

Fouille
de
Données Massives
Master 2 Informatique - SISE

Guillaume Metzler

Université Lumière Lyon 2
Laboratoire ERIC, UR 3083, Lyon

guillaume.metzler@univ-lyon2.fr

Automne 2022

Les Séparateurs à Vaste Marge (SVM)

A propos des SVM I

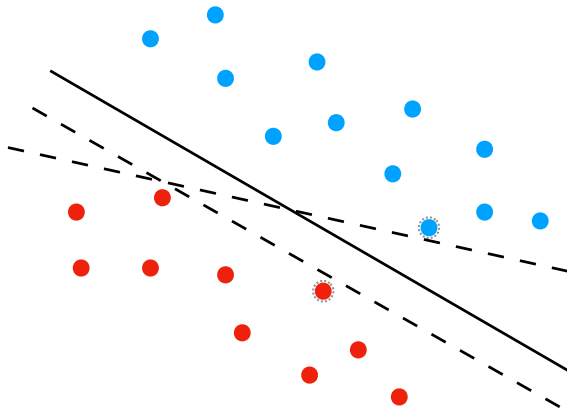
Certainement l'un des algorithmes les plus répandues en apprentissage machine [Vapnik and Cortes, 1995] pour effectuer des tâches de classification binaires.

Idée : apprendre une hypothèse h dont le but est de prédire l'étiquette d'une donnée. Elle se présente sous la forme d'un hyperplan (affine) qui va séparer l'espace en deux : $\{-1, +1\}$:

$$h(\mathbf{x}) = \text{sign}[\langle \mathbf{w}, \mathbf{x} \rangle + b] = \begin{cases} -1 & \text{if } \langle \mathbf{w}, \mathbf{x} \rangle + b < 0, \\ +1 & \text{if } \langle \mathbf{w}, \mathbf{x} \rangle + b > 0. \end{cases}$$

Problème : plusieurs plans peuvent être solutions et séparer parfaitement nos données.

A propos des SVM II



Quel séparateur choisir et pourquoi ?

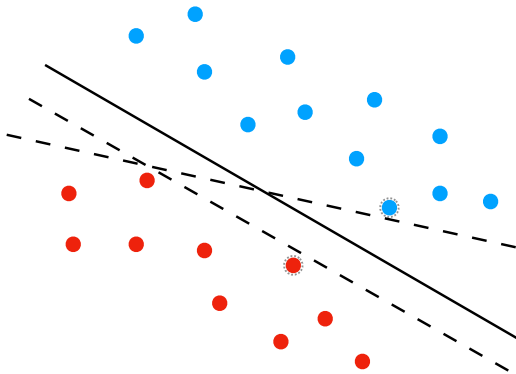
A propos des SVM III

L'idée présentée par [Boser et al., 1992] est de choisir le modèle qui se trouve à la plus grande distance entre les deux classes.

Mais pourquoi choisir un tel modèle ?

Il suffit de se rappeler de notre objectif en Machine Learning, *on souhaite avoir un modèle qui est performant et avec de bonnes capacités en généralisation*

A propos des SVM IV



Cela est vérifié pour la droite en trait plein, mais pas pour celles en trait pointillé. On retient un modèle qui présente **la plus grande marge**.

A propos des SVM V

Comment est-ce que l'on peut définir cette notion de marge? Nous avons dit que c'est la distance qui sépare les exemples de classe opposée.

Si on reprend les équations de notre SVM, on pourra définir la marge comme la distance entre les deux droites d'équations :

$$\langle \mathbf{w}, \mathbf{x} \rangle + b = \pm \rho.$$

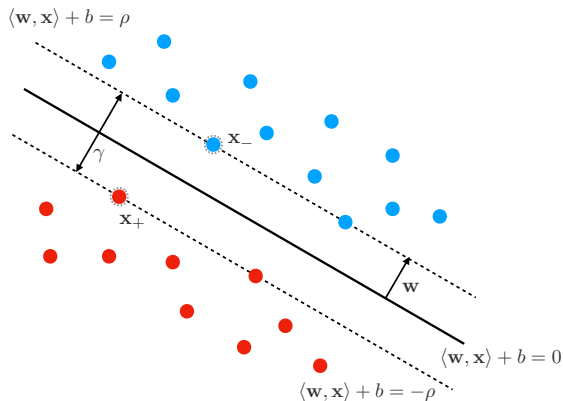
Ce que l'on peut également réécrire

$$\langle \mathbf{w}, \mathbf{x} \rangle + b = \pm 1,$$

en normalisant les paramètres.

