Machine à vecteurs supports

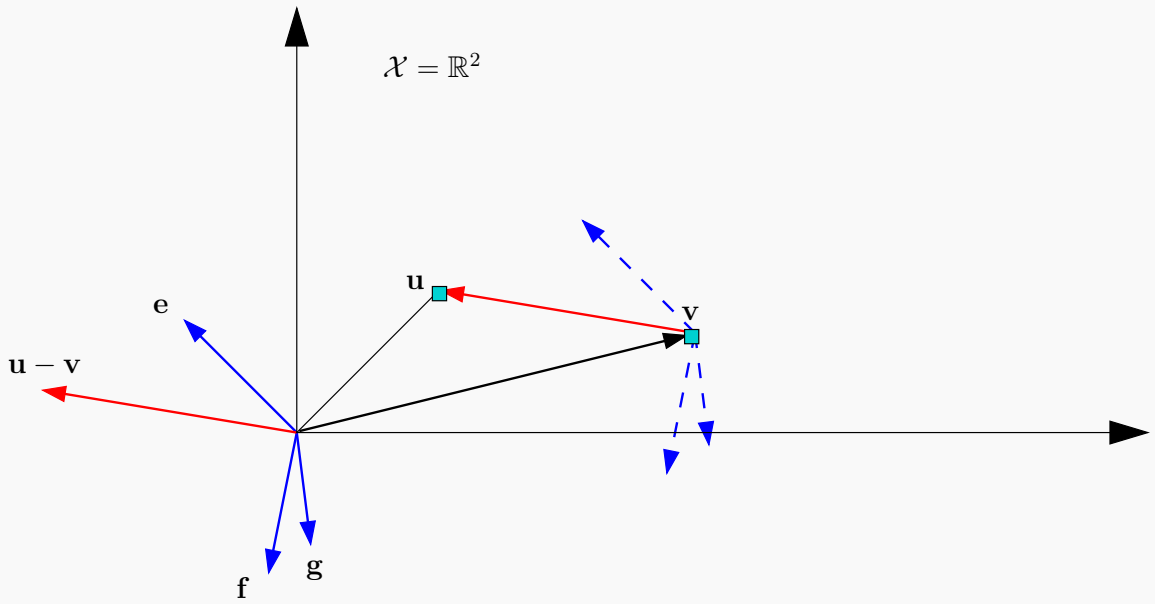
Gilles Cohen

25 février 2018

Vecteurs et produit scalaire 2/3

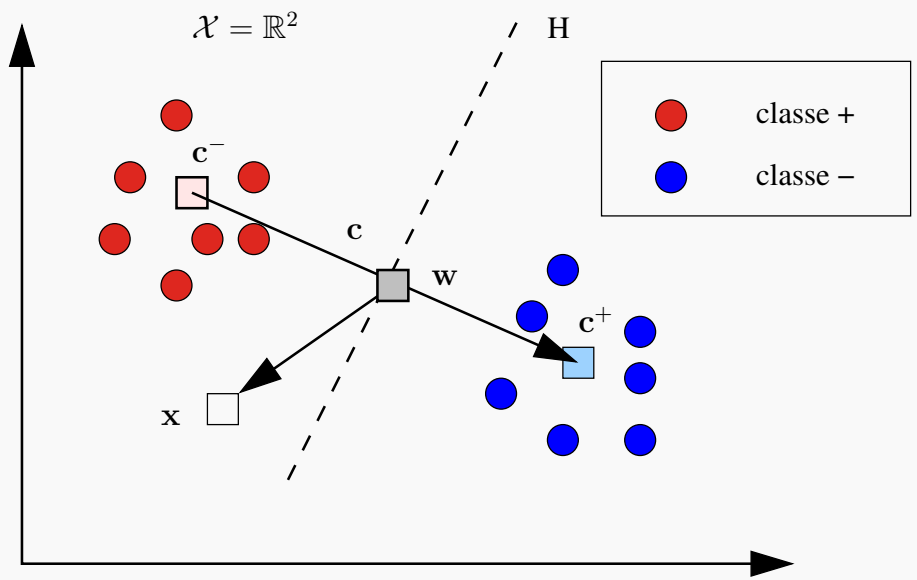
- Produit scalaire $\langle ., . \rangle : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$:
  - symétrique : $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$
  - bilinéaire : $\langle \lambda \mathbf{u}_1 + \gamma \mathbf{u}_2, \mathbf{v} \rangle = \lambda \langle \mathbf{u}_1, \mathbf{v} \rangle + \gamma \langle \mathbf{u}_2, \mathbf{v} \rangle$
  - positif : $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$
  - défini : $\langle \mathbf{u}, \mathbf{u} \rangle = 0 \Rightarrow \mathbf{u} = 0$
- Un produit scalaire
  - Donne une structure à $\mathcal{X}$
  - peut être vu comme une “similarité”
  - défini une norme $\|.\|$ sur $\mathcal{X}$: $\sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$

Vecteurs et produit scalaire 3/3



- $\langle \mathbf{u} - \mathbf{v}, \mathbf{e} \rangle > 0$: $\mathbf{u} - \mathbf{v}$ et $\mathbf{e}$ pointent dans la même direction
- $\langle \mathbf{u} - \mathbf{v}, \mathbf{f} \rangle = 0$: $\mathbf{u} - \mathbf{v}$ et $\mathbf{f}$ sont orthogonaux
- $\langle \mathbf{u} - \mathbf{v}, \mathbf{g} \rangle < 0$: $\mathbf{u} - \mathbf{v}$ et $\mathbf{g}$ pointent dans des directions opposées

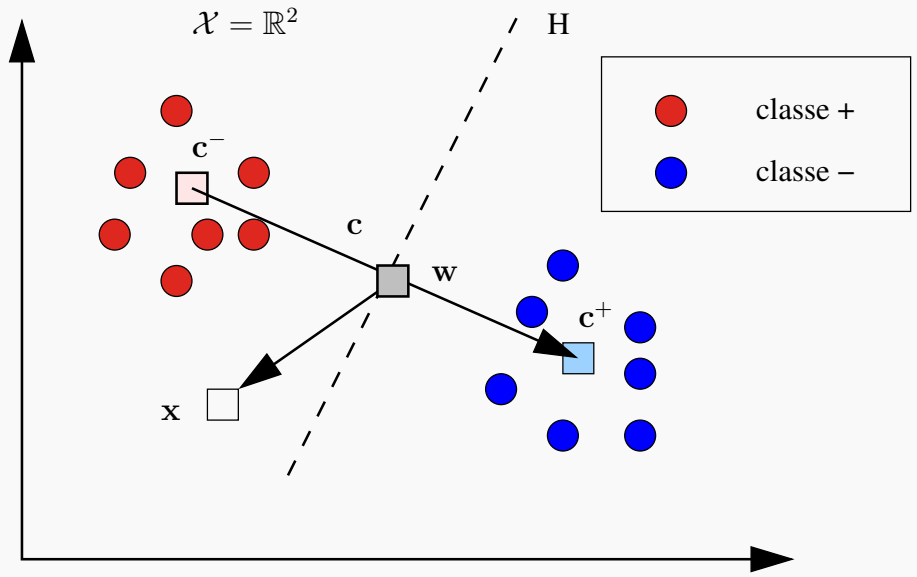
Un classifieur linéaire simple 1/3



- $\mathbf{c}^+ = \frac{1}{n^+} \sum_{i:y_i=+1} \mathbf{x}_i$
- $\mathbf{c}^- = \frac{1}{n^-} \sum_{i:y_i=-1} \mathbf{x}_i$
- $\mathbf{c} = \frac{1}{2}(\mathbf{c}^+ + \mathbf{c}^-)$
- $\mathbf{w} = \mathbf{c}^+ - \mathbf{c}^-$

- Classifier les exemples $\mathbf{x}$ selon leur distance par rapport à la moyenne c^+ ou c^- des classes :
  - $\forall \mathbf{x} \in \mathcal{X}$, on prend le signe de $\langle \mathbf{w}, \mathbf{x} - \mathbf{c} \rangle$, si on pose $h(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} - \mathbf{c} \rangle$, on a le classifieur $f(\mathbf{x}) = signe(h(\mathbf{x}))$
  - l'hyperplan $H(\mathbf{w})$ en pointillé est la surface de décision

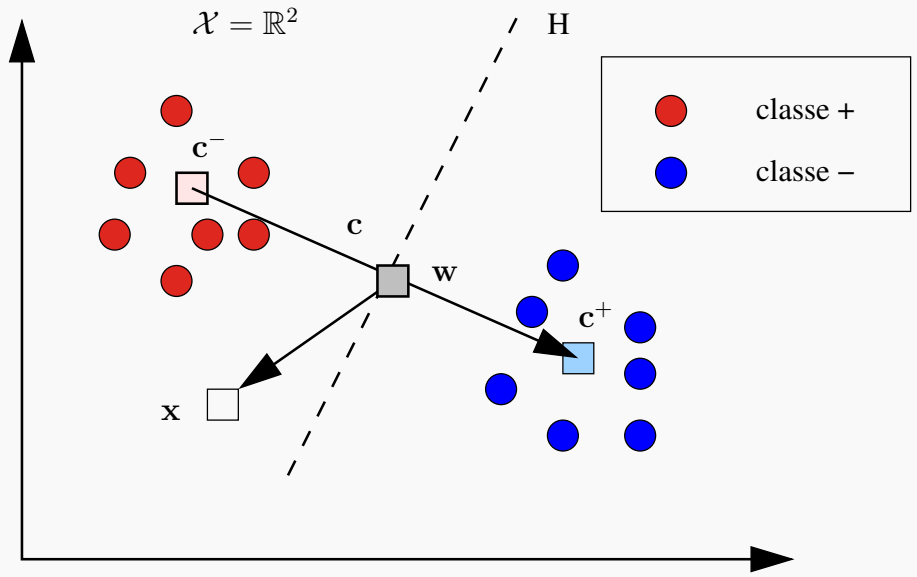
Un classifieur linéaire simple 2/3



- $\mathbf{c}^+ = \frac{1}{n^+} \sum_{i:y_i=+1} \mathbf{x}_i$
- $\mathbf{c}^- = \frac{1}{n^-} \sum_{i:y_i=-1} \mathbf{x}_i$
- $\mathbf{c} = \frac{1}{2}(\mathbf{c}^+ + \mathbf{c}^-)$
- $\mathbf{w} = \mathbf{c}^+ - \mathbf{c}^-$

$$
\begin{aligned}
\text{On évalue } h(\mathbf{x}) &= \langle \mathbf{w}, \mathbf{x} - \mathbf{c} \rangle \\
&= \langle (\mathbf{x} - (\mathbf{c}^+ + \mathbf{c}^-)/2), (\mathbf{c}^+ - \mathbf{c}^-) \rangle \\
&= \langle \mathbf{x}, \mathbf{c}^+ \rangle - \langle \mathbf{x}, \mathbf{c}^- \rangle + b \\
&= \langle \mathbf{x}, (\mathbf{c}^+ - \mathbf{c}^-) \rangle + b \\
\text{où} \qquad b &= \frac{1}{2}(\|\mathbf{c}^-\|^2 - \|\mathbf{c}^+\|)
\end{aligned}
$$

Un classifieur linéaire simple 3/3



- $\mathbf{c}^+ = \frac{1}{n^+} \sum_{i:y_i=+1} \mathbf{x}_i$
- $\mathbf{c}^- = \frac{1}{n^-} \sum_{i:y_i=-1} \mathbf{x}_i$
- $\mathbf{c} = \frac{1}{2}(\mathbf{c}^+ + \mathbf{c}^-)$
- $\mathbf{w} = \mathbf{c}^+ - \mathbf{c}^-$

- On obtient en exprimant la classification en terme de $\{\mathbf{x}_i, y_i\}$

$$h(\mathbf{x}) = \sum_{i=1,\ldots,n} \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b, \; b \in \mathbb{R}$$

$$\text{avec} \quad : \quad \alpha_{i:y_i=+1} = 1/n^+ \text{ et } \alpha_{i:y_i=-1} = 1/n^-$$

Question : Que faire si l'ensemble d'apprentissage n'est pas linéairement séparable ?

L'astuce du noyau (1/4)

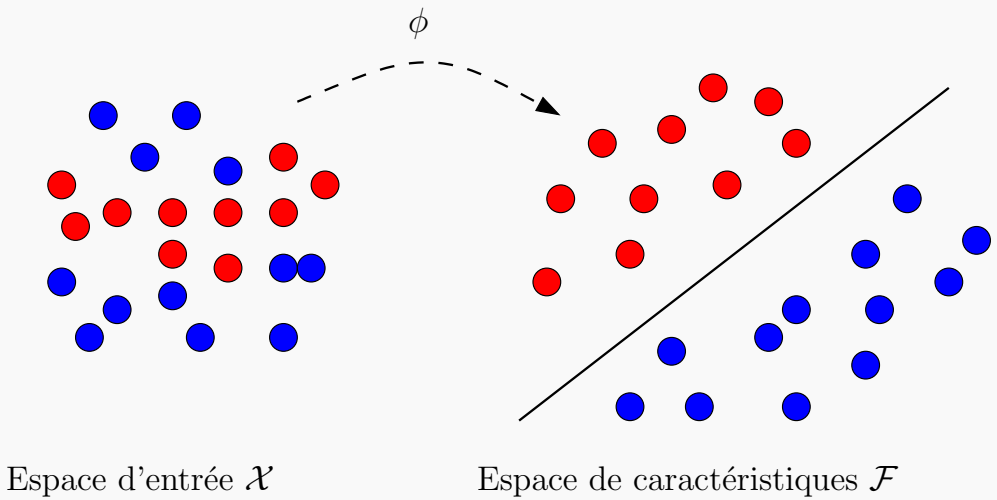
- Contexte : Un ensemble de données non linéairement séparable S_n : $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$
- Idée
  - choisir une transformation non linéaire ϕ

$$\begin{aligned} \phi : \quad & \mathcal{X} \to \mathcal{F} \\ & \mathbf{x} \to \phi(\mathbf{x}) \end{aligned}$$

    où $\mathcal{F}$ est un espace vectoriel appelé *espace de caractéristiques*
  - trouver un classifieur linéaire (i.e. un hyperplan séparateur) dans $\mathcal{F}$ pour classifier $\{(\phi(\mathbf{x}_1), y_1), \ldots, (\phi(\mathbf{x}_n), y_n)\}$

L'astuce du noyau (2/4)

- Classification linéaire dans l'espace de caractéristiques



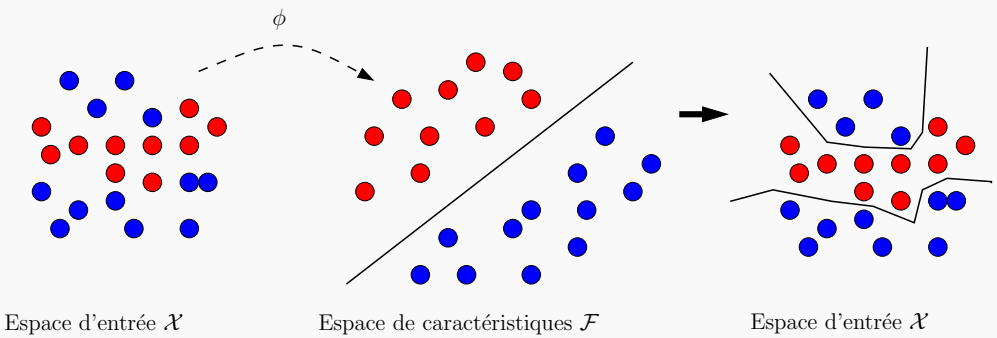
- prendre le classifieur linéaire précédent et l'implémenter dans $\mathcal{H}$

$$h(\mathbf{x}) = \sum_{i=1,\ldots,n} \alpha_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle + b$$

L'astuce du noyau (3/4)

- L'astuce du noyau peut s'appliquer si il y a une fonction $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ telle que : $k(\mathbf{u}, \mathbf{v}) = \langle \phi(\mathbf{u}), \phi(\mathbf{v}) \rangle_{\mathcal{F}}$. Dans ce cas toutes les occurences de $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle$ sont remplacées par $k(\mathbf{x}, \mathbf{x}_i)$
- Les noyaux doivent vérifier certaines propriétés pour être des noyaux valides.
  - S'assurer qu'il existe un espace $\mathcal{F}$ et une transformation $\phi : \mathcal{X} \rightarrow \mathcal{F}$ t. q. $k(\mathbf{u}, \mathbf{v}) = \langle \phi(\mathbf{u}), \phi(\mathbf{v}) \rangle_{\mathcal{F}}$
- k peut être vu comme une mesure de similarité

L'astuce du noyau 4/4



- Stratégie
  - considérer un problème non linéaire sur $\mathcal{X} \times \mathcal{Y}$
  - choisir un algorithme de classification linéaire (exp. en termes de $\langle ., . \rangle$)
  - remplacer toutes les occurences de $\langle ., . \rangle$ par $k(., .)$
- Classifieur obtenu :

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1,\ldots,n} \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b\right)$$

Matrice noyau (1/3)

- Soit $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ un noyau de Mercer
  - pour un ensemble $S_n : \{\mathbf{x}_i\}_{i=1}^n$

$$K = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \dots & k(\mathbf{x}_1, \mathbf{x}_n) \\ k(\mathbf{x}_1, \mathbf{x}_2) & k(\mathbf{x}_2, \mathbf{x}_2) & \dots & k(\mathbf{x}_2, \mathbf{x}_n) \\ \dots & \dots & \dots & \dots \\ k(\mathbf{x}_1, \mathbf{x}_n) & k(\mathbf{x}_2, \mathbf{x}_n) & \dots & k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}$$

est la matrice de Gram (ou noyau) de k

Matrice noyau (2/3)

- Une propriété de la matrice de Gram (Mercer) :
  - Soit $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ une fonction symétrique.
    k est un noyau de Mercer $\Leftrightarrow$

$$\forall S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathcal{X}, \mathbf{v}^T K \mathbf{v} \geq 0, \forall \mathbf{v} \in \mathbb{R}^n$$

- Ce qui signifie que pour tout noyau de Mercer k et tout ensemble S, la matrice de Gram K a seulement des valeurs propres non négatives i.e. elle est toujours positive semi-definie.

