SY19 - TP04
Séparateurs à Vaste Marge

Alice Ngwembou - Antoine Hars

January 9, 2014

Introduction

Ce TP a pour but d'étudier les Séparateurs à Vaste Marge. Nous verrons pour cela les différentes fonctions noyau (linéaire, polynomial et gaussien), ainsi que l'influence de paramètres, tels que la largeur de bande, sur les résultats obtenus.
Nous nous pencherons ainsi sur le cas de données non séparables linéairement, puis sur le cas de données séparables linéairement.
Ensuite nous travaillerons sur la résolution d'un problème de classification par support vector machines étudié dans le cours et pour finir nous nous attacherons à comprendre comment optimiser les paramètres de coût et de largeur de bande afin d'obtenir un modèle performant.

Exercice 1

Nous étudions, dans un premier temps, les Séparateurs à Vaste Marge dans le cas linéairement séparable.

Nous avons l'ensemble d'apprentissage dans $\mathbb{R}^2$ suivant :
$x_1 = (2, 0)$ de la classe C_1,
$x_2 = (0, 2)$, $x_3 = (-2, 2)$ et $x_4 = (-1, 3)$ de la classe C_2.

Question 1 : Donner l'équation de l'hyperplan à vaste marge. Dessiner la frontière de décision correspondante. Vous indiquez quelles exemples sont des vecteurs support.

De manière intuitive on devine que la frontière de décision serait $x = y$ pour l'ensemble d'apprentissage donné, avec les vecteurs de support x_1 et x_2.

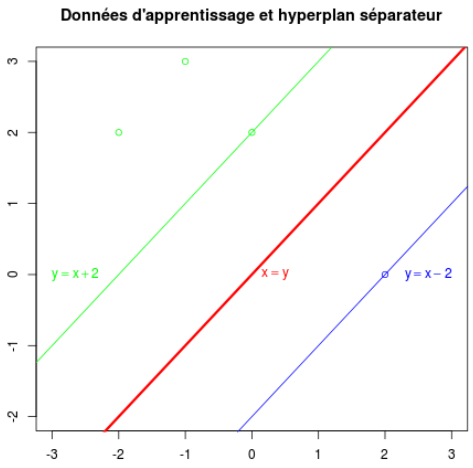


Figure 1: Visualisation de la frontière de décision (en rouge) et de la marge (en vert et en bleu)

Question 2 : Calculer la marge optimale correspondante.

La marge optimale ρ correspond à la distance la plus courte entre un vecteur et l'hyperplan séparateur. Les vecteurs (0,2) et (2,0) sont à égale distance de l'hyperplan mais sont aussi les vecteurs les plus proches en distance de l'hyperplan.
Il s'agit donc des vecteurs de support, qui sont situés au bord de la marge.

La marge optimale ρ équivaut par conséquent à la distance entre $x_1 = (2, 0)$ et son point projeté sur l'hyperplan qui a pour coordonnées $(1, 1)$.
Donc $\rho = \sqrt{2}$.

Question 3 : Déterminer la région de R2 pour laquelle une nouvelle observation venant dans l'ensemble d'apprentissage, et appartenant à C_1, est sans effet sur la solution (vous donnez l'équation et vous montrez sur la figure). Idem pour C_2.

Une nouvelle observation sera sans effet sur la solution si elle se situe sur ou au dehors de la marge, c'est à dire si elle vérifie :

$$y >= x + 2 \text{ pour } C_2$$
$$y <= x - 2 \text{ pour } C_1$$

En effet dans le cas contraire, la distance la plus courte d'un vecteur à l'hyperplan serait plus petite. Par conséquent la marge ρ serait incorrecte et devrait être modifiée.

Question 4 : Donner (en quelques lignes seulement) le code R qui permet d'aboutir à la solution que vous nommez "model".

```
# Matrice des points
obs = matrix(c(2,0,-2,-1,0,2,2,3), ncol=2)

# Vecteur des classes
classes = c(1,-1,-1,-1)
classes = t(classes)

# Methode SVM
model = svm(obs,as.factor(classes), scale = FALSE, type = NULL, kernel = "linear")
```

Question 5 : Vous expliquez comment on peut confirmer les résultats intuitifs de la première question à partir du model obtenu (vous montrez les instructions qui permettent de donner les multiplicateurs de Lagrange et leurs valeurs numériques).

