SVM et machines à noyaux

Stéphane Canu
stephane.canu@litislab.eu

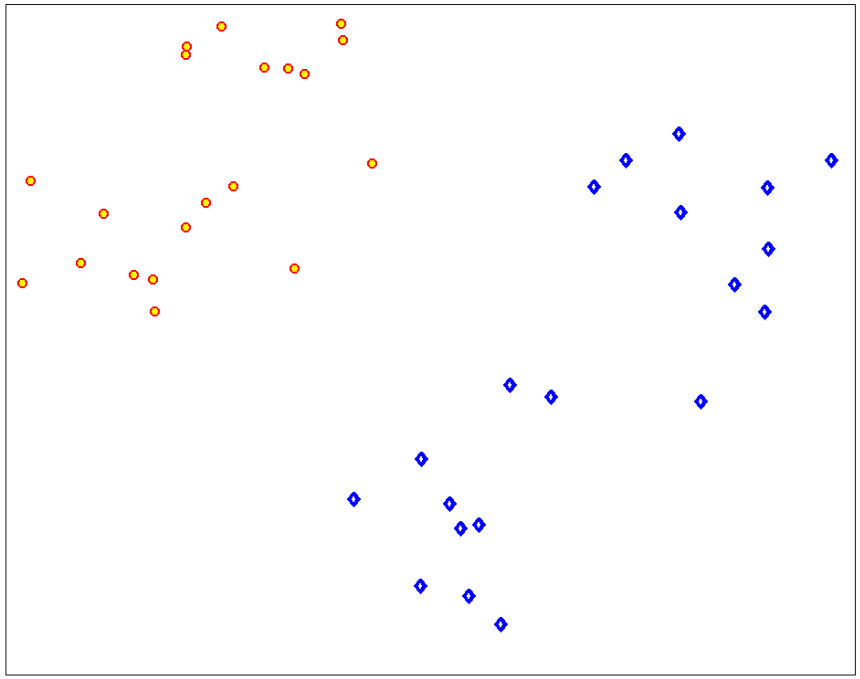
Ecole d'été du GRETSI - Peyresq

June 27, 2010

Plan

Le problème de discrimination linéaire

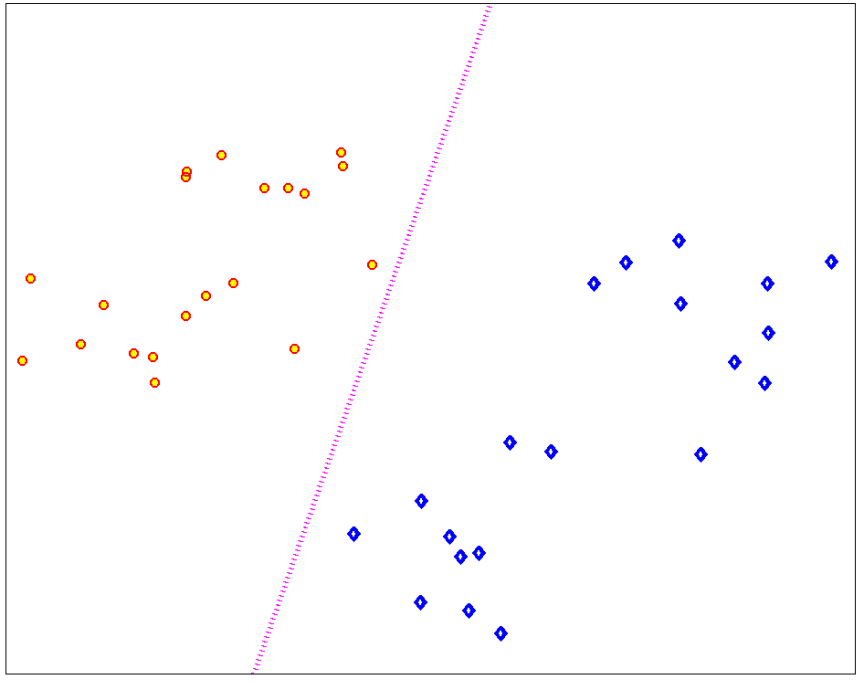
Trouver une droite que sépare les bleus des rouges



