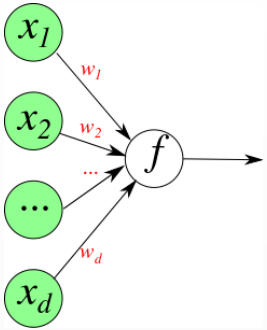
Prédicteurs linéaires

(avant les réseaux de neurones)

Pascal Germain

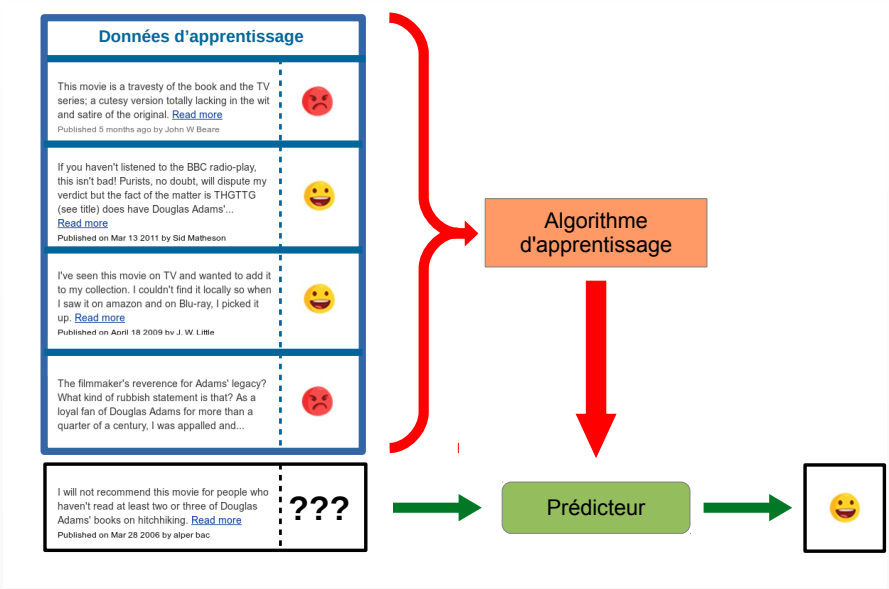
2019


x_1
x_2
...
x_d
w_1
w_2
...
w_d
f

Plan

Plan

Exemple : Déterminer l'appréciation d'un film à partir d'un commentaire



Représentation des données

Ensemble d'apprentissage :

$$S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\},,$$

avec

Régression

$$\mathbf{x}_i \in \mathbb{R}^d \text{ et } y_i \in \mathbb{R}.$$

Classification binaire

$$\mathbf{x}_i \in \mathbb{R}^d \text{ et } y_i \in \{-1, +1\} \qquad (\text{ ou } y_i \in \{0, 1\})$$

Prédicteur linéaire

Régression

$$f_{\mathbf{w},b}(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} - b$$

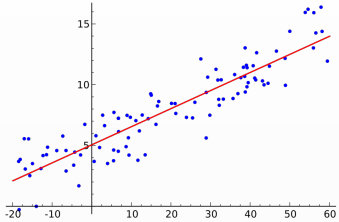
Classification binaire

$$f_{\mathbf{w},b}(\mathbf{x}) = \operatorname{sgn}\big[\mathbf{w} \cdot \mathbf{x} - b\big] = \begin{cases} +1 & \text{si } \mathbf{w} \cdot \mathbf{x} - b > 0 \\ -1 & \text{sinon.} \end{cases}$$

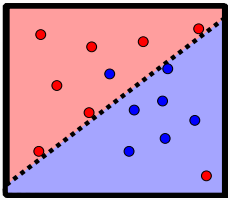
Note : $f_{\mathbf{w},b}(\mathbf{x}) = f_{\mathbf{w}',b'}(\mathbf{x})$ avec $\mathbf{w}' = c\,\mathbf{w}$ et $b' = c\,b$ pour tout $c > 0$.

Interprétation des prédicteurs

Régression : Un prédicteur est une surface qui relie les exemples



Classification : Un prédicteur est une frontière de décision qui sépare les exemples



Plan

Plan

Classification binaire

Ensemble d'apprentissage :

$$S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\},,$$

avec $\mathbf{x}_i \in \mathbb{R}^d$ et $y_i \in \{-1, +1\}$.