Dans R on utilise la commande `model$SV` pour obtenir les vecteurs de support.

```
# Affichage des vecteurs supports
> model$SV
  X1 X2
1  2  0
2  0
```

On remarque que les vecteurs de support trouvés de manière intuitive et ceux retournés par R sont les mêmes, ce qui semble confirmer notre intuition.

Concernant les multiplicateurs de Lagrange, on utilise la commande `model$coefs` qui retourne le produit des coefficients de Lagrange par les facteurs de classe (-1 ou 1).
On obtient le résultat suivant :

```
# Affichage des coefficients de lagrange*traininglevels
> model$coefs
     [,1]
[1,]  0.25
[2,] -0.25
```

Comme la valeur 1 est associée au vecteur $(2, 0)$ et -1 au vecteur $(0, 2)$ alors les valeurs de coefficients de Lagrange sont 0,25 et -0,25.

Exercice 2

Dans ce deuxième exercice nous étudions la méthode d'optimisation dans le cas non linéairement séparable.

Question 1 :

ξ est porté au carré dans la fonction coût, son signe n'influe donc pas sur l'optimisation de $\frac{1}{2}\sum_{i=1}^{n}\xi_i^2$

Question 2 :

On ne prend pas en compte les dernières contraintes $\sum_{i=1}^{n}\mu_i\xi_i$, ce qui donne :

$$L(\mathbf{w}, w_0, \xi, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \tfrac{1}{2}|\mathbf{w}|^2 + \tfrac{c}{2}\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\alpha_i[y_i(\mathbf{w'x_i} + w_0) - (1-\xi_i)]$$

Question 3 :

De l'expression précédente, on obtient les dérivées partielles $\frac{\partial L}{\partial \mathbf{w}}$, $\frac{\partial L}{\partial w_0}$ et $\frac{\partial L}{\partial \xi_i}$ suivantes :

$$\begin{gathered}\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n}\alpha_i y_i \mathbf{x_i} \\ \frac{\partial L}{\partial w_0} = -\sum_{i=1}^{n}\alpha_i y_i \\ \frac{\partial L}{\partial \xi_i} = c\xi_i - \alpha_i\end{gathered}$$

Au point selle, ces dérivées sont nulles, par conséquent on obtient :

$$\begin{gathered}\frac{\partial L}{\partial \mathbf{w}} = 0 \Longleftrightarrow \mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x_i} \\ \frac{\partial L}{\partial w_0} = 0 \Longleftrightarrow \sum_{i=1}^{n}\alpha_i y_i = 0 \\ \frac{\partial L}{\partial \xi_i} = 0 \Longleftrightarrow \xi_i = \frac{\alpha_i}{c}\end{gathered}$$

Question 4 :

Pour obtenir le Lagrangien dual $W(\boldsymbol{\alpha})$ on développe le Lagrangien de la **question 2** :

$$L(\mathbf{w}, w_0, \xi, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \tfrac{1}{2}|\mathbf{w}|^2 + \tfrac{c}{2}\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\alpha_i[y_i(\mathbf{w'x_i} + w_0) - (1-\xi_i)]$$

$$L(\mathbf{w}, w_0, \xi, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \tfrac{1}{2}|\mathbf{w}|^2 + \tfrac{c}{2}\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\alpha_i y_i \mathbf{w'x_i} - \sum_{i=1}^{n}\alpha_i y_i w_0 + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i\xi_i$$

