Datamining 2: Classification supervisée Partie 2: les SVM

M2 STD

September 30, 2015

SVM intro : le cas de classes linéairement séparables

On dit que nos deux classes sont linéairement séparables ssi il existe un vecteur $w \in \mathbb{R}^d$ et un réel w_0 tel quel que

$$\langle X_i, w \rangle + w_0 > 0 \Leftrightarrow Y_i = 1$$

$$\langle X_i, w \rangle + w_0 < 0 \Leftrightarrow Y_i = -1$$

Autrement dit si pour tout i :

$$(\langle X_i, w \rangle + w_0) Y_i > 0$$

SVM intro : le cas de classes linéairement séparables

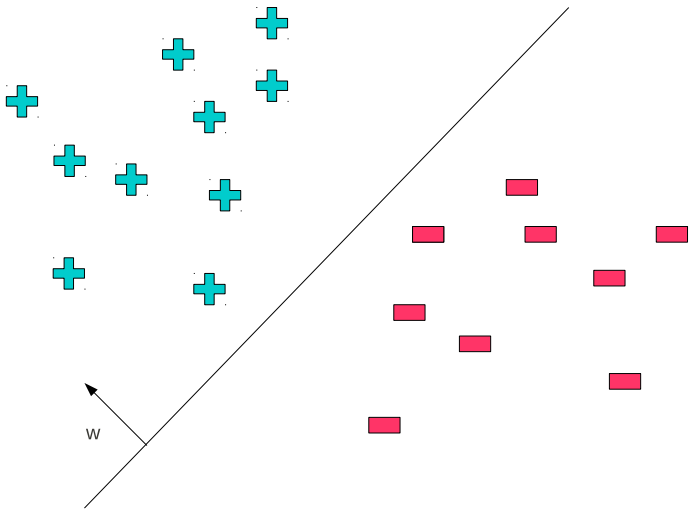


Figure: Illustration du cas linéairement séparable: w est orthogonal au plan séparateur, w_0 dépend de l'origine choisie du repere

SVM intro : le cas de classes linéairement séparables

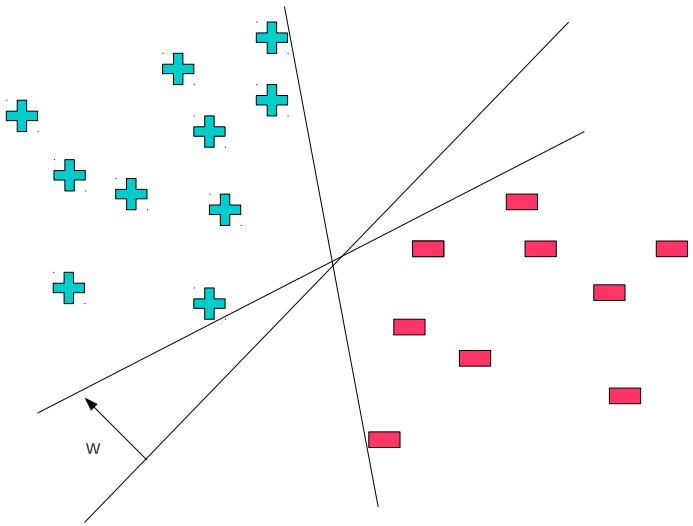


Figure: Bien sur il n'existe pas une seule séparation linéaire

définition de la marge

La marge est la distance du plan séparateur a son observation la plus proche soit $m = \min_i \frac{(\langle X_i, w \rangle + w_0) Y_i}{||w||}$

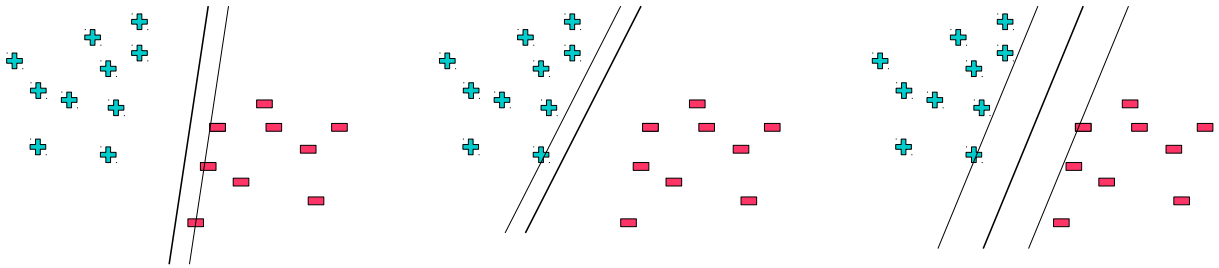


Figure: quelques marges

Théoreme

Le meilleur classifieur linéaire en terme de "généralisation" sera celui qui maximise la marge (Vapnik) (Il y a des vrais math dans ce théoreme)