Matrice noyau (3/3)

Une matrice $K \in \mathbb{R}^{n \times n}$ est semi définie positive si pour tous les $v \in \mathbb{R}^n$ on a $\mathbf{v}^T K \mathbf{v} \geq 0$.

- Preuve :

$$
\begin{aligned}
\mathbf{v}^T K \mathbf{v} &= \sum_{i,j=1}^{n} v_i v_j k(x_i, x_j) \\
&= \sum_{i,j=1}^{n} v_i v_j \langle \phi(x_i), \phi(x_j) \rangle \\
&= \left\langle \sum_{i=1}^{n} v_i \phi(x_i), \sum_{i=1}^{n} v_j \phi(x_j) \right\rangle \\
&= \left\| \sum_{i=1}^{n} v_i \phi(x_i) \right\|^2 \\
&\geq 0.
\end{aligned}
$$

Noyau courant

- Linéaire

$$k(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle$$

- Polynomial

$$k(x, x') = (\langle \mathbf{x}, \mathbf{z} \rangle)^d \text{ ou } (c + \langle \mathbf{x}, \mathbf{z} \rangle)^d$$

- Gaussien

$$k(\mathbf{x}, \mathbf{z}) = e^{-\|\mathbf{x}-\mathbf{x}'\|^2/2\sigma^2}$$

- Laplacien

$$k(\mathbf{x}, \mathbf{z}) = e^{\|\mathbf{x}-\mathbf{z}\|/\sigma}$$

Noyau polynomial

Noyau polynomial : $k(\mathbf{x}, \mathbf{z})) = \langle \mathbf{x}, \mathbf{z} \rangle^p$ Example dans $\mathbb{R}^2$:
Trouver $\phi(\mathbf{x})$ tel que : $\langle \mathbf{x}, \mathbf{z} \rangle^2 = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$

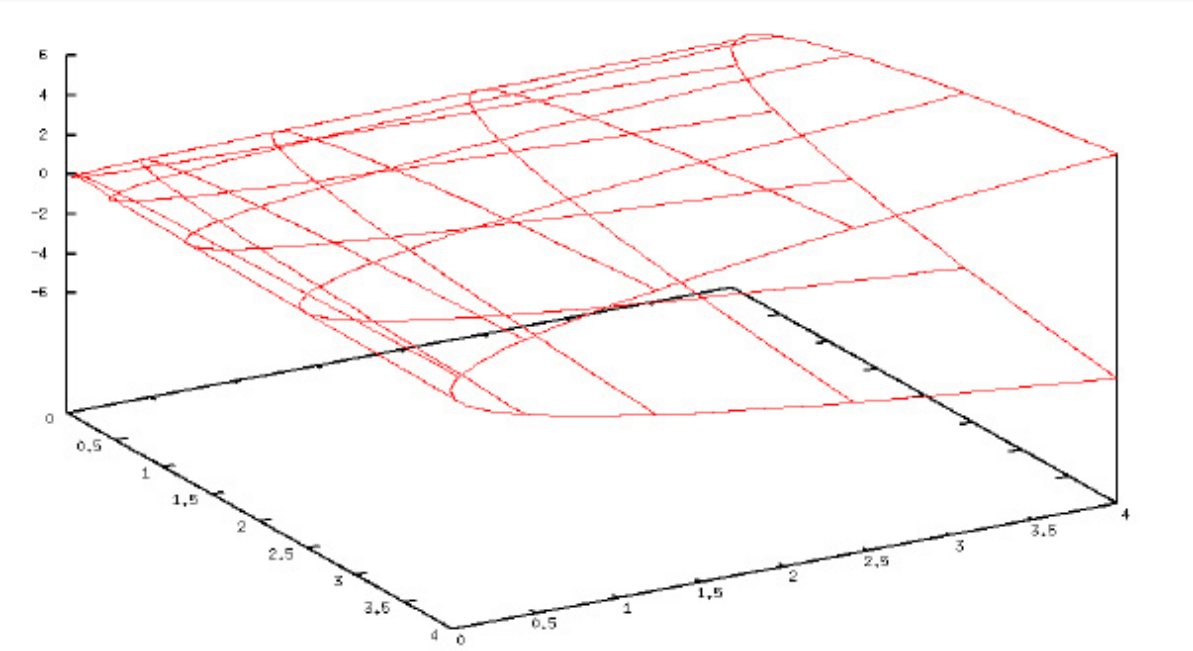
$$\begin{aligned} k(\mathbf{x}, \mathbf{z}) &= \langle \mathbf{x}, \mathbf{z} \rangle^2 \\ &= (x_1 z_1 + x_2 z_2)^2 = x_1^2 z_1^2 + 2 x_1 z_1 x_2 z_2 + x_2^2 z_2^2 \\ &= \langle (x_1^2, \sqrt{2} x_1 x_2, x_2^2), (z_1^2, \sqrt{2} z_1 z_2, z_2^2) \rangle \end{aligned}$$

$$\phi(\mathbf{x}) = \begin{pmatrix} x_1^2 \\ \sqrt{2} x_1 x_2 \\ x_2^2 \end{pmatrix}$$

- Le produit scalaire $\langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$ peut-être calculé dans $\mathbb{R}^2$ l'espace d'origine au moyen du noyau $\langle \mathbf{x}, \mathbf{z} \rangle^2$ sans avoir à se projeter dans $\mathbb{R}^3$

Géométrie de l'espace transformé

La fonction $\phi(x_1, x_2)^T = (x_1^2, x_2^2, \sqrt{2}x_1x_2)^T)$ transforme $\mathbb{R}^2$ en la surface suivante :



Noyau Gaussien

$$\kappa(\mathbf{x}, \mathbf{z}) = \exp(-\|\mathbf{x} - \mathbf{z}\|^2/\sigma), \text{ où } \sigma \in \mathbb{R}^+$$

$\forall \mathbf{x}, \mathbf{z} \in \mathbb{R}$

$$\begin{aligned}
\exp(-\|\mathbf{x} - \mathbf{z}\|^2/\sigma) &= \\
&= \exp(-\mathbf{x}^2 - \frac{\mathbf{z}^2}{\sigma} + 2\frac{\mathbf{xz}}{\sigma}) \\
&= \exp(-\frac{\mathbf{x}^2}{\sigma} - \frac{\mathbf{z}^2}{\sigma}) \left\{ 1 + \frac{2\mathbf{xz}}{\sigma 1!} + \frac{(2\mathbf{xz})^2}{\sigma 2!} + \dots \right\} \\
&= \phi(\mathbf{x}^T \mathbf{z}) \\
\text{où } \phi(\mathbf{x}) &= \exp(-\mathbf{x}^2/\sigma) \left[1, \sqrt{\frac{2}{\sigma 1!}} x, \sqrt{\frac{(2)^2}{\sigma 2!}} x^2, \dots \right]
\end{aligned}$$

Quelques calculs

Nombre de termes de $\phi(\mathbf{x})$

$$\begin{pmatrix} d+2 \\ 2 \end{pmatrix} = \frac{(d+2)(d+1)}{2} \simeq \frac{d^2}{2}$$

- 1 terme constant
- d termes linéaires
- d termes quadratiques auto multiplicatifs
- $d(d-1)/2$ termes quadratiques croisés

$$\phi(\mathbf{x}) = \begin{bmatrix} 1 \\ \sqrt{2}\mathbf{x}_1 \\ \sqrt{2}\mathbf{x}_2 \\ \dots \\ \sqrt{2}\mathbf{x}_d \\ \mathbf{x}_1^2 \\ \mathbf{x}_2^2 \\ \dots \\ \mathbf{x}_d^2 \\ \sqrt{2}\mathbf{x}_1\mathbf{x}_2 \\ \sqrt{2}\mathbf{x}_1\mathbf{x}_3 \\ \dots \\ \sqrt{2}\mathbf{x}_1\mathbf{x}_d \\ \dots \sqrt{2}\mathbf{x}_{d-1}\mathbf{x}_d \end{bmatrix}$$

Simplification calcul du produit scalaire (1/3)

$$\langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle = \begin{bmatrix} 1 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ \dots \\ \sqrt{2}x_d \\ x_1^2 \\ x_2^2 \\ \dots \\ x_d^2 \\ \sqrt{2}x_1x_2 \\ \sqrt{2}x_1x_3 \\ \dots \\ \sqrt{2}x_1x_d \\ \dots \sqrt{2}x_{d-1}x_d \end{bmatrix} \times \begin{bmatrix} 1 \\ \sqrt{2}z_1 \\ \sqrt{2}z_2 \\ \dots \\ \sqrt{2}z_d \\ z_1^2 \\ z_2^2 \\ \dots \\ z_d^2 \\ \sqrt{2}z_1z_2 \\ \sqrt{2}z_1z_3 \\ \dots \\ \sqrt{2}z_1z_d \\ \dots \sqrt{2}z_{d-1}z_d \end{bmatrix} = \begin{cases} 1 \\ 2\sum_{i=1}^{d} x_1z_1 \\ + \\ \sum_{i=1}^{d}(x_1z_1)^2 \\ + \\ 2\sum_{i=1}^{d}\sum_{j=i+1}^{d} x_ix_jz_iz_j \\ \dots \\ \sqrt{2}\mathbf{x}_1\mathbf{x}_d \\ \dots \sqrt{2}\mathbf{x}_{d-1}\mathbf{x}_d \end{cases}$$

Simplification calcul du produit scalaire (2/3)

Considérons maintenant :

$$
\begin{aligned}
(\mathbf{xz}+1)^2 &= (\mathbf{xz})^2 + 2\mathbf{xz} + 1 \\
&= \left(\sum_{i=1}^{d} x_i z_i\right)^2 + 2\sum_{i=1}^{d} x_i z_i + 1 \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d} x_i z_i x_j z_j + 2\sum_{i=1}^{d} x_i z_i + 1 \\
&= \sum_{i=1}^{d} (x_i z_i)^2 + 2\sum_{i=1}^{d}\sum_{j=i+1}^{d} x_i z_i x_j z_j 2\sum_{i=1}^{d} x_i z_i + 1
\end{aligned}
$$

Ce qui donne la définition de $\langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$.

- calcul explicite de $\phi(\mathbf{x}), \phi(\mathbf{z})\rangle$ coût $O(d^2)$
- calcul de $(\mathbf{xz}+1)^2$ coût $O(d)$

Simplification calcul du produit scalaire (3/3)

$\forall p (\mathbf{xz} + 1)^p$ coût $O(d)$

p	Calcul explicite	$n = 100$	Coût noyau	$n = 100$
2	$d^2 n^2/4$	$2500n^2$	$dn^2/2$	$50n^2$
3	$d^3 n^2/12$	$83000n^2$	$dn^2/2$	$50n^2$
4	$d^4 n^2/48$	$1960000n^2$	$dn^2/2$	$50n^2$

Pour $p = 5$ $d = 100$ chaque produit scalaire nécessite 103 opérations, au lieu de 75 millions.

Quelques propriétés des noyaux

Si $\kappa_1, \kappa_2, \ldots$ sont des noyaux de Mercer (dp) alors les suivants le sont également

- $\alpha\kappa_1$ $\alpha \geq 0$ est un noyau de Mercer
- $\kappa_1 + \kappa_2$
- $\kappa_1.\kappa_2$

D'autres opérations pour construire des noyaux à partir d'autres noyaux : produit tensoriel, convolutions,...

Astuce du noyau

- Tout algorithme qui n'utilise que le produit scalaire entre les vecteurs d'entrée peut se généraliser à un algorithme calculant la fonction noyau entre ces vecteurs. Les calculs sont fait implicitement en ne passant pas directement par la fonction ϕ : C'est l'astuce noyau (kernel trick)

Discriminant linéaire de Fisher (Rappel)

Projeter les données selon $\mathbf{w} \in \mathbb{R}^d$, l'analyse discriminante de Fisher consiste à maximiser le critère de Fisher suivant :

$$\begin{aligned} J(\mathbf{w}) &= \frac{[\mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)]^2}{\mathbf{w}^T S_W \mathbf{w}} \\ &= \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} \end{aligned}$$

où $S_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$, $\boldsymbol{\mu_i}$ est la moyenne de $C_i,\ i = 1, 2$ et S_W est défini par :

$$S_W = \sum_{\mathbf{x} \in C_1} (\mathbf{x} - \mathbf{m}_1)(\mathbf{x} - \mathbf{m}_1)^T + (\mathbf{x} - \mathbf{m}_2)(\mathbf{x} - \mathbf{m}_2)^T$$

$\mathbf{w} \propto S_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \Rightarrow$ On en déduit le vecteur de projection optimal :

$$\mathbf{w} = S_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

Discriminant non linéaire de Fisher (1/3)

Soit $\phi : \mathcal{X} \rightarrow \mathcal{F}$, on applique le DLF dans $\mathcal{F}$ ce qui revient à maximiser

$$J(\mathbf{w}) = \frac{\mathbf{w} S_B^\phi \mathbf{w}}{\mathbf{w} S_W^\phi \mathbf{w}}$$

où $\mathbf{w} \in \mathcal{F}, S_B^\phi = (\mathbf{m}_1^\phi - \mathbf{m}_2^\phi)(\mathbf{m}_1^\phi - \mathbf{m}_2^\phi)^T$ et

$$S_W^\phi = \sum_{j=1}^{2} \sum_{i=1}^{n_i} (\phi(\mathbf{x}_i) - \mathbf{m}_j^\phi)(\phi(\mathbf{x}_i) - \mathbf{m}_j^\phi)^T \text{ avec } \mathbf{m}_j^\phi = \frac{1}{n_i} \sum_{i=1}^{n_i} \phi(\mathbf{x}_i^j)$$

Discriminant non linéaire de Fisher (2/3)

La théorie des noyaux reproduisants dit que $\mathbf{w} = \sum_{i=1}^{n} \alpha_i \phi(\mathbf{x}_i)$ d'où

$$\begin{aligned} \mathbf{w}^T \mathbf{m}_j^\phi &= \frac{1}{n_j} \sum_{i=1}^{n} \sum_{k=1}^{n_j} \alpha_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_k^j) \rangle \\ &= \frac{1}{n_j} \sum_{i=1}^{n} \sum_{k=1}^{n_j} \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}_k^j) \\ &= \boldsymbol{\alpha} \boldsymbol{\mu}_j \end{aligned}$$

où $\boldsymbol{\mu}_j^i = \frac{1}{n_j} \sum_{k=1}^{n_j} \kappa(\mathbf{x}_i, \mathbf{x}_k^j)$.
D'où

$$\begin{aligned} \mathbf{w}^T S_B^\phi \mathbf{w} &= \boldsymbol{\alpha}^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\alpha} \\ &= \boldsymbol{\alpha}^T M \boldsymbol{\alpha} \end{aligned}$$

Discriminant non linéaire de Fisher (3/3)

$$\begin{aligned} \mathbf{w}^T S_W^\phi \mathbf{w} &= \boldsymbol{\alpha}^T \sum_{j=1,2} K_j (I - L_j) K_j^T \boldsymbol{\alpha} \\ &= \boldsymbol{\alpha}^T N \boldsymbol{\alpha} \end{aligned}$$

où K_j la matrice $n \times n_j$ noyau de la classe j, avec $(K_j)_{pq} = \kappa(x_p, x_q^j)$, I la matrice identité et L_j est la matrice avec toutes les entrées n_j^{-1}. D'où

$$J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^T M \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T N \boldsymbol{\alpha}}$$

Une fois $\boldsymbol{\alpha}^*$ obtenu (vecteur propre de $N^{-1}M$), la projection d'un nouvel exemple $\mathbf{x}_t$ selon $\mathbf{w}$ est donnée par :

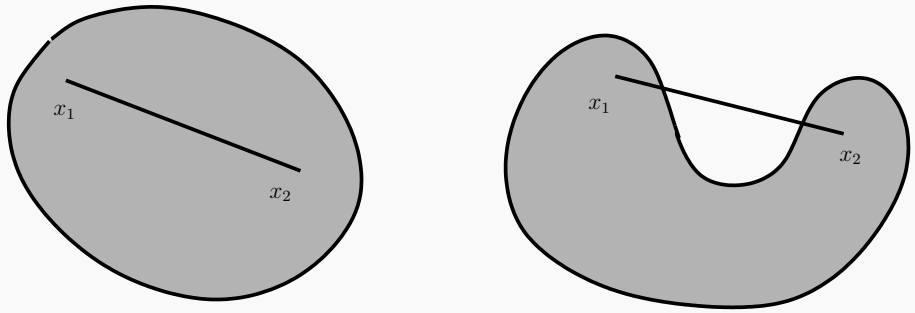
$$\langle \mathbf{w}, \phi(x_t) \rangle = \sum_{i=1}^{n} \alpha_i \kappa(x_i, x_t)$$

Ensemble convexe

Ensemble convexe

Un ensemble $S \subseteq \mathbb{R}^n$ est *convexe* si le segment de droite qui joint deux points de l'ensemble appartient à l'ensemble. Formellement, si $\mathbf{x}_1, \mathbf{x}_2 \in S$ alors :

$$\lambda \mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2 \in S, \ \forall \lambda \in [0, 1]$$



(a) Convexe

(b) Non convexe

Dans le cas (b) (non convexe) la droite joignant $\mathbf{x}_1$ et $\mathbf{x}_2$ ne se trouve pas entièrement dans l'ensemble.

