Dupuis Octave
Science des Données
Apprentissage statistique & machine learning

TP/TD Optimisation Classique 2021 - Majeure Science des Données

Première Partie : Cas linéaire :

La méthode de la SVM revient à chercher un hyperplan de R^n qui « sépare » au mieux tous les points de notre ensemble. C'est-à-dire que cet hyperplan devra maximiser la **marge** ie *la demi-largeur de la bande symétrique de largeur maximale autour de l'hyperplan qui ne contient aucun point.*

On rappelle qu'un hyperplan de R^n peut être vu comme le noyau d'une forme linéaire non nulle, dès lors on écrit pour décrire notre hyperplan :

$$f: \begin{cases} R^n \to R \\ x \to w^T x + b \end{cases}$$

avec $w \in R^n$ et $b \in R$, les paramètres du modèle que l'on cherchera à déterminer.

La distance d'un point x_k à l'hyperplan que l'on vient de décrire, s'exprime de la manière suivante :

$$d(x_k, Hyperplan) = \frac{l_k(w^T x_k + b)}{\|w\|}$$

Les points vérifiant $l_k(w^T x_k + b) \geq 1$ sont situés hors d'une marge de demi-largeur $\frac{1}{\|w\|}$, que l'on voudra donc maximiser. Ce qui revient à minimiser $\|w\|$. On a ici introduit l_k, le label du point qui prend comme valeur 1 ou -1 dans la mesure où la classification que l'on effectue ici est binaire.

Dès lors on considère la fonction objectif :

$$\forall k \in (1, \dots, p) Min\ \frac{1}{2}\|w\|^2, avec\ la\ contrainte\ \left(1 - l_k(w^T x_k + b)\right) \leq 0$$

Le lagrangien de ce problème s'exprime de la manière suivante en supposant que les (α_k) sont tous positifs :

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 + \sum_{k=1}^{p} \alpha_k (1 - l_k(w^T x_k + b)$$

Cela nous permet d'écrire la fonction duale $H(\alpha) = L(x, b, \alpha)$.

On cherche le minimum de cette fonction. On utilise dès lors le gradient de L selon w et selon b. On obtient dès lors pour ce point extrême :

$$\nabla_w L(w, b, \alpha) = w_\alpha - \sum_{k=1}^{p} \alpha_k\, l_k x_k = 0$$

$$\nabla_b L(w, b, \alpha) = \sum_{k=1}^{p} \alpha_k\, l_k = 0$$

Après développement, on résume le problème dual suivant :

- La fonction objectif : Max($H(\alpha) = \sum_{k=1}^{p} \alpha_k - \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} \alpha_i\alpha_j\, l_i l_j {x_i}^T x_j$)

Sous forme matricielle : $H(\alpha) = -\frac{1}{2}\alpha^T A\alpha + u^T\alpha$, où $u = (1,1,\dots,1) \in R^p$ et $(A_{i,j}) = l_i l_j {x_i}^T x_j$

- Les contraintes : $\sum_{k=1}^{p} \alpha_k\, l_k = 0$ et $\forall k \in (1,\dots,p), \alpha_k \geq 0$.

La résolution de ce problème donne une valeur α^*, qui nous permet de trouver les valeurs w et b :

$$w = \sum_{k=1}^{p} l_k \alpha_k^* x_k$$

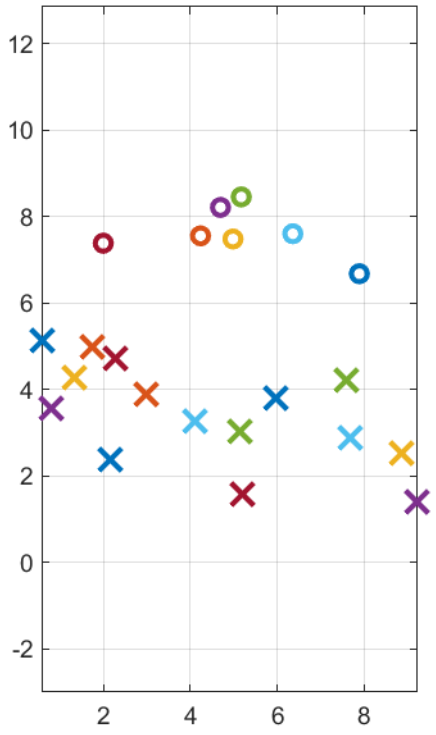
On peut ensuite déterminer b à l'aide des contraintes actives. Celles-ci correspondent aux α^*_k qui sont non nuls, ce qui correspond à une nullité des contraintes du problème primal. Il suffit donc

simplement de résoudre les équations :

