# The online Perceptron Algorithm and Linear Support Vector Machines (Séparateurs linéaires à Vaste Marge)

Alain Rakotomamonjy

Université de Rouen - Criteo AI Lab

19 novembre 2021

# Plan


1. Discrimination linéaire
   - Formulation
   - Séparateur linéaire et maximisation de la marge
   - Perceptron
2. Résolution du problème SVM
   - Problème primal et Lagrangien
   - Problème dual de SVM
3. SVM pour les problèmes non-séparable linéairement
4. SVM en pratique

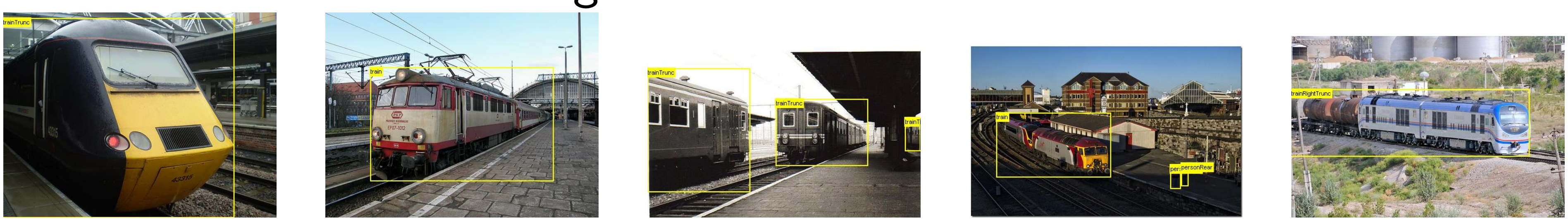
# Séparateur linéaire

## But

- $\mathcal{D} = \{(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}\}_{i=1 \cdots n}$ : ensemble de points étiquetés
- Construire à partir de $\mathcal{D}$ une fonction $f : \mathcal{X} \to \{-1, 1\}$ ou $f : \mathcal{X} \to \mathbb{R}$ qui permet de prédire la classe $-1$ ou $1$ d'un point $x \in \mathcal{X}$

Image contenant un bus

Image contenant un train



# Séparateur linéaire

## But

- $\mathcal{D} = \{(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}\}_{i=1\cdots n}$ : ensemble de points étiquetés
- Construire à partir de $\mathcal{D}$ une fonction $f : \mathcal{X} \to \{-1, 1\}$ ou $f : \mathcal{X} \to \mathbb{R}$ qui permet de prédire la classe $-1$ ou $1$ d'un point $x \in \mathcal{X}$

## Fonction de décision (scoring function)

- On suppose l'espace des entrées $\mathcal{X} = \mathbb{R}^d$
- Fonction de décision : $f : \mathbb{R}^d \to \mathbb{R}$ telle que si

$$
\begin{array}{ll}
f(x) < 0 & \text{affecter } x \text{ à la classe } -1 \\
f(x) > 0 & \text{affecter } x \text{ à la classe } 1
\end{array}
$$

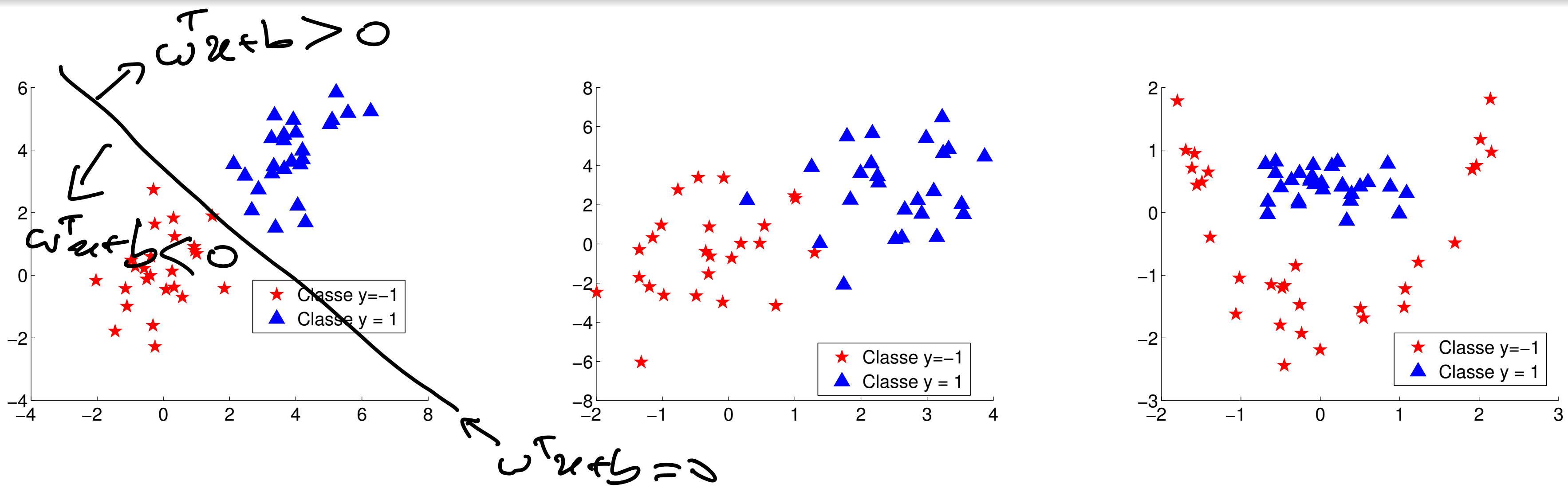
- Fonction de décision linéaire :

