SVM
Support Vector Machines
Séparateurs à Vaste Marge

Introduction

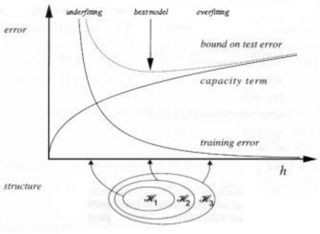
Suite aux travaux du mathématicien Vladimir Vapnik, Boser, Guyon et Vapnik ont proposé l'algo SVM en 1992. Sa popularité a augmenté ensuite pour devenir le principal algo de classif.

Théorie de Vapnik-Chervonenkis → VC dimension
Borne sur la relation entre capacité d'apprentissage et performance

Introduction

Travaux de Vapnick en théorie de l'apprentissage statistique

<u>Principe d'induction</u>
Minimisation du Risque Structurel (SRM)
Espace d'hypothèses de capacité réduite et performant



$h \leq \min([r^2a^2], n) + 1$

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\left(\frac{h(\log(2l/h)+1) - \log(\eta/4)}{l}\right)}$$

Introduction

Botaniste avec mémoire photographique.
On lui présente un nouvel arbre, deux réactions extrêmes :
– « Ce n'est pas un arbre car il n'a pas le même nombre de feuilles que ceux que je connais. »
– « Il y a du vert, c'est donc un arbre. »

Notations

D : base d'apprentissage contenant ℓ exemples

$D = \{ (x_i, y_i) \}_{i \text{ de } 1 \text{ à } l}$

Avec

x_i un vecteur contenant les caract. de la forme
$y_i \in \{-1, +1\}$

Hyperplan séparateur

- Si les 2 classes linéairement séparables alors il existe un ensemble d'hyperplans séparateurs
- Un hyperplan h peut être représenté par
  - Un vecteur w perpendiculaire à h
  - Une constante b
  - y = h(x) = sgn(<w, x> + b)

•Comment trouver h à partir de D ?

Hyperplan séparateur

- Par exemple, Perceptron
  - $R \leftarrow \max ||xi||^2$, $w \leftarrow b \leftarrow 0$
  - Tant que tous les ex. ne sont pas bien classés faire
    - Parcourir les ex
      - Si y_i ($<w, x_i> + b$) ≤ 0
      - Alors $w \leftarrow w + \eta\, y_i\, x_i$

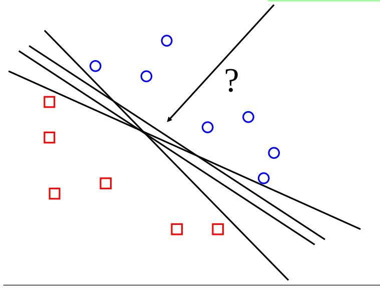
        $b \leftarrow b + \eta\, y_i\, R^2$

- Ou dans sa version duale avec $w = \Sigma_j\, \alpha_j\, y_j\, x_j$
  - $R \leftarrow \max ||xi||^2$, $\alpha \leftarrow b \leftarrow 0$
  - Tant que tous les ex ne sont pas bien classés faire
    - Parcourir les ex
      - Si y_i ($\Sigma_j\, \alpha_j\, y_j <x_j , x_i> + b$) ≤ 0
      - Alors $\alpha_i \leftarrow \alpha_i + \eta$

        $b \leftarrow b + \eta\, y_i\, R^2$

Hyperplan optimal

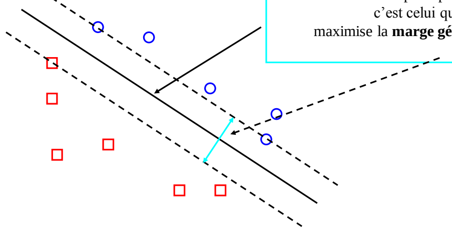
Quel hyperplan choisir parmi ceux séparant linéairement les données ?



Hyperplan optimal

Quel hyperplan choisir parmi ceux séparant linéairement les données ?

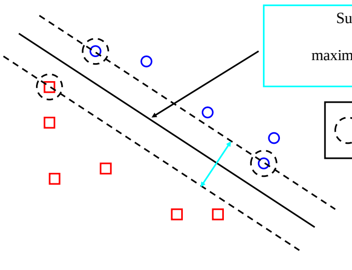
Suivant le principe SRM, c'est celui qui maximise la **marge géométrique**



Hyperplan optimal

Quel hyperplan choisir parmi ceux séparant linéairement les données ?