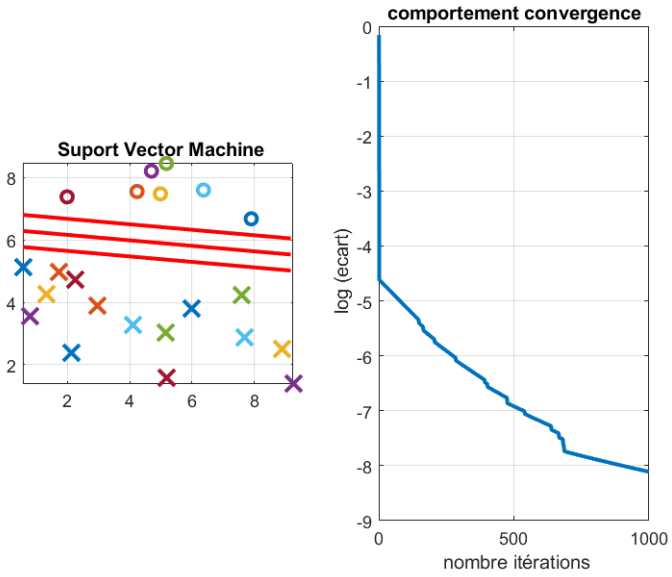
$$\left(1 - l_k(w^T x_k + b)\right) = 0,\ \forall k\ \alpha^*_k \neq 0$$

Suivant le nombre d'équations, on obtient plusieurs fois la valeur de b. On prend pour le b final, la moyenne de ces valeurs afin de minimiser les erreurs. Cette méthode permet de même d'améliorer le résultat obtenu.

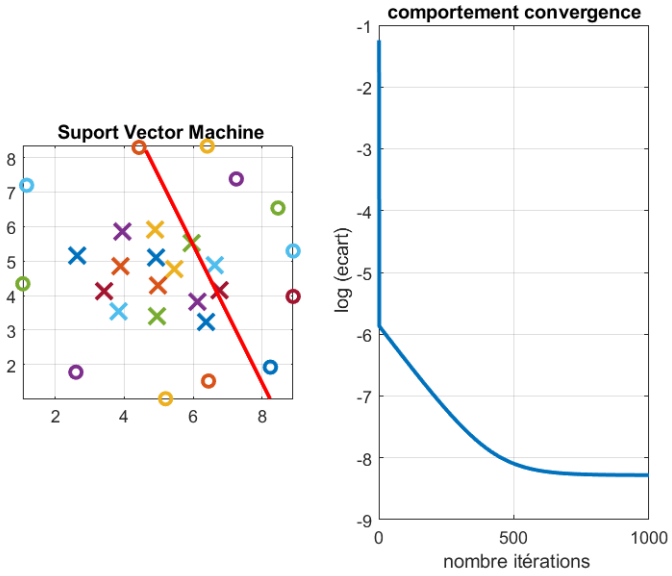
Cette méthode permet de classifier des individus, dans des cas relativement simples comme suit :



On obtient alors une séparation des individus suivant le plan en rouge :



Cependant, pour des situations plus complexes, lorsque les individus ne sont pas séparables par des plans (voir ci-dessous), cette méthode n'est pas efficace.