Fonction convexe

Fonction convexe

Soit $f : S \to \mathbb{R}$, où S est un ensemble convexe dans $\mathbb{R}^n$. La fonction f est convexe sur S si $\forall \mathbf{x}_1, \mathbf{x}_2 \in S$ et $\forall \lambda \in [0, 1]$

$$f(\lambda \mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2) \leq \lambda f(\mathbf{x}_1) + (1 - \lambda) f(\mathbf{x}_2)$$

- Le segment de droite reliant 2 points $f(\mathbf{x}_1)$ et $f(\mathbf{x}_2)$ se trouve entièrement sur ou au dessus de la fonction f.
- L'ensemble des points qui se trouve sur ou au dessus de f est convexe
- La fonction f est concave sur S si $-f$ est convexe sur S

Fonction quadratique

Définition (Fonction quadratique)

Une fonction $f : \mathbb{R}^n \to \mathbb{R}$ sera dite quadratique si elle peut s'écrire

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q\mathbf{x} + g^T + c$$

où Q est une matrice symétrique $n \times n$, $g \in \mathbb{R}^n$ et $c \in \mathbb{R}$. On a alors

$$\nabla f(\mathbf{x}) = Q\mathbf{x} + g \text{ et } \nabla^2 f(\mathbf{x}) = Q$$

Fonction Affine

Définition (Fonction affine)

Une fonction $f : \mathbb{R}^n \to \mathbb{R}^m$ est affine si elle est de la forme : linéaire plus constante

$$f(\mathbf{x}) = A\mathbf{x} + b$$

Problème d'optimisation primal

- On utilise la notation

$$(P)\begin{cases}\min & f(\mathbf{x}) \\ \text{s.c.} & g_i(\mathbf{x}) \leq 0\ , i = 1, \ldots, k \\ & h_i(\mathbf{x}) = 0\ , i = 1, \ldots, m\end{cases}$$

Définitions

- Un point $\mathbf{x} \in \mathbb{R}^n$ est dit admissible s'il vérifie toutes les contraintes
- Soit $g : \mathbb{R}^n \to \mathbb{R}^k$. Pour $1 \leq i \leq k$ une contrainte d'inégalité $g_i(\mathbf{x}) \leq 0$ est dite active en $\mathbf{x}^*$ si $g_i(\mathbf{x}^*) = 0$ et inactive en $\mathbf{x}^*$ si $g_i(\mathbf{x}^*) < 0$

Fonction Lagrangienne

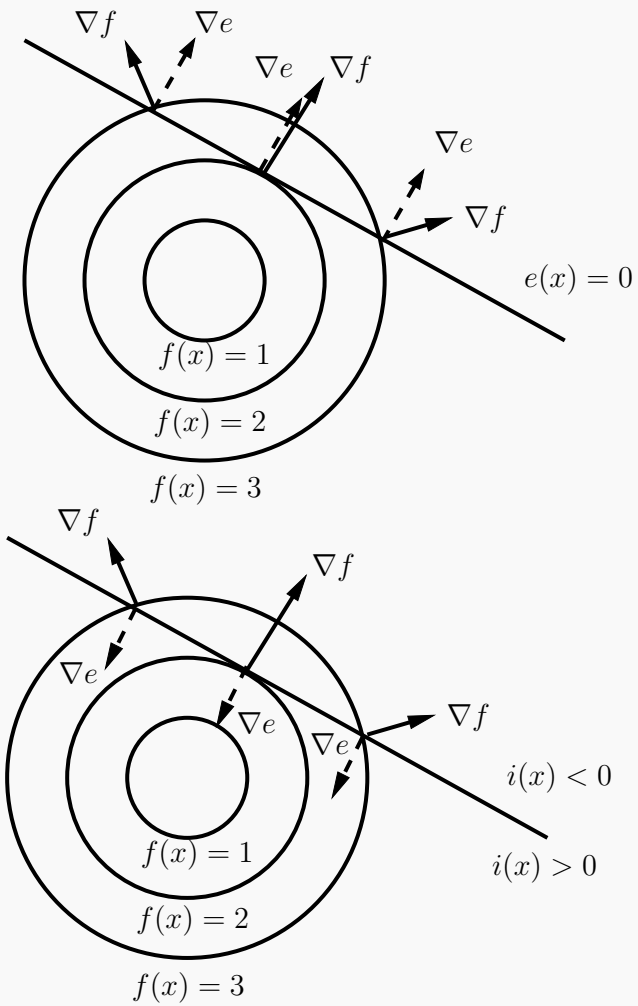
Définition

La fonction $\mathcal{L} : \mathbb{R}^{n+k+m} \to \mathbb{R}$ définie par

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{x}) + \sum_{i=1}^{k} \alpha_i g_i(\mathbf{x}) + \sum_{i=1}^{m} \beta_i h_i(\mathbf{x})$$

est appelée Lagrangien ou fonction Lagrangienne et les $\alpha_i \in \mathbb{R}^k$ et $\beta_i \in \mathbb{R}^m$ sont appelés les *multiplicateurs de Lagrange*.

Idée du Lagrangien (1/3)



A l'optimum :

- Contrainte d'égalité :

$$\begin{cases} \nabla f(\mathbf{x}) + \alpha \nabla e(\mathbf{x}) = 0 \\ e(\mathbf{x}) = 0 \end{cases}$$

- Contrainte d'inégalité :

$$\begin{cases} \nabla f(\mathbf{x}) + \alpha \nabla i(\mathbf{x}) = 0 \\ i(\mathbf{x}) = 0 \\ \alpha > 0 \end{cases}$$

Idée du Lagrangien (2/3)

Contraintes d'égalités : On cherche un petit changement en $\mathbf{x}$ qui fera décroitre f tout en gardant $e = 0$:

$$df = f_{\mathbf{x}} d\mathbf{x} < 0$$
$$de = e_{\mathbf{x}} d\mathbf{x} = 0$$

où $e_{\mathbf{x}} = \frac{\partial e}{\partial \mathbf{x}}$ est le jacobien : les lignes de e_x sont les transposés des gradients de e_i avec $\mathbf{x}, i = 1, \ldots, p$. Ces gradients sont perpendiculaires à la surface $e(\mathbf{x}) = 0$. A un minimum, ∇f ne doit avoir aucune composante parallèle à la surface $e(\mathbf{x}) = 0$; autrement un $d\mathbf{x}$ pourrait être trouvé pour avoir $df < 0$ tout en conservant $de = 0$. Alors ∇f doit être perpendiculaire à $e(\mathbf{x}) = 0$ pour un minimum.

Idée du Lagrangien (3/3)

∇f doit donc être une combinaison linéaire de m gradients de contraintes ∇e_i :

$$\nabla f = -\sum_{i=1}^{p} \alpha_i \nabla e_i$$

Les conditions nécessaires pour un point stationnaire sont :

$$e(\mathbf{x}) = 0$$
$$\nabla f + [e_{\mathbf{x}}]^T \alpha = 0$$

Si on défini le Lagrangien $\mathcal{L} = f(\mathbf{x}) + \alpha^T e(\mathbf{x})$ la condition d'optimalité peut s'écrire : $e(\mathbf{x}) = 0$ et $\nabla \mathcal{L} = 0$

Exemple (Théorème de Lagrange)

Trouver la boite avec le volume le plus grand étant donné une certaine surface. On note les cotés de la boite $u, v.w$.
Problème d'optimisation :

$$(P) \begin{cases} \min & -uvw \\ \text{s.c.} & wu + uv + vw = c \end{cases}$$

- Le lagrangien : $\mathcal{L} = -uvw + \alpha(uv + wu + vw - c)$
- Conditions nécessaires : $\frac{\partial \mathcal{L}}{\partial w} = -uv + \alpha(u + v)$; $\frac{\partial \mathcal{L}}{\partial u} = -wv + \alpha(w + v)$; $\frac{\partial \mathcal{L}}{\partial v} = -uw + \alpha(u + w)$;
- La résolution des équations donne : $\alpha v(w - u) = \alpha w(u - v) = 0$, ce qui donne comme solution : $w = v = u = (c/3)^{\frac{1}{2}}$

Fonction duale

Définition (Fonction duale)

Soit le problème d'optimisation (P) et sa fonction de Lagrange $\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ la fonction $\theta : \mathbb{R}^{k+m} \rightarrow \mathbb{R}$ définie par

$$\theta(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \inf_{\mathbf{x} \in X} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

est la fonction duale de (P). Les paramètres $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$ sont appelés variables duales et les variables $\mathbf{x}$ variables primales.

Problème dual de Lagrange

- Le problème dual de Lagrange correspondant à (P) est donné par

$$(D) \begin{cases} \max & \theta(\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ \text{s.c.} & \alpha_i \geq 0 \ , i = 1, \ldots, k \end{cases}$$

  où $\theta(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \inf_{\mathbf{x} \in X} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$

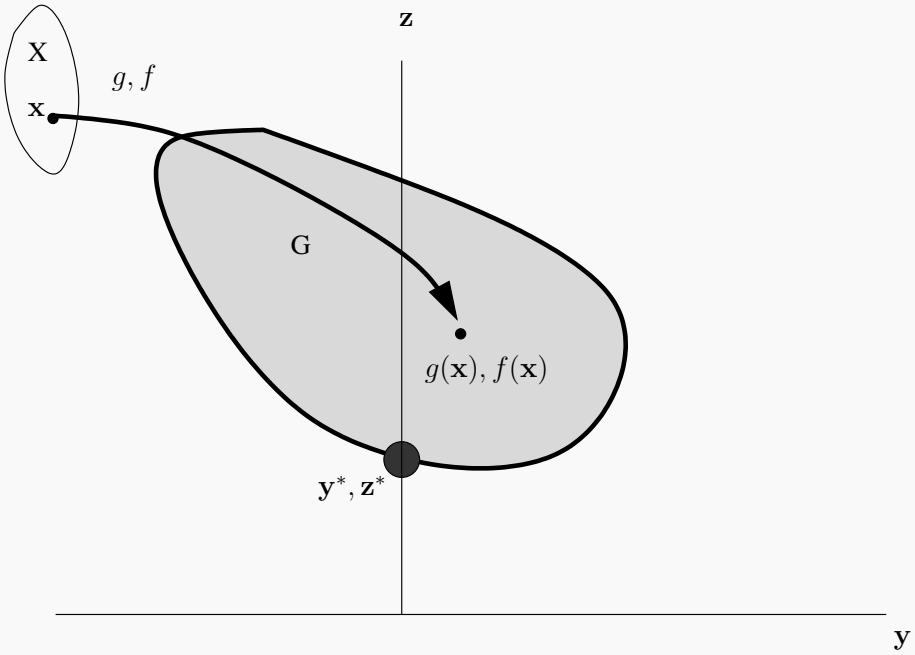
- La valeur de la fonction objectif à la solution optimale est appelée la *valeur du problème*.
- La *différence* entre les valeurs des problèmes primal et dual s'appelle le *saut de dualité*

Interprétation géométrique de la dualité (1/4)

Problème primal P

$$(P) \begin{cases} \min & f(\mathbf{x}) \\ \text{s.c.} & g(\mathbf{x}) \leq 0 \\ \mathbf{x} \in X, \end{cases}$$

où $f : \mathbb{R}^n \to \mathbb{R}$ $g : \mathbb{R}^n \to \mathbb{R}$



On définit l'ensemble suivant dans $\mathbb{R}^2$

$$G = \{(y, z) : y = g(\mathbf{x}), \mathbf{z} = f(\mathbf{x}) \text{ pour } \mathbf{x} \in X\}$$

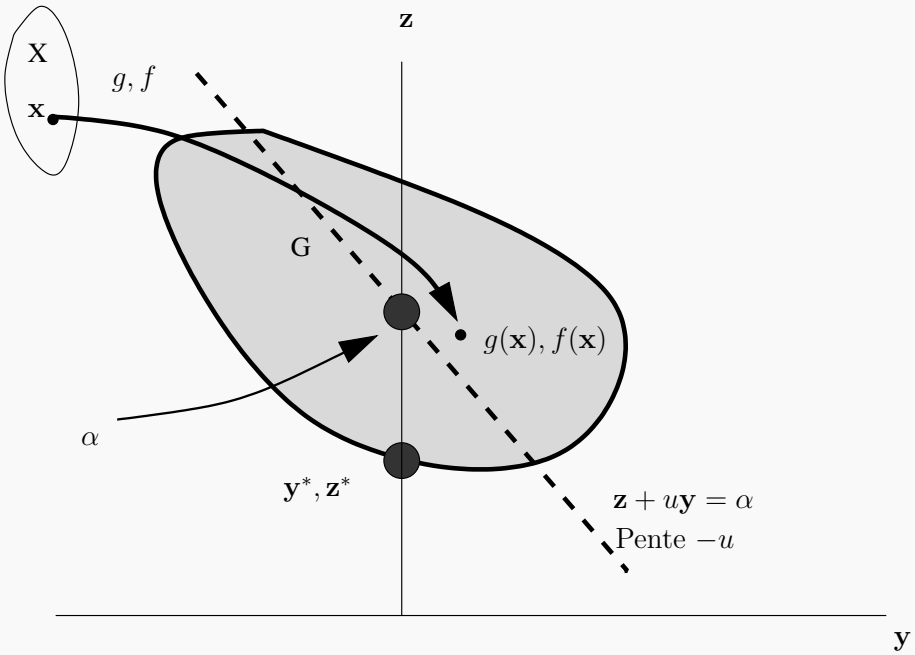
(P) trouver $\mathbf{x}^* \in G, y \leq 0$ d'ordonnée minimale $\to (y^*, \mathbf{z}^*)$

Interprétation géométrique de la dualité (2/4)

Problème dual (D)

$$(D) \begin{cases} \max & \theta(\mathbf{u}) \\ \text{s.c.} & \mathbf{u} \geq 0 \end{cases}$$

où $\theta(\mathbf{u}) = \inf\{f(\mathbf{x}) + ug(\mathbf{x}) : \mathbf{x} \in X\}$



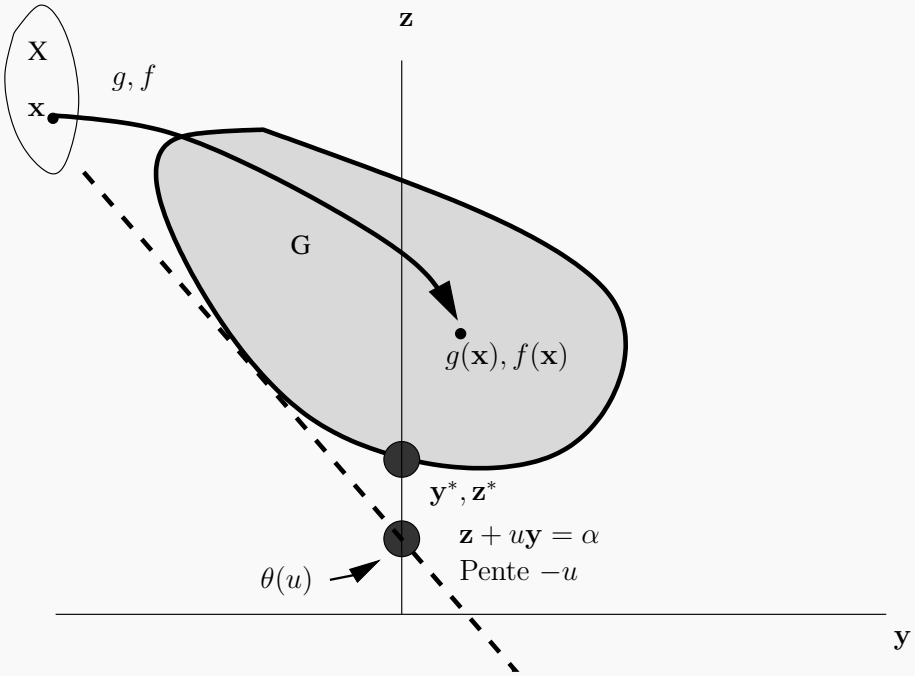
Soit $u \geq 0$, (D) est équivalent à minimiser $\mathbf{z} + u\mathbf{y}$ sur $(\mathbf{y}, \mathbf{z}) \in G$. L'équation $\mathbf{z} + u\mathbf{y} = \alpha$ est l'équation d'une droite de pente $-u$ qui coupe l'axe $\mathbf{z}$ en α.