Suivant le principe SRM, c'est celui qui maximise la **marge géométrique**

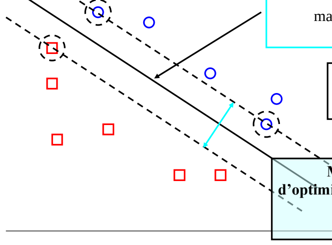


Les vecteurs supports

Hyperplan optimal

Quel hyperplan choisir parmi ceux séparant linéairement les données ?

Suivant le principe SRM, c'est celui qui maximise la **marge géométrique**



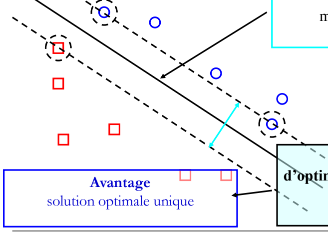
Les vecteurs de support

Maximisation d'un problème d'optimisation quadratique sous contraintes

Hyperplan optimal

Quel hyperplan choisir parmi ceux séparant linéairement les données ?

Suivant le principe SRM, c'est celui qui maximise la **marge géométrique**



Les vecteurs de support

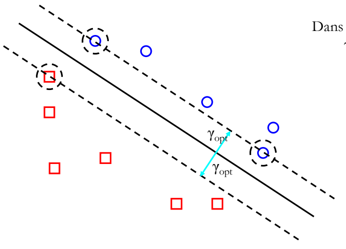
Maximisation d'un problème d'optimisation quadratique sous contraintes

Avantage
solution optimale unique

Hyperplan optimal

$y_i (<w, x_i> + b) \geq \gamma_{opt}$
or w et b sont définis à une constante multiplicative près
Donc $y_i (<w, x_i> + b) \geq 1$ ==> exemple à l'ext. de la marge

Dans ce cas, on a :
$\gamma_{opt} = 1 / ||w||^2$



Hyperplan optimal

Pb : maximiser la marge → minimiser $f(w, b) = <w, w> = ||w||^2$
Tel que $y_i (<w, x_i> + b) \geq 1$ avec $1 \leq i \leq l$

Dans le dual, $||w||^2 = <w, w> = \Sigma_{i,j} \alpha_i \alpha_j y_i y_j <x_j , x_i>$
Le pb devient :
Minimiser $f(\alpha, b) = \Sigma_{i,j} \alpha_i \alpha_j y_i y_j <x_j , x_i>$
Tel que $y_i (\Sigma_j \alpha_j y_j <x_j , x_i> + b) \geq 1$ et $\alpha_i \geq 0$ avec $1 \leq i \leq l$

Fonction à minimiser convexe sous contraintes linéaires → minimum global que l'on peut déterminer grâce à un solveur QP (quadratic programming) ou SMO (plus loin)

Résolution hyperplan

Quadratic programming with linear constraints

$$\boxed{\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \end{aligned}}$$

$$\Updownarrow$$

Lagrangian Function

$$\boxed{\begin{aligned} &\text{minimize} \quad L_p(\mathbf{w}, b, \alpha_i) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 \right) \\ &\text{s.t.} \quad \alpha_i \geq 0 \end{aligned}}$$

Résolution hyperplan

$$\boxed{\begin{aligned} &\text{minimize} \quad L_p(\mathbf{w}, b, \alpha_i) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 \right) \\ &\text{s.t.} \quad \alpha_i \geq 0 \end{aligned}}$$

$$\frac{\partial L_p}{\partial \mathbf{w}} = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial L_p}{\partial b} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

Ces équations sont obtenues grâce au théorème du Kuhn -Tucker

Résolution hyperplan

$$\boxed{\begin{aligned} &\text{minimize} \quad L_p(\mathbf{w}, b, \alpha_i) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 \right) \\ &\text{s.t.} \quad \alpha_i \geq 0 \end{aligned}}$$

$$\Updownarrow$$

Lagrangian Dual Problem

$$\boxed{\begin{aligned} &\text{maximize} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ &\text{s.t.} \quad \alpha_i \geq 0 \text{ , and } \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned}}$$

Problèmes non-linéairement séparables

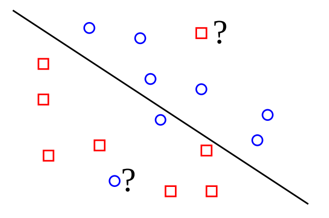
Pourquoi a-t-on des problèmes non-linéairement séparables ?
- Les x_i sont bruités
- Erreurs de classif sur y_i
- Problème de nature non linéaire

- 2 solutions : marge souple et astuce du noyau

Marge souple (soft margin)

Problèmes non-linéairement séparables

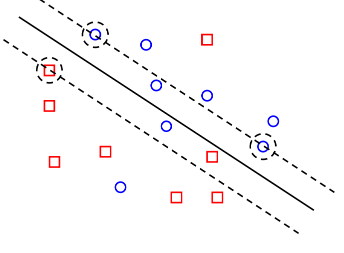
Que faire si le problème n'est pas linéairement séparable ?



