Apprentissage Machine / Statistique

Support Vector Machines (SVM)

PHILIPPE BESSE

INSA de Toulouse
Institut de Mathématiques

Principes généraux

- Séparateur à Vaste Marge (SVM)
- Machine à Vecteurs Support (MVS)
- Apprentissage en discrimination : $\{-1, 1\}$
- Etendu à $m > 2$ et $\mathbb{R}$
- Hyperplan de marge optimale pour la généralisation
- Vapnik (1998) et VC-dimension
- Contrôle de la complexité
- L'objectif, seulement l'objectif
- Coût calcul fonction de n, pas de p

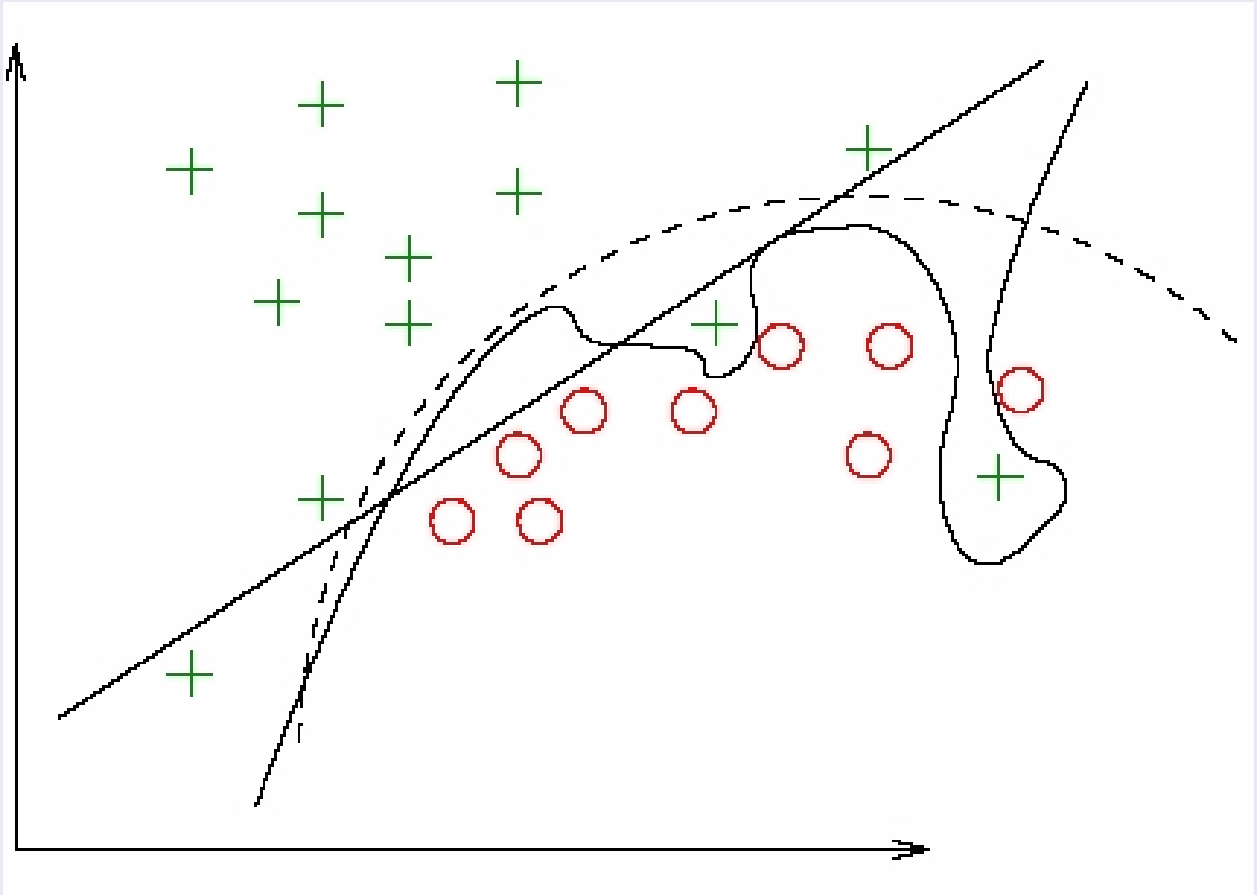
Spécificités

- Ramener la discrimination à un problème linéaire
- Problème d'optimisation sous-contrainte et support
- Utilisation d'un espace intermédiaire (*feature space*)
- Produit scalaire et noyau reproduisant

