

• Idée 1: séparation linéaire avec une **marge maximale**

- marge **fonctionnelle**:

$$\gamma_i = f(\mathbf{x}_i)y_i = (\mathbf{w}'\mathbf{x}_i + w_0)y_i$$

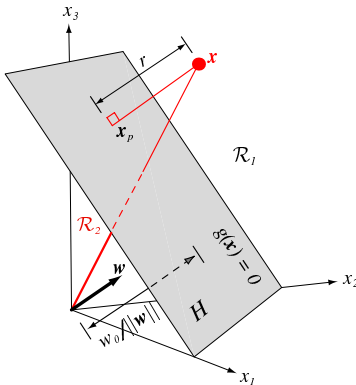
- marge **géométrique**:

$$\gamma_i^{(g)} = \frac{1}{\|\mathbf{w}\|}(\mathbf{w}'\mathbf{x}_i + w_0)y_i = \frac{1}{\|\mathbf{w}\|}\gamma_i$$

• Optimisations équivalentes

- maximiser la marge **géométrique**
- maximiser la marge **fonctionnelle** sous la **contrainte $\|\mathbf{w}\| = 1$**
- minimiser $\|\mathbf{w}\|$ sous la **contrainte $\gamma_i \geq 1$**

• Géométrie – **deux classes**



• Problème **dual** – théorème de **Kuhn-Tucker**

- minimiser par rapport à \mathbf{w} et w_0 et maximiser par rapport à α :

$$\begin{aligned} L(\mathbf{w}, w_0, \alpha) &= \frac{1}{2}\mathbf{w}'\mathbf{w} - \sum_{i=1}^n \alpha_i[\gamma_i - 1] \\ &= \frac{1}{2}\mathbf{w}'\mathbf{w} - \sum_{i=1}^n \alpha_i[(\mathbf{w}'\mathbf{x}_i + w_0)y_i - 1] \end{aligned}$$

- sous les contraintes $\alpha_i \geq 0, i = 1, \dots, n$

• Géométrie – **deux classes**

- $r =$ **distance algébrique** de \mathbf{x} et H :

$$\begin{aligned} \mathbf{x} &= \mathbf{x}_p + r \frac{\mathbf{w}}{\|\mathbf{w}\|} \\ g(\mathbf{x}) &= \mathbf{w}'\mathbf{x} + w_0 = r\|\mathbf{w}\| \\ r &= \frac{g(\mathbf{x})}{\|\mathbf{w}\|} \end{aligned}$$

• Problème d'**optimisation**:

- soit $D_n = ((x_1, y_1), \dots, (x_n, y_n))$ **linéairement séparable**
- minimiser $\|\mathbf{w}\|^2 = \mathbf{w}'\mathbf{w}$
- sous les contraintes $\gamma_i = (\mathbf{w}'\mathbf{x}_i + w_0)y_i \geq 1, i = 1, \dots, n$

• Résultat:

- l'hyperplan $(\mathbf{w}'\mathbf{x} + w_0 = 0)$ avec une **marge géométrique $\frac{1}{\|\mathbf{w}\|}$ maximale**

• Optimisation:

- les gradients:

$$\begin{aligned} \frac{\partial L(\mathbf{w}, w_0, \alpha)}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^n \alpha_i \mathbf{x}_i y_i = \mathbf{0} \\ \frac{\partial L(\mathbf{w}, w_0, \alpha)}{\partial w_0} &= \sum_{i=1}^n \alpha_i y_i = 0 \end{aligned}$$

- resubstitution: maximiser par rapport à α :

$$\begin{aligned} W(\alpha) &= \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i' \mathbf{x}_j - \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i' \mathbf{x}_j + \sum_{i=1}^n \alpha_i \\ &= \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i' \mathbf{x}_j \end{aligned}$$

- sous les contraintes $\alpha_i \geq 0, i = 1, \dots, n$ et $\sum_{i=1}^n \alpha_i y_i = 0$