Prédicteur linéaire :

$$f_{\mathbf{w},b}(\mathbf{x}) = \operatorname{sgn}\big[\mathbf{w} \cdot \mathbf{x} - b\big] = \begin{cases} +1 & \text{si } \mathbf{w} \cdot \mathbf{x} - b > 0 \\ -1 & \text{sinon.} \end{cases}$$

Note : $f_{\mathbf{w},b}(\mathbf{x}) = f_{\mathbf{w}',b'}(\mathbf{x})$ avec $\mathbf{w}' = c\,\mathbf{w}$ et $b' = c\,b$ pour tout $c > 0$.

Marge d'un prédicteur linéaire sur un exemple

Marge fonctionnelle du prédicteur $f_{\mathbf{w},b}$ sur l'exemple $(\mathbf{x}, y)$:

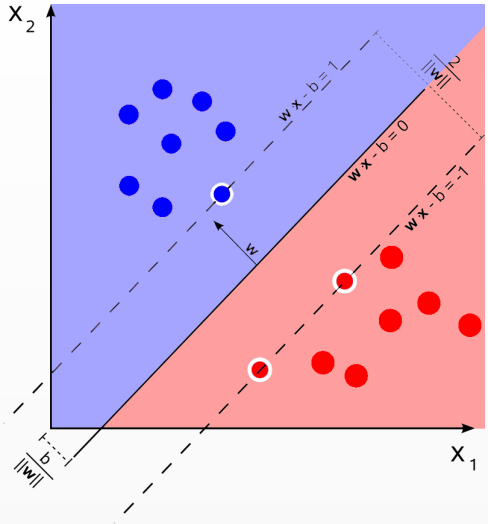
$$y\,(\mathbf{w} \cdot \mathbf{x} - b)$$

Marge géométrique du prédicteur $f_{\mathbf{w},b}$ sur l'exemple $(\mathbf{x}, y)$:

$$\frac{y\,(\mathbf{w} \cdot \mathbf{x} - b)}{\|\mathbf{w}\|}$$

Note 1 : Un exemple bien classifié ($f_{\mathbf{w},b}(\mathbf{x}) = y$) ssi sa marge est positive.

Note 2 : Avec $c > 0$, $\mathbf{w}' = c\,\mathbf{w}$ et $b' = c\,b$ les prédicteurs $f_{\mathbf{w},b}$ et $f_{\mathbf{w}',b'}$ possèdent la même marge géométrique.


x2
x1
w x - b = 1
w x - b = 0
w x - b = -1
w
2/||w||
b
b/||w||

Marge d'un prédicteur linéaire sur un ensemble d'apprentissage

Marge fonctionnelle du prédicteur $f_{\mathbf{w},b}$ sur l'ensemble S :

$$\min_{(\mathbf{x}_i,y_i)\in S} \left[y_i \left(\mathbf{w} \cdot \mathbf{x}_i - b \right) \right]$$

Marge géométrique du prédicteur $f_{\mathbf{w},b}$ sur l'ensemble S :

$$\min_{(\mathbf{x}_i,y_i)\in S} \left[\frac{y_i \left(\mathbf{w} \cdot \mathbf{x}_i - b \right)}{\|\mathbf{w}\|} \right]$$

Note : On dit qu'un ensemble S est linéairement séparable lorsqu'il existe un couple $(\mathbf{w}, b) \in \mathbb{R}^d \times \mathbb{R}$ tel que la marge sur l'ensemble S est positive.

Plan

SVM à marge rigide

Supposons que l'ensemble d'apprentissage S soit linéairement séparable.

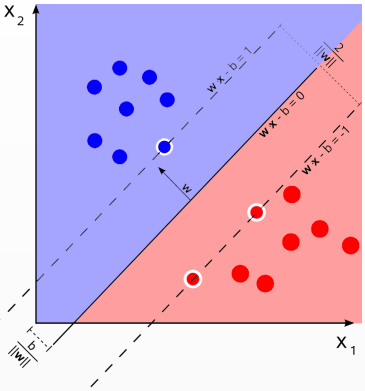
Le SVM à marge rigide trouve un prédicteur $f_{\mathbf{w},b}$ de marge géométrique maximale :

$$\max_{(\mathbf{w},b)\in\mathbb{R}^d\times\mathbb{R}} \left(\min_{(\mathbf{x}_i,y_i)\in S} \left[\frac{y_i\left(\mathbf{w}\cdot\mathbf{x}_i - b\right)}{\|\mathbf{w}\|} \right] \right)$$