Remarques

- Efficacité et flexibililté des noyaux
- Schölkopf et Smola (2002)
- `www.kernel-machines.org`

Sur-ajustement



Frontière, complexité, généralisation et VC-dimension

Notations

- Y à valeurs dans $\{-1, 1\}$
- $\boldsymbol{X} = X^1, \ldots, X^p$ les variables prédictives
- $Y = f(\boldsymbol{X})$ un modèle pour Y
- Un échantillon statistique de loi F

$$\boldsymbol{z} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$$

- Estimation de $\widehat{f}$ de f, ($\mathbb{R}^p$(ou $\mathcal{F}$) $\mapsto \{-\infty, \infty\}$)
- par minimisation de :

$$P(f(\boldsymbol{X}) \neq Y)$$

Définition de la marge

- f définie par une fonction réelle f : $\widehat{f} = \text{signe}(f)$
- L'erreur devient : $P(f(\boldsymbol{X}) \neq Y) = P(Yf(\boldsymbol{X}) \leq 0)$
- $|Yf(\boldsymbol{X})|$ est un indicateur de confiance
- $Yf(\boldsymbol{X})$ est la marge de f en $(\boldsymbol{X}, Y)$

Espace hilbertien

- $\Phi : \mathbb{R}^p (\text{ou } \mathcal{F}) \mapsto \mathcal{H}$
- $\mathcal{H}$: *feature space* de grande dimension avec produit scalaire
- Φ ramène à un problème linéaire : hyperplan séparateur
- Première approche : Φ est la fonction identité

Recherche du plan de marge maximale

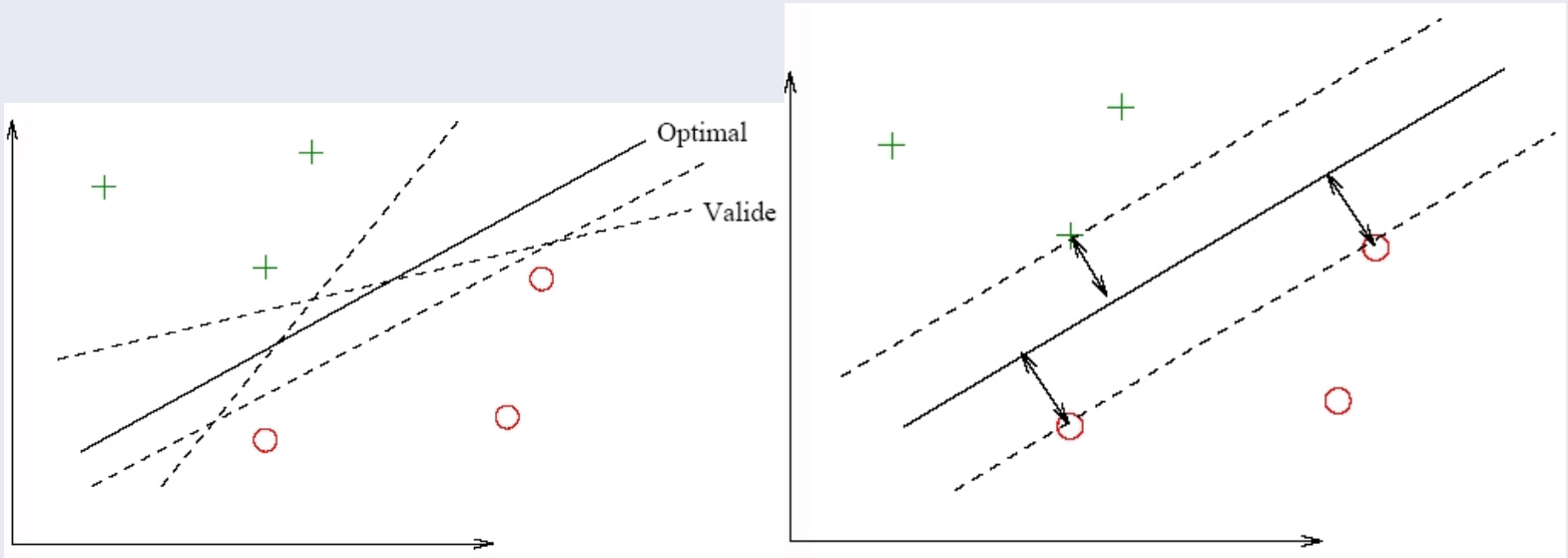
- Un hyperplan est défini à l'aide du produit scalaire de $\mathcal{H}$:

$$\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 0$$

où $\boldsymbol{w}$ est un vecteur orthogonal au plan

- Le signe de la fonction $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$ indique la position de $\boldsymbol{x}$ à prédire
- Un point est bien classé si et seulement si : $yf(\boldsymbol{x}) > 0$
- $(\boldsymbol{w}, b)$ est défini à un coef. près ; on impose : $yf(\boldsymbol{x}) \geq 1$
- Un plan $(\boldsymbol{w}, b)$ est un séparateur si : $\forall i \quad y_i f(\boldsymbol{x}_i) \geq 1$
- Distance de $\boldsymbol{x}$ au plan $(\boldsymbol{w}, b)$: $d(\mathbf{x}) = \frac{|\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b|}{\|\boldsymbol{w}\|} = \frac{|f(\boldsymbol{x})|}{\|\boldsymbol{w}\|}$
- La marge du plan a pour valeur : $\frac{2}{\|\boldsymbol{w}\|^2}$

Plan de marge maximale



Problème primal d'optimisation sous contraintes

$$\begin{cases} \min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \\ \text{avec } \forall i, y_i < \boldsymbol{w}, \boldsymbol{x}_i > +b \geq 1 \end{cases}$$

Problème dual avec multiplicateurs de Lagrange

- La solution est un point-selle $(\boldsymbol{w}^*, b^*, \boldsymbol{\lambda}^*)$ du lagrangien :

$$L(\boldsymbol{w}, b, \boldsymbol{\lambda}) = 1/2\|w\|_2^2 - \sum_{i=1}^{n} \lambda_i \left[y_i \left(< \boldsymbol{w}, \boldsymbol{x}_i > +b\right) - 1\right]$$

- Ce point-selle vérifie : $\forall i \quad \lambda_i^* \left[y_i \left(< \boldsymbol{w}^*, \boldsymbol{x}_i > +b^*\right) - 1\right] = 0$
- *Vecteurs support* : $\boldsymbol{x}_i$ avec contrainte active
- Appartiennent au plan : $y_i \left(< \boldsymbol{w}^*, \boldsymbol{x}_i > +b^*\right) = 1$

Formule duale du lagrangien

- Plan optimal : $\boldsymbol{w}^* = \sum_{i=1}^{n} \lambda_i^* y_i \boldsymbol{x}_i$ et $\sum_{i=1}^{n} \lambda_i^* y_i = 0$
- $W(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j < \boldsymbol{x}_i, \boldsymbol{x}_j >$
- Le point-selle maximise $W(\boldsymbol{\lambda})$ avec $\lambda_i \geq 0 \quad \forall i$
- Problème d'optimisation quadratique de taille n
- Hyperplan optimal : $\sum_{i=1}^{n} \lambda_i^* y_i < \boldsymbol{x}, \boldsymbol{x}_i > + b^* = 0$
- avec $b^* = -\frac{1}{2} \left[< \boldsymbol{w}^*, sv_{class+1} > + < \boldsymbol{w}^*, sv_{class-1} > \right]$
- La prévision de $\boldsymbol{x}$ est fournie par le signe de

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \lambda_i^* y_i \left\langle \boldsymbol{x}, \boldsymbol{x}_i \right\rangle + b^*$$

Cas non séparable

- Assouplissement des contraintes
- les termes d'erreur ξ_i contrôlent le dépassement :

$$y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b \geq +1 - \xi_i \quad \forall i \in \{1, \ldots, n\}$$

- La prédiction de $\boldsymbol{x}_i$ est fausse à un vecteur si $\xi_i > 1$
- La somme des ξ_i est une borne du nombre d'erreurs
- Nouveau problème de minimisation avec pénalisation par le dépassement de la contrainte :

$$\begin{cases} \min \frac{1}{2} \|\boldsymbol{w}\|^2 + \delta \sum_{i=1}^{n} \xi_i \\ \forall i, y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b \geq +1 - \xi_i \end{cases}$$

Remarques

- δ contrôle le compromis entre ajustement et généralisation
- Même forme duale mais avec les λ_i bornés par δ
- n grand : algorithmes avec décomposition de l'ensemble d'apprentissage
- Capacité de généralisation dépend du nombre de vecteurs supports mais pas de la taille de l'espace
- Si les $\boldsymbol{X}$ sont dans une boule de rayon R, l'ensemble des hyperplans de marge fixée δ a une VC-dimension bornée par $\frac{R^2}{\delta^2}$ avec $\|\boldsymbol{w}\| \leq R$
- Bornes d'erreur estimables mais trop pessimistes