• La solution

- $\mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i$
- $w_0^* = -\frac{1}{2} \left(\max_{y_i=-1} \mathbf{w}^* \mathbf{x}_i + \min_{y_i=1} \mathbf{w}^* \mathbf{x}_i \right)$

• La structure de la solution

- pour $j \in sv$

$$\gamma_j = y_j f^*(\mathbf{x}_j) = y_j \left(\sum_{i \in sv} \alpha_i^* y_i \mathbf{x}_i^T \mathbf{x}_j + w_0^* \right) = 1$$
- alors

$$\begin{aligned} \mathbf{w}^* \mathbf{w}^* &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i^* \alpha_j^* y_i y_j \mathbf{x}_i^T \mathbf{x}_j = \sum_{j \in sv} \alpha_j^* y_j \sum_{i \in sv} \alpha_i^* y_i \mathbf{x}_i^T \mathbf{x}_j = \sum_{j \in sv} \alpha_j^* (1 - y_j w_0^*) \\ &= \sum_{j \in sv} \alpha_j^* \end{aligned}$$
- la marge maximale:

$$\gamma^* = \frac{1}{\|\mathbf{w}^*\|} = \left(\sum_{i \in sv} \alpha_i^* \right)^{-1/2}$$

• Exemples de noyaux

- linéaire: $K^l(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}'$
- polynômial: $K_d^p(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \mathbf{x}' + 1)^d$
- base radiale: $K^r(\mathbf{x}, \mathbf{x}') = K(\|\mathbf{x} - \mathbf{x}'\|)$
 - gaussien: $K_g^r(\mathbf{x}, \mathbf{x}') = e^{-\gamma \|\mathbf{x} - \mathbf{x}'\|^2}$
- réseaux de neurones: $K^g(\mathbf{x}, \mathbf{x}') = \sigma(\mathbf{s} \mathbf{x}^T \mathbf{x}' + c)$
 - sigmoïde: $K_{s,c}^g(\mathbf{x}, \mathbf{x}') = \tanh(\mathbf{s} \mathbf{x}^T \mathbf{x}' + c)$

• La structure de la solution

- \mathbf{w}^* est une combinaison linéaire des points d'entraînement
- théorème de Kuhn-Tucker:

$$\alpha_i^* [(\mathbf{w}^* \mathbf{x}_i + w_0^*) y_i - 1] = 0, \quad i = 1, \dots, n$$
- si $\gamma_i > 1$ alors $\alpha_i^* = 0$
- si $\alpha_i^* > 0$ alors $\gamma_i = 1$: vecteurs de support
- $\mathbf{w}^* = \sum_{i \in sv} \alpha_i^* y_i \mathbf{x}_i$
- $f^*(\mathbf{x}) = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i^T \mathbf{x} + w_0^* = \sum_{i \in sv} \alpha_i^* y_i \mathbf{x}_i^T \mathbf{x} + w_0^*$

• Idée 2: le noyau

- l'optimisation: $W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$
- la fonction optimale: $f^*(\mathbf{x}) = \sum_{i \in sv} \alpha_i^* y_i \mathbf{x}_i^T \mathbf{x} + w_0^*$
- Remplacer $\mathbf{x}^T \mathbf{x}'$ par une fonction de noyaux $K(\mathbf{x}, \mathbf{x}')$
- l'optimisation: $W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$
- la fonction optimale: $f^*(\mathbf{x}) = \sum_{i \in sv} \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}) + w_0^*$
- matrice de Gram: $\mathbf{G}_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$
- équivalent à une transformation non-linéaire dans un espace de traits de dimension très élevée