Interprétation géométrique de la dualité (3/4)

Problème dual (D)

$$(D) \begin{cases} \max & \theta(\mathbf{u}) \\ \text{s.c.} & \mathbf{u} \geq 0 \end{cases}$$

où $\theta(\mathbf{u}) = \inf\{f(\mathbf{x}) + ug(\mathbf{x}) : \mathbf{x} \in X\}$



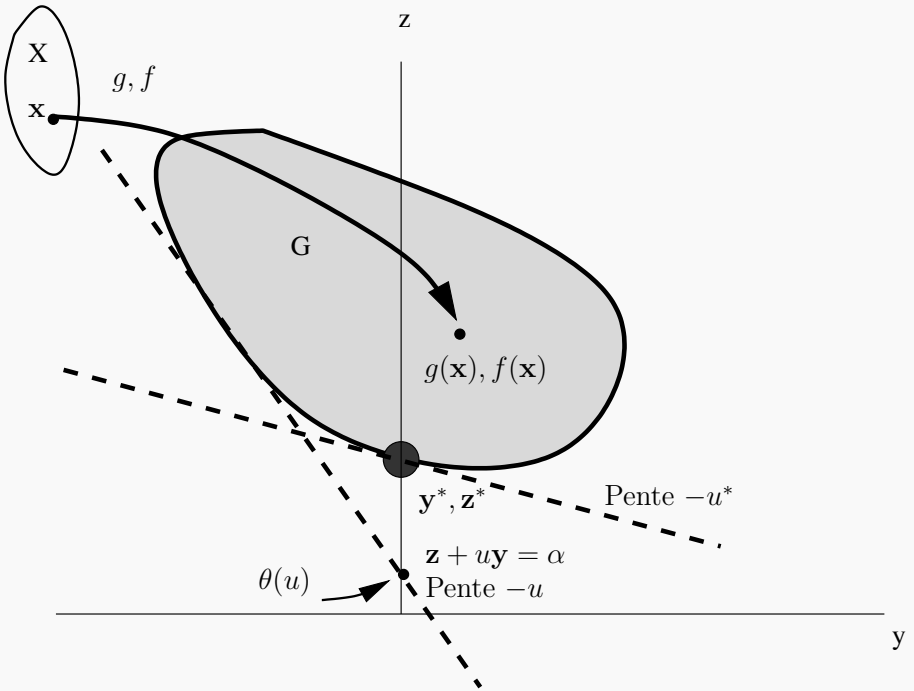
Pour minimiser $\mathbf{z} + u\mathbf{y}$ sur G on doit déplacer la droite $\mathbf{z} + u\mathbf{y} = \alpha$ parallèllement à elle même aussi loin que possible, tout en restant en contact avec G. La dernière intersection avec l'axe des $\mathbf{z}$ ainsi obtenue est la valeur $\theta(u)$ correspondant au $u \geq 0$ donné.

Interprétation géométrique de la dualité (4/4)

Problème dual (D)

$$(D) \begin{cases} \max & \theta(\mathbf{u}) \\ \text{s.c.} & \mathbf{u} \geq 0 \end{cases}$$

où $\theta(\mathbf{u}) = \inf\{f(\mathbf{x}) + ug(\mathbf{x}) : \mathbf{x} \in X\}$



Finalement, pour résoudre (D), on cherche la droite de pente $-u$ t.q. la dernière intersection avec l'axe $\mathbf{z}$, $\theta(u)$, soit maximale. La pente de cette droite est $-u^*$. La solution de (D) est donc u^*, et la valeur optimale de l'objectif dual est $\mathbf{z}^*$. La solution optimal de P est également $\mathbf{z}^*$.

Dualité faible (1/3)

Théorème (dualité faible)

Soit $\mathbf{x} \in X$ une solution admissible de (P) et (α, β) une solution admissible de (D) le dual de (P). Alors $\theta(\alpha, \beta) \leq f(\mathbf{x})$

Preuve

$$\begin{aligned} \theta(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \inf_{\mathbf{x} \in X} \mathcal{L}(\mathbf{x}, \alpha, \beta) \\ &= f(\mathbf{x}) + \sum_{i=1}^{n} \underbrace{\alpha_i}_{\geq 0} \underbrace{g_i(\mathbf{x})}_{\leq 0} + \sum_{i=1}^{n} \beta_i \underbrace{h_i(\mathbf{x})}_{=0} \leq f(\mathbf{x}) \end{aligned}$$

L'admissibilité de $\mathbf{x}$ implique $g_i(\mathbf{x}) \leq 0$ et $h_i(\mathbf{x}) = 0$, alors que l'admissibilité de (α, β) implique $\alpha_i \geq 0$

Dualité faible (2/3)

On a, comme corollaire du théorème précédent, le résultat suivant.

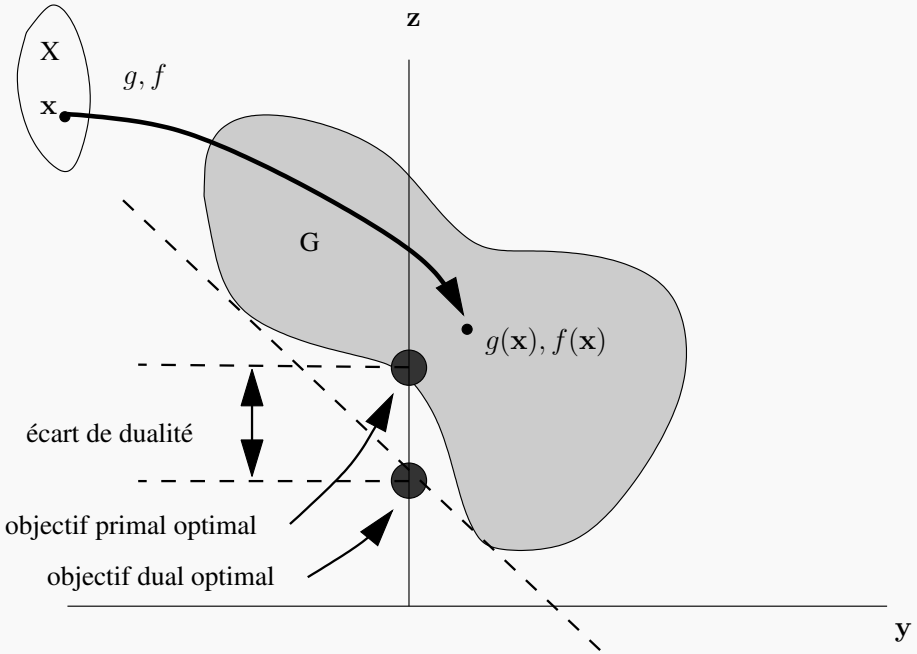
Corollaire

$$\inf\{f(\mathbf{x}) : \mathbf{x} \in X, g(\mathbf{x}) \leq 0, h(\mathbf{x}) = 0\} \geq \sup\{\theta(\alpha, \beta) : \alpha \geq 0\}$$

On notera d'après le corollaire que la valeur optimale de la fonction objectif de (P) est plus grande ou égale que celle du problème dual (D)
Si l'inégalité est stricte, on dit qu'il existe un *saut de dualité*.

Dualité faible (3/3)

La figure suivante montre un exemple d'interprétation géométrique du problème primal et dual.



Dans le cas illustré par la figure, il existe un *saut de dualité* due à la non convexité de l'ensemble G

Si certaines conditions de convexité sont satisfaites, alors il n'y a pas d'écart de dualité entre (P) et (D).

Dualité forte

Si la fonction objectif est convexe, la région admissible est convexe et les contraintes affines ($h(\mathbf{x}) = A\mathbf{x} - b$) alors le saut de dualité est nul ; c'est ce que l'on appelle la dualité forte $f(\mathbf{x}^*) = \theta(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$.

Définition (Point selle)

Un ensemble $\mathbf{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$ de solutions admissibles de (P) et (D) est appelé un *point selle* du Lagrangien si $\forall \mathbf{x} \in X,\ \forall \boldsymbol{\alpha} \geq 0$

$$\mathcal{L}(\mathbf{x}^*, \boldsymbol{\alpha}, \boldsymbol{\beta}) \leq \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \leq \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$$

Si on a un point selle $(\mathbf{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$. Alors il y a dualité forte, $\mathbf{x}^*$ résoud (P) et $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ résoud (D).

Conditions de Karush-Kuhn-Tucker

- Etant donné un problème d'optimisation (P) sur un domaine convexe $X \subseteq \mathbb{R}^n$ avec des contraintes affines.
  - Les conditions nécessaires et suffisantes d'optimalité pour un point $\mathbf{x}^*$ sont l'existence de α^*, β^* tels que

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\mathbf{x}^*, \alpha^*, \beta^*)}{\partial \mathbf{x}} &= 0 \\
\frac{\partial \mathcal{L}(\mathbf{x}^*, \alpha^*, \beta^*)}{\partial \beta} &= 0 \\
\alpha_i^* g_i(\mathbf{x}^*) &= 0 \quad i = 1, \ldots, k \qquad \text{condition complémentaire} \\
g_i(\mathbf{x}^*) &\leq 0 \quad i = 1, \ldots, k \\
\alpha_i^*(\mathbf{x}^*) &\geq 0 \quad i = 1, \ldots, k
\end{aligned}
$$

Problème d'optimisation convexe

Le théorème principal en analyse convexe, et la raison fondamentale de l'intérêt des fonctions convexes, est qu'un minimum local sur une fonction convexe définie sur un ensemble convexe est également un minimum global.

Théorème

Soit $X \subseteq \mathbb{R}^n$ un ensemble convexe et $f : X \to \mathbb{R}$ une fonction convexe. Soit un point $\mathbf{x}^* \in X$, supposons qu'il y a une boule $S(\mathbf{x}^*, \epsilon) \subset X$ tel que pour tout $\mathbf{x} \in S(\mathbf{x}^*, \epsilon)$ on a $f(\mathbf{x}^*) \leq f(\mathbf{x})$. Alors $f(\mathbf{x}^*) \leq f(\mathbf{x})$ pour tout $\mathbf{x} \in X$

Problème d'optimisation convexe (Preuve)

Soit $x \in X$. Puisque f est convexe, pour tout $\lambda \in [0,1]$ on a $f(x^* + (1-\lambda)x) \leq \lambda f(x^*) + (1-\lambda)f(x)$. On remarque qu'il existe $\bar{\lambda} \in [0,1]$ tel que $\bar{\lambda}x^* + (1-\bar{\lambda})x = \bar{x} \in S(x^*, \epsilon)$. On considére $\bar{x}$: par convexité de f on a $f(\bar{x})\bar{\lambda}f(x^*) + (1-\bar{\lambda})f(x)$. On obtient :

$$f(x) \geq \frac{f(\bar{x}) - \lambda f(x^*)}{1-\lambda}$$

Puisque $\bar{x} \in S(x^*, \epsilon)$, on a $f(\bar{x}) \geq f(x^*)$, donc

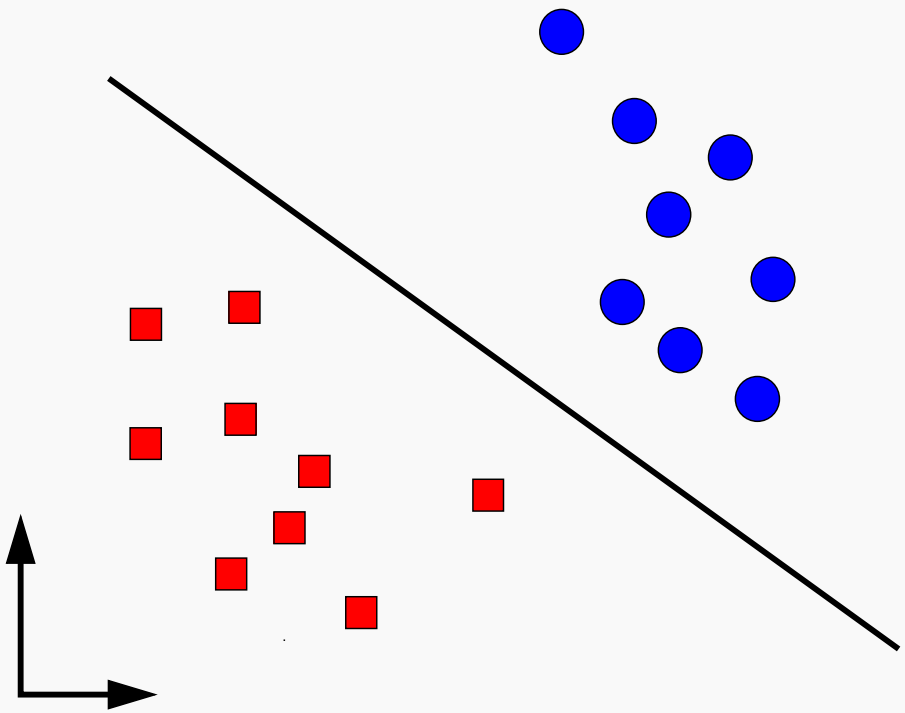
$$f(x) \geq \frac{f(x^*) - \lambda f(x^*)}{1-\lambda} = f(x^*)$$

Qu'est ce qu'une MVS ?

Définition

- selon l'informaticien : la MVS est un classifieur linéaire à marge maximale dans un espace à noyau
- selon le statisticien : la MVS est un estimateur non-paramétrique. Il est basé sur une minimisation du risque empirique *régularisé* sur un espace fonctionnel de Hilbert et avec une fonction de perte linéaire par morceaux.

Ensemble linéairement séparable

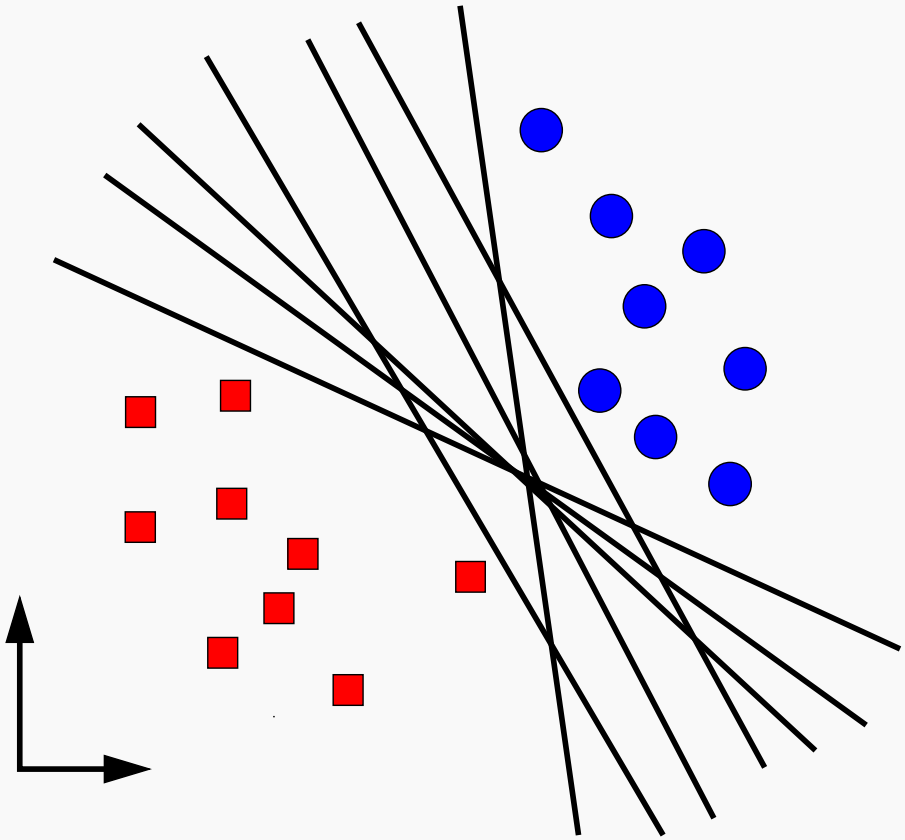


Un ensemble d'exemples étiquetés $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ est dit *linéairement séparable*, si il existe $\mathbf{w}, b$ tels que :

$$\gamma_i = y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) > 0, \quad \forall_i$$

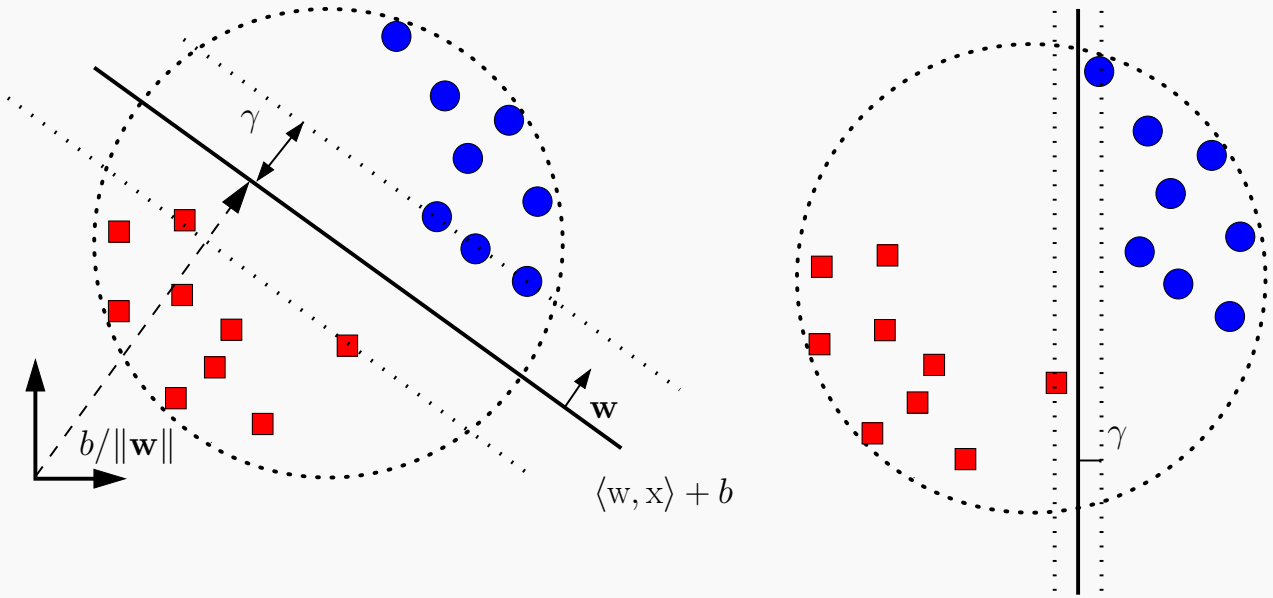
Ce qui signifie qu'il existe un hyperplan tel que tous les exemples d'apprentissage positifs sont dans un demi-plan et tous les négatifs dans l'autre.