Produit scalaire et noyau

- $\Phi : \mathbb{R}^p (\text{ou } \mathcal{F}) \mapsto \mathcal{H}$
- $\mathcal{H}$muni d'un produit scalaire et de plus grande dimension
- Le problème de minimisation et la solution :

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

font intervenir $\boldsymbol{x}$ et $\boldsymbol{x}'$ par l'intermédiaire de produits scalaires :

$$\langle \boldsymbol{x}, \boldsymbol{x}' \rangle$$

Astuce du noyau

- Il est inutile d'expliciter Φ
- Il suffit de calculer les produits scalaires dans $\mathcal{H}$
- Fonction noyau $k : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$ symétrique :

$$k(\boldsymbol{x}, \boldsymbol{x'}) = \left\langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x'}) \right\rangle_{\mathcal{H}}$$

- Le noyau définit une notion de distance

Exemple trivial

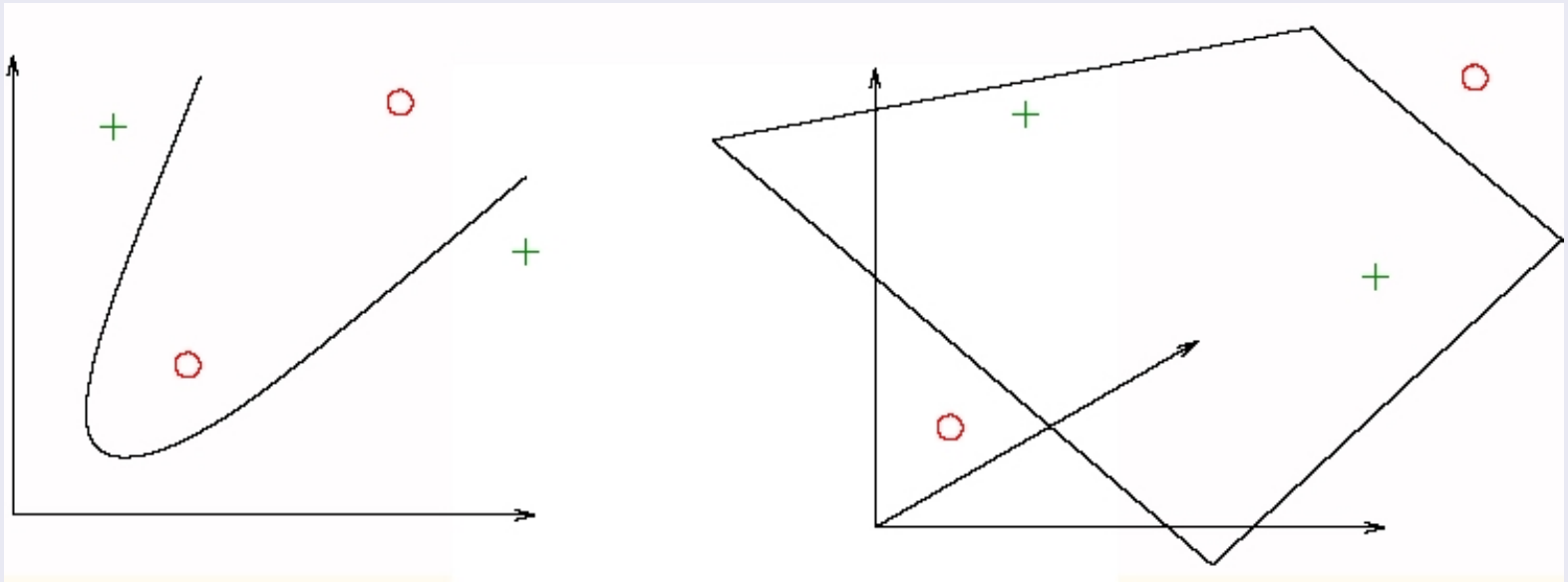
- $\boldsymbol{x} = (x_1, x_2)$ dans $\mathbb{R}^2$
- $\Phi(\boldsymbol{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$
- $\mathcal{H}$ de dimension 3 et de produit scalaire :

$$\begin{aligned}
\left\langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x'}) \right\rangle &= x_1^2 x_1'^2 + 2x_1x_2x_1'x_2' + x_2^2x_2'^2 \\
&= (x_1x_1' + x_2x_2')^2 \\
&= \left\langle \boldsymbol{x}, \boldsymbol{x'} \right\rangle^2 \\
&= k(\boldsymbol{x}, \boldsymbol{x'})
\end{aligned}$$