$$f(x) = w^\top x + b, \qquad w \in \mathbb{R}^d, b \in \mathbb{R}$$

# Définition

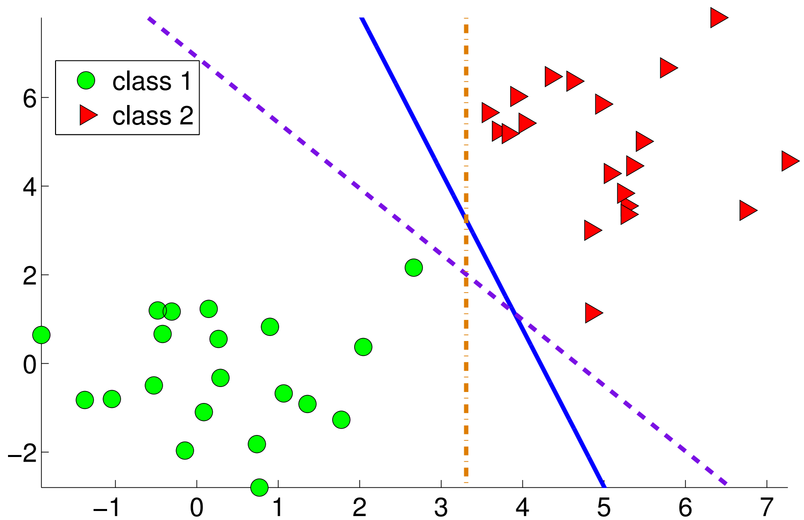
## Problème linéairement séparable

*Les points $\{(x_i, y_i)\}$ sont linéairement séparables si il existe un hyperplan qui permet de discriminer correctement l'ensemble des données. Dans le cas contraire, on parle d'exemples non séparables linéairement.*



# Discrimination linéaire en 2D

Trouver une fonction linéaire séparant les points des classes 1 et 2




- Frontière de décision : $w^\top x + b = 0$
- Plusieurs frontières possibles
- Toutes les fonctions de décision se valent-elles ?

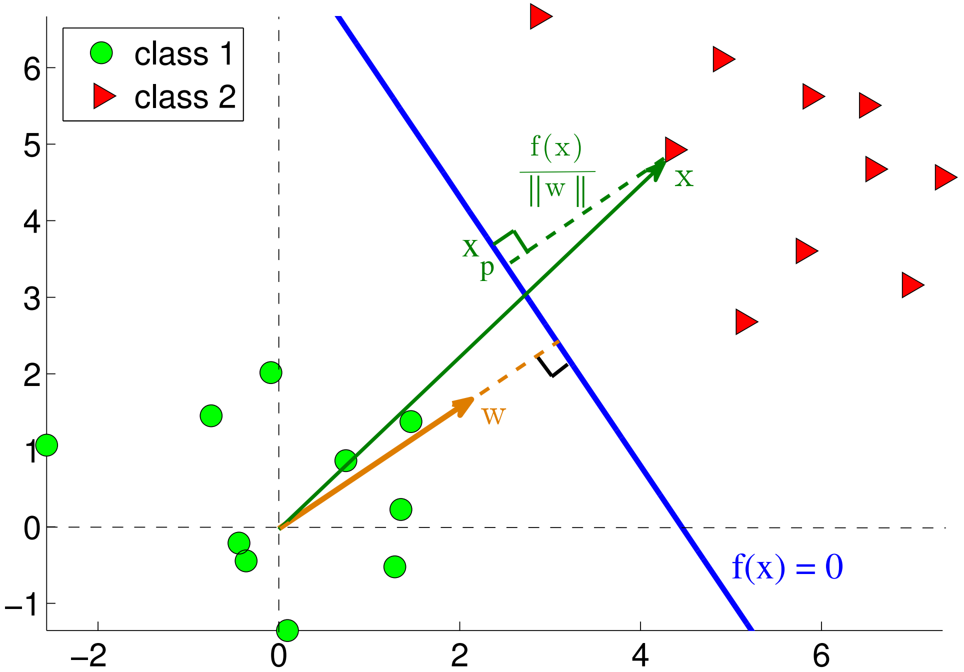
**Quelle solution choisir ?**

Choisir celle qui maximise la marge entre les points des classes

# Notion de géométrie

## Distance d'un point à la frontière de décision

Soit $H(w,b) = \{z \in \mathbb{R}^d \mid f(z) = w^\top z + b = 0\}$ un hyperplan et soit $x \in \mathbb{R}^d$. La distance du point $x$ à l'hyperplane $H$ est $d(x,H) = \frac{|w^\top x + b|}{\|w\|} = \frac{|f(x)|}{\|w\|}$




proof:

$$x = x_p + \frac{\omega}{\|\omega\|} \times d$$

take the dot prod of $x$ with $\omega$

$$\omega^T x = \omega^T x_p + \underbrace{\omega^T \frac{\omega}{\|\omega\|} d}_{= \|\omega\| d}$$

$$\|\omega\| d = \omega^T x - \omega^T x_p$$

$$= (\omega^T x + b) - \underbrace{(\omega^T x_p + b)}_{=0}$$

$$d = \frac{\omega^T x + b}{\|\omega\|}$$

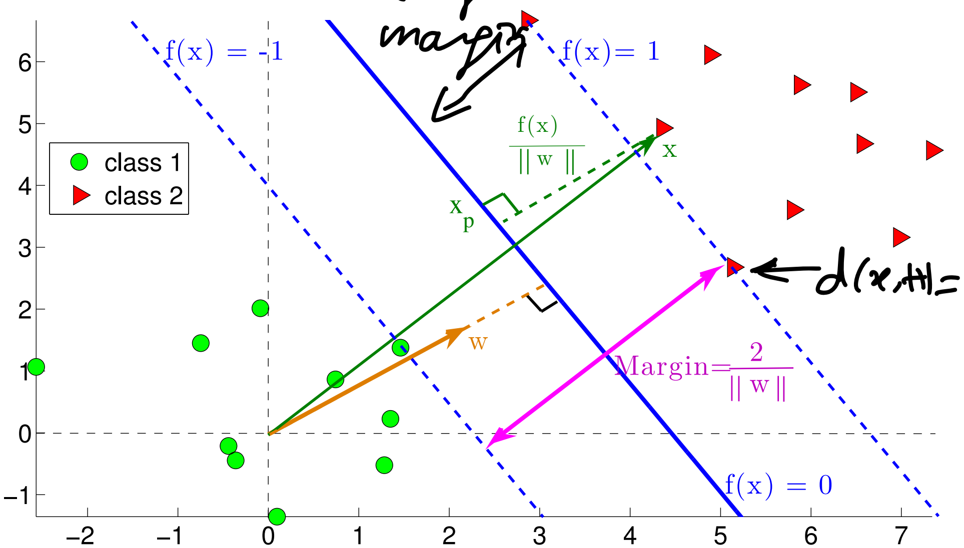
# La marge

## Hyperplan canonique

- Un hyperplan est dit canonique par rapport aux données $\{x_1, \cdots, x_N\}$ si $\min_{x_i} |w^\top x_i + b| = 1$