• Discrimination linéaire généralisée

- linéaire:

$$f^s(\mathbf{x}) = \mathbf{w}^s \mathbf{x} + w_0^s = \sum_{j=1}^d w_j^s x^{(j)} + w_0^s = \sum_{i \in sv} \alpha_i^s y_i \mathbf{x}_i^T \mathbf{x} + w_0^s$$
- généralisée:

$$f^s(\mathbf{x}) = \sum_{j=1}^D w_j^s \phi^{(j)}(\mathbf{x}) + w_0^s = \sum_{i \in sv} \alpha_i^s y_i \sum_{j=1}^D \phi^{(j)}(\mathbf{x}_i) \phi^{(j)}(\mathbf{x}) + w_0^s$$
- produit scalaire dans l'espace de traits:

$$\phi^{(j)}(\mathbf{x}) \phi^{(j)}(\mathbf{x}') = \sum_{l=1}^D \phi^{(l,j)}(\mathbf{x}) \phi^{(l,j)}(\mathbf{x}')$$

- Discrimination linéaire **généralisée**

- exemple:

$$\begin{aligned}\phi^{(j)}(\mathbf{x}) &= \sqrt{2}x^{(j)}, \quad j = 1, \dots, d, \\ \phi^{(i+d)}(\mathbf{x}) &= x^{(i)}x^{(i)}, \quad i, j = 1, \dots, d \\ \phi^{(d^2)}(\mathbf{x}) &= 1\end{aligned}$$

- produit scalaire dans l'espace des traits:

$$\begin{aligned}\phi'(\mathbf{x})\phi(\mathbf{x}') &= \sum_{j=1}^{d^2} \phi^{(j)}(\mathbf{x})\phi^{(j)}(\mathbf{x}') \\ &= \sum_{j=1}^d \sqrt{2}x^{(j)}\sqrt{2}x'^{(j)} + \sum_{i=1}^d \sum_{j=1}^d x^{(i)}x^{(i)}x'^{(i)}x'^{(j)} + 1 \\ &= 2\mathbf{x}'\mathbf{x}' + (\mathbf{x}'\mathbf{x}')^2 + 1 \\ &= (\mathbf{x}'\mathbf{x}' + 1)^2 = K_2^2(\mathbf{x}, \mathbf{x}')\end{aligned}$$

Machines à vecteurs de support

- Théorème de **Mercer**:

- $\phi'(\mathbf{x})\phi(\mathbf{x}') = K(\mathbf{x}, \mathbf{x}')$

- Conditions

- **diagonaliser** la matrice de Gram: $\mathbf{G}_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{V}\mathbf{A}\mathbf{V}$
- valeurs propres: λ_i , vecteurs propres: \mathbf{v}_i
- condition **suffisante**: \mathbf{G} est **positif semi-défini** ($\forall i: \lambda_i > 0$) pour tout $\mathcal{X}_n = (\mathbf{x}_1, \dots, \mathbf{x}_n)$
- D peut être ∞ !!!

Machines à vecteurs de support

- **Problème soluble 1**: minimiser $\|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i^2$

- on peut supprimer la contrainte de positivité des ξ_i
- C est un **hyper-paramètre** réglé par la validation croisée (par exemple)

- Caractérisation des noyaux

- $K(\mathbf{x}, \mathbf{x}') = \phi'(\mathbf{x})\phi(\mathbf{x}') = \sum_{j=1}^D \phi^{(j)}(\mathbf{x})\phi^{(j)}(\mathbf{x}')$

- commutativité: $K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}', \mathbf{x})$

- inégalité de **Cauchy-Schwartz**:

$$\begin{aligned}K(\mathbf{x}, \mathbf{x}')^2 &= (\phi'(\mathbf{x})\phi(\mathbf{x}'))^2 \\ &\leq \|\phi(\mathbf{x})\|^2 \|\phi(\mathbf{x}')\|^2 \\ &= \phi'(\mathbf{x})\phi(\mathbf{x}) \cdot \phi'(\mathbf{x}')\phi(\mathbf{x}') \\ &= K(\mathbf{x}, \mathbf{x})K(\mathbf{x}', \mathbf{x}')\end{aligned}$$

Machines à vecteurs de support

- Idée 3: les **variables d'écart** (slack variables)

- cas **non-séparable**

- Permettre des **erreurs**:

- minimiser $\|\mathbf{w}\|^2 = \mathbf{w}'\mathbf{w}$
- sous les **contraintes** $\gamma_i = (\mathbf{w}'\mathbf{x}_i + w_0)y_i \geq 1 - \xi_i$, $i = 1, \dots, n$
- où $\xi_i \geq 0$, $i = 1, \dots, n$
- minimiser l'erreur \equiv minimiser le nombre de points avec $\xi_i > 0$: **NP-difficile**

Machines à vecteurs de support

- **Problème dual**

- minimiser par rapport à \mathbf{w} , ξ et w_0 et maximiser par rapport à α :

$$L(\mathbf{w}, w_0, \xi, \alpha) = \frac{1}{2}\mathbf{w}'\mathbf{w} + \frac{C}{2} \sum_{i=1}^n \xi_i^2 - \sum_{i=1}^n \alpha_i [(\mathbf{w}'\mathbf{x}_i + w_0)y_i - 1 + \xi_i]$$

- sous les contraintes $\alpha_i \geq 0$, $i = 1, \dots, n$

- les gradients:

$$\begin{aligned}\frac{\partial L(\mathbf{w}, w_0, \alpha)}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^n \alpha_i \mathbf{x}_i y_i = \mathbf{0} \\ \frac{\partial L(\mathbf{w}, w_0, \alpha)}{\partial \xi} &= C\xi - \alpha = \mathbf{0} \\ \frac{\partial L(\mathbf{w}, w_0, \alpha)}{\partial w_0} &= \sum_{i=1}^n \alpha_i y_i = 0\end{aligned}$$

• Problème dual

- resubstitution: maximiser par rapport à α :

$$\begin{aligned} W(\alpha) &= \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{2C} \alpha' \alpha - \frac{1}{C} \alpha' \alpha \\ &= \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{2C} \alpha' \alpha \\ &= \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \left(K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{C} \delta_{i,j} \right) \end{aligned}$$

- sous les contraintes $\alpha_i \geq 0$, $i = 1, \dots, n$ et $\sum_{i=1}^n \alpha_i y_i = 0$

- **Problème soluble 2:** minimiser $\|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i$

• Problème dual

- minimiser par rapport à \mathbf{w} , ξ et w_0 et maximiser par rapport à α :

$$L(\mathbf{w}, w_0, \xi, \alpha, \mathbf{r}) = \frac{1}{2} \mathbf{w}' \mathbf{w} + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i [(\mathbf{w}' \mathbf{x}_i + w_0) y_i - 1 + \xi_i] - \sum_{i=1}^n r_i \xi_i$$

- sous les contraintes $\alpha_i \geq 0, r_i \geq 0$ $i = 1, \dots, n$

• Problème dual

- resubstitution: maximiser par rapport à α :

$$W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

- sous les contraintes $\alpha_i \geq 0, r_i \geq 0, C - \alpha_i - r_i = 0$ $i = 1, \dots, n$ et $\sum_{i=1}^n \alpha_i y_i = 0$
- \equiv sous les contraintes $0 \leq \alpha_i \leq C$ (contraintes de boîte)

• La solution

- $\mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i = \sum_{i \in \text{SV}} \alpha_i^* y_i \mathbf{x}_i$
- $f^*(\mathbf{x}) = \sum_{i=1}^n \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}) + w_0^* = \sum_{i \in \text{SV}} \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}) + w_0^*$
- w_0^* est choisi tel que $\gamma_i = f(\mathbf{x}_i) y_i \geq 1 - \xi_i = 1 - \frac{\alpha_i^*}{C}$
- la marge obtenue: $\gamma = \sum_{i \in \text{SV}} \alpha_i^* - \frac{1}{C} \alpha'^* \alpha^*$

• Problème dual

- les gradients:

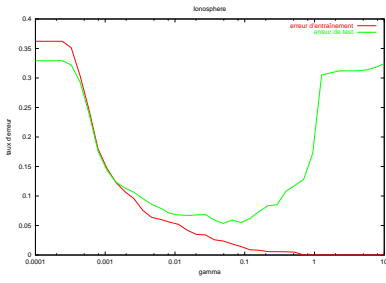
$$\begin{aligned} \frac{\partial L(\mathbf{w}, w_0, \alpha)}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^n \alpha_i \mathbf{x}_i y_i = \mathbf{0} \\ \frac{\partial L(\mathbf{w}, w_0, \alpha)}{\partial \xi} &= C - \alpha_i - r_i = 0 \\ \frac{\partial L(\mathbf{w}, w_0, \alpha)}{\partial w_0} &= \sum_{i=1}^n \alpha_i y_i = 0 \end{aligned}$$

• La solution

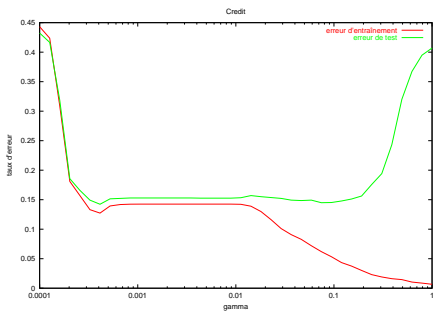
- $\mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i = \sum_{i \in \text{SV}} \alpha_i^* y_i \mathbf{x}_i$
- $f^*(\mathbf{x}) = \sum_{i=1}^n \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}) + w_0^* = \sum_{i \in \text{SV}} \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}) + w_0^*$
- w_0^* est choisi tel que $\gamma_i = f(\mathbf{x}_i) y_i = 1$ pour tous $i : 0 < \alpha_i^* < C$
- la marge obtenue: $\gamma = \sum_{i, j \in \text{SV}} \alpha_i^* \alpha_j^* y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$

• Expériences

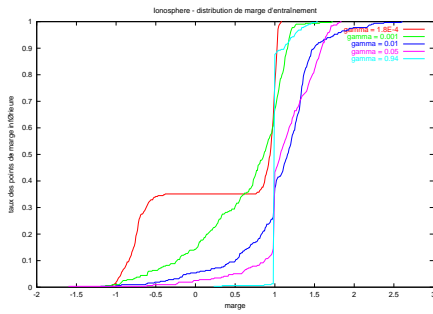
- 4 données de UCI, test croisé en 10 blocs, contrainte de $L_1, C = 1$, noyau gaussien



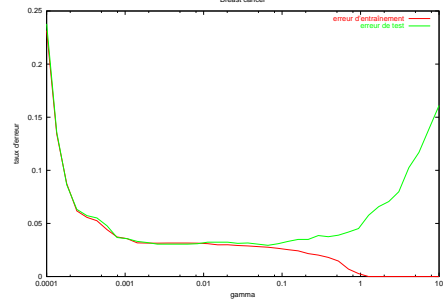
• Expériences



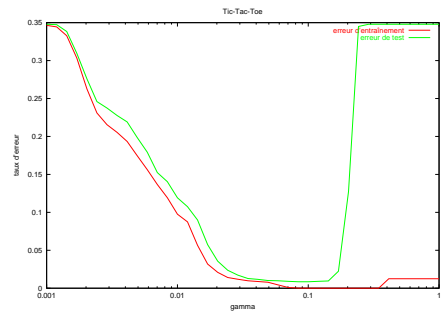
• Distribution de la marge d'entraînement