Il existe une multitude de solutions $(\mathbf{w}, b)$ à ce problème... Prenons la solution telle que la marge fonctionnelle sur l'ensemble S vaut 1, c'est-à-dire :

$$\min_{(\mathbf{x}_i,y_i)\in S} \left[\frac{y_i\left(\mathbf{w}\cdot\mathbf{x}_i - b\right)}{\|\mathbf{w}\|} \right] = \frac{1}{\|\mathbf{w}\|}$$

SVM à marge rigide



Étant donné :

$$S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}\,.$$

Minimiser :	$\frac{1}{2}\|\mathbf{w}\|^2$
sous contraintes :	$y_i\,(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1$
	pour $i = 1, \ldots, n\,.$

On nomme *vecteurs de supports* les exemples dont la marge fonctionnelle est 1.

Plan

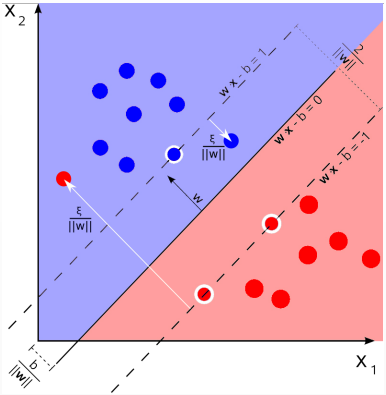
SVM à marge floue

Pour d'adapter à la situation où S n'est pas linéairement séparable, on introduit des variables d'écarts.

$$\xi_i = \max\left\{0, 1 - y_i(\mathbf{w} \cdot \mathbf{x}_i - b)\right\}$$

$$\begin{aligned} \textbf{Minimiser :} \quad & \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{sous contraintes :} \quad & y_i\left(\mathbf{w} \cdot \mathbf{x}_i - b\right) \geq 1 - \xi_i \;\text{ et }\; \xi_i \geq 0 \\ & \text{pour } i = 1, \ldots, n\,. \end{aligned}$$

SVM à marge floue



La valeur de chaque ξ_i s'interprète ainsi :

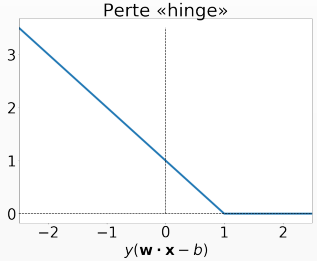
- $\xi_i > 1$: L'exemple est mal classifié.
- $0 < \xi_i < 1$: L'exemple est bien classifié, mais il est situé à l'intérieur de la marge du SVM.
- $\xi_i = 0$: L'exemple est bien classifié et il est situé à l'extérieur de la marge du SVM.

SVM à marge floue

Minimiser : $\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \ell_{\text{hinge}}(f_{\mathbf{w},b}(\mathbf{x}_i), y_i)$

avec

$$\ell_{\text{hinge}}(\hat{y}, y) = \max\left\{0, 1 - \hat{y} \times y\right\}$$



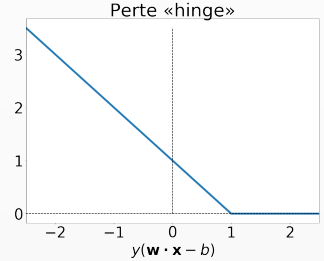
Plan

Plan

SVM $\Leftrightarrow$ Minimisation de la perte «hinge» régularisée

Minimiser : $$F(\mathbf{w}) = C \underbrace{\sum_{i=1}^{n} \ell_{\text{hinge}}(f_{\mathbf{w},b}(\mathbf{x}_i), y_i)}_{\text{perte empirique}} + \underbrace{\frac{1}{2}\|\mathbf{w}\|^2}_{\text{régularisation}}$$

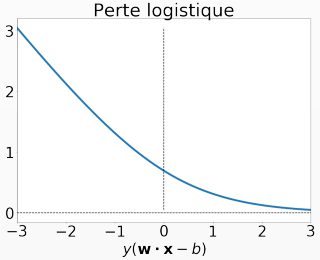
$$\ell_{\text{hinge}}(\hat{y}, y) = \max\left\{0, 1 - \hat{y} \times y\right\}$$