$d(x, H) = \frac{|w^T x + b|}{\|w\|}$

half margin $= \frac{1}{\|w\|}$




$\leftarrow d(x, H) = \frac{1}{\|w\|}$

## Marge

La **marge géométrique** est $M = \frac{2}{\|w\|}$

## Hyperplan canonique optimal

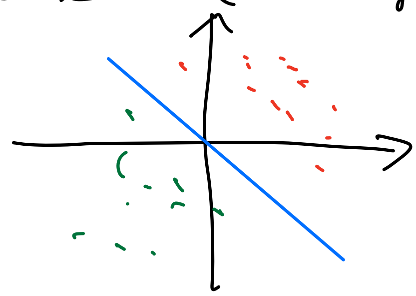
- Maximiser la marge
- Classer correctement chaque point i.e. $\forall i, \quad y_i f(x_i) \geqslant 1$

$y_i f(x_i) > 0$

$|f(x_i)| \geqslant 1$

# The Perceptron Algorithm (on-line)

For homogenous linear classifier
$f(x) = w^T x$ (no $b$ for now)



$t \leftarrow 0$
$w_0 \leftarrow 0$
Repeat
- receive $x_t$
- predict $\hat{y}_t = \text{sign}(w_t^T x_t)$
- receive $y_t \in \{-1, 1\}$
- if $y_t \neq \hat{y}_t$ then
  - $w_{t+1} \leftarrow w_t + y_t x_t$
- else
  - $w_{t+1} \leftarrow w_t$

## Thm (Block, Novikoff)

Assume $\|x_t\| < R$ for all $t$, $y_t \in \{-1, 1\}$
Assume there exist canonical hyperplane $w^*$ classifying data perfectly passing through the origin with half margin $\rho = \frac{1}{\|w^*\|}$
then, the number of mistakes $t_e$ of perceptron is at most $\frac{R^2}{\rho^2}$

proof:
Step 1) After an update (a prediction error), $w_{t+1}$ is "more aligned" to $w^*$

$$\langle w_{t+1}, w^* \rangle = \langle w_t + y_t x_t, w^* \rangle$$
$$= \langle w_t, w^* \rangle + \underbrace{y_t \langle x^t, w^* \rangle}_{\geq 1 \text{ because } w^* \text{ is canonical}}$$
$$\geq \langle w_t, w^* \rangle + 1$$

Unrolling, we get
$$\langle w_t, w^* \rangle \geq t_e$$
↑ nb of mistakes

Step 2) After an update (classification error)

$$\|w_{t+1}\|^2 = \langle w_t + y_t x_t , w_t + y_t x_t \rangle$$

$$= \|w_t\|^2 + \underbrace{2\, y_t \langle w_t , x_t \rangle}_{\leq 0} + \|y_t x_t\|^2$$

because misclassification at this step

$$\leq \|w_t\|^2 + R^2$$

$$\Rightarrow \|w_t\|^2 \leq t_e \cdot R^2$$

Step 3)

$$t_e \leq \langle w_t , w^* \rangle \leq \underbrace{\|w_t\| \cdot \|w^*\|}_{\text{Cauchy-Schwartz}} \leq \sqrt{t_e} \cdot R \cdot \|w^*\|$$

$$\|w^*\| \text{ } \| \text{ } \sqrt{t_e}\, \frac{R}{\rho}$$

$$\Rightarrow \sqrt{t_e} \leq \frac{R}{\rho} \Rightarrow t_e \leq \frac{R^2}{\rho^2}$$

<u>Exercise:</u>

Rewrite perceptron algo for non homogeneous hyperplanes

$w^T x + b =$

# The Perceptron Algorithm as a "SGD"

## online Learner

Perceptron Update

$$\boxed{\begin{array}{l} \text{if } y_t \langle w_t, x_t \rangle < 0 \\ \qquad w_{t+1} \leftarrow w_t + y_t x_t \\ \text{else} \\ \qquad w_{t+1} \leftarrow w_t \end{array}}$$

SGD update

$$\boxed{w_{t+1} \leftarrow w_t - \alpha \nabla_w \ell(w_t^T x, y)}$$

predicted score of $x_t$

Let $s_t = w_t^T x$

$$\begin{array}{l} \text{if } y_t s_t < 0 \\ \qquad w_{t+1} \leftarrow w_t + y_t x_t \\ \text{else} \\ \qquad w_{t+1} \leftarrow w_t \end{array}$$

$$\ell^{perceptron}(s_t, y) = \begin{cases} 0 & \text{if } y_t s_t \geq 0 \\ -y \cdot s_t & \text{otherwise} \end{cases}$$