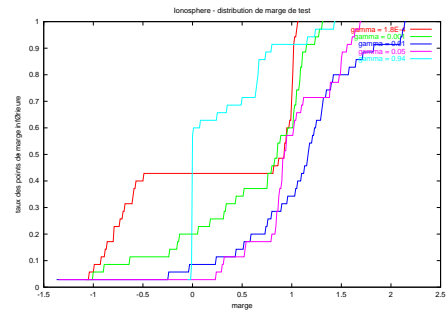
• Expériences



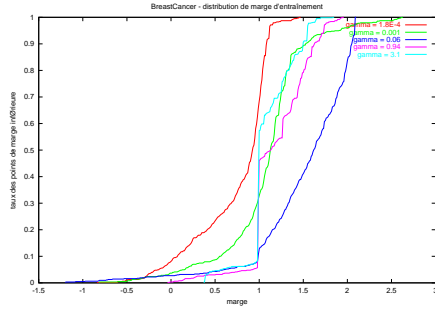
• Expériences



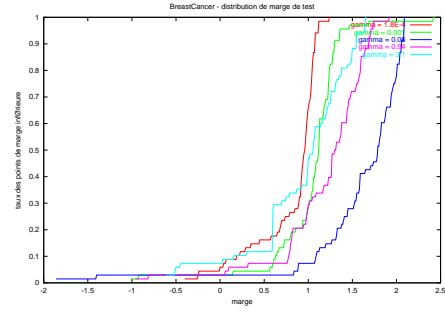
• Distribution de la marge de test



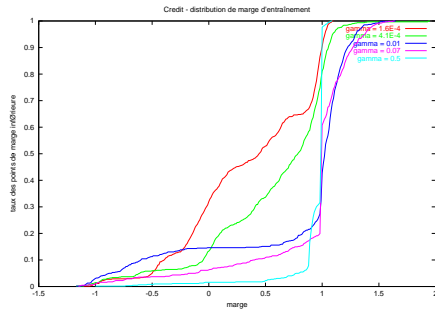
• Distribution de la marge d'entraînement



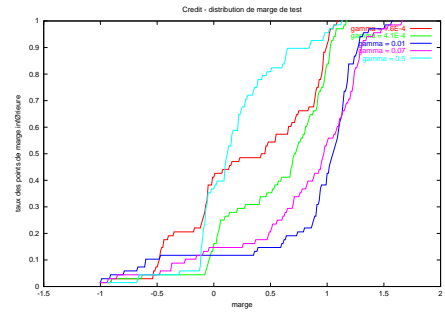
• Distribution de la marge de test



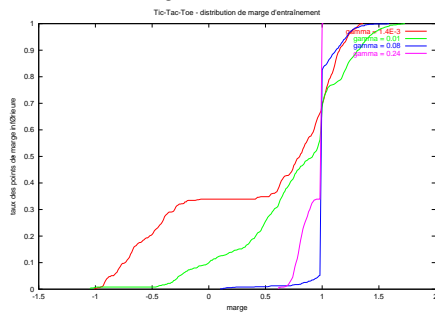
• Distribution de la marge d'entraînement



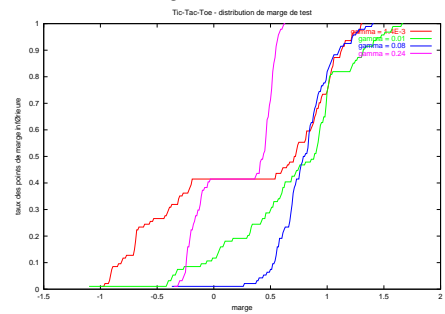
• Distribution de la marge de test



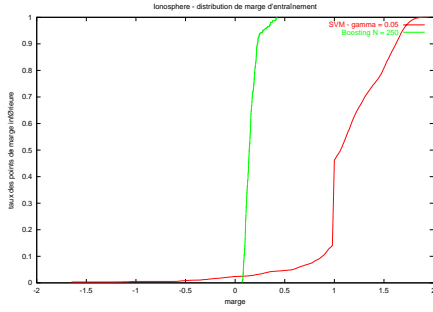
• Distribution de la marge d'entraînement



