin{gathered}\forall w, \lambda \in \mathbb{R}^{2m}, L(\omega, \lambda, z) = \\ = \frac{1}{2}\|w\|^2 + C \times \sum_i z_i - \sum_i \lambda_i \times z_i + \sum_i \lambda_i \times (1 - y_i \omega^T x_i) \\ = \frac{1}{2}\|w\|^2 + C\mathbf{1}^T z - C\lambda^T z + \mathbf{1}^T \lambda - \sum_i \lambda_i \times y_i \omega^T x_i \\ = \frac{1}{2}(\|w - \sum_i \lambda_i y_i x_i\|_2^2 - \|\sum_i \lambda_i y_i x_i\|_2^2) + (C\mathbf{1} - \lambda)^T z + \mathbf{1}^T \lambda\end{gathered}$$

Minimisons L par rapport à ω. Comme le langrangien est convexe en ω, il faut annuler le gradient :

$$\begin{gathered}\nabla_\omega L(\omega, \lambda, z) = \frac{1}{2}(2\omega - 2\sum_i \lambda_i y_i x_i) + 0 = 0 \\ \Leftrightarrow \omega = \sum_i \lambda_i y_i x_i \ (1)\end{gathered}$$

Minimisons L par rapport à z. Comme le langrangien est convexe en z, il faut annuler le gradient :

$$\begin{gathered}\nabla_z L(\omega, \lambda, z) = 0 + C\mathbf{1}{1_{z>0}}^T - \lambda {1_{z>0}}^T = 0 \\ \Leftrightarrow mC - \sum_i \lambda_i = 0 \text{ si } z_i > 0 \\ \Leftrightarrow mC = \sum_i \lambda_i \text{ si } z_i > 0 \ (2)\end{gathered}$$

Le minimum en L par rapport à z a une valeur finie ssi $C\mathbf{1} - \lambda = 0$. On obtient le problème dual en injectant les valeurs de ω et de z dans le lagrangien :

$$\begin{gathered}max_{\lambda \in \mathbb{R}^{+m}} -\frac{1}{2}\|\sum_i \lambda_i y_i x_i\|_2^2 + \mathbf{1}^T \lambda \text{ par } (1) \\ \text{avec } \forall i, 0 \leq \lambda_i \leq C \text{ si } z_i > 0 \text{ (vient de (2))}\end{gathered}$$

On obtient la solution optimale du primal (ω^*, z^*) à partir de celle du dual λ^* :

$$(1)\ \omega^* = \sum_i \lambda_i^* y_i x_i$$

3 Utilisation de l'astuce du noyau (*kernel trick*)

Pour pouvoir trouver efficacement la solution au problème avec la méthode de Newton, il faut s'affranchir de la contrainte quadratique sur la dimension des échantillons. On note X la matrice de *design* (des échantillons), et la matrice du noyau $K = X^T X$, avec $K \geq 0$. On montre alors que le problème dual peut se réécrire de la façon suivante :

$$max\ -\frac{1}{2}\lambda^T diag(y) K diag(y)\lambda + \mathbf{1}^T\lambda$$
$$\text{avec } \forall i, 0 \leq \lambda_i \leq C$$

On remarque que la dimension m des échantillons n'intervient plus, et que donc la complexité de la résolution du problème ne dépend que du nombre d'échantillons.