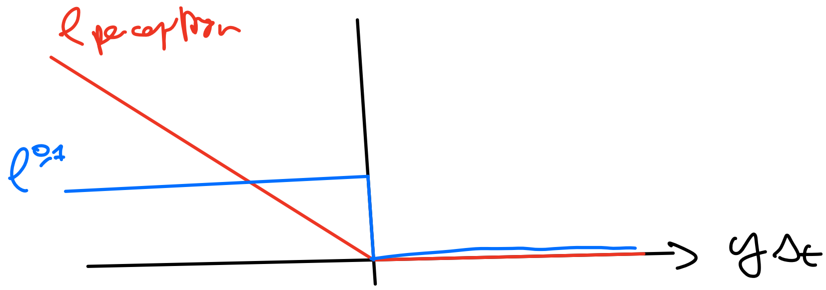
$\Downarrow$

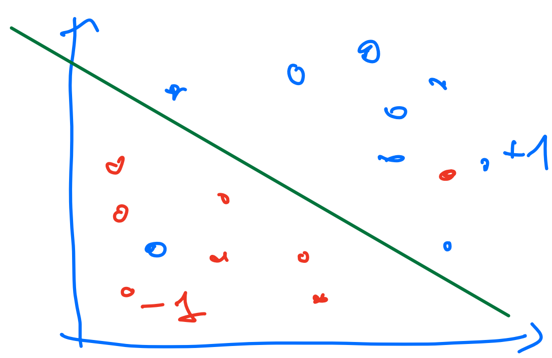
if $\alpha = 1$

Applying SGD here gives.

$$w_{t+1} \leftarrow w_t - \alpha \begin{cases} 0 & \text{if } y_t s_t > 0 \\ -y_t x_t & \text{otherwise} \end{cases}$$

$$\ell^{perceptron}(s_t, y) = \max(0, -y s_t)$$



# Marge et borne de généralization

### Borne VC

Risque sur une classe de fonction $\mathcal{H}$. Avec une prob $1-\delta$

$$R(h) \leq R_{emp}(h) + C\sqrt{\frac{D(\log(2N/D)+1)+\log(4\delta)}{N}}$$

où $D$ est la VC dimension de $\mathcal{H}$

### VC dim de la classe des fonctions linéaires à marge $\rho$

Soit $\mathcal{H}$ la classe de fonction $f(x) = w^\top x + b$ à une marge $\rho$ des exemples d'apprentissage alors

$$D \leq 1 + \min\left(d, \frac{R^2}{\rho^2}\right)$$

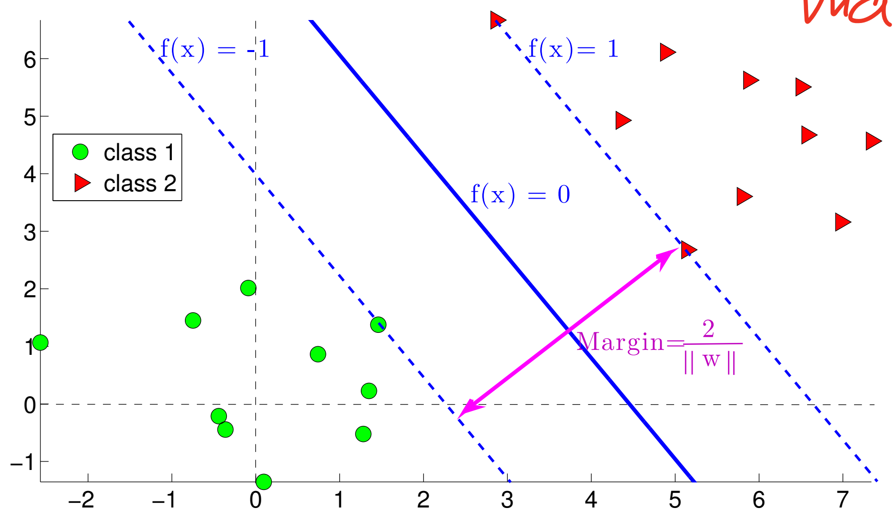
$R$, rayon d'une boule contenant les données d'apprentsisages.

# Formulation du problème de maximisation de marge

## Séparateur à vaste marge (SVM) : formulation

- $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}\}_{i=1}^n$ : points linéairement séparables
- Objectif : trouver une fonction de décision $f(x) = w^\top x + b$ qui maximise la marge et discrimine correctement les points de $\mathcal{D}$

$$\begin{array}{lll} \min_{w,b} & \frac{1}{2}\|w\|^2 & \text{maximisation de la marge} \\ \text{s.c.} & y_i(w^\top x_i + b) \geq 1 \quad \forall i = 1, \cdots, n & \text{tous les points bien classés} \end{array}$$




$\max \rho = \frac{1}{\|\omega\|}$

$\Updownarrow$

$\min \|\omega\|$

# Le Lagrangien du problème SVM

min objective$(\omega)$
s.t. $h_i(\omega) \leq 0$

$L(\omega) = obj(\omega) + \sum \alpha_i h_i(\omega)$

## Problème primal de SVM

$$
\begin{array}{ll}
\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} & \frac{1}{2}\|w\|^2 \\
\text{s.c.} & y_i(w^\top x_i + b) \geq 1 \quad \forall i = 1, \cdots, n
\end{array}
$$

$h_i(\omega) = 1 - y_i(\omega^T x_i + b)$

- On introduit des multiplicateurs de Lagrange $\alpha_i \geq 0$ associés aux $n$ contraintes d'inégalités i.e. $n$ paramètres $\alpha_i$
- Lagrangien

$$
L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i (y_i(w^\top x_i + b) - 1)
$$

$\frac{dL}{d\omega} = \omega - \sum \alpha_i y_i x_i = 0$ $\qquad$ $\frac{dL}{db} = \sum \alpha_i y_i = 0$

# Le problème dual

$L = \frac{1}{2} \langle \sum_i \alpha_i y_i x_i , \sum_j \alpha_j y_j x_j \rangle$
$- \sum_i \alpha_i ( y_i ( \sum_j \alpha_j y_j x_j^\top x_i + b ) - 1 )$

- Condition de stationnarité

$$\frac{\partial L(w, b, \alpha)}{\partial b} = 0 \qquad \frac{\partial L(w, b, \alpha)}{\partial w} = 0$$

Soit :