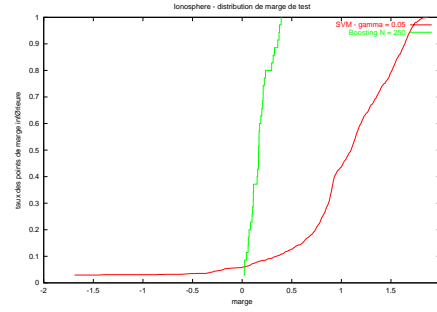
• Distribution de la marge de test



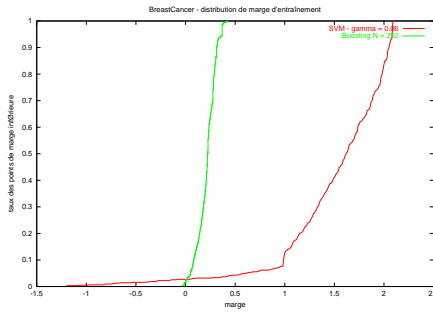
• Comparaison avec boosting (marge d'entraînement)



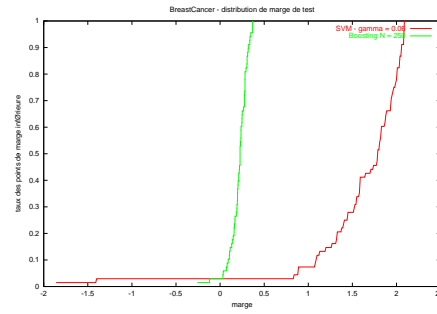
• Comparaison avec boosting (marge de test)



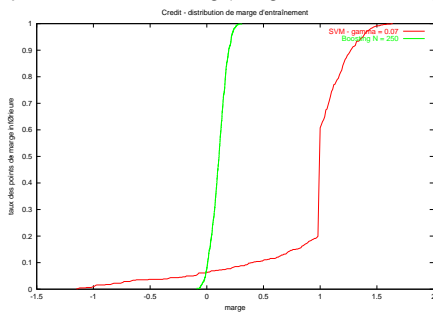
• Comparaison avec boosting (marge d'entraînement)



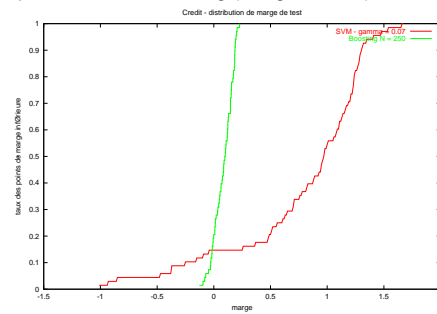
• Comparaison avec boosting (marge de test)



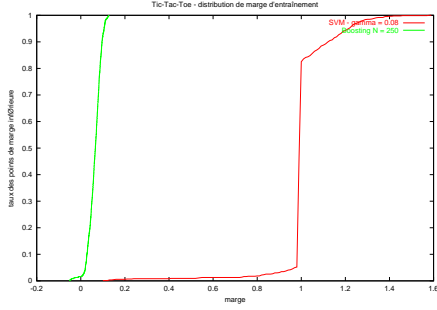
• Comparaison avec boosting (marge d'entraînement)



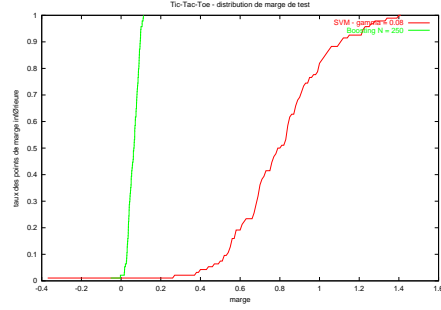
• Comparaison avec boosting (marge de test)



• Comparaison avec boosting (marge d'entraînement)



• Comparaison avec boosting (marge de test)



• Erreur de généralisation

- dimension VC effective:

$$\min\left(\frac{R^2}{\gamma^2}, d\right) + 1$$

- espérance de l'erreur de généralisation:

$$\mathbb{E}[R(f_n)] \leq \frac{1}{n+1} \mathbb{E}\left[\frac{R_{n+1}^2}{\gamma_{n+1}}\right]$$

• Estimation LOO

- erreur empirique:

$$\hat{R}(f, D_n) = \hat{R}(f) = \frac{1}{n} \sum_{i=1}^n I_{\{f(X_i)Y_i < 0\}}$$