Or on a $\xi_i = \frac{\alpha_i}{c}$, d'où :

$$W(\boldsymbol{\alpha}) = \tfrac{1}{2}|\mathbf{w}|^2 + \tfrac{c}{2}\sum_{i=1}^{n}(\tfrac{\alpha_i}{c})^2 - \sum_{i=1}^{n}\alpha_i y_i \mathbf{w'x_i} - \sum_{i=1}^{n}\alpha_i y_i w_0 + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\frac{\alpha_i^2}{c}$$

$$W(\boldsymbol{\alpha}) = \tfrac{1}{2}|\mathbf{w}|^2 + \tfrac{1}{2}\sum_{i=1}^{n}\frac{\alpha_i^2}{c} - \sum_{i=1}^{n}\alpha_i y_i \mathbf{w'x_i} - \sum_{i=1}^{n}\alpha_i y_i w_0 + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\frac{\alpha_i^2}{c}$$

$$W(\boldsymbol{\alpha}) = \tfrac{1}{2}|\mathbf{w}|^2 - \sum_{i=1}^{n}\alpha_i y_i \mathbf{w'x_i} - \sum_{i=1}^{n}\alpha_i y_i w_0 + \sum_{i=1}^{n}\alpha_i - \tfrac{1}{2}\sum_{i=1}^{n}\frac{\alpha_i^2}{c}$$

Or on a $\sum_{i=1}^{n}\alpha_i y_i = 0 \Longleftrightarrow \sum_{i=1}^{n}\alpha_i y_i w_0 = 0$, d'où :

$$W(\boldsymbol{\alpha}) = \tfrac{1}{2}|\mathbf{w}|^2 - \sum_{i=1}^{n}\alpha_i y_i \mathbf{w'x_i} + \sum_{i=1}^{n}\alpha_i - \tfrac{1}{2}\sum_{i=1}^{n}\frac{\alpha_i^2}{c}$$

Or on a $\mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x_i}$, d'où :

$$W(\boldsymbol{\alpha}) = \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x_i x_j} - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x_i x_j} + \sum_{i=1}^{n}\alpha_i - \tfrac{1}{2}\sum_{i=1}^{n}\frac{\alpha_i^2}{c}$$

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\alpha_i - \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x_i x_j} - \tfrac{1}{2}\sum_{i=1}^{n}\frac{\alpha_i^2}{c}$$

On trouve donc bien un $W(\boldsymbol{\alpha})$ de la forme

$$\begin{gathered}W(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\alpha_i - \tfrac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j[y_i y_j(\mathbf{x_i}.\mathbf{x_j}) + \tfrac{\delta_{ij}}{c}] \\ \text{où } \delta_{ij} = 1 \text{ si } i = j \\ \text{et } \delta_{ij} = 0 \text{ si } i \neq j\end{gathered}$$

qui est un problème d'optimisation quadratique.

Question 5 :

Le problème dual consiste à maximiser $W(\boldsymbol{\alpha})$ par rapport à $\boldsymbol{\alpha}$ sous les contraintes suivantes

$$\begin{array}{rcll} \alpha_i & \geq & 0, & i = 1, .., n \\ \xi_i & = & \frac{\alpha_i}{c}, & i = 1, .., n \\ \sum_{i=1}^{n} \alpha_i y_i & = & 0 & \end{array}$$

Soit $\boldsymbol{\alpha^*}$ la solution de notre problème. Les conditions de Kuhn et Tucker s'écrivent :

$$\begin{array}{rcll} \alpha_i^*[y_i(\mathbf{w'}^*\mathbf{x_i} + w_0^*) - (1 - \xi_i^*)] & = & 0, & i = 1, .., n \\ \xi_i & = & \frac{\alpha_i}{c}, & i = 1, .., n \end{array}$$

Étant donné que les vecteurs de support ont un $\alpha_i^* > 0$ alors :

$$\begin{array}{rcl} \alpha_i^* & > & 0 \\ \text{et} \quad c & > & 0 \\ \iff \frac{\alpha_i^*}{c} & > & 0 \\ \iff \xi_i^* & > & 0 \end{array}$$

Par conséquent les vecteurs de support vérifient $\xi_i^* > 0$.