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

- Problème dual : problème de programmation quadratique

En remplacant ces valeurs dans le Lagrangien, on obtient :

$$\max_{\{\alpha_i\}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j$$

$$\text{s.c.} \quad \alpha_i \geq 0, \quad \forall i = 1, \cdots, n$$

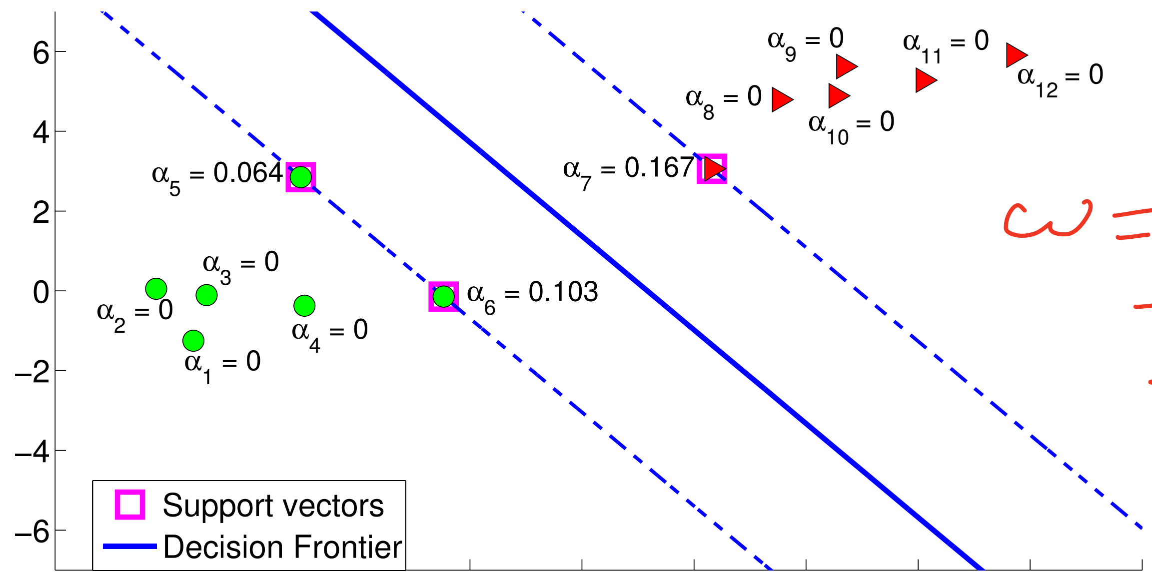
$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

constraints : $h_i(\omega) \leq 0$
$\alpha_i \cdot h_i(\omega) = 0$

- Complementary slackness : $\alpha_i ( y_i ( \omega^\top x_i + b ) - 1 ) = 0$

## Les vecteurs supports

- Résoudre le dual pour trouver les $n$ paramètres $\{\alpha_i\}_{i=1}^n$
- On obtient deux types de paramètres $\alpha_i$
  - Pour un point $x_j$, si $y_j(w^\top x_j + b) > 1$ alors $\alpha_j = 0$
  - Pour un point $x_i$, si $y_i(w^\top x_i + b) = 1$ alors $\longrightarrow \alpha_i \geq 0$
- Solution : $w = \sum_{i=1}^n \alpha_i y_i x_i$. $w$ n'est défini que par les points tels que $y_i(w^\top x_i + b) = 1$. On les appelle **vecteurs supports**




$\omega = 0.16 . x_7$
$- 0.064 . x_5$
$- 0.103 . x_6$

# SVM linéairement séparable en pratique

## Calcul de $w$

- Utiliser les données $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ pour résoudre le dual

  $\longrightarrow$ On obtient les paramètres $\{\alpha_i\}_{i=1}^n$
- En déduire la solution $w = \sum_{i=1}^n \alpha_i y_i x_i$

## Calcul de $b$

- Les $\alpha_i > 0$ correspondent aux points supports qui vérifient la relation

$$y_i(w^\top x_i + b) = 1$$

- En déduire la valeur de $b$

## La fonction de décision

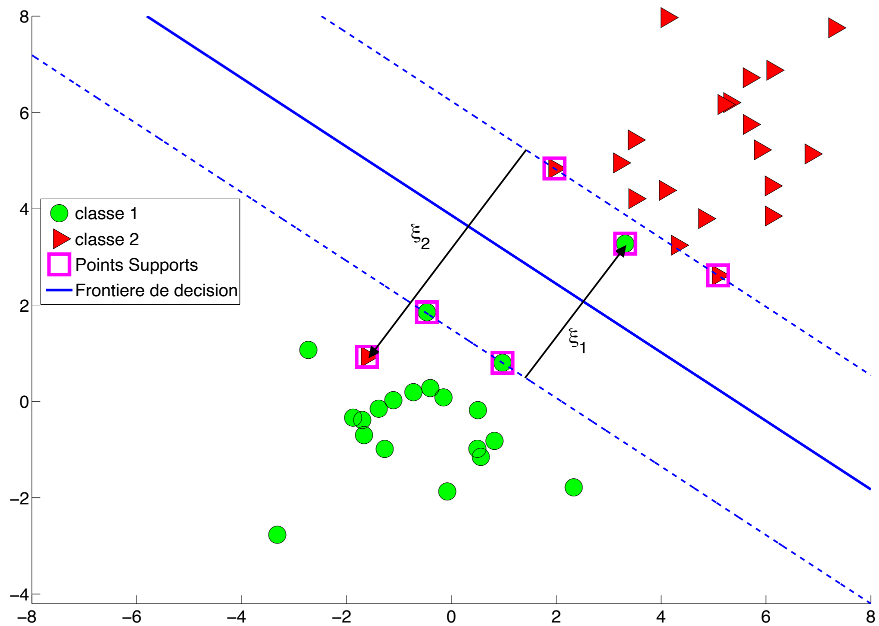
$$f(x) = w^\top x + b = \sum_{i=1}^n \alpha_i y_i x_i^\top x + b$$

# Cas non séparable

Que se passe-t-il si les données ne sont pas linéairement séparable ?