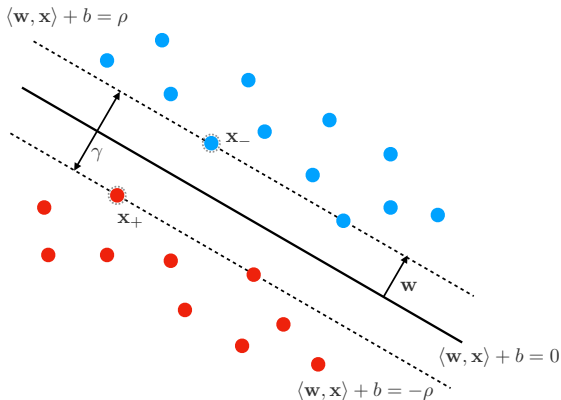
A propos des SVM VI

On considère maintenant deux points x_- et x_+ qui se trouvent sur les deux droites (ou plus généralement hyperplans) représentés ci-dessous



A propos des SVM VII

Sur ce même dessin, on définit également la marge γ de notre séparateur linéaire.



A propos des SVM VIII

Chacun des vecteurs \mathbf{x}_+ et \mathbf{x}_- peuvent se décomposer comme la somme d'un vecteur qui se *colinéaire* à \mathbf{w} (noté \mathbf{x}^w) et d'un vecteur qui sera orthogonal à \mathbf{w}^{\perp} .

Ce qui nous amène aux relations suivantes :

$$\begin{aligned}
 h(\mathbf{x}_+) - h(\mathbf{x}_-) &= 2, \\
 \langle \mathbf{w}, \mathbf{x}_+ \rangle + b - (\langle \mathbf{w}, \mathbf{x}_- \rangle + b) &= 2, \\
 \langle \mathbf{w}, \mathbf{x}_+^w \rangle + \underbrace{\langle \mathbf{w}, \mathbf{x}_+^{w\perp} \rangle}_{=0} + b - (\langle \mathbf{w}, \mathbf{x}_-^w \rangle + \underbrace{\langle \mathbf{w}, \mathbf{x}_-^{w\perp} \rangle}_{=0} + b) &= 2, \\
 \langle \mathbf{w}, \mathbf{x}_+^w \rangle - \langle \mathbf{w}, \mathbf{x}_-^w \rangle &= 2. \quad (1)
 \end{aligned}$$

A propos des SVM IX

De plus, la marge γ , *i.e.* la distance entre nos deux hyperplans, est définie comme le coefficient de la projection du vecteur $\mathbf{x}_+ - \mathbf{x}_-$ sur le vecteur unitaire $\frac{\mathbf{w}}{\|\mathbf{w}\|_2}$, *i.e.*

$$\gamma = \frac{\langle \mathbf{x}_+ - \mathbf{x}_-, \mathbf{w} \rangle}{\|\mathbf{w}\|_2}.$$

On va maintenant réécrire ce produit scalaire et réutiliser la définition de nos deux hyperplans.

A propos des SVM X

Ainsi, nous avons :

$$\gamma = \frac{\langle \mathbf{x}_+ - \mathbf{x}_-, \mathbf{w} \rangle}{\|\mathbf{w}\|_2},$$

$$= \frac{1}{\|\mathbf{w}\|_2} \underbrace{\langle \mathbf{x}_+ - \mathbf{x}_-, \mathbf{w} \rangle},$$

↓ en développant et en conservant uniquement la partie colinéaire

$$= \frac{1}{\|\mathbf{w}\|} \underbrace{\langle \mathbf{w}, \mathbf{x}_+^{\mathbf{w}} \rangle - \langle \mathbf{w}, \mathbf{x}_-^{\mathbf{w}} \rangle},$$

↓ en utilisant l'équation (1)

$$\gamma = \frac{2}{\|\mathbf{w}\|_2}.$$

A propos des SVM XI

Ainsi, maximiser notre marge γ revient à minimiser la quantité $\frac{\|\mathbf{w}\|_2}{2}$ ou encore $\frac{\|\mathbf{w}\|_2^2}{2}$.

L'ajout d'un carré sur la norme est plus pour des raisons "pratiques" et permet de s'affranchir de la racine carré dans la résolution du problème d'optimisation.

Le problème de minimisation associé est appelé *hard margin SVM* et est défini par :

A propos des SVM XII

Définition 1.1: Hard Margin SVM

Soit $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ un ensemble de m exemples (individus). Alors le meilleur *Séparateur à Vaste Marge* que l'on puisse obtenir est la solution du problème d'optimisation suivant :

$$\begin{aligned} \min_{(\mathbf{w}, b) \in \mathbb{R}^{d+1}} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1, \quad \text{for all } i = 1, \dots, m. \end{aligned}$$

où la contrainte d'inégalité n'est rien d'autre qu'une écriture "synthétique" de celle présentée en équation (3).

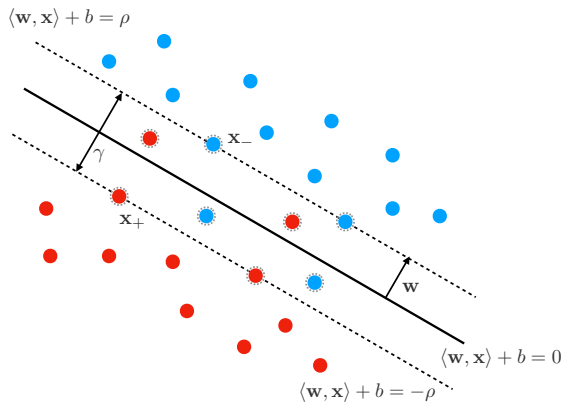
A propos des SVM XIII

Ce problème peut facilement être résolu par des procédés classiques et cela sans grande difficultés.