En général

- Le produit scalaire dans $\mathcal{H}$ ne nécessite pas d'expliciter Φ
- Le plongement dans $\mathcal{H}$ peut rendre possible la séparation linéaire

Feature space



Rôle de l'espace intermédiaire dans la séparation des données

Définition

Une fonction $k(.,.)$ symétrique est un noyau si, pour tous les $\boldsymbol{x}_i$ possibles, la matrice de terme général $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ est une matrice définie positive

- Elle définit une matrice de produit scalaire
- Dans ce cas, 'il existe un espace $\mathcal{H}$ (Hilbert à noyau reproduisant) et une fonction Φ tels que :

$$k(\boldsymbol{x}, \boldsymbol{x}') = \left\langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}') \right\rangle$$

Attention

Condition d'existence, pas constructive et difficile à vérifier

Noyaux classiques

- Linéaire

$$k(\boldsymbol{x}, \boldsymbol{x'}) = \langle \boldsymbol{x}, \boldsymbol{x'} \rangle$$

- Polynômial

$$k(\boldsymbol{x}, \boldsymbol{x'}) = (c + \langle \boldsymbol{x}, \boldsymbol{x'} \rangle)^d$$

- Radial gaussien

$$k(\boldsymbol{x}, \boldsymbol{x'}) = e^{-\frac{\|x - x'\|^2}{2\sigma^2}}$$

Noyaux spécifiques

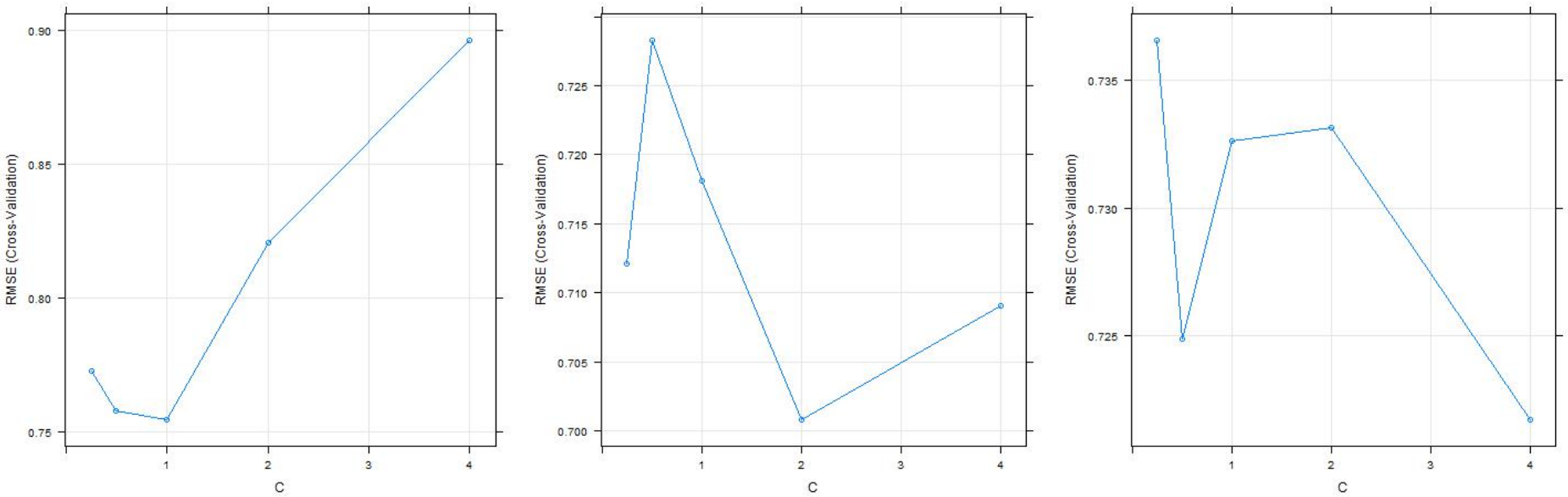
- Travail : construction d'un noyau adapté : reconnaissance de séquences, de caractères, l'analyse de textes, de graphes...
- Grande flexibilité entraîne une bonne efficacité
- Choix de noyau, des paramètres par validation croisée
- Paradoxe : les SVM à noyaux gaussiens dans le cas séparable ou à pénalité variable, dont de VC-dimension infinie

