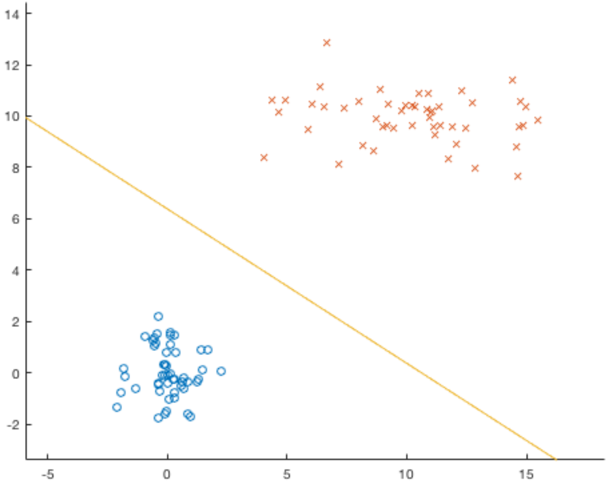
Support Vectors Machines

Jules Pondard - Joseph de Vilmarest

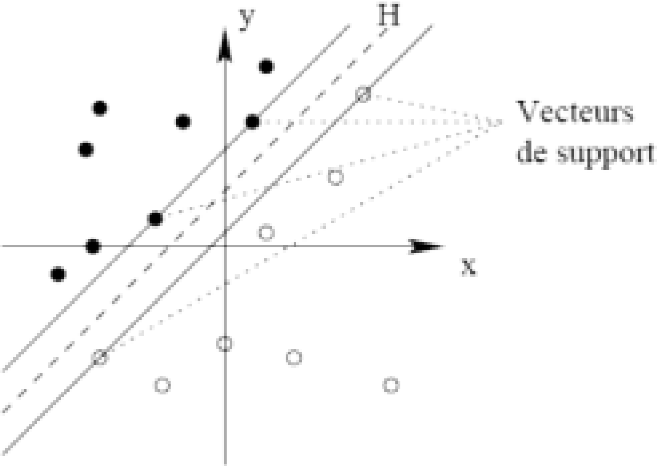
Optimisation Convexe et Combinatoire, M1 ENS

12 Janvier 2017


14
12
10
8
6
4
2
0
-2
-5
0
5
10
15

- entrée : n points $x_i \in \mathbb{R}^n$, d'étiquettes $y_i \in \{-1, 1\}$
- sortie : $w \in \mathbb{R}^n$ tel que le signe de $w^T x_i$ soit l'étiquette de x_i

SVM : cas séparable



SVM : cas séparable

On maximise l'écart entre deux hyperplans parallèles qui s'appuient sur des points de chaque catégorie.
Ainsi on maximise $\frac{1}{\|w\|_2}$ où $w \in \mathbb{R}^n$ avec $y_i(w^T x_i) \geq 1$ pour $1 \leq i \leq n$.
Cela revient à minimiser $\|w\|_2$ ou encore $\frac{1}{2}\|w\|_2^2$ sous les mêmes contraintes.

SVM : cas non séparable (marge souple)

minimiser $\frac{1}{2}\|w\|_2^2 + C1^T z$
sous les contraintes $y_i(w^T x_i) \geq 1 - z_i$ et $z_i \geq 0$ pour $1 \leq i \leq n$

On minimise

$$f(w,x) = t\left(\frac{w^T w}{2} + C1^T z\right) - \sum_{i=1}^{m} log(z_i) - \sum_{i=1}^{m} log(y_i(w^T x_i) - 1 + z_i)$$

Méthode de Newton : $x_{k+1} = x_k - \nabla^2 f^{-1} \nabla f$

Pour le gradient ∇f,

- $\frac{\partial f}{\partial w_i} = tw_i - \sum_{k=1}^{m} \frac{y_k x_k^i}{y_k(w^T x_k) - 1 + z_k}$
- $\frac{\partial f}{\partial z_i} = tC - \frac{1}{z_i} - \frac{1}{y_i(w^T x_i) - 1 + z_i}$

Pour la hessienne $\nabla^2 f$,

- $\frac{\partial^2 f}{\partial w_i \partial w_j} = t\delta_{i=j} + \sum_{k=1}^{m} \frac{x_k^i x_k^j}{(y_i(w^T x_i) - 1 + z_i)^2}$
- $\frac{\partial^2 f}{\partial w_i \partial z_j} = \frac{y_j x_j^i}{(y_j(w^T x_j) - 1 + z_j)^2}$
- $\frac{\partial^2 f}{\partial z_i \partial z_j} = \delta_{i=j}\left(\frac{t}{z_i^2} + \frac{1}{(y_i(w^T x_i) - 1 + z_i)^2}\right)$

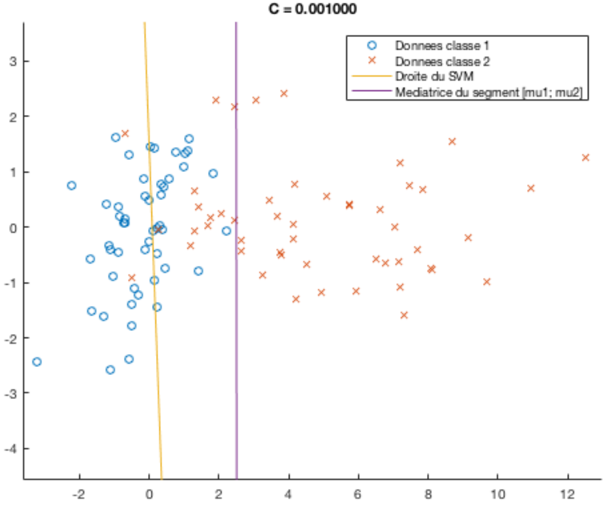
CVX est nettement plus rapide quand on augmente la dimension ou le nombre de points.
En effet $\nabla^2 f$ est de taille $(m+n) \times (m+n)$.