On pourra par exemple regarder les packages **e1071** de  ou encore la librairie **libsvm** ou encore **linearSVC** sous Python.

Mais le problème présenté ici est un cas idyllique où l'on a supposé que nos données sont linéairement séparables, mais cela n'arrive jamais en pratique.

A propos des SVM XIV



A propos des SVM XV

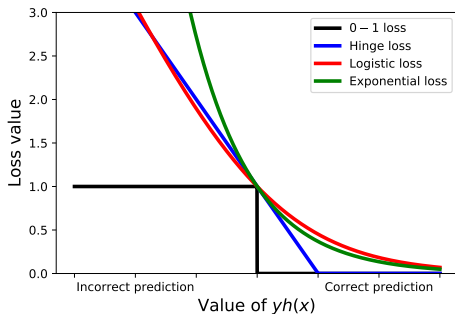
On voit que certains points violent la contrainte

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1.$$

Ce qui peut signifier deux choses :

- votre point se trouve du mauvais côté de l'hyperplan.
- votre point se trouve du bon côté de l'hyperplan mais à l'intérieur de la marge.

A propos des SVM XVI



A propos des SVM XVII

Ainsi, dans cette situation, le problème d'optimisation que nous avons précédemment présenté n'admettrait pas de solutions car il est impossible de mettre tous les points du bon côté

$$\begin{aligned} \min_{(\mathbf{w}, b) \in \mathbb{R}^{d+1}} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1, \quad \text{for all } i = 1, \dots, m. \end{aligned}$$

On doit donc **relaxer** ou **relâcher** notre problème de façon à lui autoriser à faire des erreurs.

A propos des SVM XVIII

Ces dernières vont prendre la forme de variables dites *slacks* $\xi = (\xi_1, \dots, \xi_m)$ que l'on va inclure dans le problème d'optimisation. Ces variables prennent des valeurs positives si les contraintes ne sont pas respectées et nulle dans le cas contraire.

Ces variables vont directement injectées dans le problème d'optimisation de façon à autoriser assouplir la contrainte et l'objectif sera alors de trouver un **compromis** entre la *maximisation de la marge du modèle* et la *minimisation du taux d'erreur*.

A propos des SVM XIX

Définition 1.2: Soft Margin SVM

Soit $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ un échantillon de taille m et notons $\xi = (\xi_1, \dots, \xi_m)$ le vecteur des variables dites *slack*. Alors le meilleur séparateur linéaire que l'on puisse trouver au sens de l'algorithme SVM est solution du problème d'optimisation suivant :

$$\begin{aligned} \min_{\xi \in \mathbb{R}^m, (\mathbf{w}, b) \in \mathbb{R}^{d+1}} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 + \frac{C}{m} \sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i, \quad \text{for all } i = 1, \dots, m, \\ & \xi_i \geq 0, \quad \text{for all } i = 1, \dots, m. \end{aligned} \tag{2}$$

A propos des SVM XX

Comparée à la Définition 1.1, un terme en $\frac{C}{m} \sum_{i=1}^m \xi_i$ a été ajouté et il représente le coût de violation des contraintes.

Le juste équilibre entre le contrôle de l'erreur et la maximisation de la marge va se faire avec un hyper-paramètres C qui devra alors être tuné lors du processus d'apprentissage.

Travailler avec cette formulation peut se révéler cependant difficile, et il n'est pas rare de réécrire le problème.

A propos des SVM XXI

Proposition 1.1: Formulation équivalente du Soft Margin SVM

Soit $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ un ensemble de données de taille m et soit $\xi = (\xi_1, \dots, \xi_m)$ le vecteur des variables slacks. Si les contraintes sont prises en compte et directement injectées dans le problème d'optimisation, alors le problème 2 peut se réécrire :

$$\min_{(\mathbf{w}, b) \in \mathbb{R}^{d+1}} \frac{1}{2} \|\mathbf{w}\|_2^2 + \frac{C}{m} \sum_{i=1}^m [1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)]_+,$$

où $[x]_+ = \max(0, x)$, *i.e.* représente la hinge loss.

A propos des SVM XXII

Démonstration

Pour montrer ce résultat, on a uniquement besoin de se focaliser sur les variables *slacks*, qui, rappelons le, sont des variables positives d'après la deuxième contrainte de notre définition des *Soft SVM*.