$$D(x) = \text{sign}\left(\mathbf{v}^\top \mathbf{x} + a\right)$$

Le problème de discrimination linéaire

Trouver une droite que sépare les bleus des rouges



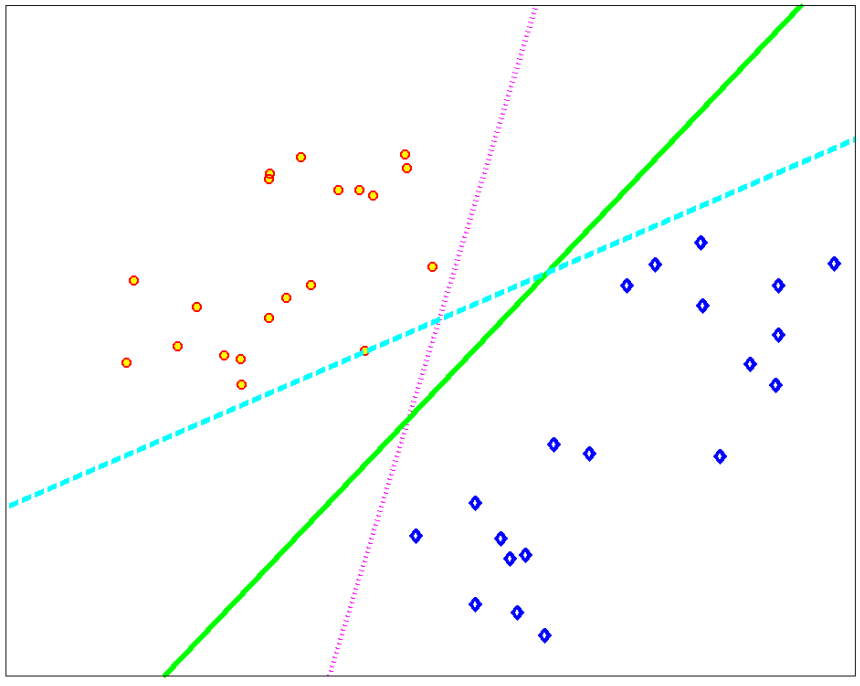
$$D(x) = \text{sign}\left(\mathbf{v}^\top \mathbf{x} + a\right)$$

la frontière de décision:

$$\mathbf{v}^\top \mathbf{x} + a = 0$$

Le problème de discrimination linéaire

Trouver une droite que sépare les bleus des rouges



$$D(x) = \text{sign}\left(\mathbf{v}^\top \mathbf{x} + a\right)$$

la frontière de décision:

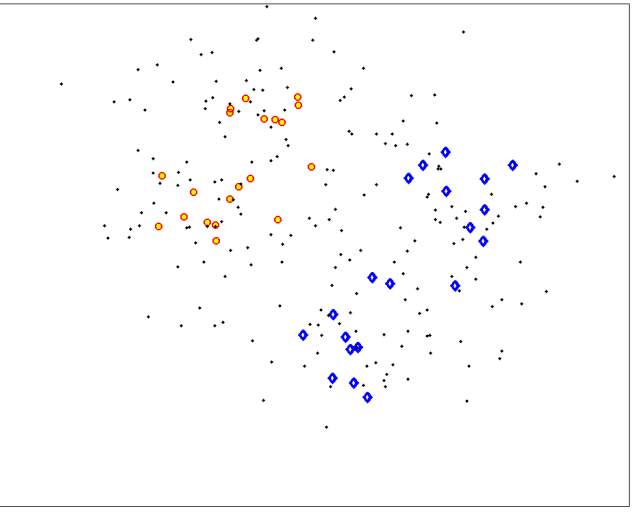
$$\mathbf{v}^\top \mathbf{x} + a = 0$$

Il y plusieurs solutions...
Le problème est mal posé

Quelle solution choisir ?

Ce problème n'est pas tout à fait celui qui nous intéresse

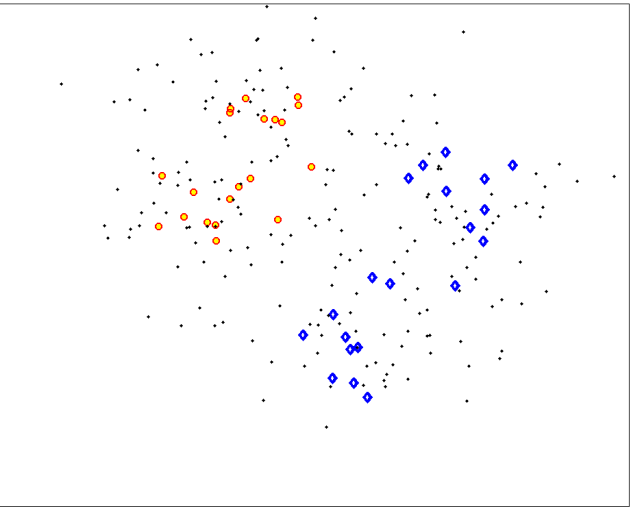
$\{(\mathbf{x}_i, y_i);\ i = 1 : n\}$ un échantillon tiré suivant $\mathbb{P}(\mathbf{x}, y)$ inconnue