## Relacher les contraintes

- Relâcher $y_i(w^\top x_i + b) \geq 1$
- Accepter $y_i(w^\top x_i + b) \geq 1 - \xi_i$ avec $\xi_i \geq 0$ le terme ''d'erreur''
- Inclure la somme des "erreurs" $\sum_{i=1}^{n} \xi_i$ dans le problème de SVM

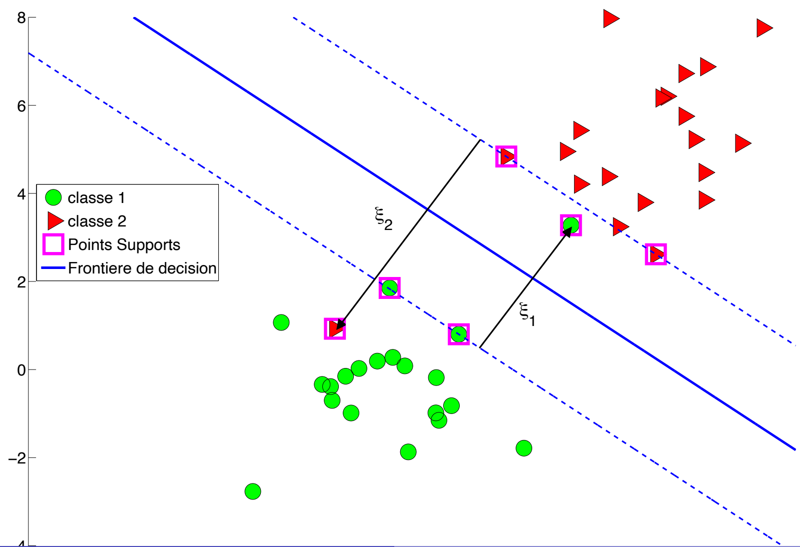


# Cas non séparable : formulation

## SVM : cas non-séparable

$$
\begin{array}{lll}
\min_{w,b,\{\xi_i\}} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i & \\
\text{s.c.} & y_i(w^\top x_i + b) \geq 1 - \xi_i & \forall i = 1, \cdots, n \\
 & \xi_i \geq 0 & \forall i = 1, \ldots, n
\end{array}
$$

- $C > 0$ : paramètre de régularisation (compromis entre erreur et marge)
- $C$ est à fixer par l'utilisateur !



# Cas non séparable : le problème dual

## Le lagrangien

$$L(w, b, \xi, \alpha, \nu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i (y_i(w^\top x_i + b) - 1 + \xi_i) - \sum_{i=1}^{n} \nu_i \xi_i$$

avec $\alpha_i \geq 0$, $\nu_i \geq 0$, pour tout $i = 1, \cdots, n$

## Conditions d'optimalité de stationnarité

$$\frac{\partial L(w, b, \xi_i, \alpha)}{\partial b} = 0 \qquad \frac{\partial L(w, b, \xi_i, \alpha)}{\partial w} = 0 \qquad \frac{\partial L(w, b, \xi_i, \alpha)}{\partial \xi_k} = 0$$

ce qui donne

$\sum_i^n \alpha_i y_i = 0$ $\qquad$ $w = \sum_i^n \alpha_i y_i x_i,$ $\qquad$ $C - \alpha_i - \nu_i = 0, \; \forall i = 1, \cdots, n$

# Cas non séparable : la solution

## Le problème dual

$$\max_{\{\alpha_i\}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j$$
$$\text{s.c.} \quad 0 \leq \alpha_i \leq C, \quad \forall\, i = 1, \cdots, n$$
$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

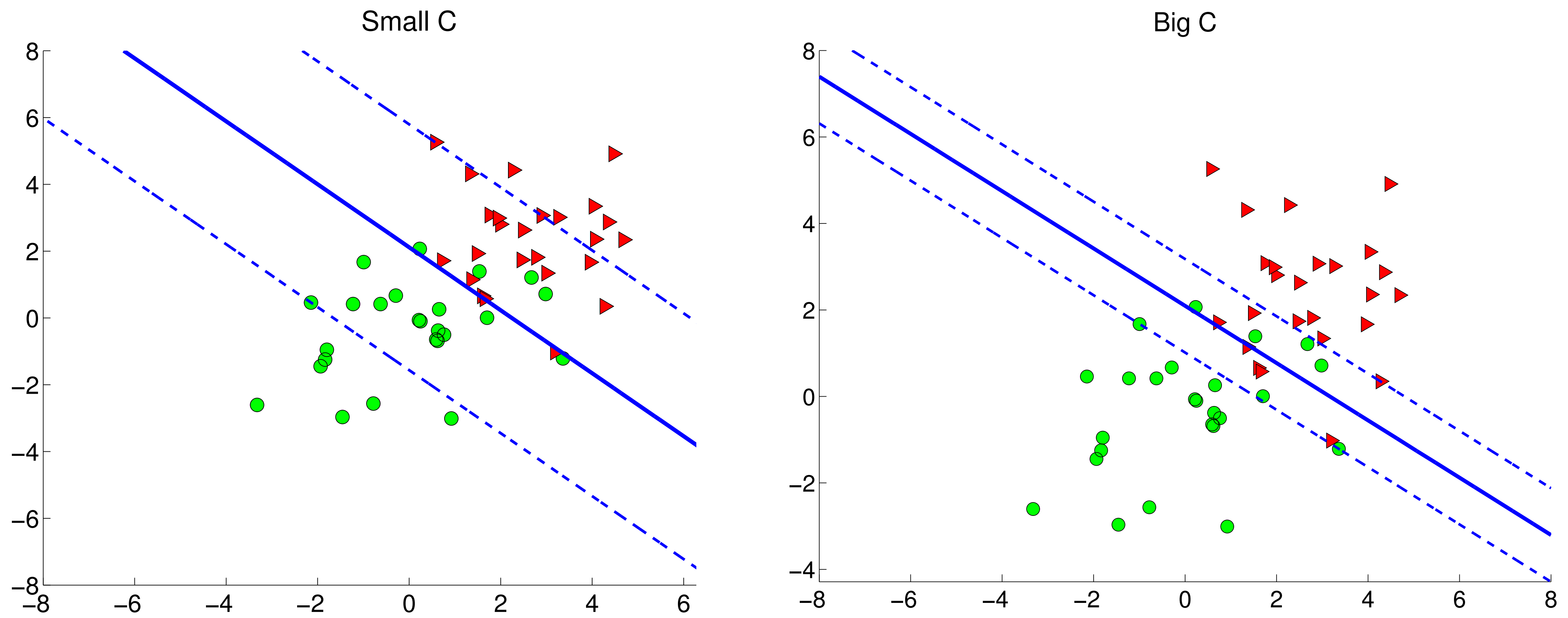
## Theorem [Solution d'un SVM linéaire : cas non séparable]