Si on se concentre sur la première contrainte, on remarque ξ_i est nulle si et seulement si $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$, sinon elle est égale à la différence, *i.e.*

$$\forall i = 1, \dots, m, \xi_i = \begin{cases} 0 & \text{si } 1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \leq 0, \\ 1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) & \text{sinon.} \end{cases}$$

Ce que l'on peut synthétiser par

$$\forall i = 1, \dots, m, \xi_i = [1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)]_+$$

A propos des SVM XXIII

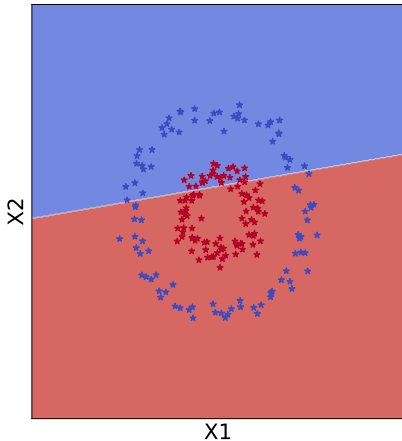
Bien que ce problème soit convexe, il reste très compliqué à optimiser car la fonction dont on cherche le minimum n'est pas lisse.

De plus, la résolution du problème est d'autant plus complexe que nos données possèdent de descripteurs. Cette complexité est en $\mathcal{O}(d^3)$, ce qui rend le problème très long à résoudre en grande dimension.

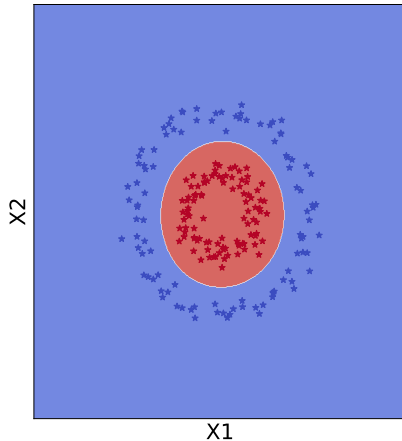
En outre, cela reste un séparateur qui est **linéaire**, son expressivité ou son **pouvoir discriminant** reste donc limité quand notre jeu de données est plus complexe.

A propos des SVM XXIV

SVM with linear kernel



SVM with Gaussian Kernel



SVM non linéaire I

L'objectif serait d'apprendre un séparateur qui *a priori* n'est plus forcément linéaire qui serait linéaire mais dans un autre espace ...

Dit autrement, il faudrait être en mesure de trouver une transformation des données qui permette de pouvoir résoudre le problème avec un algorithme de séparation linéaire.

Pour cela on a besoin de considérer le problème "dual" de notre problème d'optimisation initial.

SVM non linéaire II

Proposition 1.2: Problème Dual SVM

Soit $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ un échantillon de m exemples. La formulation duale du problème de Soft Margin SVM présenté en Définition 1.2 est :

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & -\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^m \alpha_i, \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq \frac{C}{m} \quad \forall i = 1, \dots, m, \\ & \sum_{i=1}^m y_i \alpha_i = 0. \end{aligned} \tag{3}$$

Le vecteur $\boldsymbol{\alpha}$ est le vecteur des variables lagrangiennes (ou duales).

SVM non linéaire III

Démonstration

On repart de la définition de départ, afin d'obtenir la version duale de ce problème, nous allons considérer une seule fonction objectif. Cette dernière va regrouper à la fois la quantité que l'on va initialement minimiser mais on y ajoute les contraintes.

Cette nouvelle fonction s'appelle le Lagrangien

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2} \|\mathbf{w}\|_2^2 + \sum_{i=1}^m \alpha_i (1 - \xi_i - y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b)) - \sum_{i=1}^m \beta_i \xi_i.$$

ou α and β sont les variables duales associées aux deux contraintes du problème (primal).

SVM non linéaire IV

On rappelle que si le problème primal est convexe (de même que les contraintes) alors les solutions du problème vérifient les égalités suivantes :

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 0, \quad \frac{\partial \mathcal{L}}{\partial b}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 0,$$

et

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\xi}}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 0.$$

Elles vont nous permettre de trouver une expression des variables primales (*i.e.* \mathbf{w} , b et $\boldsymbol{\xi}$) comme une fonction des variables duales (*i.e.* $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$). C'est ce que l'on appelle les conditions de KKT (Karush-Kuhn-Tucker)

SVM non linéaire V

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 0, \quad \Longleftrightarrow \quad \mathbf{w} - \sum_{i=1}^m y_i \alpha_i \mathbf{x}_i = 0 \quad \Longleftrightarrow \quad \mathbf{w} = \sum_{i=1}^m y_i \alpha_i \mathbf{x}_i \quad (4)$$

$$\frac{\partial \mathcal{L}}{\partial b}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 0, \quad \Longleftrightarrow \quad \sum_{i=1}^m \alpha_i y_i = 0 \quad (5)$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 0 \quad \Longleftrightarrow \quad \frac{C}{m} - \alpha_i - \beta_i = 0, \quad \Longleftrightarrow \quad \alpha_i, \beta_i \geq 0. \quad (6)$$

On utilise enfin ces résultats et on injecte tout dans l'expression du Lagrangien