Nous cherchons à bien classer des points issus de cette même distribution : minimiser $\mathbb{P}(\text{erreur})$

Ce problème n'est pas tout à fait celui qui nous intéresse

$\{(\mathbf{x}_i, y_i);\ i = 1 : n\}$ un échantillon tiré suivant $\mathbb{P}(\mathbf{x}, y)$ inconnue



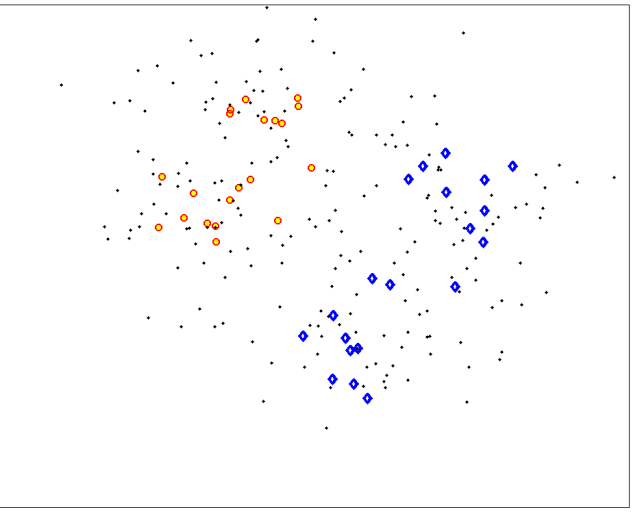
Nous cherchons à bien classer des points issus de cette même distribution : minimiser $\mathbb{P}(\text{erreur})$

A la recherche d'une méthode « universelle »

- être fidèle aux données : (faire **peu** d'erreurs)
- garantir que $\mathbb{P}(\textit{erreur})$ ne s'éloigne pas trop
- non linéaire
- passe à l'échelle - complexité algorithmique

Ce problème n'est pas tout à fait celui qui nous intéresse

$\{(\mathbf{x}_i, y_i);\ i = 1 : n\}$ un échantillon tiré suivant $\mathbb{P}(\mathbf{x}, y)$ inconnue



Nous cherchons à bien classer des points issus de cette même distribution : minimiser $\mathbb{P}(\text{erreur})$

A la recherche d'une méthode « universelle »

- être fidèle aux données : (faire **peu** d'erreurs)
- garantir que $\mathbb{P}(\textit{erreur})$ ne s'éloigne pas trop
- non linéaire
- passe à l'échelle - complexité algorithmique

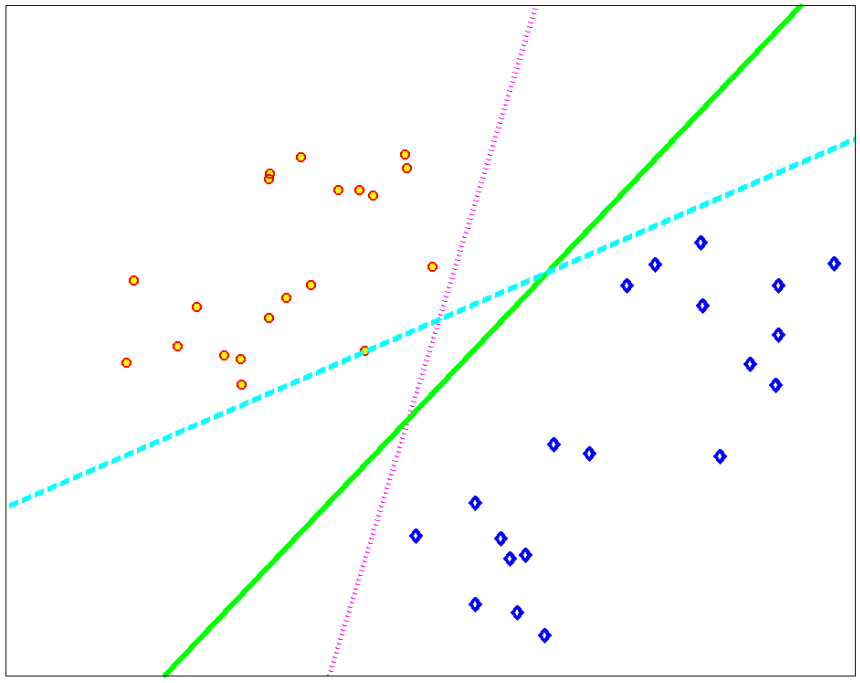
avec une grande probabilité :

$$\mathbb{P}(\text{erreur}) < \underbrace{\widehat{\mathbb{P}}(\text{erreur})}_{=0 \text{ ici}} + \varphi\Big(\underbrace{\|\mathbf{v}\|}_{\frac{1}{\text{marge}}} \Big)$$