Marge souple

Que faire si le problème n'est pas linéairement séparable ?

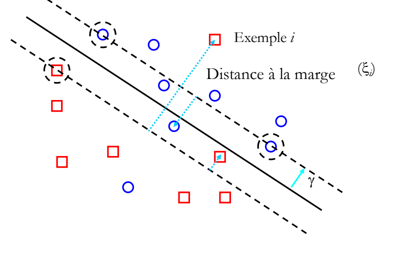
Trouver un compromis entre marge géométrique maximale et nombre d'exemples mal-classés



Marge souple

Que faire si le problème n'est pas linéairement séparable ?

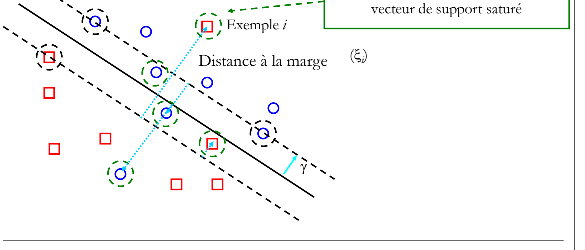
Trouver un compromis entre marge géométrique maximale et nombre d'exemples mal-classés



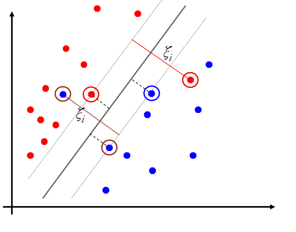
Marge souple

Que faire si le problème n'est pas linéairement séparable ?

Trouver un compromis entre marge géométrique maximale et nombre d'exemples mal-classés



Marge souple (2[e] exemple)



Marge souple

- Formulation

$$\text{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

Tel que

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

- Paramètre C : compromis entre fidélité et régularité.

Résolution marge souple

- Formulation: (Lagrangian Dual Problem)

$$\text{maximize} \quad \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$$

Tel que

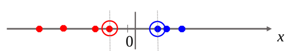
$$0 \leq \alpha_i \leq C$$

$$\sum_{i=1}^{n}\alpha_i y_i = 0$$

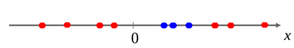
Astuce du noyau (kernel trick)

Astuce du noyau

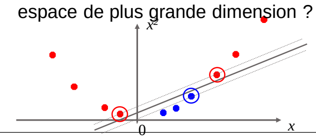
- Les bases de données bruitées sont bien séprées



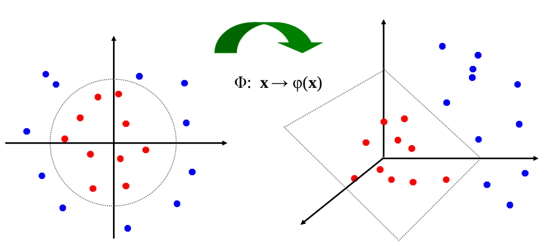
- Mais que faire si le problème est trop difficile ?



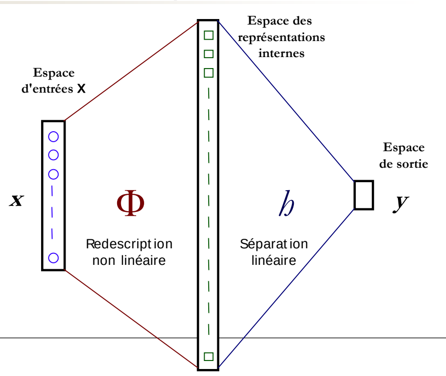
- Pourquoi ne pas projeter (*mapper*) les données dans un espace de plus grande dimension ?



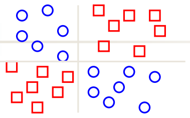
Astuce du noyau



SVM et redescription

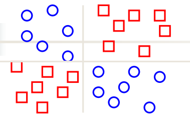


Exploitation de fonctions noyaux



Aucun séparateur linéaire convenable même avec pénalisation ?

Exploitation de fonctions noyaux



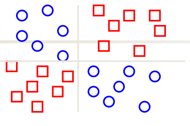
Aucun séparateur linéaire convenable même avec pénalisation ?