Algorithme

On cherge a résoudre :

$$\begin{cases} \text{maximiser (en } w \text{ et } w_0 \text{)} : \min_i \dfrac{(\langle X_i, w\rangle + w_0) Y_i}{||w||} \\ \text{sachant : } (\langle X_i, w\rangle + w_0) Y_i > 0 \end{cases}$$

Equivalent a :

$$\begin{cases} \text{minimiser (en } w \text{ et } w_0 \text{)} : \dfrac{1}{2}||w||^2 \\ \text{sachant : } (\langle X_i, w\rangle + w_0) Y_i \geq 1 \end{cases}$$

Algorithme

$$\begin{cases} \text{minimiser (en } w \text{ et } w_0 \text{)} : \frac{1}{2}||w||^2 \\ \text{sachant : } (\langle X_i, w\rangle + w_0) Y_i \geq 1 \end{cases}$$

Equivalent a minimiser le Lagrangien
$L(w, w_0, \alpha) = \frac{1}{2}||w||^2 - \sum \alpha_k (\langle X_i, w\rangle + w_0) Y_i - 1)$

Algorithme

On cherche minimiser le Lagrangien
$L(w, w_0, \alpha) = \frac{1}{2}||w||^2 - \sum \alpha_k(\langle X_i, w\rangle + w_0)Y_i - 1)$.
L'annulation des dérivées partielles donne:

$$\begin{cases} \sum \alpha_k Y_k X_k = w \text{ en dérivant par rapport a } w \\ \sum \alpha_k Y_k = 0 \text{ en dérivant par rapport a } w_0 \end{cases}$$

On réinjecte dans le Lagrangien.... et au final le probleme revient a:

$$\begin{cases} \text{minimiser } : L(\alpha) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j Y_i Y_j \langle X_i X_j \rangle \\ \text{sachant : } \alpha_i \geq 0 \text{ et } \sum \alpha_i Y_i = 0 \end{cases}$$

Algorithme.. résumé

On résoud (numériquement)

$$\begin{cases} \text{minimiser } : L(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j Y_i Y_j \langle X_i X_j \rangle \\ \text{sachant : } \alpha_i \geq 0 \text{ et } \sum \alpha_i Y_i = 0 \end{cases}$$

QUI NE DEPEND (en X) QUE DES $\langle X_i X_j \rangle$ et on réinjecte dans les équations précédentes pour obtenir w et w_0... Remarque : on n'a aucun parametre a regler pour l'instant...

SVM intro : le cas de "presque" classes linéairement séparables

L'hypothese de classe linéairement séparables est fortes ! On peut avoir besoin de la "relacher" un peu les contraintes

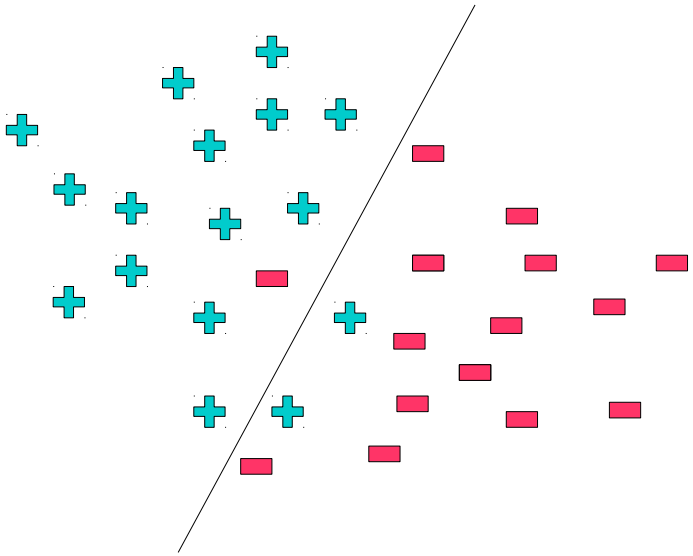


Figure: Le cas "presque" linéairement séparable

SVM intro: les marges souples

marges dures (linéairement séparable)

$$\begin{cases} \text{minimiser (en } w \text{ et } w_0 \text{)} : \frac{1}{2}||w||^2 \\ \text{sachant : } (\langle X_i, w \rangle + w_0) Y_i \geq 1 \end{cases}$$

marges souples ("presque" linéairement séparable)