Vapnik's Book, 1982

Le problème de discrimination linéaire

Trouver une droite que sépare les bleus des rouges



$$D(x) = \text{sign}\left(\mathbf{v}^\top \mathbf{x} + a\right)$$

la frontière de décision:

$$\mathbf{v}^\top \mathbf{x} + a = 0$$

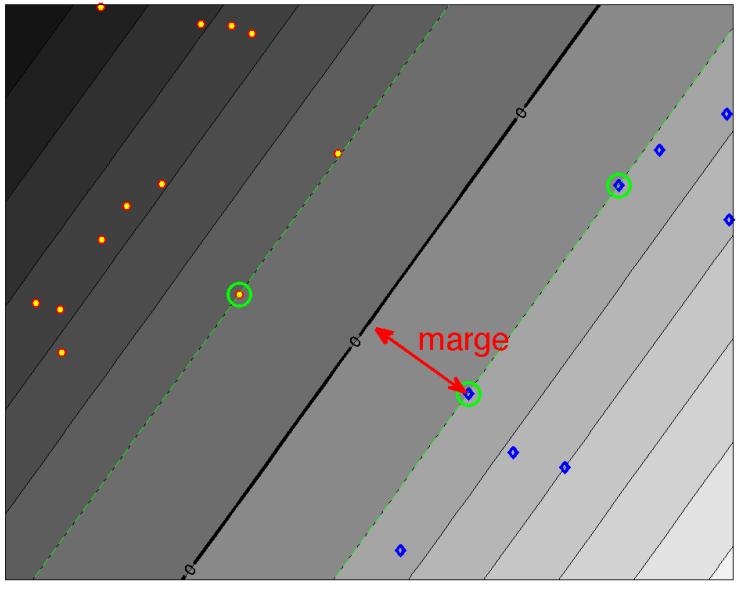
Il y plusieurs solutions...
Le problème est mal posé

Quelle solution choisir ?

$\Rightarrow$ Celle qui aura la plus grande marge

Maximiser la « confiance » = maximiser la marge

La frontière de décision : $\Delta(\mathbf{v}, a) = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{v}^\top \mathbf{x} + a = 0\}$



Maximiser la marge

$$\max_{\mathbf{v},a} \underbrace{\min_{i\in[1,n]} \text{dist}(\mathbf{x}_i, \Delta(\mathbf{v}, a))}_{\text{marge : } m}$$

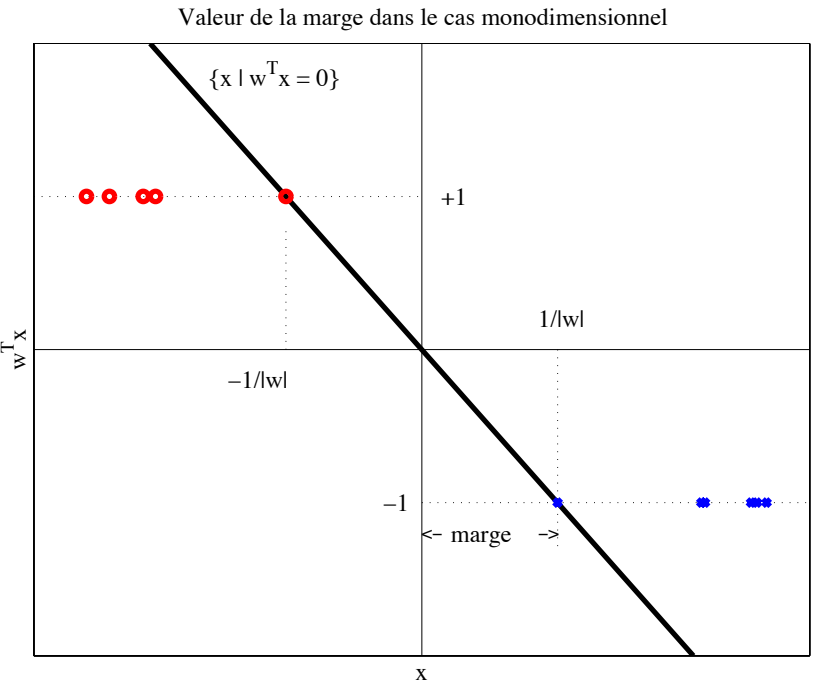
Maximiser la marge

$$\begin{cases} \max\limits_{\mathbf{v},a} & m \\ \text{avec} & \min\limits_{i=1,n} |\mathbf{v}^\top \mathbf{x}_i + a| \geq m \end{cases}$$

Le problème est toujours mal posé

si $(\mathbf{v}, a)$ est une solution, $\forall k \in \mathbb{R}^+$ $(k\mathbf{v}, ka)$ l'est aussi...