Utilisation de fonctions noyaux *K(.,.)*

Fonction symétrique de deux variables qui retourne un scalaire correspondant à **une distance entre deux exemples**

Espace de redescription

Exploitation de fonctions noyaux



Aucun séparateur linéaire convenable même avec pénalisation ?

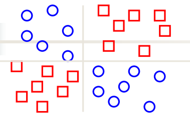
Utilisation de fonctions noyaux *K(.,.)*

Fonction symétrique de deux variables qui retourne un scalaire correspondant à **une distance entre deux exemples**

Noyau gaussien

$$K(\mathbf{x}_i,\mathbf{x}_j)=\exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma^2}\right)$$

Exploitation de fonctions noyaux



Aucun séparateur linéaire convenable même avec pénalisation ?

Utilisation de fonctions noyaux *K(.,.)*

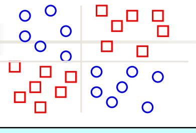
Fonction symétrique de deux variables qui retourne un scalaire correspondant à **une distance entre deux exemples**

Noyau gaussien

$$K(\mathbf{x}_i,\mathbf{x}_j)=\exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma^2}\right)$$

Introduction d'un paramètre propre au noyau

Exploitation de fonctions noyaux



Aucun séparateur linéaire convenable même avec pénalisation ?

Utilisation de fonctions noyaux *K(.,.)*

Fonction symétrique de deux variables qui retourne un scalaire correspondant à **une distance entre deux exemples**

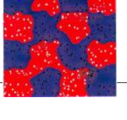
Noyau gaussien

$$K(\mathbf{x}_i,\mathbf{x}_j)=\exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma^2}\right)$$

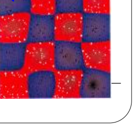
Introduction d'un paramètre propre au noyau

Frontières de décision non-linéaires

σ petit



σ grand



Exemple des 2 spirales séparées par un hyperplan (noyau gaussien)

Formulation sans noyau

- Formulation: (Lagrangian Dual Problem)

$$\text{maximize} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boxed{\mathbf{x}_i^T \mathbf{x}_j}$$

Tel que

$$0 \le \alpha_i \le C$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

Formulation duale complète des SVM

$$\underset{\alpha}{\text{Max}} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right)$$

tel que $0 \le \alpha_i \le C$ et $\sum_{i=1}^{m} \alpha_i y_i = 0$

Théorie Lagrangienne

Formulation duale complète des SVM

$$\underset{\alpha}{\text{Max}} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \underline{K(x_i, x_j)} \right)$$

tel que $0 \le \alpha_i \le C$ et $\sum_{i=1}^{m} \alpha_i y_i = 0$

Théorie Lagrangienne

La recherche de l'hyperplan dans l'**espace de redescription** est réalisée de **façon implicite**!

Formulation duale complète des SVM

$$\underset{\alpha}{\text{Max}} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \underline{K(x_i, x_j)} \right)$$

tel que $0 \le \alpha_i \le \underline{C}$ et $\sum_{i=1}^{m} \alpha_i y_i = 0$

Théorie Lagrangienne

La recherche de l'hyperplan dans l'**espace de redescription** est réalisée de **façon implicite**!

constante de régularisation

Formulation duale complète des SVM

$$\underset{\alpha}{\text{Max}} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \underline{K(x_i, x_j)} \right)$$

tel que $0 \le \alpha_i \le \underline{C}$ et $\sum_{i=1}^{m} \alpha_i y_i = 0$

Théorie Lagrangienne

La recherche de l'hyperplan dans l'**espace de redescription** est réalisée de **façon implicite**!

constante de régularisation

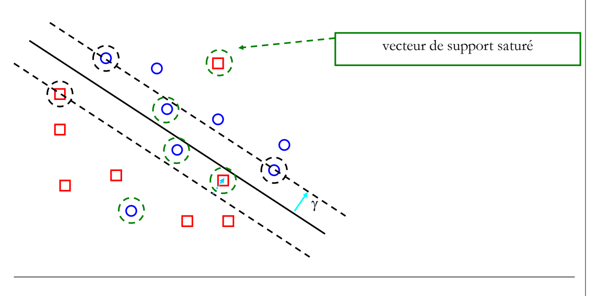
la solution est parcimonieuse

- Si $\alpha_i > 0$ alors l'exemple i est appelé un **vecteur de support**.