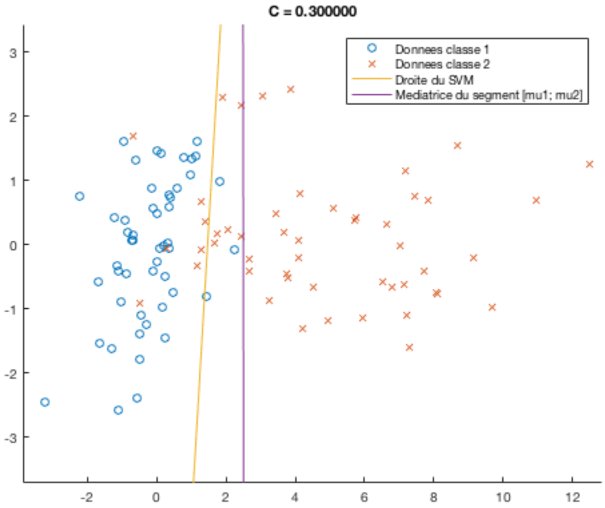
Coordinate Descent Method

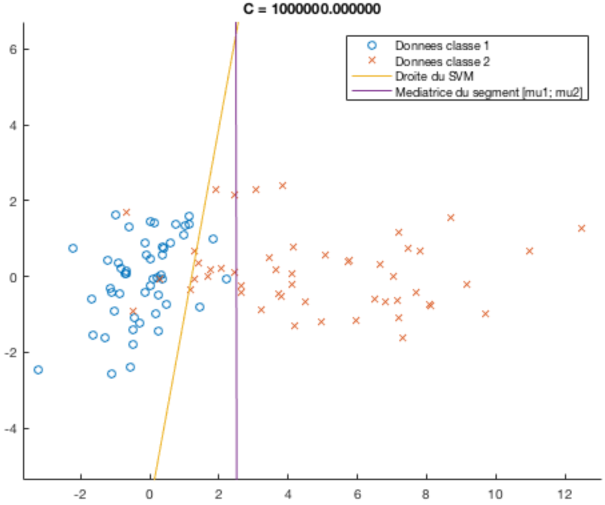
- Descente de Gradient Stochastique : on minimise selon une donnée à chaque étape.
- Coordinate Descent Method : on minimise selon une direction à chaque étape

On a testé sur différents exemples en 2D.

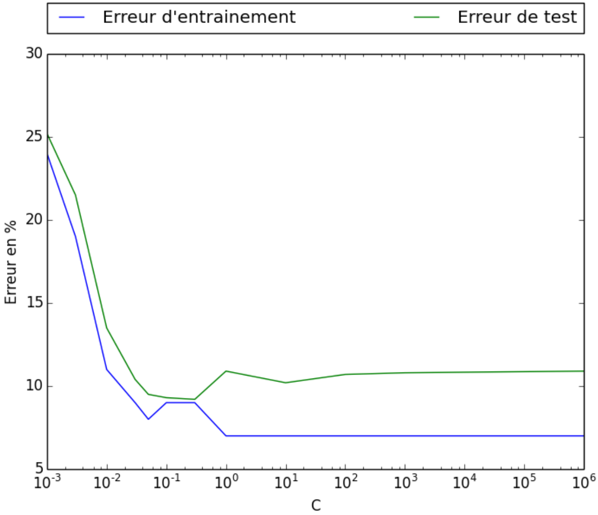
Ici : deux gaussiennes d'espérances $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ et $\begin{pmatrix} 7 \\ 0 \end{pmatrix}$ et de covariances $\Sigma_1 = I_2$ et $\Sigma_2 = \begin{pmatrix} 8 & 0 \\ 0 & 1 \end{pmatrix}$.


C = 0.001000
Donnees classe 1
Donnees classe 2
Droite du SVM
Mediatrice du segment [mu1; mu2]
3
2
1
0
-1
-2
-3
-4
-2
0
2
4
6
8
10
12

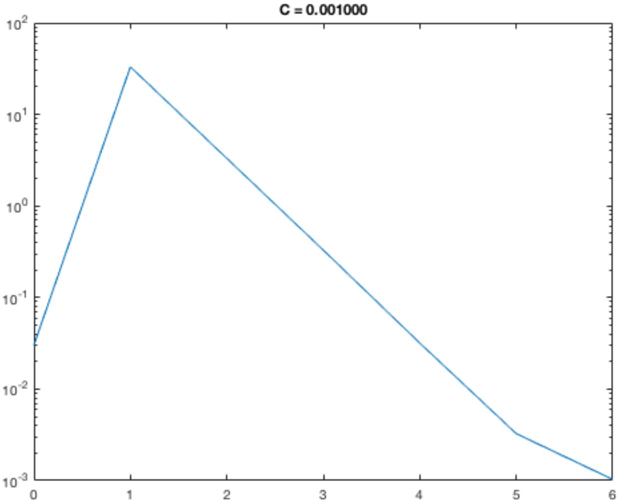

C = 0.300000
Donnees classe 1
Donnees classe 2
Droite du SVM
Mediatrice du segment [mu1; mu2]
3
2
1
0
-1
-2
-3
-2
0
2
4
6
8
10
12


C = 1000000.000000
Donnees classe 1
Donnees classe 2
Droite du SVM
Mediatrice du segment [mu1; mu2]
6
4
2
0
-2
-4
-2
0
2
4
6
8
10
12

Analyse selon C



Duality gap



Conclusion