Marge, norme et régularité



Maximiser la marge

$$\begin{cases} \max\limits_{\mathbf{v},a} & m \\ \text{avec} & \min\limits_{i=1,n} |\mathbf{v}^\top \mathbf{x}_i + a| \geq m \\ & \|\mathbf{v}\|^2 = 1 \end{cases}$$

si le min est plus grand, tout le monde est plus grand ($y_i \in \{-1, 1\}$)

$$\begin{cases} \max\limits_{\mathbf{v},a} & m \\ \text{avec} & y_i(\mathbf{v}^\top \mathbf{x}_i + a) \geq m, \quad i = 1 : n \\ & \|\mathbf{v}\|^2 = 1 \end{cases}$$

changement de variable : $\mathbf{w} = \frac{\mathbf{v}}{m}$ et $b = \frac{a}{m} \implies \|\mathbf{w}\| = \frac{\|\mathbf{v}\|}{m}$

$$\begin{cases} \max\limits_{\mathbf{w},b} & m \\ \text{avec} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \quad ; \; i = 1, n \\ \text{et} & m = \frac{1}{\|\mathbf{w}\|} \end{cases}$$

$$\begin{cases} \min\limits_{\mathbf{w},b} & \|\mathbf{w}\|^2 \\ \text{avec} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \\ & i = 1, n \end{cases}$$

Le problème des SVM linéaires

Le problème des SVM linéaires (dans le primal)

soit $\{(\mathbf{x}_i, y_i);\ i = 1 : n\}$ un ensemble de données étiquetés avec $\mathbf{x}_i \in \mathbb{R}^d, y_i \in \{1, -1\}$.
Un séparateur à vaste marge linéaire (SVM) est un discriminateur de la forme : $D(x) = \text{sign}(\mathbf{w}^\top \mathbf{x} + b)$ où $\mathbf{w} \in \mathbb{R}^d$ et $b \in \mathbb{R}$ sont donnés par la résolutation du problème suivant :

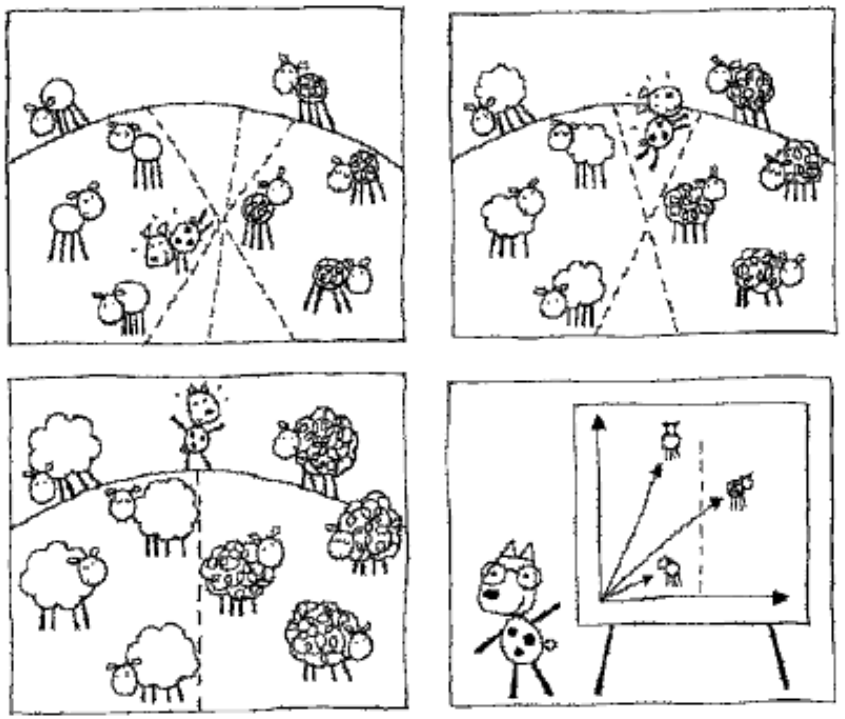
$$\begin{cases} \min\limits_{\mathbf{w},b} & \frac{1}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\mathbf{w}^\top \mathbf{w} \\ \text{avec} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \\ & i = 1, n \end{cases}$$

C'est un programe quadratique de la forme

$$\begin{cases} \min\limits_{\mathbf{z}} & \frac{1}{2}\mathbf{z}^\top A\mathbf{z} - \mathbf{d}^\top \mathbf{z} \\ \text{avec} & B\mathbf{z} \leq \mathbf{e} \end{cases}$$

$\mathbf{z} = (\mathbf{w}, b)^\top$, $\mathbf{d} = (0, \ldots, 0)^\top$, $A = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$, $B = -[\mathbf{y}X, \mathbf{y}]$ et $\mathbf{e} = -(1, \ldots, 1)^\top$

Le problème de discrimination linéaire



...the story of the sheep dog who was herding his sheep, and serendipitously invented the large margin classification and Sheep Vectors ...