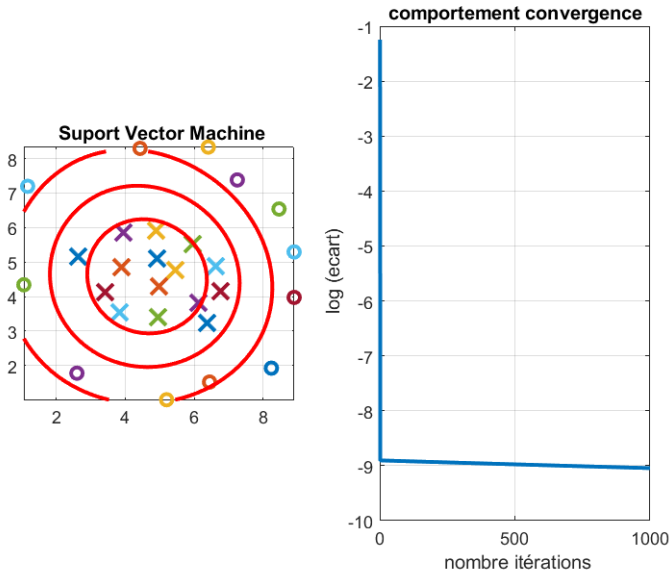
Deuxième partie : cas non linéaire, noyaux :

Dans ce cas, les points ne sont pas directement séparables par un hyperplan. Le principe ici va être de projeter les points dans un espace de dimension supérieure où les futurs points seront séparables, à l'aide d'un noyau. Un noyau est une application linéaire, symétrique définie et positive.

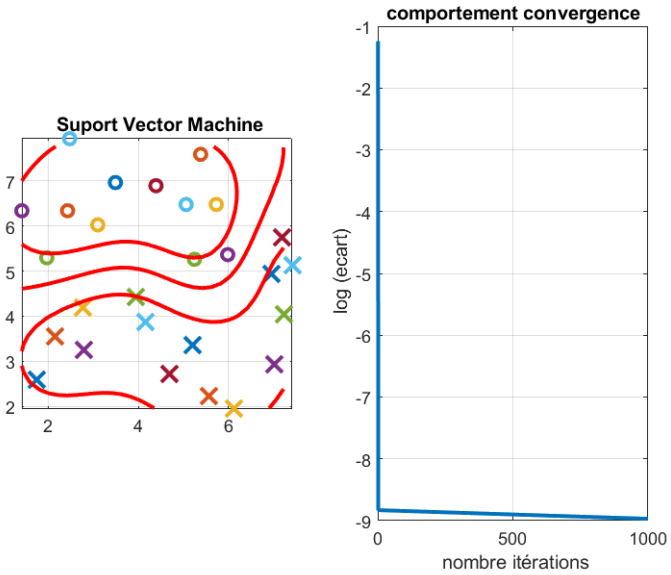
La transformation précédemment évoquée nécessite un formalisme rigoureux autour de la création de la définition d'un espace de Hilbert.

Ce qu'il convient de retenir est que la méthode de calcul reste sensiblement la même, si ce n'est l'introduction d'un noyau pour le calcul du produit scalaire.

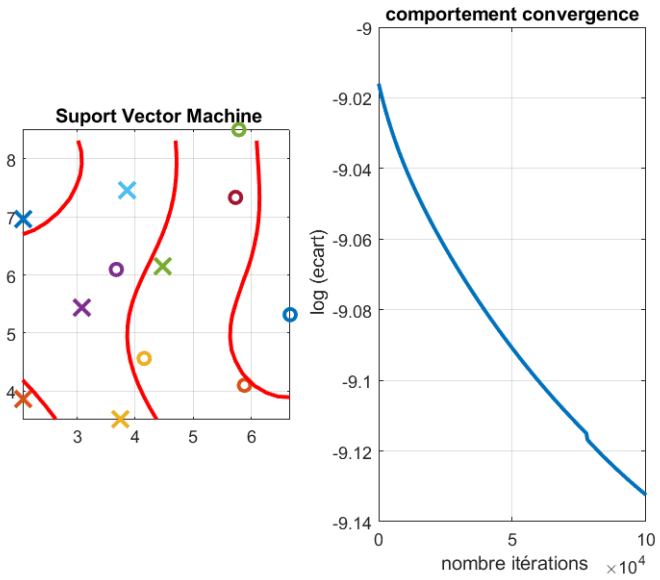
En choisissant le bon noyau et en ajustant ses paramètre, on parvient à résoudre le problème de la partie précédente :