Pour une constante C (plus C est grande moins on tolère les erreurs de classement) donnée on

$$\begin{cases} \text{minimiser (en } w \text{ et } w_0 \text{)} : \frac{1}{2}||w||^2 + C \sum \varepsilon_i \\ \text{sachant : } (\langle X_i, w \rangle + w_0) Y_i \geq 1 - \varepsilon_i \text{ et } \varepsilon_i \geq 0 \end{cases}$$

marges dures (linéairement séparable)

$$\begin{cases} \text{minimiser } : L(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j Y_i Y_j \langle X_i X_j \rangle \\ \text{sachant : } \alpha_i \geq 0 \text{ et } \sum \alpha_i Y_i = 0 \end{cases}$$

marges souples ("presque" linéairement séparable)

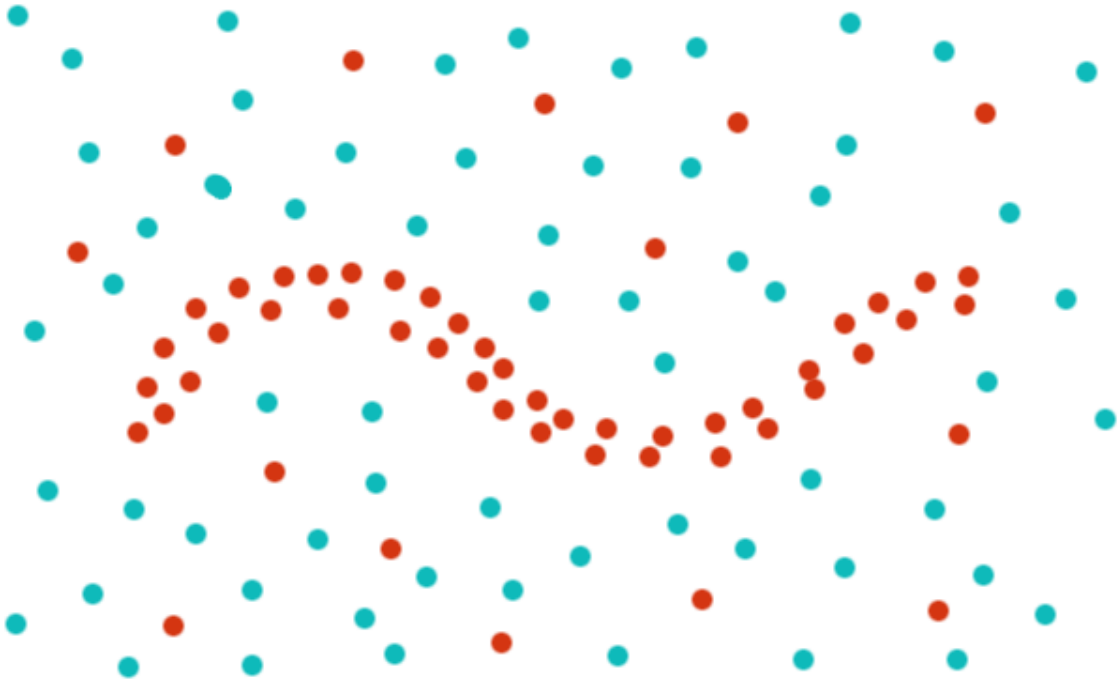
Pour une constante C (plus C est grande moins on tolère les erreurs de classement) donnée on

$$\begin{cases} \text{minimiser } : L(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j Y_i Y_j \langle X_i X_j \rangle \\ \text{sachant : } 0 \leq \alpha_i \leq C \text{ et } \sum \alpha_i Y_i = 0 \end{cases}$$

choix de C (validation, validation croisée...)

SVM : le cas non séparable linéairement

Dans notre exemple "jouet" les classes ne sont pas séparables linéairement...



SVM : le cas non séparable linéairement

Il existe certainement une fonction f qui envoie les X_i sur des Y_i qui sont "presque" linéairement séparables... (Dessin au tableau)
Alors on résoud le problème suivant:

$$\begin{cases} \text{minimiser } : L(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j Y_i Y_j \langle f(X_i) f(X_j) \rangle \\ \text{sachant : } 0 \leq \alpha_i \leq C \text{ et } \sum \alpha_i Y_i = 0 \end{cases}$$

SVM : le "kernel trick"

Il suffit d'avoir les $\langle f(X_i)f(X_j)\rangle = K(X_i, X_j)$ pour résoudre notre probleme... Avec un peu de chance choisir K

a. Gaussien $K(X_i, X_j) = \exp\left(-\frac{||X_i - X_j||^2}{2\sigma^2}\right)$

b. Polynomial $K(X_i, X_j) = (\langle X_i, X_j\rangle)^p$

Va marcher...

Un truc de fou

Il est étonnant mais vrai qu'on a souvent... beaucoup de chance... et que cela fonctionne !

Attention néanmoins on a désormais beaucoup de parametres a selectionner (forme du noyau, parametre du noyau et parametre de marge)