- $f^{(i)}$ est entraîné sur

$$D^{(i)} = ((X_1, Y_1), \dots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \dots, (X_n, Y_n))$$

- erreur LOO:

$$\hat{R}_{LOO}(f) = \frac{1}{n} \sum_{i=1}^n I_{\{f^{(i)}(X_i)Y_i < 0\}}$$

• Estimation LOO

- $\hat{R}_{LOO}(f)$ est presque non-biaisé:

$$\mathbb{E}\hat{R}_{LOO}(f_{n+1}) = \mathbb{E}R(f_n)$$

- erreur LOO:

$$\begin{aligned} \hat{R}_{LOO}(f) &= \frac{1}{n} \sum_{i=1}^n I_{\{f^{(i)}(X_i)Y_i < 0\}} = \frac{1}{n} \sum_{i=1}^n I_{\{-f^{(i)}(X_i)Y_i \geq 0\}} \\ &= \frac{1}{n} \sum_{i=1}^n I_{\{-f(X_i)Y_i + (f(X_i) - f^{(i)}(X_i))Y_i \geq 0\}} \\ &\geq \frac{1}{n} \sum_{i=1}^n I_{\{U^{(i)} - 1 \geq 0\}} \end{aligned}$$

- où $U^{(i)} \geq (f(X_i) - f^{(i)}(X_i))Y_i$

• Nombre de vecteurs de support

- $U^{(i)} = 0$ pour les vecteurs de non-support

- $\hat{R}_{LOO}(f) \leq \frac{n_{SV}}{n}$

• Borne de Jaakkola-Haussler

• $U^{(i)} \leq \alpha_i K(X_i, X_i)$ pour les vecteurs de non-support

• $\hat{R}_{LOO}(f) \leq \frac{1}{n} \sum_{i=1}^n I_{\{\alpha_i K(X_i, X_i) \geq 1\}}$

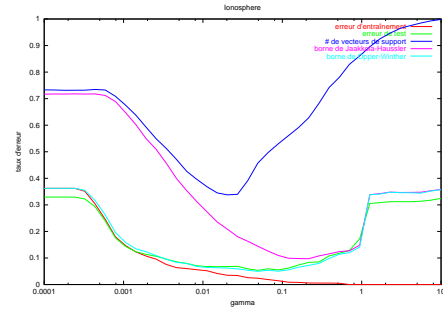
• Borne de Opper-Winther

• matrice de Gram des vecteurs de support: $G_{SVij} = K(X_i, X_j)$

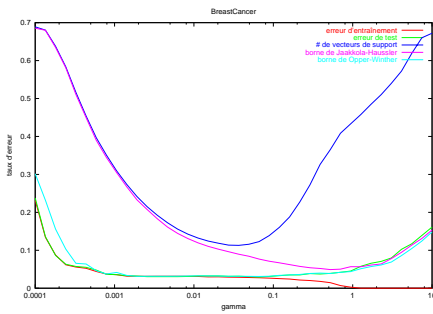
• $U^{(i)} = \frac{\alpha_i}{(G_{SV}^{-1})_{ii}}$

• $\hat{R}_{LOO}(f) \leq \frac{1}{n} \sum_{i=1}^n I_{\left\{ \frac{\alpha_i}{(G_{SV}^{-1})_{ii}} \geq 1 \right\}}$

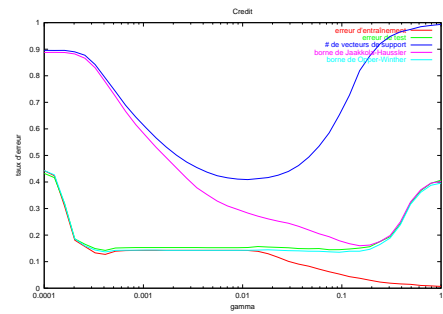
• Expériences



• Expériences



• Expériences



• Expériences