Lequel est le meilleur ?



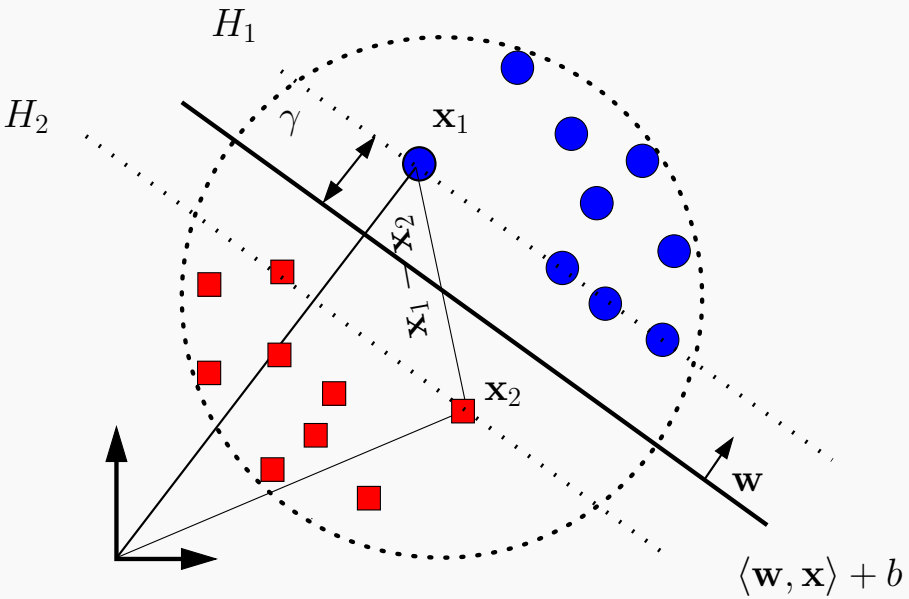
Il y a une infinité de séparatrices linéaires !

Réponse : La plus grande marge



$\Rightarrow$ Celui qui a la plus grande marge

Comment calculer l'hyperplan



- Prendre $\mathbf{x}_1$ et $\mathbf{x}_{-1}$ correspondants aux vecteurs se trouvant sur les frontières définissant les marges
- Il est aisé de voir que l'hyperplan doit se trouver à mi-distance entre $\mathbf{x}_1$ et $\mathbf{x}_{-1}$.
- Et donc la marge sera simplement la moitié de la distance perpendiculaire entre $\mathbf{x}_1$ et $\mathbf{x}_{-1}$ projeté sur la normale au plan :

$$\gamma = \frac{1}{2} \frac{\langle \mathbf{w}, \mathbf{x}_1 - \mathbf{x}_2 \rangle}{\|\mathbf{w}\|}$$

Elimination du degré de liberté

- Pour tout $c \neq 0$, on a

$$\{x : \langle w, x \rangle + b = 0\} = \{x : \langle cw, x \rangle + cb = 0\}$$

- ainsi (cw, cb) défini le même hyperplan que (w, b).
- l'hyperplan est en forme canonique relativement à l'ensemble de points $X = \{x_1, \ldots, x_r\}$ si

$$\min_{x_i \in X} |\langle w, x_i \rangle + b| = 1.$$

Marge de l'hyperplan canonique optimal

- Pour trouver la valeur de cette expression, on effectue les manipulations suivantes :

$$\langle \mathbf{w}, \mathbf{x}_1 \rangle + b = +1 \quad \langle \mathbf{w}, \mathbf{x}_{-1} \rangle + b = -1$$

- Donc

$$2 = (\langle \mathbf{w}, \mathbf{x}_1 \rangle + b) - (\langle \mathbf{w}, \mathbf{x}_{-1} \rangle + b) = \langle \mathbf{w}, \mathbf{x}_1 - \mathbf{x}_{-1} \rangle$$

- Ainsi, l'expression de la marge est

$$\text{marge} \quad = \quad \frac{1}{\|\mathbf{w}\|}$$

Maximisation de la marge

Maximiser la marge correspond à minimiser la norme de $\mathbf{w}$ pour une marge $\gamma = 1$ sous la contrainte que tous les exemples d'apprentissage sont correctement classifiés

$$(P) \begin{cases} \min & \|w\|^2 \\ \text{s.c.} & y_i[\langle w, x_i \rangle + b] \geq 1 \end{cases}$$

on construit d'abord le Lagrangien :

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

puis on minimise $\mathcal{L}(\mathbf{w}, \alpha, b)$ par rapport aux variables primales $\mathbf{w}$ et b afin d'obtenir le dual.

Stationnarité du Lagrangien

Les conditions à l'optimum exigent que les dérivées par rapport à $\mathbf{w}$ et b soient nulles :

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0$$

Pour obtenir le problème d'optimisation dual il faut substituer les conditions stationnaires $\mathbf{w}^*$ et b^* dans $\mathcal{L}$. Les variables duales α_i ont comme contrainte d'être ≥ 0 ($\alpha_i \geq 0$).

Stationnarité du Lagrangien

En resubstituant $\partial\mathcal{L}/\partial\mathbf{w} = 0, \partial\mathcal{L}/\partial\mathbf{b} = 0$ dans $\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})$ on obtient :

$$\begin{aligned}
\mathcal{L}(\mathbf{w}, b, \alpha) &= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i (y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1) \\
&= \sum_{i=1}^{n} \alpha_i - \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\
&\quad + \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\
&= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle
\end{aligned}$$

Problème dual

Ce qui conduit à la formulation duale du problème :

$$(D) \begin{cases} \max\limits_{\boldsymbol{\alpha}} & \theta_{\mathcal{D}}(\boldsymbol{\alpha}) = \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{2} \sum\limits_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ \text{s.c.} & \sum\limits_{i=1}^{n} \alpha_i y_i = 0 \\ & \alpha_i \geq 0 \quad i = 1, \ldots, n \end{cases}$$

On remarque que le problème dual est indépendant de la dimension de $\mathbf{x}$ mais est dépendant du nombre d'observations. C'est une propriété très intéressante, spécialement lorsque $\mathcal{X}$ est de grande dimension et que le nombre d'observations reste petit.

Supports vecteurs

Les conditions de **Karush-Kuhn-Tucker** pour la SVM à marge maximale sont

$$\begin{aligned} \alpha_i^*(y_i f^*(\mathbf{x}_i) - 1) &= 0, \quad i = 1, \ldots, n \\ \alpha_i^* &\geq 0, \quad i = 1, \ldots, n \\ y_i f^*(\mathbf{x}_i) - 1 &\geq 0, \quad i = 1, \ldots, n \end{aligned}$$

La première est appelée la condition complémentaire de KKT. Ces conditions impliquent

$$\begin{aligned} y_i f^*(\mathbf{x}_i) = 1 &\Rightarrow \alpha_i^* \geq 0 \quad \text{Vecteurs supports} \\ y_i f^*(\mathbf{x}_i) \neq 1 &\Rightarrow \alpha_i^* = 0 \quad \text{Pas vecteurs supports} \end{aligned}$$

Fonction de décision optimale

- vecteur de poids : $\mathbf{w}^* = \sum_{i:\alpha_i^*>0} \alpha_i^* \mathbf{x}_i y_i$
- marge : $\gamma^* = \frac{1}{\|\mathbf{w}^*\|} = \left(\sum_{i=1}^{n} \alpha_i^* \right)^{-1/2}$
- le biais b^* : $b^* = -\frac{1}{2} \left[\min_{y_i=1}(\langle \mathbf{w}^*, \mathbf{x}_i \rangle) + \max_{y_i=-1}(\langle \mathbf{w}^*, \mathbf{x}_i \rangle) \right]$
- Seuls les α_i correspondant aux points les plus proches sont non-nuls. Ceux sont les vecteurs supports.
- Fonction de décision (parcimonie) :
  $f(\mathbf{x}, \boldsymbol{\alpha}^*, b^*) = \sum_{i \in SV} \underbrace{y_i \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle}_{\langle \mathbf{w}^*, \mathbf{x} \rangle} + b^*$

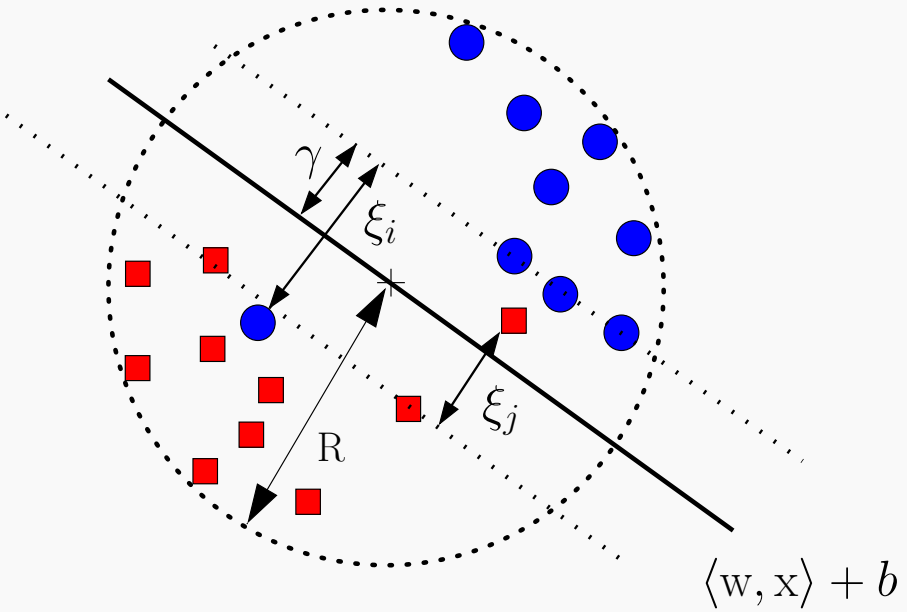
Conclusion pour le classifieur à marge max.

- Trouver l'hyperplan optimal, qui correspond à la plus grande marge
- Peut être résolu facilement en utilisant la formulation duale
- La solution est creuse (sparse) : le nombre de vecteurs support peut-être très petit en comparaison de la taille de l'ensemble d'apprentissage
- Seulement les vecteurs supports sont importants pour la prédiction de futurs exemples. Tous les autres exemples peuvent être oubliés (parcimonie)

Problèmes non séparables

- Précédemment on a fait l'hypothèse que les classes étaient séparables
- Ce qui est rarement le cas en pratique
- Un problème séparable est "facile" la plupart des classifieurs y arrivent très bien
- On a besoin d'étendre la SVM au cas non séparable
- **Idée de base :**
  - avec le recouvrement des classes on ne peut pas appliquer la recherche d'une marge
  - mais on peut appliquer la recherche d'une marge souple
  - pour la plupart des points il y a une marge, mais il y a quelques "outliers" qui traversent, ou qui sont plus proche de la frontière de décision que de la marge.

Marge souple



- **Objectif :** Trouver un "bon" hyperplan séparateur pour le cas non séparable.
- **Problème :** $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$ n'est pas satisfait pour tous les i
- **Solution :** Relacher les contraintes de marge en introduisant des variables ressort (slack).
- **Rappel :** Minimiser le nombre d'erreurs est NP-dur.

$$\begin{aligned} \text{On veut : } \langle \mathbf{w}, \mathbf{x}_i \rangle + b &\geq 1 - \xi_i \quad \forall y_i = +1, \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b &\leq -1 + \xi_i \quad \forall y_i = -1, \\ \xi_i &\geq 0 \quad i = 1, 2, \ldots, n \end{aligned}$$

Marge souple

Les variables ressort mesurent la déviation par rapport à la condition idéale

- Pour $0 \leq \xi \leq 1$, les exemples se trouvent du bon coté de l'hyperplan mais dans la région de la marge maximale
- Pour $\xi > 1$, les exemples se trouvent du mauvais coté de l'hyperplan
- Le problème n'est pas bien défini $\Rightarrow$ en prenant ξ_i arbitrairement grand, n'importe quel $\mathbf{w}$ conviendra
- On a besoin de pénaliser les grandes valeurs de ξ_i

On a vu que l'on a une erreur si $\xi_i > 1$. Ainsi

$$\text{nb erreurs} \leq \sum_{i=1}^{n} I(\xi_i > 1)$$

Marge souple

Solution possible :

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} I(\xi_i > 1) \right\} \textbf{non convexe !}$$

Suggestion : On remplace $I(\xi_i > 1)$ par ξ_i, puisque $I(\xi_i > 1) \leq \xi_i$

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i^{\sigma} \right\} \textbf{convexe !}$$

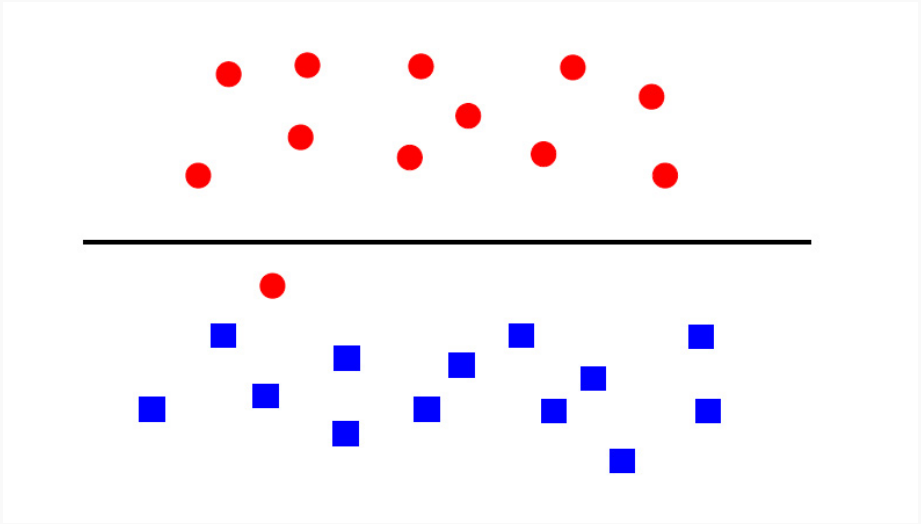
On considère habituellement les cas $\sigma = 1$ (norme 1) et $\sigma = 2$ (norme 2).

Marge souple

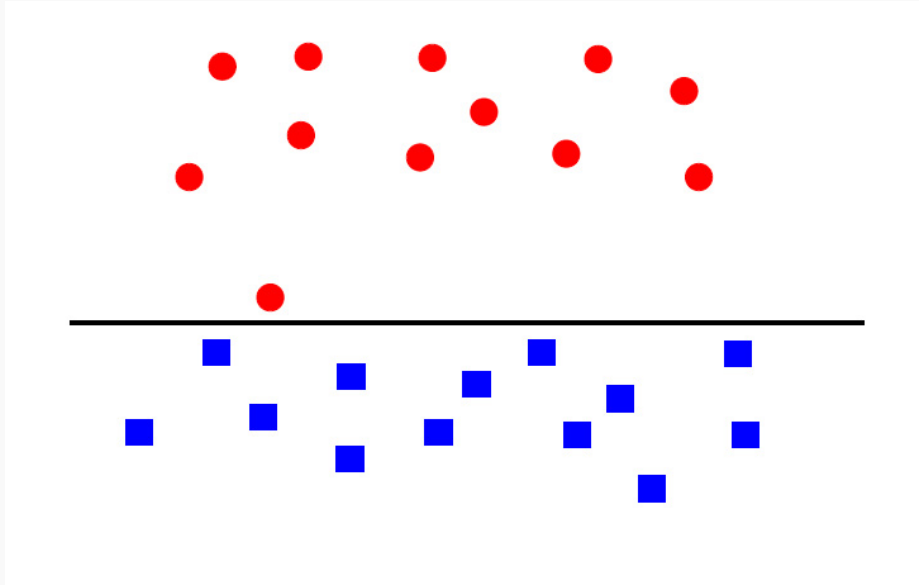
Compromis : La constante C borne le nombre d'erreurs tolérées.