SVM non linéaire VI

$$\begin{aligned}
& \mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= \underbrace{\frac{1}{2} \|\mathbf{w}\|_2^2}_{\text{blue}} + \sum_{i=1}^m \alpha_i \left(1 - \xi_i - y_i \left(\underbrace{\langle \mathbf{w}, \mathbf{x}_i \rangle + b}_{\text{red}} \right) \right) - \sum_{i=1}^m \beta_i \xi_i, \\
&\downarrow \text{ using Equation (4)} \\
&= \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^m \xi_i \left(\underbrace{\frac{C}{m} - \alpha_i - \beta_i}_{\text{green}} \right) + \sum_{i=1}^m \alpha_i \\
&\quad - \sum_{i=1}^m \alpha_i y_i \left(\underbrace{\sum_{j=1}^m y_j \alpha_j \mathbf{x}_j, \mathbf{x}_i}_{\text{red}} \right) + b \underbrace{\sum_{i=1}^m \alpha_i y_i}_{\text{orange}}, \\
&\downarrow \text{ using Equations (5) and (6)}
\end{aligned}$$

SVM non linéaire VII

$$\begin{aligned}
&= \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^m \xi_i \times 0 + \sum_{i=1}^m \alpha_i \\
&\quad - \sum_{i=1}^m \sum_{j=1}^m y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_j, \mathbf{x}_i \rangle + b \times 0,
\end{aligned}$$

↓ computing the difference

$$= -\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^m \alpha_i.$$

On doit maximiser le Lagrangien par rapport aux variables duales. Enfin les équations(5) and (6) conduisent aux deux contraintes du problème dual.

SVM non linéaire VIII

Garder à l'esprit que le problème dual est un problème strictement concave par rapport aux variables duales. Ainsi, il n'admet qu'une seule et unique solution.

On peut réécrire le problème précédent un peu plus simplement en définissant une matrice dite de *Gram*.

$$\max_{\alpha} -\frac{1}{2}\alpha^T \mathbf{G}\alpha + \sum_{i=1}^m \alpha_i,$$

où \mathbf{G} est la matrice définie par $G_{ij} = y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$. Elle est semi-définie positive, donc notre problème va admettre une seule et unique solution.

SVM non linéaire IX

Remarques

- La résolution du problème dual conduit à trouver une solution dans un espace différent : *espace des individus au lieu de l'espace des variables*.
- La solution fournie par le problème dual est la même que celle donnée par le problème primal ! (on parle de dualité forte)
- Si on connaît la solution du problème dual, il est possible de retrouver les valeurs de w et b correspondantes à l'aide des relation de KKT.

Cette approche sera intéressante intéressante lorsque l'on dispose d'un nombre relativement raisonnable de données par rapport au nombre de variables. Mais cela ne résout pas le problème des données non linéairement séparables. On va devoir employer ce que l'on appel **l'astuce du noyau**.

SVM non linéaire X

Au lieu d'utiliser le produit scalaire standard entre deux exemples, on va définir une fonction $K(\cdot, \cdot)$ qui prendra en entrée deux vecteurs et qui va retourner une valeur réelle.

Une telle fonction s'appelle un noyau et la matrice \mathbf{K} associée doit vérifier les propriétés suivantes : (i) symétrique et (ii) semi-définie positive, *i.e.*

$$(i) \quad \forall (\mathbf{x}, \mathbf{x}') \in \mathbb{R}^d \times \mathbb{R}^d, \text{ nous avons : } K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}', \mathbf{x}),$$

$$(ii) \quad \forall (\mathbf{x}_i, \mathbf{x}_j) \in \mathbb{R}^d \times \mathbb{R}^d \text{ et } \forall \mathbf{c} \in \mathbb{R}^d, \text{ nous avons :}$$

$$\mathbf{c}^T \mathbf{K} \mathbf{c} = \sum_{i=1}^m \sum_{j=1}^m c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0.$$

SVM non linéaire XI

Ces propriétés de la matrice \mathbf{K} (ou de la fonction K) sont importantes et conduisent au résultat suivant [Mercer, 1909].

Théorème 1.1: Théorème de Mercer

Soit \mathcal{X} un compact de \mathbb{R}^d et soit K une forme bilinéaire symétrique semi-définie positive, i.e. un noyau. Alors il existe une base orthogonale $(\Phi_j)_{j \in \mathbb{N}}$ et une suite $(\lambda_j)_{j \in \mathbb{N}}$, où $\lambda_j \geq 0$ pour tout j , telle que :

$$K(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{\infty} \lambda_j \Phi_j(\mathbf{x}) \Phi_j(\mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle,$$

où $\Phi(\mathbf{x}) = (\sqrt{\lambda_1} \Phi_1(\mathbf{x}), \dots, \sqrt{\lambda_j} \Phi_j(\mathbf{x}), \dots)$ est la représentation du vecteur \mathbf{x} dans un nouvel espace.

SVM non linéaire XII

En introduisant K , dans la formulation duale 3, on obtient une formulation généralisée du problème :

$$\begin{aligned} \max_{\alpha} \quad & -\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^m \alpha_i, \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq \frac{C}{m}, & \text{for all } i = 1, \dots, m, \\ & \sum_{i=1}^m \alpha_i = 0, \end{aligned}$$

Il existe une grande série de noyaux que l'on pourrait employer pour nos tâches de classification (ou même d'estimation de densité), voir [Genton, 2002] pour une liste non exhaustive de tels noyaux.