*Soit un problème de SVM linéaire non-séparable de fonction de décision $f(x) = w^\top x + b$. Le vecteur $w$ est défini par $w = \sum_{i=1}^{n} \alpha_i y_i x_i$ où les coefficients $\alpha_i$ sont solution du problem dual ci-dessus.*

Qu'est-ce qui a changé ? Rien sauf les contraintes sur $\alpha_i$ qui sont maintenant $0 \leq \alpha_i \leq C$.

# Illustrations

Résolution d'un SVM pour $C = 0.01$ petit et $C = 1000$ grand




Le choix de $C$ influence la solution : $C$ petit $\rightarrow$ marge grande ; $C$ grand $\rightarrow$ marge petite

# En pratique

## Elements d'entrée

Données étiquetées : $\{(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}\}_{i=1}^n$

## Méthodologie

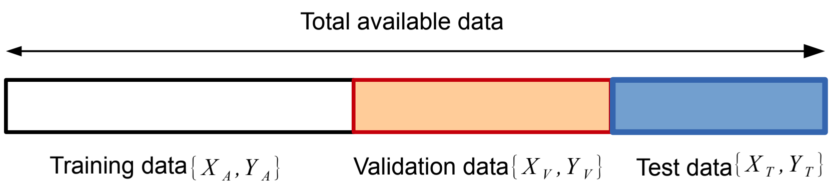
1. Centrer les données : $\{x_i\}_{i=1}^n \longrightarrow \{x_i = x_i - \bar{x}\}_{i=1}^n$
2. Fixer le paramètre $C > 0$ du SVM
3. Utiliser un solveur pour résoudre le problème dual et obtenir les $\alpha_i \neq 0$, les points supports $x_i$ correspondants et le biais $b$
4. En déduire la fonction de décision : $f(x) = \sum_{i \in SV} \alpha_i y_i x_i^\top x + b$
5. Evaluer l'erreur de généralisation du SVM obtenu (validation croisée ...). Recommencer à partir de l'étape 2 si elle n'est pa satisfaisante.

# Solveurs de SVM

- LibSVM http://www.csie.ntu.edu.tw/~cjlin/libsvm/
- ScikitLearn (Python)
  http://scikit-learn.org/stable/modules/svm.html

  ...

# Réglage du paramètre $C$ : une procédure pratique




- Ensemble d'apprentissage : pour calculer $w$ et $b$
- Ens. de validation : évaluer l'erreur de classification pour différents $C$
- Ens. de test : évaluation du "meilleur modèle"

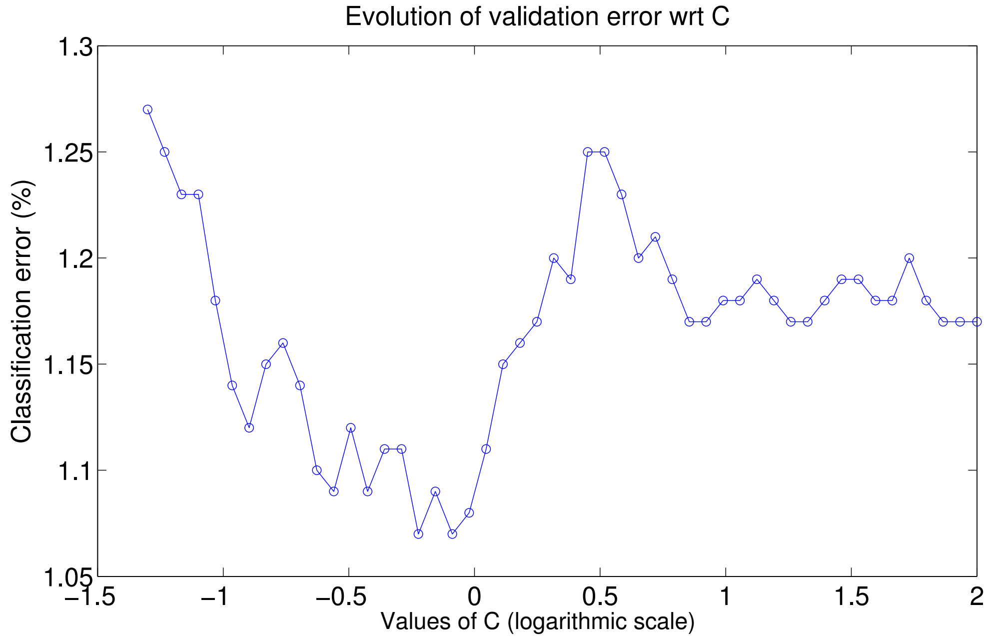
**Sélection de modèle : réglage de $C$**