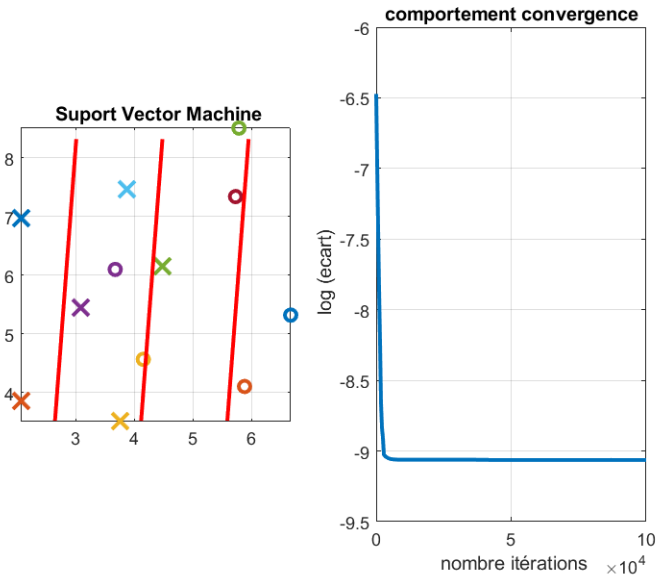
Cette méthode permet de résoudre d'autres types de problèmes de classifications qui ne pourrait pas être résolus linéairement :



Cependant, nous sommes jusqu'à présent dans des cas encore simples : les données sont séparables facilement et de façon parfaite. Mais si ce n'est pas le cas (cf figure ci-dessous), cette méthode n'est plus efficace, avec un souci qui s rapproche du surapprentissage (en voulant séparer les données au mieux, on obtient un modèle faux).



Dans ce dernier cas, c'est un modèle linéaire qui conviendrai, cependant, à cause des points aberrants, l'algorithme ne parvient pas à trouver le plan adéquat :



Troisième partie : SVM à marge souple (cas non séparable) :

On utilise le même raisonnement que dans la première partie en introduisant une variable d'écart $\{\xi_i\}, \forall i \in (1, \dots, p)$. Cette dernière permet cette fois à un point d'appartenir à une marge assouplie. Plus ξ_i est grande, moins le point est bien classé.

En termes de mise en équation, la contrainte revient à :

$$\forall\, k \in (1, \dots, p): l_k(< w^T, x_{k>} + b) \geq 1 - \xi_i$$

En notant $C \geq 0$ un terme quantifiant l'importance accordée au respect des contraintes. Le Lagrangien du problème est :

$$L(w,b,\xi,\beta) = \frac{1}{2}\|w\|^2 + C\sum_{k=1}^{p}\xi_k + \sum_{k=1}^{p}\alpha_k(1 - \xi_k - l_k(< w^T, x_{k>} + b)) - \sum_{k=1}^{p}\beta_k\xi_k$$

En exploitant l'annulation du Lagrangien pour les variables w, b et ξ, il vient :

$$\begin{cases} w = \sum_{k=1}^{p} \alpha_k l_k x_k \\ \sum_{k=1}^{p} \alpha_k l_k = 0 \\ C - \alpha_k - \beta_k = 0 \end{cases}$$

Le calcul de la fonction dual donne :

$$H(\alpha) = -\frac{1}{2}\alpha^T A\alpha + C\sum_{k=1}^{p}\xi_k + \sum_{k=1}^{p}\alpha_k(1 - \xi_k) - \sum_{k=1}^{p}\xi_k(C - \alpha_k)$$

$= -\frac{1}{2}\alpha^T A\alpha + \alpha^T u$, en exploitant la relation entre α et β.