SVM non linéaire XIII

Exemples de noyaux

- **Noyau linéaire** : $K(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle$, c'est le noyau le plus classique mais aux capacités limitées.
- **Noyau polynomial** : $K(\mathbf{x}, \mathbf{x}') = (\langle \mathbf{x}, \mathbf{x}' \rangle + c)^p$, où c est une constante
- **Noyau Gaussien** : $K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\sigma^2}\right)$, où σ est un hyper-paramètre que l'on doit tuner.

Il contrôle l'importance que l'on va donner à la similarité entre deux exemples. Plus cette valeur est grande, moins on accorde d'importance à la distance entre les points, on aura tendance à lisser la frontière de décision.

SVM non linéaire XIV

Quelques exemples

On considère deux vecteurs \mathbf{x} et \mathbf{z} de \mathbb{R}^2 et noyau polynomial de degré 2. On va essayer de réécrire ce noyau comme un produit scalaire entre deux vecteurs *i.e.* comme une fonction Φ , telle que $K(\mathbf{x}, \mathbf{z}) = \Phi(\mathbf{x})^T \Phi(\mathbf{z})$.

$$\begin{aligned} K(\mathbf{x}, \mathbf{z}) &= (x_1 z_1 + x_2 z_2 + c)^2, \\ &= c^2 + x_1^2 z_1^2 + x_2^2 z_2^2 + 2c x_1 z_1 + 2c x_2 z_2 + 2x_1 z_1 x_2 z_2, \\ &= \left(c, x_1^2, \sqrt{2c} x_1, \sqrt{2c} x_2, \sqrt{2c} x_1 x_2 \right)^T \left(c, z_1^2, \sqrt{2c} z_1, \sqrt{2c} z_2, \sqrt{2c} z_1 z_2 \right) \\ &= \Phi(\mathbf{x})^T \Phi(\mathbf{z}), \end{aligned}$$

où $\Phi : \mathbf{x} \mapsto (c, x_1^2, \sqrt{2c} x_1, \sqrt{2c} x_2, \sqrt{2c} x_1 x_2)$.

En utilisant ce noyau, on projette implicitement nos données dans un espace de dimension 5.

SVM non linéaire XV

Phase de prédiction

Avec le SVM linéaire (hard or soft) dans sa version primale, la prédiction du label d'un nouvel exemple \mathbf{x}' se fait via le calcul suivant :

$$h(\mathbf{x}') = \text{sign}(\langle \mathbf{w}, \mathbf{x}' \rangle + b).$$

Si on emploie a version dite "kernelisée", *i.e.* en passant par la résolution du problème duale, notre prédicteur va prendre la forme suivante :

$$h(\mathbf{x}') = \text{sign} \left(\sum_{i=1}^m \alpha_i y_i K(\mathbf{x}', \mathbf{x}_i) \right).$$

Noter que dans ce dernier cas, il est nécessaire de calculer la similarité du nouveau point à l'ensemble des points d'apprentissage !

SVM non linéaire XVI

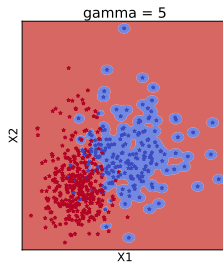
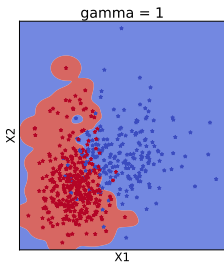
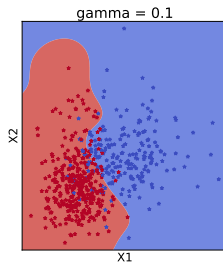
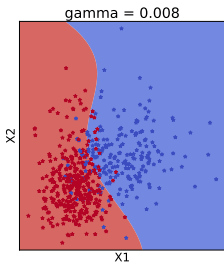
Quelques remarques

La librairie *sklearn* abrite une fonction **SVM** qui vous permettra d'implémenter des SVM gaussiens avec une approche "linéaire", "polynomiale" ou encore "gaussienne"

```
from sklearn import svm
# SVM avec noyau linéaire
svm.SVC(C = 1, kernel = 'linear')
# SVM avec noyau polynomial
svm.SVC(C = 1, degree = 2 , kernel = 'poly')
# SVM avec noyau gaussien
svm.SVC(C=1, kernel='rbf', gamma=0.1)
```

Attention $\gamma = 1/\sigma^2$.

SVM non linéaire XVII

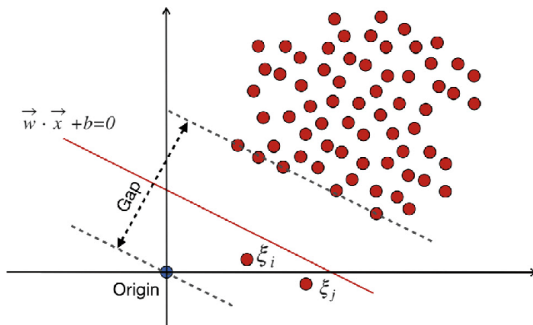


One Class SVM

Une version non supervisée I

Regardons maintenant une méthode que l'on appelle les One Class SVM, qui permettent de faire de la détection d'anomalies.