(drawing by Ana Martin Larranaga)

from *Learning with Kernels*, B. Schölkopf and A. Smolla, MIT Press, 2002.

Formulation duale des SVM linéaires - Le lagrangien

$$\begin{cases} \min\limits_{\mathbf{w},b} & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{avec} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \\ & i = 1, n \end{cases}$$

on recherche un point selle du lagrangien $\max\limits_{\alpha} \min\limits_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \alpha)$ avec des multiplicateurs de lagrange $\alpha_i \geq 0$

$$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1\right)$$

α_i traduit l'influence de la contrainte et donc l'influence du point (x_i, y_i)

Conditions d'optimalité

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

Les gradients :

$$\begin{cases} \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, b, \alpha) & = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \\ \dfrac{\partial \mathcal{L}(\mathbf{w}, b, \alpha)}{\partial b} & = \sum_{i=1}^{n} \alpha_i \; y_i \end{cases}$$

on écrit les conditions d'optimalité :

$$\begin{cases} \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, b, \alpha) & = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \\ \dfrac{\partial \mathcal{L}(\mathbf{w}, b, \alpha)}{\partial b} & = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \alpha_i \; y_i = 0 \end{cases}$$

Formulation duale des SVM linéaires

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

Optimalité : $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \qquad \sum_{i=1}^{n} \alpha_i \, y_i = 0$

$$\begin{aligned}\mathcal{L}(\alpha) &= \tfrac{1}{2} \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_j \alpha_i y_i y_j \mathbf{x}_j^\top \mathbf{x}_i}_{\mathbf{w}^\top \mathbf{w}} - \textstyle\sum_{i=1}^{n} \alpha_i y_i \underbrace{\sum_{j=1}^{n} \alpha_j y_j \mathbf{x}_j^\top}_{\mathbf{w}^\top} \mathbf{x}_i - b \underbrace{\sum_{i=1}^{n} \mathbf{x}_i y_i}_{=0} + \textstyle\sum_{i=1}^{n} \alpha_i \\ &= -\frac{1}{2} \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_j \alpha_i y_i y_j \mathbf{x}_j^\top \mathbf{x}_i + \sum_{i=1}^{n} \alpha_i\end{aligned}$$

Formulation duale des SVM linéaires

$$\begin{cases} \min\limits_{\alpha \in \mathbf{R}^n} & \frac{1}{2}\alpha^\top G \alpha - \mathbf{e}^\top \alpha \\ \text{avec} & \mathbf{y}^\top \alpha = 0 \\ \text{et} & 0 \leq \alpha_i \qquad\qquad i = 1, n \end{cases}$$

avec G une matrice symétrique $n \times n$ de terme général $G_{ij} = y_i y_j \mathbf{x}_j^\top \mathbf{x}_i$

L'exemple des moindres carrés

Le modèle linéaire

$$y_i = \sum_{j=1}^{d} \beta_j x_{ij} + \varepsilon_i \quad , \qquad i = 1, n$$

n observations et d variables; $d < n$

$$\min_{\beta} = \sum_{i=1}^{n} \left(\sum_{j=1}^{d} x_{ij}\beta_j - y_i \right)^2 = \|X\beta - Y\|^2$$

Solution: $\widetilde{\beta} = (X^\top X)^{-1} X^\top Y$

$$f(\mathbf{x}) = \mathbf{x}^\top \underbrace{(X^\top X)^{-1} X^\top Y}_{\widetilde{\beta}}$$

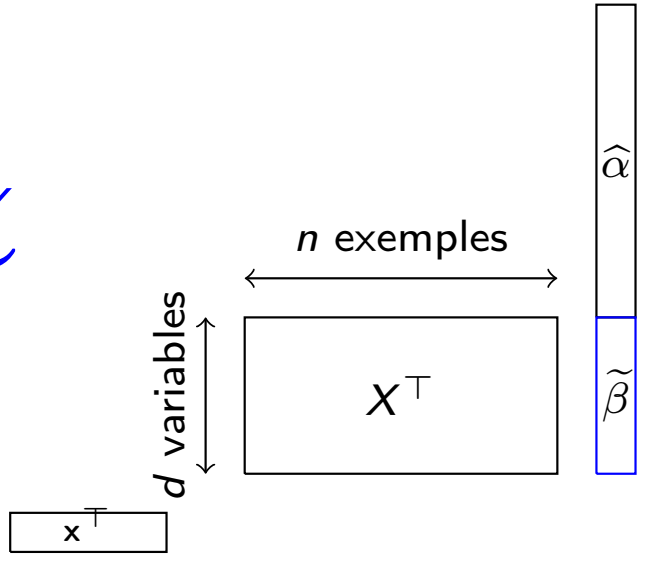
Quelle est l'influence de chacun des exemples (les lignes de X) ?