$i \in SV$ (ensemble des Vecteurs de Support)

- $\alpha_i < C$: Vecteur de support (« sur » la marge géométrique)
- $\alpha_i = C$: Vecteur de support saturé (« dans » ou « hors » de la marge géométrique)

Formulation duale complète des SVM



Exemples de fonctions Kernel

- Linear kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$
- Polynomial kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i^T \mathbf{x}_j)^p$
- Gaussian (Radial-Basis Function (RBF)) kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$$

- Sigmoid:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\beta_0 \mathbf{x}_i^T \mathbf{x}_j + \beta_1)$$

- In general, functions that satisfy *Mercer's condition* can be kernel functions.

Conditions KKT

- Karush-Kuhn-Tucker ont défini pour l'algo SVM les conditions nécessaire et suffisantes d'optimalité suivantes :
- $\alpha_i = 0 \quad \Rightarrow y_i u_i > 1$
  $0 < \alpha_i < C \quad \Rightarrow y_i u_i = 1$
  $\alpha_i = C \quad \Rightarrow y_i u_i < 1$
  avec $u_i = \sum_{j=1}^{l} \alpha_j y_j K(xi, xj) + b$
- Les vecteurs supports sont les exemples dont $\alpha_i \neq 0$

Calcul de b et fonction de décision

- On prend un exemple dont $0 < \alpha_i < C$
  Alors $b^* = y_i - \sum_{j=1}^{l} \alpha_j^* y_j K(xi, xj)$
- La fonction de décision est :
  sgn ($\sum_{j=1}^{l} \alpha_j^* y_j K(x, xj)$ + b*)

SMO

Sequential Minimal Optimization

- Résolution du problème quadratique (convexe) sous contraintes linéaires sans solveur QP (Quadratic Programming) car complexité en O (l^3)
- Plusieurs solutions : SimpleSVM, SMO…
- Pour SMO : décomposition du pb en sous-problèmes jusqu'à prendre en compte que 2 exemples. Il faut juste qu'au moins un des exemples viole une condition KKT.
  Résolution analytique pour 2 α par itération

Algo SMO

- Tant que tous les exemples d'apprent. ne respectent pas KKT à ε (10^{-3}) près faire :
  - Choisir 2 exemples x_1 et x_2
  - Optimiser α_1 et α_2
  - Calculer le seuil b

SVM multi-classe

- 1 classe contre toutes
  1 SVM pour chaque classe
- Par paire (1 contre 1) puis vote ou élimination
  1 SVM pour chaque paire de classe (n(n-1)/2)

Importance de la sélection de modèle

$$Q = \max_{\alpha} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right)$$

tel que $0 \le \alpha_i \le \boxed{C}$ et $\sum_{i=1}^{m} \alpha_i y_i = 0$

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\boxed{2\sigma^2}} \right)$$

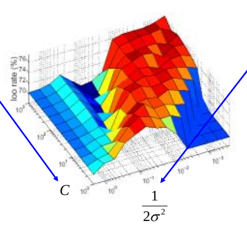
C et σ sont les hyper-paramètres des SVM

Importance de la sélection de modèle

$$Q = \max_{\alpha} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right)$$

tel que $0 \le \alpha_i \le \boxed{C}$ et $\sum_{i=1}^{m} \alpha_i y_i = 0$

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\boxed{2\sigma^2}} \right)$$



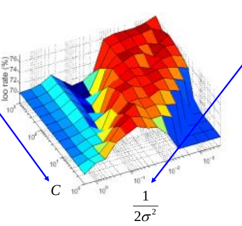
Importance de la sélection de modèle

$$Q = \max_{\alpha} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right)$$

tel que $0 \le \alpha_i \le \boxed{C}$ et $\sum_{i=1}^{m} \alpha_i y_i = 0$

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\boxed{2\sigma^2}} \right)$$

Techniques de validation croisée
- LOO (*Leave One Out*)
- *k* parties (*k fold*)
- *Boostrap*



SVM et régression

- Fonction de perte : $|y - f(\boldsymbol{x})|_\varepsilon \; = \; \max\{0, |y - f(\boldsymbol{x})| - \varepsilon\}$
  - **Régression linéaire :** $\boxed{f(\boldsymbol{x}) \; := \; (\boldsymbol{w} \cdot \boldsymbol{x}) \; + \; w_0}$
  - **Soit à minimiser :** $\frac{1}{2} \| \boldsymbol{w} \|^2 \; + \; C \sum_{i=1}^{m} |y_i - f(\boldsymbol{x}_i)|_\varepsilon$
  - **Généralisation :** $\boxed{f(\boldsymbol{x}) \; = \; \sum_{i=1}^{m} (\alpha_i^\star - \alpha_i) \, K(\boldsymbol{x}_i, \boldsymbol{x}) \; + \; w_0}$