On peut voir cette tâche comme un problème de classification binaire mais non supervisé. C'est donc une version non supervisée de l'algorithme des SVM



Une version non supervisée II

Dans son fonctionnement, ou plutôt dans sa représentation et sa fonction, on reconnaît bien l'algorithme des SVM.

L'objectif est très différent de ce que nous avons vu jusqu'à présent. Il s'agit ici de séparer les individus en deux classes avec d'un côté, les exemples "normaux" et d'un autre côté les exemples "aberrants" ou "anormaux"

Cela peut notamment servir, quand on ne peut le voir visuellement, à supprimer des données qui ont une distribution qui ne coïncide pas avec la distribution de la majorité des données (imaginons des données qui se trouvent dans les queues d'une gaussienne multi-variée par exemple).

Mais regardons déjà un problème plus ancien ... vraiment ancien ... le problème du cercle minimum [Sylvester, 1857]

Une version non supervisée III

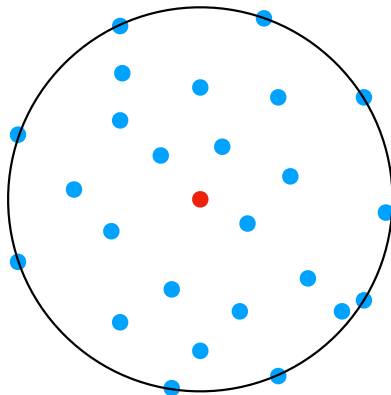
Formulation du problème

Etant donné un ensemble de m points $S = \{\mathbf{x}_i\}_{i=1}^m$ non étiquetés, trouvez le centre \mathbf{c} et le rayon R de la plus petite sphère contenant l'ensemble des points de S .

On peut résoudre cette tâche en résolvant le problème d'optimisation suivant :

$$\begin{aligned} \min_{\mathbf{c}, R} \quad & R^2, \\ \text{s.t.} \quad & \|\mathbf{x}_i - \mathbf{c}\|^2 \leq R^2, \quad \forall i = 1, \dots, m, \end{aligned}$$

Une version non supervisée IV



Une version non supervisée V

Formulation équivalente

Les deux formulations suivantes sont équivalentes [Elzinga and Hearn, 1972] :

$$\begin{aligned} \min_{\mathbf{c}, R} \quad & R^2, \\ \text{s.t.} \quad & \|\mathbf{x}_i - \mathbf{c}\|^2 \leq R^2, \quad \forall i = 1, \dots, m, \end{aligned}$$

et

$$\begin{aligned} \min_{\rho, \mathbf{w}} \quad & \frac{1}{2} \|\mathbf{w}\|^2 - \rho, \\ \text{s.t.} \quad & \mathbf{w}^T \mathbf{x}_i \geq \rho + \frac{1}{2} \|\mathbf{x}_i\|^2, \quad \forall i = 1, \dots, m, \end{aligned}$$

où $\rho = \frac{1}{2} (\|\mathbf{c}\|^2 - R^2)$ et $\mathbf{c} = \mathbf{w}$

Exercice : montrer l'équivalence entre les deux formulations

Une version non supervisée VI

Cette dernière formulation est plus connue sous le nom de **SVDD : Support Vector Data Description** [Tax and Duin, 1999, Tax and Duin, 2004]

$$\begin{aligned} \min_{\rho, \mathbf{w}} \quad & \frac{1}{2} \|\mathbf{w}\|^2 - \rho, \\ \text{s.t.} \quad & \mathbf{w}^T \mathbf{x}_i \geq \rho + \frac{1}{2} \|\mathbf{x}_i\|^2, \quad \forall i = 1, \dots, m, \end{aligned}$$

Dans le cas où l'on suppose toutes nos données \mathbf{x}_i de norme constante, *i.e.* $\|\mathbf{x}_i\|^2 = \beta$, on retrouve alors la formulation des One Class SVM [Scholkopf and Smola, 2001]

$$\begin{aligned} \min_{\mathbf{w}, \rho'} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 - \rho', \\ \text{s.t.} \quad & \mathbf{w}^T \mathbf{x}_i \geq \rho', \quad \forall i = 1, \dots, m, \end{aligned}$$

où $\rho' = \rho + \frac{1}{2} \|\mathbf{x}_i\|^2$.