L'influence des exemples

pour une nouvelle observation $\mathbf{x}$

$$f(\mathbf{x}) = \mathbf{x}^\top (X^\top X)(X^\top X)^{-1} \underbrace{(X^\top X)^{-1} X^\top Y}_{\widetilde{\beta}}$$

$$= \mathbf{x}^\top X^\top \underbrace{X(X^\top X)^{-1}(X^\top X)^{-1} X^\top Y}_{\widehat{\alpha}}$$



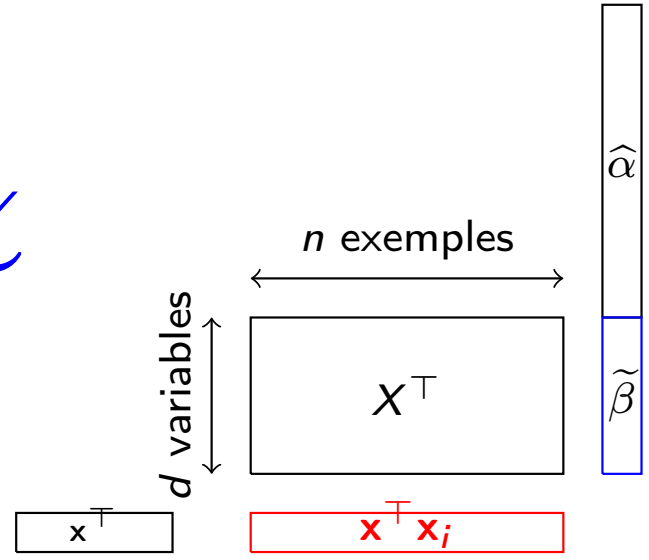
$$f(\mathbf{x}) = \sum_{j=1}^{d} \widetilde{\beta}_j x_j$$

L'influence des exemples

pour une nouvelle observation $\mathbf{x}$

$$f(\mathbf{x}) = \mathbf{x}^\top (X^\top X)(X^\top X)^{-1} \underbrace{(X^\top X)^{-1} X^\top Y}_{\widetilde{\beta}}$$

$$= \mathbf{x}^\top X^\top \underbrace{X(X^\top X)^{-1}(X^\top X)^{-1} X^\top Y}_{\widehat{\alpha}}$$



$$f(\mathbf{x}) = \sum_{j=1}^{d} \widetilde{\beta}_j x_j = \sum_{i=1}^{n} \widehat{\alpha}_i \, (\mathbf{x}^\top \mathbf{x}_i)$$

des variables aux exemples

$$\underbrace{\widehat{\alpha} = X(X^\top X)^{-1} \widetilde{\beta}}_{n \text{ examples}} \quad \text{et} \quad \underbrace{\widetilde{\beta} = X^\top \widehat{\alpha}}_{d \text{ variables}}$$

et si $d \geq n$!

SVM primal vs. dual

Primal

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{avec} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \\ & i = 1, n \end{cases}$$

- $d + 1$ inconnues
- n contraintes
- QP classique
- parfait si $d << n$

Dual

$$\begin{cases} \min\limits_{\alpha\in\mathbb{R}^n} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top\alpha \\ \text{avec} & \mathbf{y}^\top\alpha = 0 \\ \text{et} & 0 \leq \alpha_i \qquad\qquad i = 1, n \end{cases}$$

- n inconnues
- G matrice des influences de chaque couple de points
- n contraintes de boites
- plus facile à résoudre
- à utiliser si $d > n$

SVM primal vs. dual

Primal

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{avec} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \\ & i = 1, n \end{cases}$$