Question 6 :

On a le Lagrangien dual :

$$\begin{array}{c} W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j [y_i y_j (\mathbf{x_i}.\mathbf{x_j}) + \frac{\delta_{ij}}{c}] \\ \text{où } \delta_{ij} = 1 \text{ si } i = j \\ \text{et } \delta_{ij} = 0 \text{ si } i \neq j \\ \iff W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j [(\mathbf{x_i}.\mathbf{x_j}) + \frac{\delta_{ij}}{c y_i y_j}] \end{array}$$

Or on a $K(\mathbf{x_i}, \mathbf{x_j}) = [(\mathbf{x_i}.\mathbf{x_j}) + \frac{\delta_{ij}}{c y_i y_j}]$ tel que :

$$\begin{array}{llll} \text{si} \quad i = j, & \delta_{ij} = 0 & \text{donc} & K(\mathbf{x_i}, \mathbf{x_j}) = \mathbf{x_i}.\mathbf{x_j} \\ \text{si} \quad i \neq j, & \delta_{ij} = 1 & \text{donc} & K(\mathbf{x_i}, \mathbf{x_j}) = \mathbf{x_i}.\mathbf{x_j} + \frac{1}{c} \\ & \text{et } y_i y_j = 1 & & \end{array}$$

Exercice 3

Question 1 :

Le Lagrangien à maximiser par rapport à $\boldsymbol{\alpha}$ est le suivant :

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x_i}, \mathbf{x_j})$$

avec $K(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i}.\mathbf{x_j} + 1)^2$ les contraintes

$$0 \leq \alpha_i \leq c = 200, i = 1, .., 5$$
$$\sum_{i=1}^{5} \alpha_i y_i = 0$$

Question 2 :

On cherche à déduire la valeur α correcte parmi les différents choix proposés :

(a) Vu que $\alpha_i \geq 0$ pour $i = 1, ..., n$, cette proposition n'est pas la bonne solution car nous avons un $\alpha_2 = -1$.

(b) Nous avons vu que $\sum_{i=1}^{n} \alpha_i y_i = 0$ mais avec cette proposition, cette somme vaut 1. Donc la proposition (b) ne convient pas.

(c) Cette proposition est la solution que nous validons car nous avons des $\sum_{i=1}^{n} \alpha_i y_i = 0$ et $0 < \alpha_i < \gamma$.

(d) Cette proposition ne convient pas malgré le fait qu'elle respecte l'égalité $\sum_{i=1}^{n} \alpha_i y_i = 0$ car dans ce cas de figure, nous avons $\alpha_1 > \gamma$ et $\alpha_4 > \gamma$. α doit respecter : $0 < \alpha_i < \gamma$.

Question 3

Afin de vérifier notre hypothèse, nous exécutons le code suivant :

```
# Matrice des points
obs = matrix(c(1,2,6,4,5))

# Classes
classes = c(1,1,1,-1,-1)
classes = t(classes)

# Methode SVM
model = svm(obs, as.factor(classes), degree = 2, scale = FALSE, kernel = "polynomial",
coef0 = 1, cost = 100, gamma = 1)

> model$SV
     [,1]
[1,]    2
[2,]    6
[3,]    5

> model$coefs
          [,1]
[1,]  2.499180
[2,]  4.831745
[3,] -7.330924
```

La fonction `svm` est paramétrée avec la fonction noyau polynomiale :

$$K(\mathbf{x_i}, \mathbf{x_j}) = (\texttt{gamma} * \mathbf{x_i}' * \mathbf{x_i} + \texttt{coef0})^{\texttt{degree}}$$

avec `gamma = 1`, `coef0 = 1` and `degree = 2`.

Les résultats renvoyés par `model$SV` correspondent bien à la proposition déduite dans la question précédente, à savoir que x_2, x_3 et x_5 sont des vecteurs propres.

`model$coefs` nous renvoie les valeurs des coefficients α_i des vecteurs propres multipliés par les facteurs de classe (-1 ou 1). Comme x_2 et x_3 ont pour facteur 1 et x_5 a pour facteur -1, on retrouve bien les valeurs de la réponse (c) de la question précédente pour les coefficients de Lagrange des vecteurs propres.