- Grand C – pénalise les erreurs
- Petit C – impose une grande marge, pénalise la complexité

- Petit valeur de C



- Grande valeur de C



Marge souple

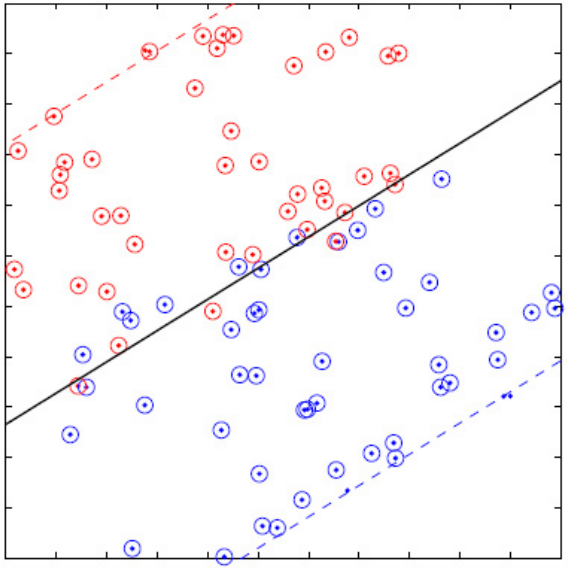
Problème d'optimisation :

$$(P) \begin{cases} \min & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi^{\sigma} \\ \text{s.c.} & y_i[\langle w, x_i\rangle + b] \geq 1 - \xi_i \quad \forall i \\ & \xi_i \geq 0 \quad \forall i \end{cases}$$

Marge souple

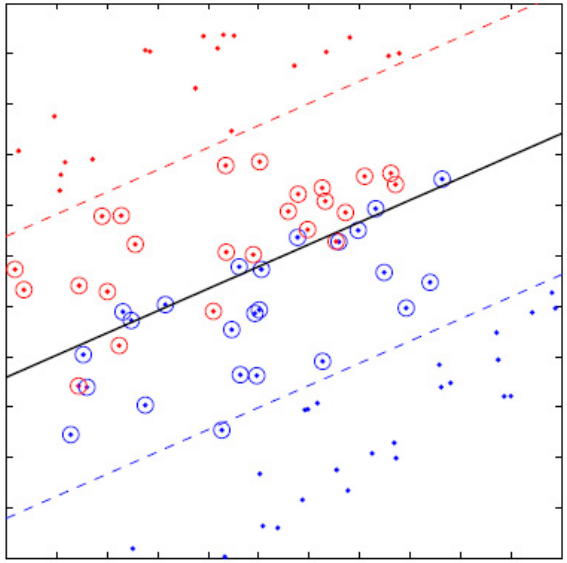
$C = 0.1$

95% SV



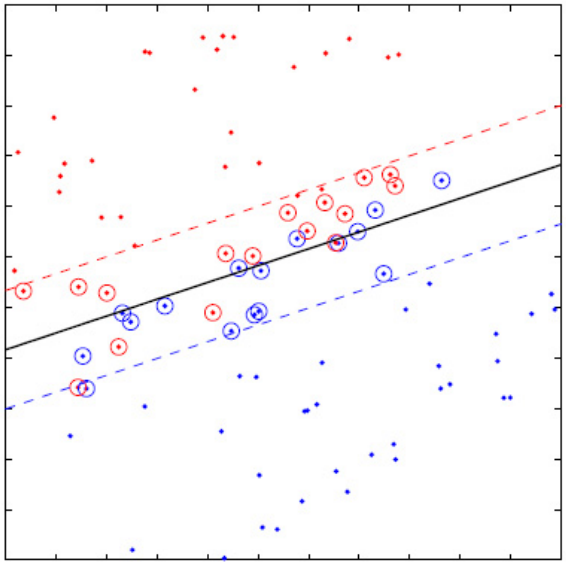
$C = 1$

48% SV



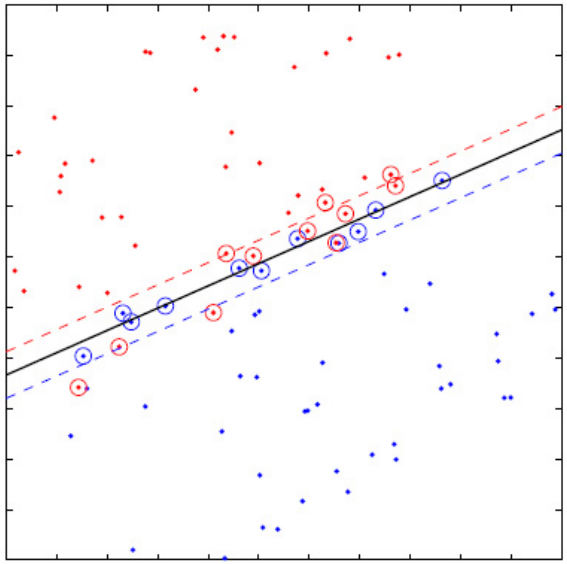
$C = 10$

32% SV



$C = 1000$

22% SV



La SVM norme 2

$$(P_{L2}) \begin{cases} \min & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi^2 \\ \text{s.c.} & y_i[\langle w, x_i\rangle + b] \geq 1 - \xi_i \quad \forall i \\ & \xi_i \geq 0 \quad \forall i \end{cases}$$

On remarque que $\xi_i < 0$ n'est jamais solution et que la contrainte de positivité sur les ξ_i est redondante, elle peut donc être retirée. Ce qui donne :

$$(P_{L2}) \begin{cases} \min & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi^2 \\ \text{s.c.} & y_i[\langle w, x_i\rangle + b] \geq 1 - \xi_i \quad \forall i \end{cases} \tag{1}$$

La SVM norme 2

Fonction de Lagrange :

$$\begin{aligned}\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) &= \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 \\ &\quad - \sum_{i=1}^{n}\alpha_i(y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) - 1 + \xi_i)\end{aligned}$$

Les conditions à l'optimum exigent que les dérivées par rapport à $\mathbf{w}$,b et $\boldsymbol{\xi}$ soient nulles :

$$\begin{aligned}\frac{\partial \mathcal{L}(\mathbf{w}, b, \xi, \boldsymbol{\alpha})}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i+1}^{n}\alpha_i y_i \mathbf{x}_i = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i \\ \frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha})}{\partial \boldsymbol{\xi}} &= C\boldsymbol{\xi} - \boldsymbol{\alpha} = 0 \\ \frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha})}{\partial b} &= -\sum_{i+1}^{n}\alpha_i y_i = 0\end{aligned}$$

Le problème dual de la SVM norme 2

En resubsituant dans la fonction de Lagrange $\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha})$ $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$, $\sum_i y_i \alpha_i = 0$ et $\xi_i = \frac{\alpha_i}{C}$ on obtient :

$$\begin{aligned}
\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) &= \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_i \left(\frac{\alpha_i}{C}\right)^2 + \sum_i \alpha_i \left[1 - \frac{\alpha_i}{C} - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)\right] \\
&= \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \frac{1}{2}\sum_{i=1}^{n} \frac{\alpha_i^2}{C} + \sum_{i=1}^{n} \alpha_i \\
&\quad - \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \underbrace{\sum_{i=1}^{n} \alpha_i y_i b}_{=0} \\
&= -\frac{1}{2}\left(\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \frac{1}{C}\sum_{i=1}^{n} \alpha_i^2\right) + \sum_{i=1}^{n} \alpha_i \\
&= -\frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \left(\langle \mathbf{x}_i, \mathbf{x}_j \rangle + \frac{\delta_{ij}}{C}\right) + \sum_{i=1}^{n} \alpha_i \; ; \; \delta_{ij} = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases}
\end{aligned}$$

Le problème dual de la SVM norme 2

Le problème dual est donc

$$(D_{L2}) \begin{cases} \max & \mathcal{L}_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i . \mathbf{x}_j \rangle \\ \text{s.c.} & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & \forall_i, \quad 0 \leq \alpha_i \leq C \end{cases}$$

La condition complémentaire de KKT correspondante est :

$$\alpha_i [y_i (\langle \mathbf{x}_i, \mathbf{w} \rangle) + b) - 1 + \xi_i] = 0, i = 1, \ldots, n$$

La fonction objectif duale est la même que pour la SVM séparable, avec $\frac{1}{C} \times \mathbf{I}$ ajouté à la matrice de Gram.

$$K'(\mathbf{x}, \mathbf{z}) = K(\mathbf{x}, \mathbf{z}) + \frac{1}{C} \delta(\mathbf{z})$$

La SVM norme 1

Problème primal :

$$(P_{L1}) \begin{cases} \min & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi \\ \text{s.c.} & y_i[\langle w, x_i\rangle + b] \geq 1 - \xi_i \quad \forall i \\ & \xi_i \geq 0 \quad \forall i \end{cases}$$

Fonction de Lagrange :

$$\begin{aligned} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\quad - \sum_{i=1}^{n}\alpha_i(y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) - 1 + \xi_i) \\ &\quad - \sum_{i=1}^{n}\beta_i\xi_i \end{aligned}$$

Problème dual de la SVM norme 1

Les conditions à l'optimum exigent que les dérivées par rapport à $\mathbf{w}$ et b soient nulles :

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

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

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \xi, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial b} = -\sum_{i+1}^{n} \alpha_i y_i = 0$$

En resubsituant dans la fonction de Lagrange $\mathcal{L}(\mathbf{w}, b, \xi, \boldsymbol{\alpha}, \boldsymbol{\beta})$

Problème dual de la SVM norme 1

$$
\begin{aligned}
\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \beta_i \xi_i \\
&\quad - \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^{n} \alpha_i (1 - \xi_i) \\
&= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\
&\quad + \sum_{i=1}^{n} (\underbrace{C - \alpha_i - \beta_i}_{=\frac{\partial \mathcal{L}}{\partial \xi_i} = 0}) \xi_i \\
&= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle
\end{aligned}
$$

Contraintes du problème dual

La fonction objectif dual est la même que pour la SVM séparable. La seule différence est la contrainte

$$C - \alpha_i - \beta_i = 0 \quad \forall i$$

avec $\beta_i \geq 0$ cela implique

$$\alpha_i \leq C$$

La condition complémentaire de KKT

$$\begin{aligned} \alpha_i(y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \xi_i) &= 0 \quad i = 1, \ldots, n \\ \xi_i(\alpha_i - C) &= 0 \end{aligned}$$

impliquent que les variables de relâchement non nulles surviennent lorsque $\alpha_i = C$

Problème dual de la SVM norme 1

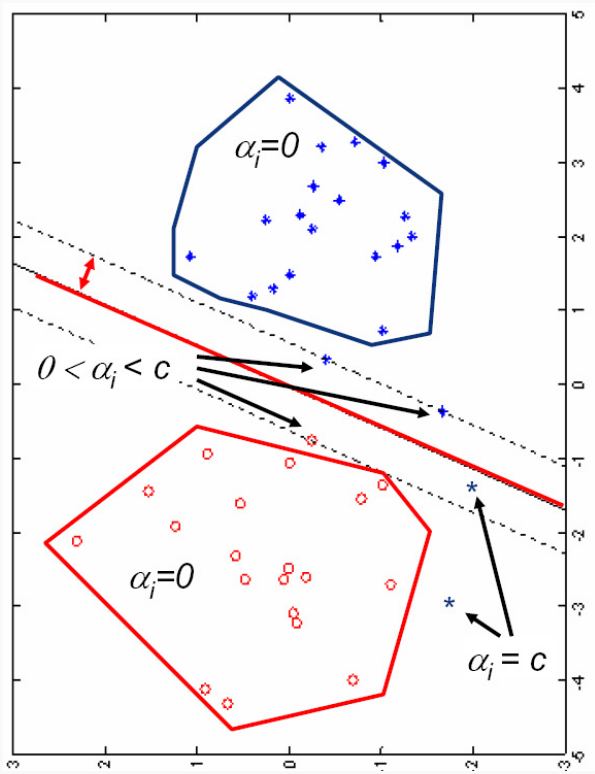
- En résumé :
  - $\mathbf{w}^* = \sum_i \alpha_i y_i \mathbf{x}_i,\ \sum_i y_i \alpha_i = 0,\ \beta_i = C - \alpha_i \quad (1)$
  - $\mathcal{L}(\mathbf{w}^*, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_i \alpha_i$
  - Puisque les β_i sont des multiplicateurs de Lagrange, $\beta_i \geq 0$, (1) signifie que $\alpha_i \leq C$
  - Egalement des conditions de KKT
    1. $\beta_i > 0 \Leftrightarrow \xi_i = 0, \quad \beta_i = 0 \Leftrightarrow \xi_i > 0, \quad (\beta_i \xi_i = 0)$
    2. $\alpha_i [1 - \xi_i - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)] = 0$
  - Puisque $\alpha_i = C - \beta_i$ 1) implique que
    - a) $\xi_i = 0$ et $\alpha_i < C$ ou b) $\xi_i > 0$ et $\alpha_i = C$
    - de 2), et cas a), on a $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1$, i.e. $\mathbf{x}_i$ est sur la marge
    - de 2), et cas b), on a $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1 - \xi_i$, i.e. $\mathbf{x}_i$ est un "outlier"
    - puis, comme auparavant, les points sont correctement classifiés lorsque $\alpha_i = 0$ ($\beta_i = C \Leftrightarrow \xi_i = 0$)

Problème dual de la SVM norme 1

Le problème dual est donc

$$(D_{L1}) \begin{cases} \max & \mathcal{L}_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i . \mathbf{x}_j \rangle \\ \text{s.c.} & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & \forall_i, 0 \leq \alpha_i \leq C \end{cases}$$

- La seule différence par rapport à la marge dure est la contrainte de "boite" sur les α_i
- Géométriquement on a celà : $\longrightarrow$



Les vecteurs supports

- Sont les points pour lesquels $\alpha_i > 0$
- Comme auparavant la fonction de décison est :

$$h(\mathbf{x}) = \text{sgn}\left(\sum_{i \in SV} y_i \alpha_i^* \langle \mathbf{x}_i, \mathbf{x} \rangle + b^*\right)$$
$$\text{où} \qquad SV = \{i | \alpha_i^* > 0\}$$

- et b tel que :

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1, \forall \mathbf{x}_i \; s.c. \; 0 < \alpha_i < C$$

- la contrainte de "boite" sur les multiplicateurs de Lagrange empêche un simple "outlier" d'avoir un grand impact

La SVM à marge souple

- On notera que C contrôle l'importance des cas atypiques (outliers)
  - Une grande valeur de C implique qu'une plus grande importance est donnée à la minimisation du nombre de cas atypiques.
  - un plus grand nombre sera à l'intérieur de la marge
- norme 1 vs norme 2
  - la norme 1 a tendance à limiter de façon plus drastique la contribution des cas atypiques
  - ceci rend la norme 1 légèrement plus robuste, et on a tendance à l'utiliser plus en pratique
- Problèmes communs
  - la façon de régler le paramètre C n'est pas vraiment trés intuitive
  - habituellement on utilise la validation croisée, par rapport à C aux paramètres du noyau

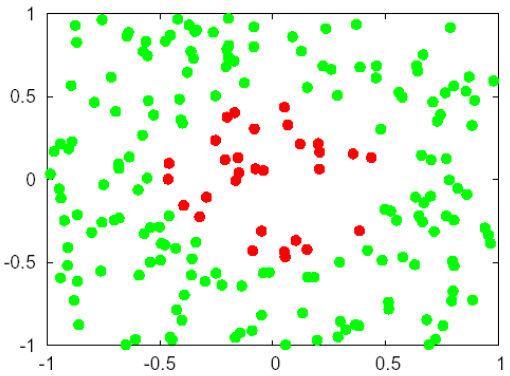
SVM non linéaire - Fonction noyau

Espace intermédiaire

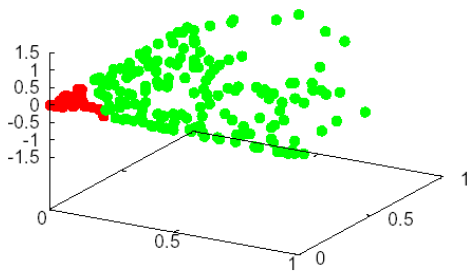
Au lieu de chercher un hyperplan dans l'espace intermédiaire des entrées, on passe dans un espace de représentation intermédiaire (*feature space*) de grande dimension.

$$\Phi : \quad \mathbb{R}^d \rightarrow \mathcal{F}$$

$$\mathbf{x} \Rightarrow \Phi(\mathbf{x})$$



$([x]_1, [x]_2)$



$([x]_1^2, [x]_2^2, \sqrt{2}[x]_1[x]_2)$

SVM non linéaire - Fonction noyau

Espace intermédiaire

On doit donc résoudre

$$(D_{L2}) \begin{cases} \max & \mathcal{L}_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle \\ \text{s.c.} & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & \forall_i, 0 \leq \alpha_i \leq C \end{cases}$$

et la solution à la forme :

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i^* y_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle + b$$