Cas de la régression

- Y est quantitative
- La fonction se décompose : $f(\boldsymbol{x}, \boldsymbol{w}) = \sum_{i=1}^{\infty} w_i v_i(\boldsymbol{x})$
- Fonction coût issue de la robustesse :

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

- $|.|_\epsilon$ fonction paire, continue, identiquement nulle sur $[0, \epsilon]$ et qui croît linéairement sur $[\epsilon, +\infty]$
- γ contrôle l'ajustement
- Même principe de résolution
- Noyaux de splines ou encore noyau de Déricl et

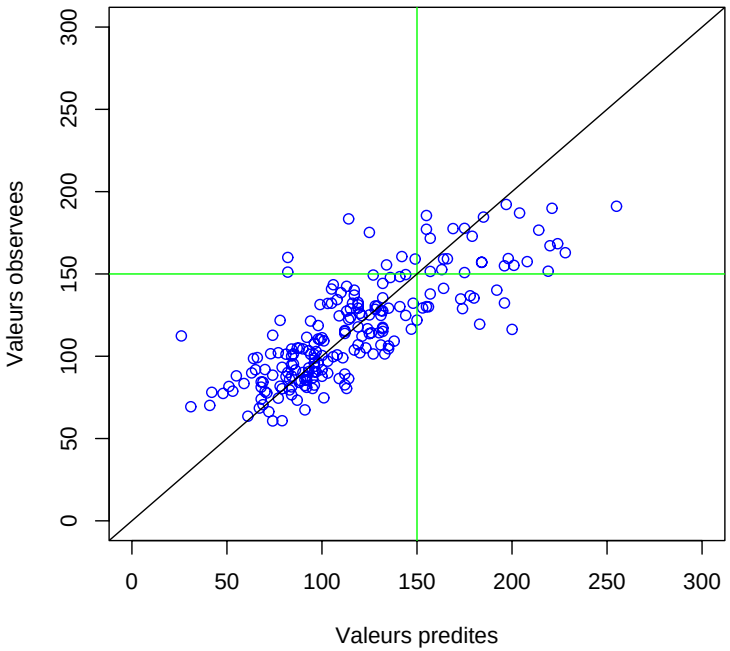


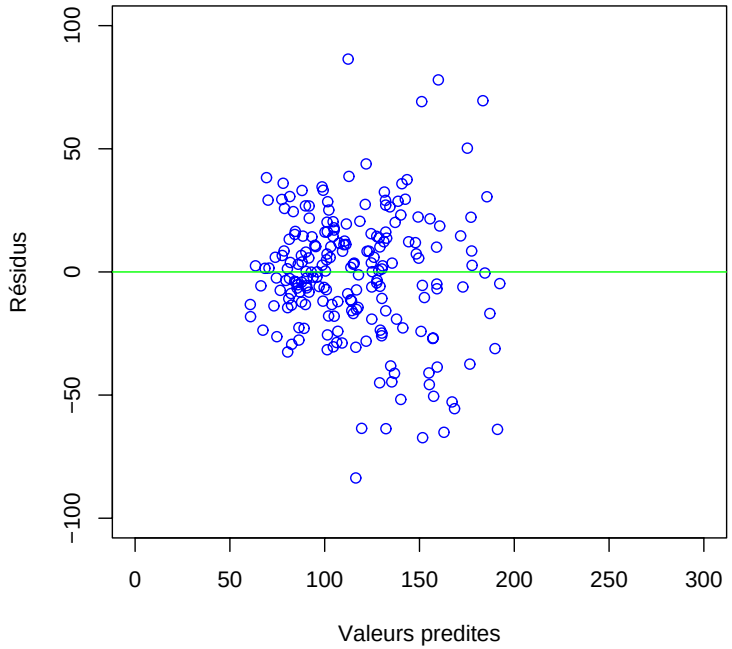
Cookies : optimisation des SVM avec noyau linéaire

Exemple de discrimination

	Cancer du sein	
	benign	malignant
benign	83	1
malignant	3	50
	Taux de 3%	

	Dépassement du seuil d'ozone	
	FALSE	TRUE
FALSE	161	13
TRUE	7	27
	Taux de 9,6%(régression)	
	et 12% (discrimination)	





Ozone : Valeurs observées et résidus du test en fonction des valeurs prédites