Régression logistique ⇔ Minimisation de la perte logistique régularisée

$$\textbf{Minimiser :} \qquad F(\mathbf{w}) = C \underbrace{\sum_{i=1}^{n} \ell_{\text{logist}}(f_{\mathbf{w},b}(\mathbf{x}_i), y_i)}_{\text{perte empirique}} + \underbrace{\frac{1}{2}\|\mathbf{w}\|^2}_{\text{régularisation}}$$

$$\ell_{\text{hinge}}(\hat{y}, y) = \ln(1 + e^{-\hat{y} \times y})$$

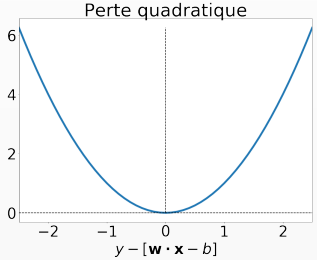


Plan

Moindre carrés $\Leftrightarrow$ Minimisation de la perte quadratique

Minimiser : $$F(\mathbf{w}) = \underbrace{\sum_{i=1}^{n} \ell_{\text{quad}}(f_{\mathbf{w},b}(\mathbf{x}_i), y_i)}_{\text{perte empirique}}$$

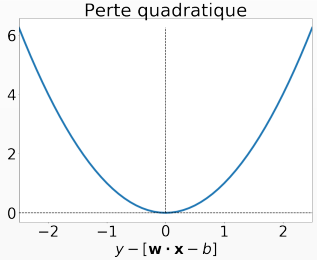
$$\ell_{\text{quad}}(\hat{y}, y) = (y - \hat{y})^2$$



Regression de ridge $\Leftrightarrow$ Minimisation de la perte quadratique régularisée

Minimiser : $$F(\mathbf{w}) = C \underbrace{\sum_{i=1}^{n} \ell_{\text{quad}}(f_{\mathbf{w},b}(\mathbf{x}_i), y_i)}_{\text{perte empirique}} + \underbrace{\frac{1}{2}\|\mathbf{w}\|^2}_{\text{régularisation}}$$

$$\ell_{\text{quad}}(\hat{y}, y) = (y - \hat{y})^2$$



Plan

Le noyau linéaire

Étant donné :

$$S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}\,.$$

Représentons le vecteur $\mathbf{w} \in \mathbb{R}^d$ par un vecteur de variables duales $\boldsymbol{\alpha} \in \mathbb{R}^n$, tel que

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i \, \mathbf{x}_i\,.$$

Nous pouvons réécrire le prédicteur $f_{\mathbf{w},b}$:

$$\begin{aligned} f_{\mathbf{w},b}(\mathbf{x}) = \operatorname{sgn}\left[\mathbf{w} \cdot \mathbf{x} + b\right] &= \operatorname{sgn}\left[\sum_{i=1}^{n} \alpha_i \, \mathbf{x}_i \cdot \mathbf{x} + b\right] \\ &= \operatorname{sgn}\left[\sum_{i=1}^{n} \alpha_i \, k(\mathbf{x}_i, \mathbf{x}) + b\right], \end{aligned}$$

où $k(\mathbf{x}_i, \mathbf{x}) = \mathbf{x}_i \cdot \mathbf{x}$ est la fonction de **noyau linéaire**.

Le noyau linéaire

$$
\begin{aligned}
\|\mathbf{w}\|^2 = \mathbf{w} \cdot \mathbf{w} &= \left(\sum_{i=1}^{n} \alpha_i \, \mathbf{x}_i \right) \cdot \left(\sum_{i=1}^{n} \alpha_i \, \mathbf{x}_i \right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \, \mathbf{x}_i \cdot \mathbf{x}_j \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \, k(\mathbf{x}_i, \mathbf{x}_j)
\end{aligned}
$$

Astuce du noyau

Une fonction $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ est un *noyau* ssi il existe une transformation $\phi : \mathbb{R}^d \to \mathbb{R}^D$ telle que

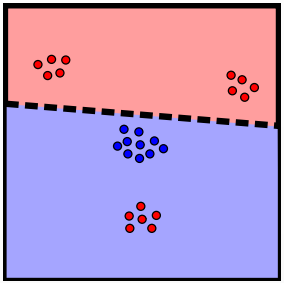
$$k(\mathbf{x}_i, \mathbf{x}) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x})$$

Autrement dit, un noyau calcule à un **produit scalaire dans un espace augmenté**.

Quelques exemples de noyaux

Noyau linéaire

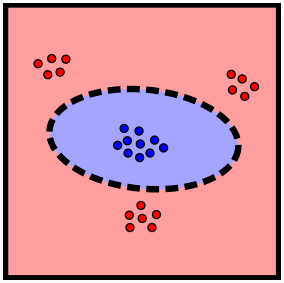
$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$



Noyau polynomial

(degré 2)

$k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^2$



Noyau gaussien

$k(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|}$