SVM non linéaire - Fonction noyau

Propriétés

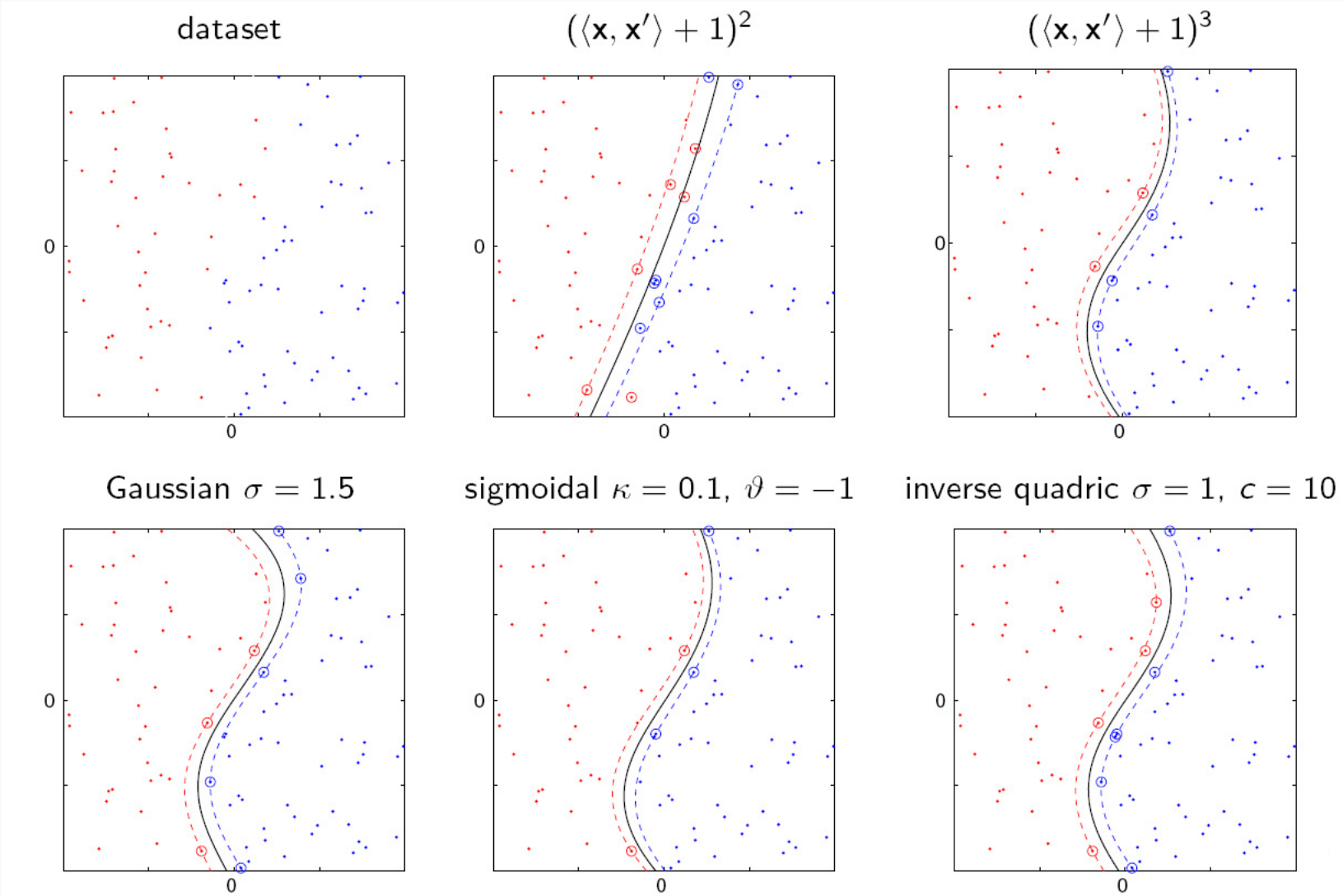
Le problème et sa solution ne dépendent que des produits scalaires $\phi(\mathbf{x}_1).\phi(\mathbf{x}_2)$. On choisit une fonction $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$

- Elle représente un produit scalaire dans l'espace de représentation intermédiaire. Elle traduit donc la répartition des exemples dans cet espace.

$$k(\mathbf{x}_1.\mathbf{x}_2) = \langle \phi(\mathbf{x}_1).\phi(\mathbf{x}_2) \rangle$$

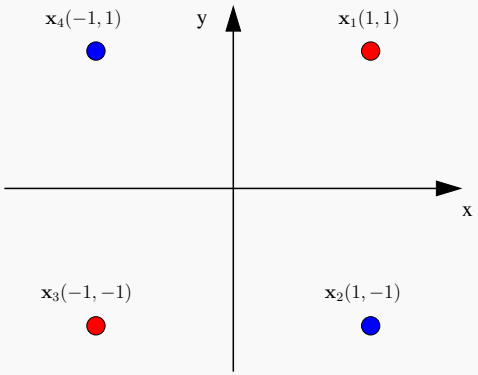
- Lorsque k est bien choisie, on n'a pas besoin de calculer la représentation des exemples dans cet espace pour calculer cette fonction.
- Le noyau matérialise une notion de proximité adaptée au problème.

Effet du choix du noyau – Un exemple



Exemple numérique (1/4)

- Problème du ou exclusif avec SVM non linéaire



- Données
  - classe 1 : $\mathbf{x}_1 = (1,1), \mathbf{x}_3 = (-1,-1)$
  - classe 2 : $\mathbf{x}_2 = (1,-1), \mathbf{x}_4 = (-1,1)$
- Fonction noyau
  - Polynomiale d'ordre 2 : $k(\mathbf{x},\mathbf{z}) = (\langle \mathbf{x},\mathbf{z}\rangle + 1)^2$
- Solution
  - On peut montrer que la transformation implicite est de dimension 5

$$\phi(\mathbf{x}_i) = \begin{bmatrix} 1 & \sqrt{2}x_i & \sqrt{2}y_i & \sqrt{2}x_i y_i & x_i^2 & y_i^2 \end{bmatrix}$$

  - Pour atteindre la séparabilité linéaire, on utilise $C = \infty$
  - La fonction objectif pour le problème dual est

$$\mathcal{L}_D(\alpha) = (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) - \frac{1}{2}\sum_{i=1}^{4}\sum_{j=1}^{4}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j)$$

  - sous les contraintes $\begin{cases} \sum_{i=1}^{n}\alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C \;\; i = 1,\ldots,n \end{cases}$

Exemple numérique (2/4)

- Où le produit scalaire est représenté par une matrice K 4×4

$$K = \begin{bmatrix} 9 & 1 & 1 & 1 \\ 1 & 9 & 1 & 1 \\ 1 & 1 & 9 & 1 \\ 1 & 1 & 1 & 9 \end{bmatrix} \Rightarrow L_D = \sum_{i=1}^{4} \alpha_i - \frac{1}{2} \begin{pmatrix} 9\alpha_1^2 - 2\alpha_1\alpha_2 - 2\alpha_1\alpha_3 + 2\alpha_1\alpha_4 + 9\alpha_2^2 \\ -2\alpha_2\alpha_3 - 2\alpha_2\alpha_4 + 9\alpha_3^2 - 2\alpha_3\alpha_4 + 9\alpha_4^2 \end{pmatrix}$$

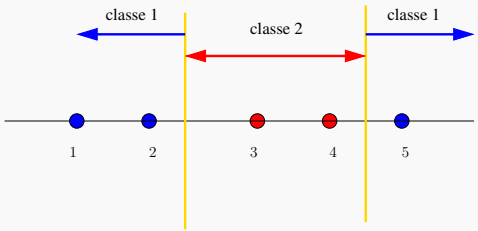
- L'optimisation $\partial \frac{L_D}{\partial \alpha_i} (i = 1, \ldots, 4)$ conduit au système d'équations suivant :

$$\begin{cases} 9\alpha_1 - \alpha_2 - \alpha_3 + \alpha_4 = 1 \\ -\alpha_1 + 9\alpha_2 + \alpha_3 - \alpha_4 = 1 \\ -\alpha_1 + \alpha_2 + 9\alpha_3 - \alpha_4 = 1 \\ \alpha_1 - \alpha_2 - \alpha_3 + 9\alpha_4 = 1 \end{cases} \Rightarrow \begin{array}{l} \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1/8 \\ \text{tous les exemples sont des vecteurs support} \end{array}$$

$$\begin{aligned} f(\mathbf{x}) &= -\frac{1}{8}(1 + x^2 + y^2 - 2x - 2y + 2xx) + \frac{1}{8}(1 + x^2 + y^2 - 2x - 2y + 2xy) \\ &\quad + \frac{1}{8}(1 + x^2 + y^2 - 2x - 2y + 2xy) - \frac{1}{8}(1 + x^2 + y^2 - 2x - 2y + 2xy) + b \\ &= -xy + b = -xy \end{aligned}$$

Exemple numérique (3/4)

- Problème 5 points en D1

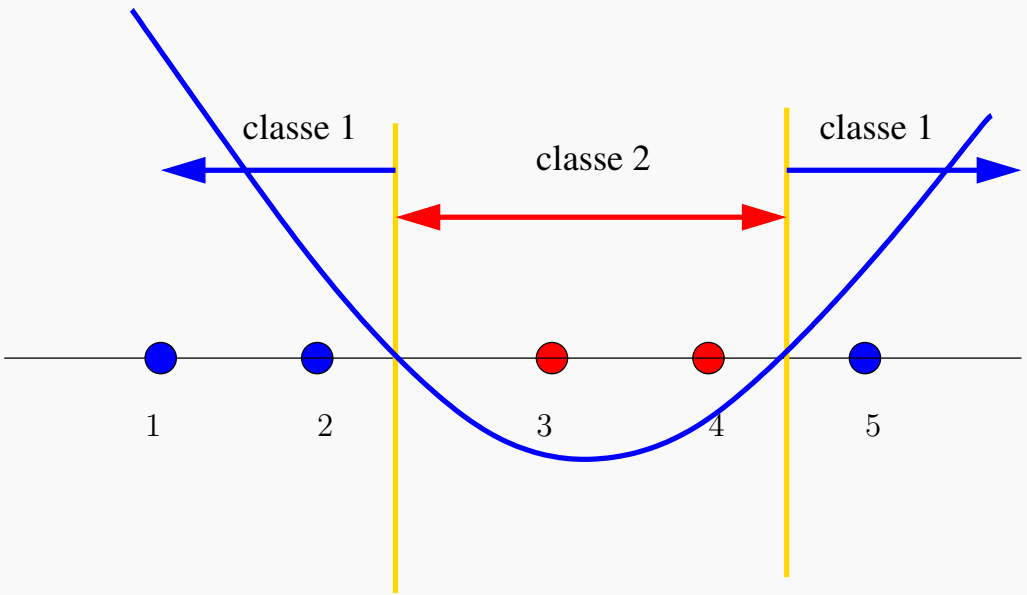


- Données
  - classe 1 : $\mathbf{x}_1 = 1, \mathbf{x}_2 = 2$ et $\mathbf{x}_5 = 6$
  - classe 2 : $\mathbf{x}_3 = 4, \mathbf{x}_4 = 5$
- Fonction noyau
  - Polynomiale 2^{nd} : $k(\mathbf{x}, \mathbf{z}) = (\langle \mathbf{x}, \mathbf{z} \rangle + 1)^2$, $C = 100$
- Solution
  - L'optimisation par rapport aux multiplieurs de Lagrange donne la solution $\alpha_1 = 0, \alpha_2 = 2.5, \alpha_3 = 0, \alpha_4 = 7.333, \alpha_5 = 4.833$
  - Les vecteurs support sont $\{\mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_5\}$
  - La fonction de décision est :

$$\begin{aligned} f(\mathbf{x}) &= 2.5(1)(2x+1)^2 + 7.333(-1)(5x+1)^2 + 4.833(1)(6x+1)^2 + b \\ &= 0.6667x^2 - 5.333x + b \end{aligned}$$

  - On trouve b en résolvant $f(2) = 1$ ou $f(5) = -1$ ou $f(6) = -1$ (condition complémentaire de KKT) (p.ex. $f(2) = 0.6667.2^2 - 5.333.2 + b = 1 \rightarrow b = 9$)

Exemple numérique (4/4)



$$f(\mathbf{x}) = 0.666x^2 - 5.333x + 1$$

Raisons des bonnes performances de SVMs

- L'espace des caractéristiques est souvent de trés grandes dimensions. Pourquoi n'avons nous pas le problème du "fléau de la dimensionnalité" ?
- Un classifieur dans un espace de grandes dimensions a beaucoup de paramètres et est trés difficile à estimer.
- Vapnik argumente que le problème fondamental ne réside pas dans le nombre de paramètres à estimer. Mais plutôt dans la capacité d'un classifieur.
- Typiquement un classifieur avec un grand nombre de paramètres est très flexible, mais il y également quelques exceptions
  - Soit $\mathbf{x}_i = 10^i$ avec $i \in \{1, \ldots, n\}$. Le classifieur $y = sign(sin(\alpha(\mathbf{x})))$ peut classifier tous les $\mathbf{x}_i$ correctement pour toutes les combinaisons d'étiquettage possibles
  - Ce classifieur à un paramètre est très flexible.

Raisons des bonnes performances de SVMs

- L'argument de Vapnik est que la capacité d'un classifieur n'est pas caractérisée pas son nombre de paramètres, mais uniquement par sa capacité.
  - C'est ce qui est formalisé par la dimension de Vapnik d'un classifieur
- La minimisation de $\|\mathbf{w}\|^2$ sujette à la condition que la marge géométrique $= 1$ a pour effet de réduire la dimension de VC du classifieur dans l'espace de caractéristiques.
- La SVM réalise une minimisation du risque structurel : le risque empirique (erreur d'apprentissage) , plus un terme lié à la capacité de généralisation du classifieur, est minimisé
- La fonction de perte est analogue à celle de la regression pénalisée (ridge regression). Le terme $\frac{1}{2}\|\mathbf{w}\|^2$ réduit les paramètres vers 0 pour éviter le surraprentissage.

Choix de la fonction noyau

- Probablement la partie la plus délicate de l'utilisation des SVM
- La fonction noyau doit maximiser la similarité parmi les instances d'une même classe et accentuer les différences entre classes
- Un grand choix de noyaux a été proposé (Noyau de Fisher, Noyau pour chaines, ...) pour différents types de données
- En pratique un noyau polynomial de degré faible ou un noyau RBF avec une largeur de bande raisonnable est un bon choix initial pour commencer.

Classification multi-classe

- k classes
- Stratégies
  - Un contre tous
  - Un contre un (par paire)
  - DAG (Directed Acyclic Graph)
  - Fonctions objectifs multi-classe

Classification multi-classe Un-Contre-Tous

- Apprendre k SVMs $f^1, \ldots, f^k$ pour séparer chaque classe du reste
- fonction de décision

$$f_{\boldsymbol{\alpha}^j, b^j}(\mathbf{x}) = \arg\max_j g^j(\mathbf{x}) \text{ où } g^j(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i^j y_i k(\mathbf{x}, \mathbf{x}_i) + b^j$$

- Pour les exemples pour lesquels il n'y a pas de fonction de décision $g^j(\mathbf{x})$ qui l'emporte $\rightarrow$ rejet, choix aléatoire de la classe ...

Classification multi-classe Un-Contre-Un

- Apprendre $k(k-1)/2$ SVMs binaires pour chaque paire possible de classes
- Assigner à l'exemple la classe avec le plus de votes
- Pour k classes celà résulte à $k(k-1)/2$ SVMs binaires
  - Ex : Si 4 classes $\rightarrow$ 6 SVMs binaires

$y_i = 1$	$y_i = -1$	Ftion Décision
classe1	classe2	$f^{12} = \langle \mathbf{w}^{12}, \mathbf{x} \rangle + b^{12}$
classe1	classe3	$f^{13} = \langle \mathbf{w}^{13}, \mathbf{x} \rangle + b^{13}$
classe1	classe4	$f^{14} = \langle \mathbf{w}^{14}, \mathbf{x} \rangle + b^{14}$
classe2	classe3	$f^{23} = \langle \mathbf{w}^{23}, \mathbf{x} \rangle + b^{23}$
classe2	classe4	$f^{24} = \langle \mathbf{w}^{24}, \mathbf{x} \rangle + b^{24}$
classe3	classe4	$f^{34} = \langle \mathbf{w}^{34}, \mathbf{x} \rangle + b^{34}$

Classification multi-classe Un-Contre-Un

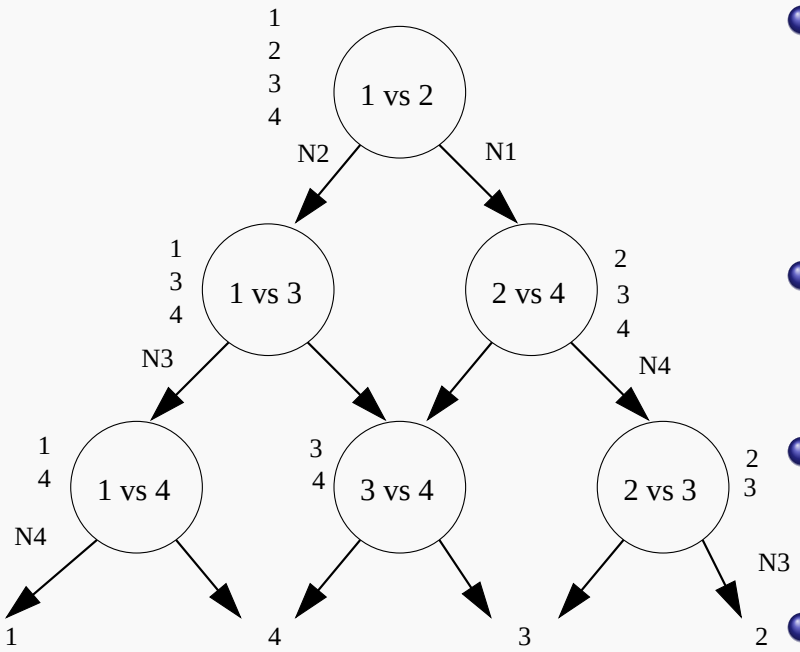
- Pour $\mathbf{x}_{new}$:

Classes	Gain
1-2	1
1-3	1
1-4	1
2-3	2
2-4	4
3-4	3

- Prendre le plus grand vote

Classe	1	2	3	4
nb votes	3	1	1	1

Classification Multi-classe DAG



Principe

- Semblable à la stratégie 1-Contre-1, mais la décision est réalisée au moyen d'un graphe dirigé acyclique.
- Le graphe a $k(k-1)/2$ noeuds internes et k feuilles.
- Chaque noeud représente une SVM binaire sur les classe i et j
- Les feuilles indiquent la classe prédite.