On procède comme précédemment en identifiant les points x_i pour lesquels α_k est non nul. Pour ces points, la contrainte est active donc :

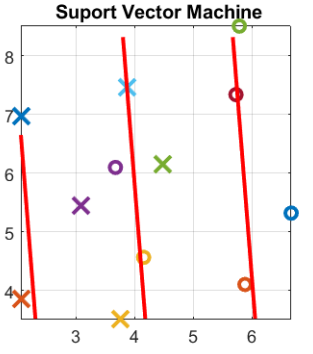
$$1 - \xi_i - l_k(< w^T, x_{k>} + b) = 0$$

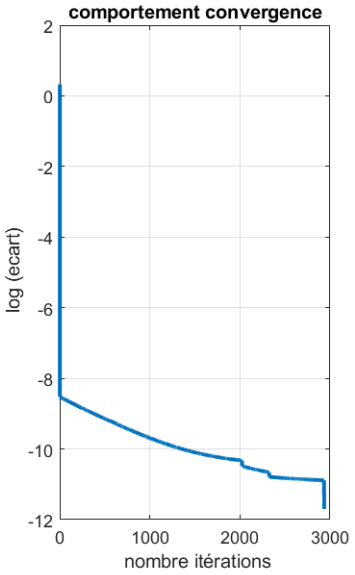
Il est impossible de calculer b à cause de la variable d'écart.

Dès lors : comme $\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0$,

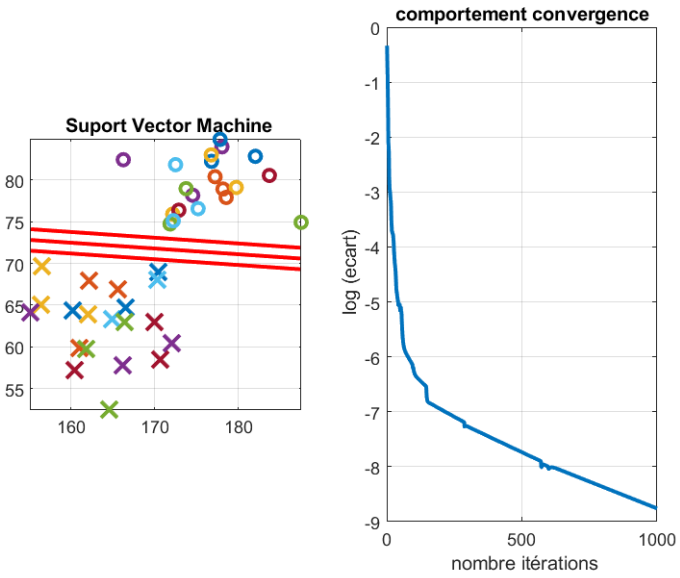
Si $\alpha_i \neq C, \beta_i \neq 0$ et donc $\xi_i = 0$, ce qui permet ainsi de calculer b.

Ce nouveau modèle permet de pallier le problème mis en évidence plus haut ; on obtient alors de meilleurs résultats :





On a désormais tous les outils nécessaires à la classifications des mensurations. Cette fois, les données ne sont plus en deux dimensions, ce qui rend impossible la représentation. Cependant, on remarque tout de même qu'en ne travaillant qu'avec les deux premères dimensions, on peut séparer parfaitement les données d'apprentissages avec un modèle linéaire :



Voici les prédictions obtenus pour les données partagées par les étudiants :

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	0	1	1	0	1	0	1	1	0	0	0	0	1

La première ligne correspond à l'indice de l'étudiant, qui correspond à son pseudo comme suit, tandis que la seconde ligne correspond à la classe de chaque étudiant.

1	Woody
2	violet
3	ICM
4	DATA_Master
5	Major
6	Gg
7	co
8	panda
9	Alex
10	Sam
11	Mat
12	Clover
13	Tigrou
14	ABC
15	Tiger

On effectue le même travail, mais cette fois en utilisant un modèle non linéaire à marge souple, et basé sur la totalistée des données :

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0	0	1	1	0	1	0	1	1	0	0	0	0	1

On observe qu'il s'agit presque des mêmes résultats, ce qui est réconfortant.