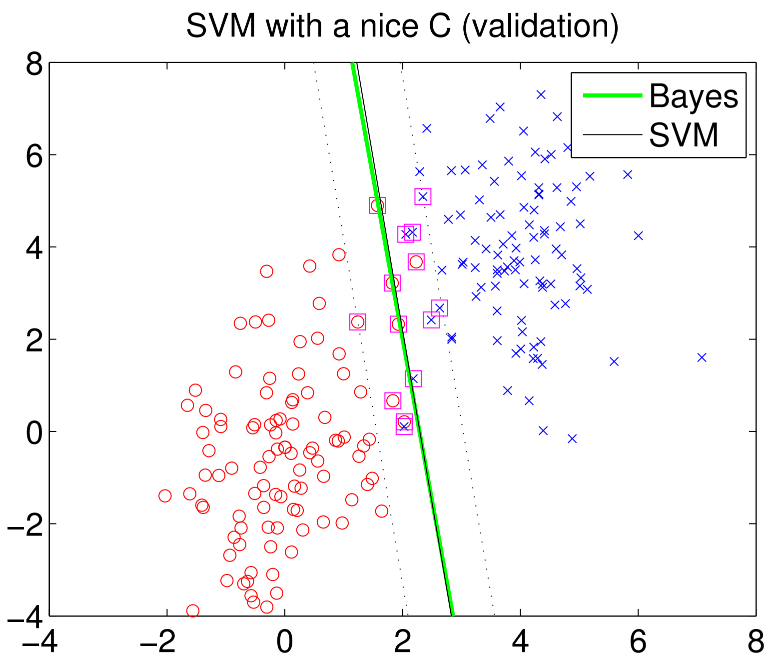
function $C \leftarrow$ `tuneC`$(\mathtt{X}, \mathtt{Y}, \mathtt{options})$

1. Split the data $(X_a, Y_a, X_v, Y_v) \leftarrow$ `SplitData`$(\mathtt{X}, \mathtt{Y}, \mathtt{options})$
2. Pour différentes valeurs de $C$
   - $(w, b) \leftarrow$ `TrainLinearSVM`$(X_a, Y_a, C, options)$
   - $error \leftarrow$ `EvaluateError`$(X_v, Y_v, w, b)$
3. $C \leftarrow \arg\min error$

# Exemple

- Les valeurs de $C$ choisies sur une échelle logarithmique
- pour chaque $C$, on apprend un SVM et on calcule son erreur de validation
- Le minimum de la courbe d'erreur correspond à la "meilleure" valeur $C^*$
- Le SVM correspond est sur la figure de droite



# Relation between soft-SVM, Hinge-loss, and Hinge-Loss Perceptron.

soft-SVM (SVM with slack variables)

$$\begin{cases} \min \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}\xi_i \\ y_i(\langle\omega,x_i\rangle+b) \geq 1-\xi_i \\ \xi_i \geq 0 \end{cases}$$

constraints on $\xi_i$:

$$\begin{cases} \xi_i \geq 0 \\ \xi_i \geq 1-y_i(\langle\omega,x_i\rangle+b) = 1-y_i s_i \end{cases} \quad \text{with } s_i := \langle\omega,x_i\rangle+b$$

$$\iff \xi_i \geq \max(0, 1-y_i s_i)$$

consider this optimisation sub-problem:

$$\begin{cases} \min \sum \xi_i \\ \text{s.t. } \xi_i \geq \max(0,1-y_i s_i) \end{cases} \xrightarrow{\text{solution}} \xi_i = \max(0,1-y_i s_i)$$

this is also the solution on $\xi_i$ to the original problem!

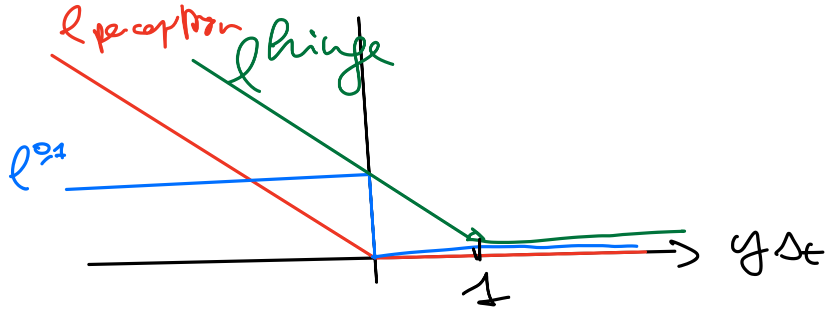
$$\boxed{\text{the soft SVM problem become: } \min_{\omega} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}\max(0, 1-y_i(\langle\omega,x_i\rangle+b))}$$

$$\Updownarrow$$

$$\min_{\omega} \frac{1}{2C}\|\omega\|^2 + \sum_{i=1}^{N} \ell^{hinge}(\langle\omega,x_i\rangle+b, y_i)$$

where $\ell^{hinge}(s_i, y_i) = \max(0, 1-y_i s_i)$

$$\ell^{perceptron}(s_t, y) = \max(0, -y s_t)$$




SGD on the objective function

$$\nabla_w \left( \frac{1}{2C}\|w\|^2 + \sum_i \ell^{hinge}(s_i, y_i)\right) = \frac{w}{C} + \sum_{i=1}^{N} \begin{cases} 0 & \text{if } y_i s_i > 1 \\ -y_i x_i & \text{else} \end{cases}$$

if $y_t \langle w_t, x_t \rangle < 1$

$\qquad w_{t+1} \leftarrow w_t + \alpha\, y_t x_t - \frac{\alpha}{C} w_t$

else

$\qquad w_{t+1} \leftarrow w_t - \frac{\alpha}{C} w_t$

SGD of soft SVM

⚠ $\ell^{hinge}(\ldots) \geq \ell^{01}(\ldots)$

# Conclusions

- Construction d'un hyperplan optimal au sens de la maximisation de la marge
- Une analyse théorique poussée montre que maximiser la marge équivaut à minimiser une borne sur l'erreur de généralisation.
- Le cas non linéaire (où on cherche une fonction de décision non-linéaire) peut être traité grâce aux noyaux.
- Généralisation possible au cas où on a plusieurs classes
- Algorithme de classification très utilisé en pratique ...