- $d + 1$ inconnues
- n contraintes
- QP classique
- parfait si $d << n$

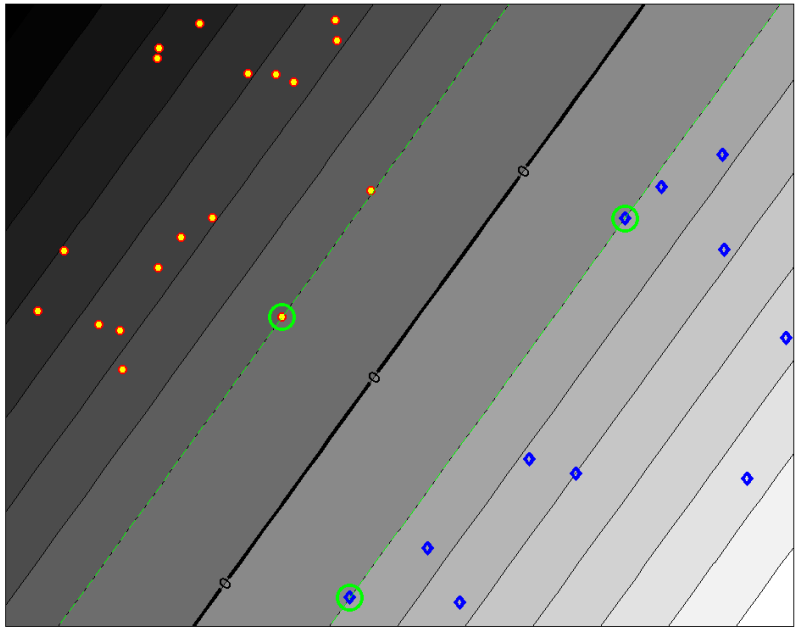
Dual

$$\begin{cases} \min\limits_{\alpha\in\mathbb{R}^n} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top\alpha \\ \text{avec} & \mathbf{y}^\top\alpha = 0 \\ \text{et} & 0 \leq \alpha_i \qquad\qquad i = 1, n \end{cases}$$

- n inconnues
- G matrice des influences de chaque couple de points
- n contraintes de boites
- plus facile à résoudre
- à utiliser si $d > n$

$$f(\mathbf{x}) = \sum_{j=1}^{d} w_j x_j + b = \sum_{i=1}^{n} \alpha_i\, y_i(\mathbf{x}^\top \mathbf{x}_i) + b$$

Méthodes de résolution du problème dual

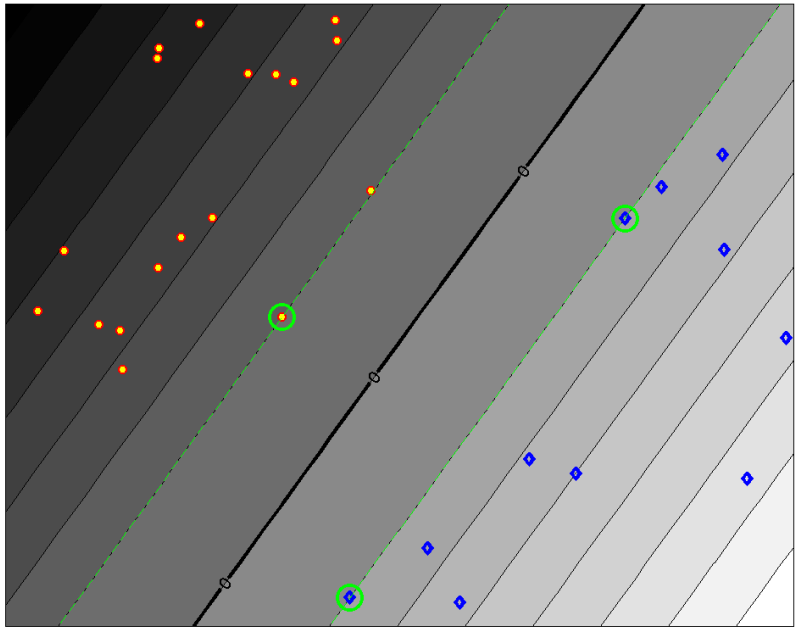


l'influence des points

- $\alpha_i = 0$ le point est inutile
- $\alpha_i \neq 0$ le point est dit support

$$f(\mathbf{x}) = \sum_{j=1}^{d} w_j x_j + b = \sum_{i=1}^{n} \alpha_i \, y_i(\mathbf{x}^\top \mathbf{x}_i) + b$$

Méthodes de résolution du problème dual



l'influence des points

- $\alpha_i = 0$ le point est inutile
- $\alpha_i \neq 0$ le point est dit support

$$f(\mathbf{x}) = \sum_{j=1}^{d} w_j x_j + b = \sum_{i=1}^{3} \alpha_i \, y_i (\mathbf{x}^\top \mathbf{x}_i) + b$$

La fonction de décision ne dépend plus que de trois points $(d+1)$

Supposons que nous connaissions les 3 points en question

Conclusion : variables ou exemples ?

- a la recherche d'une méthode d'apprentissage universelle
  - pas de modèle
- le cas linéaire - séparable
  - le cas non séparable est analogue
- double objectif : minimiser les erreur et la régularité de la solution
  - otimisation multi critère
- dualié : variable – exemple
  - quand utiliser l'une ou l'autre des formulations ?
- universalité = nonlinéarité
  - les noyaux