Une version non supervisée VII

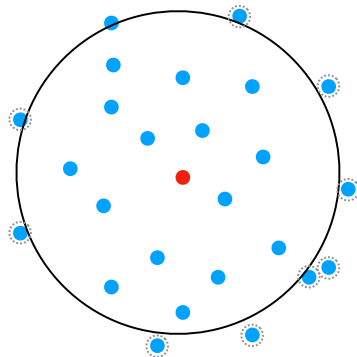
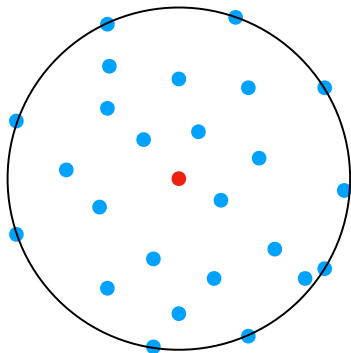
Vers de la détection d'anomalies

$$\begin{aligned} \min_{\mathbf{c}, R} \quad & R^2 + \frac{C}{m} \sum_{i=1}^m \xi_i, \\ \text{s.t.} \quad & \|\mathbf{x}_i - \mathbf{c}\|^2 \leq R^2 + \xi_i, \quad \forall i = 1, \dots, m, \\ & \xi_i \geq 0, \quad \forall i = 1, \dots, m. \end{aligned}$$

Ajout de variables dites "slacks" qui prennent en compte les erreurs de l'algorithme, *i.e.* on autorise certains points à se trouver en dehors du cercle.

Les données qui se trouvent en dehors du cercle sont considérés comme des anomalies

Une version non supervisée VIII



● data

● center

● errors

Une version non supervisée IX

One class SVM

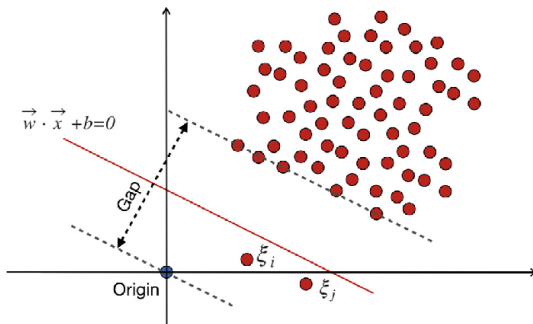
$$\begin{aligned} \min_{\mathbf{w}, \rho'} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 + \frac{1}{m\mu} \sum_{i=1}^m \xi_i - \rho', \\ \text{s.t.} \quad & \mathbf{w}^T \mathbf{x}_i \geq \rho' - \xi_i, \quad \forall i = 1, \dots, m, \\ & \xi_i \geq 0, \quad \forall i = 1, \dots, m. \end{aligned}$$

Le paramètre ρ' contrôle la distance maximale autorisée à notre hyperplan et le paramètre μ donne une borne supérieure sur le pourcentage d'anomalies dans le jeu de données mais permet aussi de contrôler de nombre de vecteurs supports (borne inférieure)

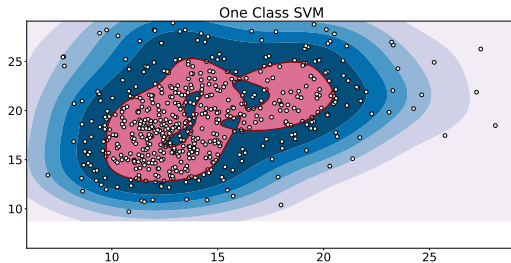
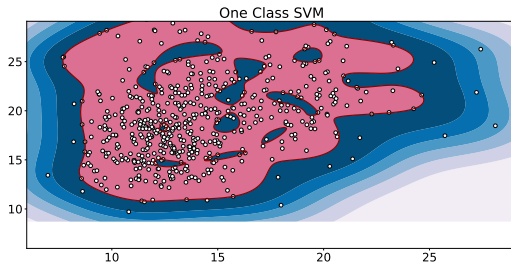
Une version non supervisée X

Implémentation

On utilisera la fonction *OneClassSVM* de la librairie **sklearn.svm**.






Une version non supervisée XI



Fuzzy SVM

Références I

-  Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992).
A training algorithm for optimal margin classifiers.
In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92*, pages 144–152. ACM.
-  Elzinga, D. J. and Hearn, D. W. (1972).
The minimum covering sphere problem.
Management Science, 19(1) :96–104.
-  Genton, M. G. (2002).
Classes of kernels for machine learning : A statistics perspective.
Journal of Machine Learning Research, 2 :299–312.

Références II



Mercer, J. (1909).

Functions of positive and negative type, and their connection with the theory of integral equations.

Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences, 209(441-458) :415–446.



Scholkopf, B. and Smola, A. J. (2001).

Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond.

MIT Press, Cambridge, MA, USA.






Sylvester, J. (1857).

A question in the geometry of situation.

Quarterly Journal of Pure and Applied Mathematics.

Références III

-  Tax, D. M. J. and Duin, R. P. W. (1999).
Data domain description using support vectors.
In *Proceedings of the European Symposium on Artificial Neural Networks*, pages 251–256.
-  Tax, D. M. J. and Duin, R. P. W. (2004).
Support vector data description.
Machine Learning Journal, 54(1) :45–66.
-  Vapnik, V. and Cortes, C. (1995).
Support-vector networks.
Machine Learning, 20 :273–297.