Fonctions objectifs multi-classe

- Similaire à la méthode Un-Contre-Tous
- Construction de k SVM binaires où la fonction de décision m sépare les exemples d'apprentissage de la classe m des autres exemples.

$$(D) \begin{cases} \min\limits_{\mathbf{w},b,\xi} & \frac{1}{2}\sum\limits_{r=1}^{k}\|\mathbf{w}_r\|^2 + C/m\sum\limits_{i=1}^{m}\sum\limits_{r\neq y_i}\xi_i^r \\ \text{s.c.} & \langle \mathbf{w}_{y_i}\phi(\mathbf{x}_i)\rangle + b_y \geq \langle \mathbf{w}_r, \phi(\mathbf{x}_i)\rangle + b_r + 2 - \xi_i^r \\ & \xi_i^r \geq 0, \end{cases}$$

où $r \in \{1,\ldots,k\} \setminus y_i$ et $y_i \in \{1,\ldots,k\}$ (classe de x_i)

$$\arg\max_{i=1\ldots k} (\mathbf{w}^i)^T\phi(\mathbf{x}) + b^i$$

SVM Régression

But : Généraliser la SVC à la régression, en préservant les propriétés suivantes :

1. formulation de l'algorithme pour le cas linéaire puis utilisation de l'astuce du noyau
2. représentation parcimonieuse de la solution en termes de SVs

Perte ϵ-insensible :

$$\ell(y, f(\mathbf{x})) := |y - f(\mathbf{x})|_\epsilon := \max\{0, |y - f(\mathbf{x})| - \epsilon\}$$

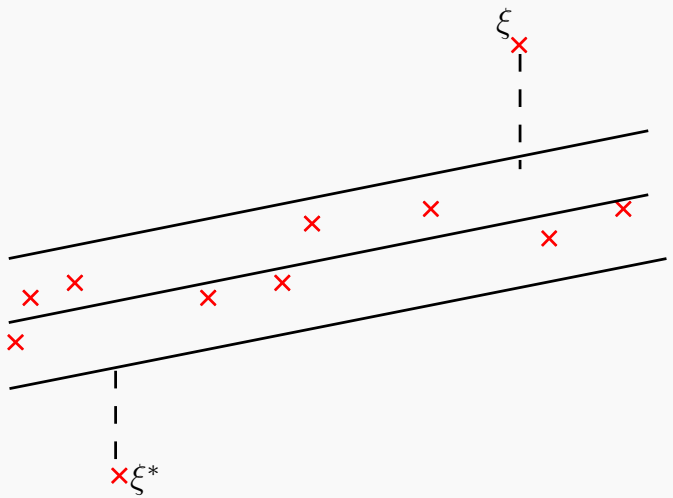
Estimer une régression linéaire $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$ en minimisant

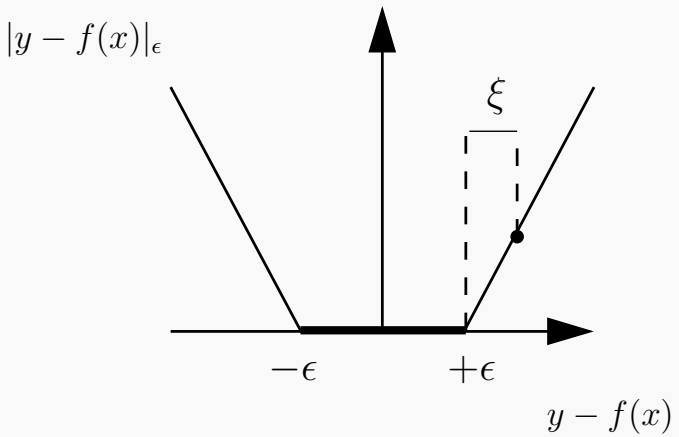
$$\frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n} |y_i - f(\mathbf{x}_i)|_\epsilon$$

SVM Régression (ϵ-SVR)

On met les données dans un hyper-tube

1. les données sur les bordures sont les SV ($\alpha_i \neq 0, \xi_i = 0$)
2. les données en dehors du tube sont considérées comme des erreurs ($\alpha_i \neq 0, \xi_i \neq 0$)





ϵ-SVR : Problème primal

Estimer une régression linéaire $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$. On transforme ce problème en problème d'optimisation contraint en introduisant les deux types de variables ressorts suivants : $f(\mathbf{x}_i) - y_i > \epsilon$ et $y_i - f(\mathbf{x}_i) > \epsilon$ notées respectivement $\boldsymbol{\xi}$ et $\boldsymbol{\xi}^*$
Le problème est alors donné par :

$$(P) \begin{cases} \min\limits_{\mathbf{w},\xi,\xi^*} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum\limits_{i=1}^{n}(\xi_i + \xi_i^*) \\ \text{s.c.} & (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - y_i \leq \epsilon + \xi_i \\ & y_i - (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \leq \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \leq 0, \quad i = 1, \ldots, n \end{cases}$$

ϵ-SVR : Problème dual (1/3)

On construit le Lagrangien à partir de la fonction objectif et des contraintes correspondantes. Cette fonction a un point selle par rapport aux variables primales et duales au point solution. On définit le Lagrangien :

$$\begin{aligned}\mathcal{L} \quad := \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}(\eta_i\xi_i - \eta_i^*\xi_i^*) \\ & - \sum_{i=1}^{n}\alpha_i(\epsilon + \xi_i + y_i - \langle \mathbf{w}, \mathbf{x}_i\rangle - b) \\ & - \sum_{i=1}^{n}\alpha_i(\epsilon + \xi_i^* - y_i - \langle \mathbf{w}, \mathbf{x}_i\rangle - b)\end{aligned}$$

où les variables duales (multiplicateurs de Lagrange) doivent satisfaire la contrainte de positivité, $\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geq 0$

ϵ-SVR : Problème dual (2/3)

D'après les conditions d'optimalités :

$$\begin{aligned}
\frac{\partial \mathcal{L}(\mathbf{w}, b, \xi_i, \xi_i^*)}{b} &= \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0 \\
\frac{\partial \mathcal{L}(\mathbf{w}, b, \xi_i, \xi_i^*)}{\mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)\mathbf{x}_i = 0 \\
\frac{\partial \mathcal{L}(\mathbf{w}, b, \xi_i, \xi_i^*)}{\xi_i} &= \frac{C}{n} - \alpha_i - \eta_i = 0 \\
\frac{\partial \mathcal{L}(\mathbf{w}, b, \xi_i, \xi_i^*)}{\xi_i^*} &= \frac{C}{n} - \alpha_i^* - \eta_i^* = 0
\end{aligned}$$

et en resubstituant dans le Lagrangien

ϵ-SVR : Problème dual (3/3)

On obtient pour $C > 0, \epsilon \geq 0$ choisi à priori, la formulation duale

$$(D) \begin{cases} \max\limits_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} & -\epsilon \sum\limits_{i=1}^{n}(\alpha_i^* - \alpha_i)y_i - \frac{1}{2}\sum\limits_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)k(\mathbf{x}_i, \mathbf{x}_j) \\ \text{s.c.} & \sum(\alpha_i^* - \alpha_i) = 0 \\ & 0 \leq \alpha_i,\ \alpha_i^* \leq C, \quad i = 1, \ldots, n \end{cases}$$

L'estimation de la regression prend alors la forme :

$$f(\mathbf{x}) = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)k(\mathbf{x}_i - \mathbf{x}) + b$$

ϵ-SVR calcul de b

Pour le calcul de b l'application des conditions de KKT qui établissent qu'au point solution le produit entre les variables duales et les contraintes doit s'annuler

$$\alpha_i(\epsilon + \xi_i - y_i + \langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 0$$
$$\alpha_i^*(\epsilon + \xi_i^* - y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 0$$

et

$$\left(\frac{C}{n} - \alpha_i\right)\xi_i = 0$$
$$\left(\frac{C}{n} - \alpha_i^*\right)\xi_i = 0$$

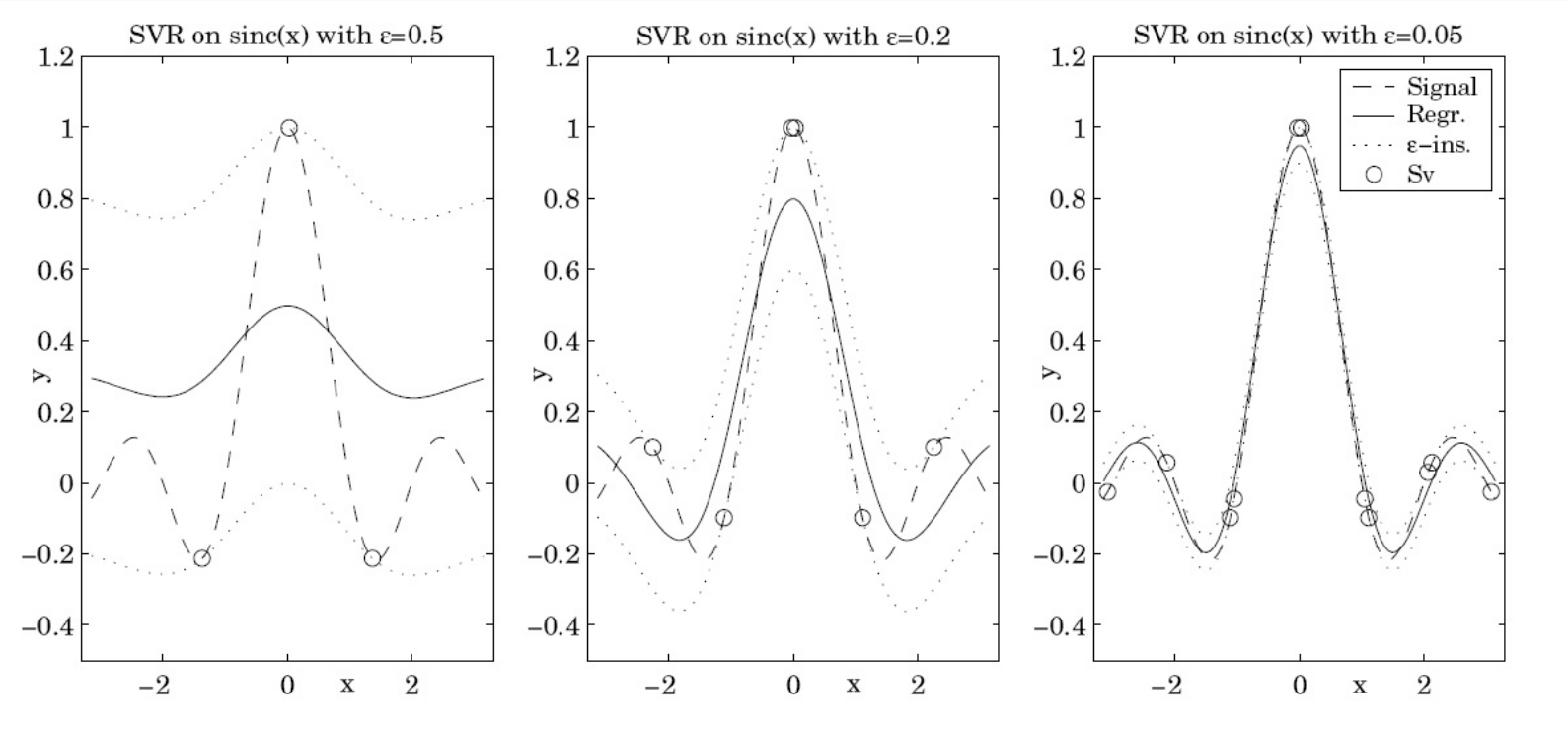
ϵ-SVR calcul de b

Ce qui nous permet de tirer plusieurs conclusions utiles :

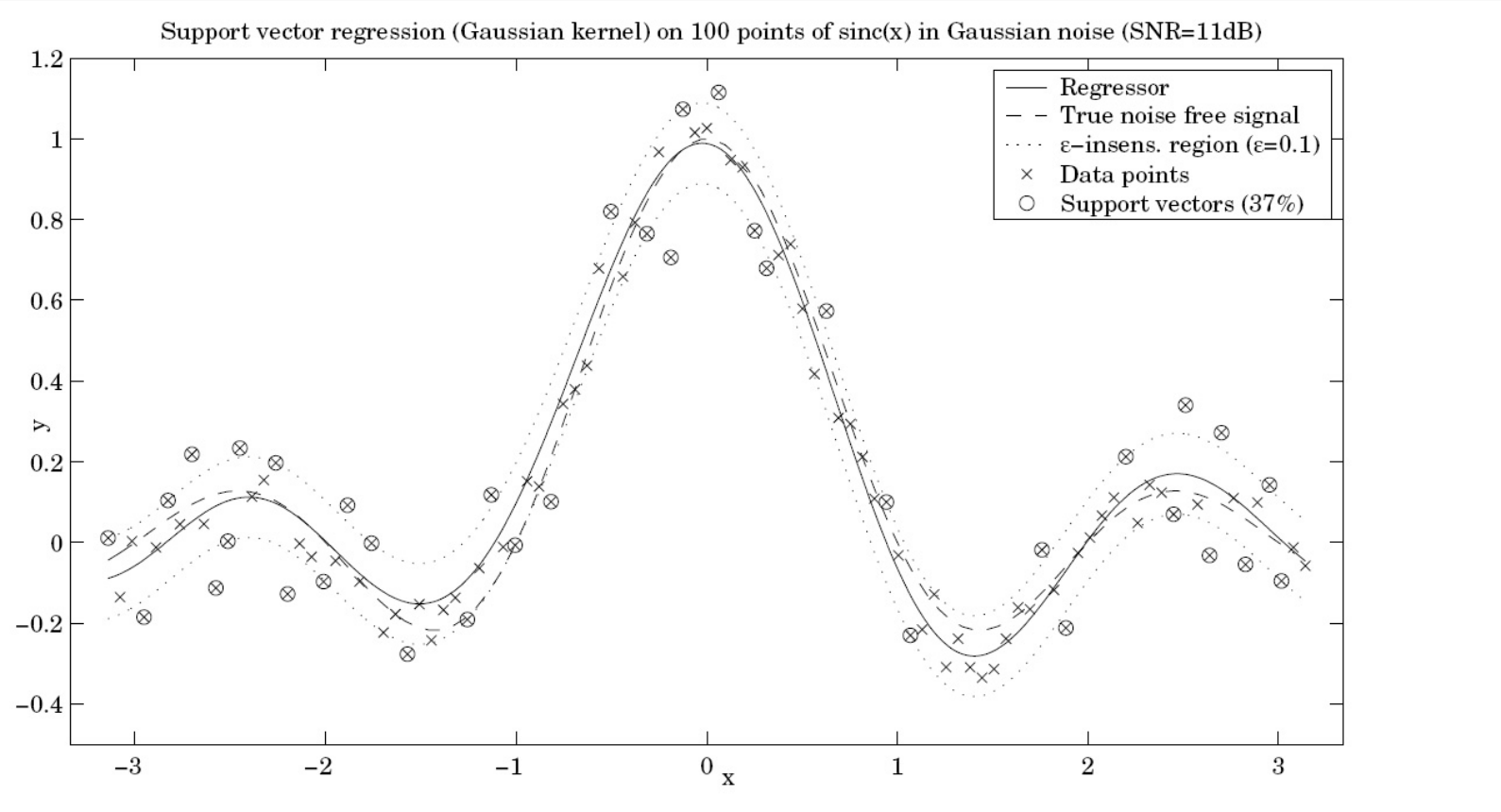
- Seulement les exemples $(\mathbf{x}_i, y_i)$ avec un $\alpha_i, \alpha_i^* = C \backslash n$ peuvent se trouver en dehors du ϵ-tube (i.e., $\xi_i, \xi_i^* > 0$) autour de f.
- On a $\alpha_i \alpha_i^* = 0$ autrement dit il n'existe pas un ensemble de variables α_i, α_i^* qui sont toutes les deux simultanément différentes de 0.
- Pour $\alpha_i, \alpha_i^* \in (0, C \backslash n)$ on a $\xi_i, \xi_i^* = 0$ et de plus le second facteur doit s'annuler. Ainsi b peut être calculé comme suit :

$$\begin{aligned} b &= y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - \epsilon \quad \text{pour } \alpha_i \in (0, C \backslash n), \\ b &= y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle + \epsilon \quad \text{pour } \alpha_i^* \in (0, C \backslash n), \end{aligned}$$

ϵ-SVR Exemple (1/2)



ϵ-SVR Exemple (2/2)



Conclusion

- L'introduction des SVMs a permis de réunir de manière productive des méthodes et des personnes venant de l'informatique, des statistiques et de l'optimisation.
- Une grande force est sa modularité qui permet de séparer clairement les tâches.
- Bien que cousine de méthodes statistiques plus anciennes, la SVM a permis d'introduire un point de vue radicalement nouveau sur ces méthodes, en faisant ressortir l'intérêt pratique d'utilisation de systématique noyaux pour des données très diverses.
- Toutefois l'utilisation généralisée de noyaux pour le traitement de données apparaît finalement comme un principe et un acquis plus fondamental que celui de l'algorithme SVM lui-même.

Références I


V. Vapnik.
Statistical learning theory.
John Wiley & Sons, 1998.


N. Cristianini,J. Shawe-Taylor.
An Introduction to Support Vector Machines.
Cambridge University Press, 2000.


B. Scholkopf,A Smola.
Learning with Kernels.
MIT Press,MA, 2002.


N. Cristianini,J. Shawe-Taylor.
Kernel Methods for Pattern Analysis.
Cambridge University Press, 2004.

Références II


B. Burges.
A tutorial on Support Vector Machines for Pattern Recognition.
Data Mining and Knowledge Discovery, 1998.


K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf
An introduction to kernel-based learning algorithms.
IEEE Neural Networks, 12(2) :181-201, May 2001.