Question 4

Nous avons :

$$\begin{array}{rrcl} & \mathbf{K}(\mathbf{x_i}, \mathbf{x_j}) & = & \mathbf{\Phi}(\mathbf{x_i})\mathbf{\Phi}(\mathbf{x_j}) \quad = \quad (\mathbf{x_i}.\mathbf{x_j} + 1)^2 \\ \text{et} & \mathbf{w} & = & \sum_{i=1}^{5} \alpha_i y_i \mathbf{\Phi}(\mathbf{x_i}) \end{array}$$

La fonction discriminante $g(\mathbf{x}) = \mathbf{w}^*.\Phi(\mathbf{x}) + w_0^*$ peut s'écrire sous la forme :

$$\begin{array}{rrcl} & g(\mathbf{x}) & = & \sum_{i \in S} \alpha_i y_i \mathbf{\Phi}(\mathbf{x_i}).\mathbf{\Phi}(\mathbf{x}) + w_0^*, \quad \text{où } S = \{i \in \{1,..,5\}, \alpha_i^* > 0\} \\ \iff & g(\mathbf{x}) & = & \sum_{i \in S} \alpha_i y_i \mathbf{K}(\mathbf{x_i}, \mathbf{x_j}) + w_0^* \end{array}$$

Alors on a finalement :

$$\begin{array}{rrcl} \iff & g(\mathbf{x}) & = & \sum_{i \in S} \alpha_i y_i (\mathbf{x_i}.\mathbf{x} + 1)^2 + w_0^* \end{array}$$

Question 5

On code les fonctions $\mathbf{K}(\mathbf{x_i}, \mathbf{x_j})$ et $g(\mathbf{x})$ de la manière suivante :

```
# Fonction noyau polynomiale K
K <- function(xi, x) {
       return ((xi * x + 1)^2)
}

# Fonction d'apprentissage g(x)
g <- function(model,obs,x) {

       n = length(x)  # Abscisses
       r = c(1:n)     # Creation de l'array des ordonnees

       for (i in 1:n) {
              # Calcul de g(x) pour chaque abscisse x[i]
              r[i] <- model$coefs[1] * K(obs[model$index[1]], x[i])
              + model$coefs[2] * K(obs[model$index[2]], x[i])
              + model$coefs[3] * K(obs[model$index[3]], x[i])
              - model$rho
       }

       # Retourne le vecteur des ordonnees g(x) calculees
       return (r)
}

# Abscisses x
> x = seq( from = 0, to = 6, by = 0.05)

# Resultats g(x)
> res = g(model, obs, x)
```

On utilise un intervalle de 0 à 7, avec un pas de 0.05 pour le tracé de notre fonction $g(\mathbf{x})$, ce qui nous permet d'afficher la courbe suivante :

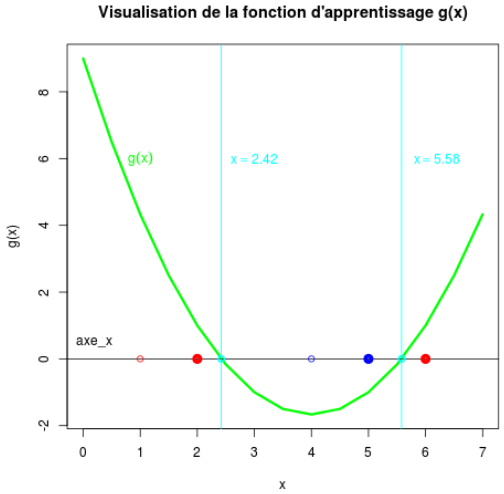


Figure 2: Visualisation de l'axe des $\mathbf{x}$ avec $\mathbf{x} \in \mathbb{R}$ (en noir), de la fonction discriminante $g(\mathbf{x})$ (en vert), des vecteurs de support x_2, x_3 et x_5 (en gras), des autres vecteurs d'apprentissage (de taille normale) et des classes C_1 (points rouges) et C_2 (points bleus).

Nos données d'apprentissage sont dans l'ensemble à une dimension de $\mathbb{R}$.
Par conséquent, on peut les représenter sur la droite des x, chaque vecteur étant un point.

Étant donné que l'équation de notre hyperplan correspond à :

$$g(\mathbf{x}) = \sum_{i \in S} \alpha_i y_i (\mathbf{x_i}.\mathbf{x} + 1)^2 + w_0^* = 0,$$

les points qui vérifient $g(\mathbf{x}) = 0$ représentent notre hyperplan dans $\mathbb{R}$.

Grâce à la fonction `uniroot` on trouve les deux racines de la fonction $g(\mathbf{x})$ à environ $\mathbf{x} = 2.42$ et $\mathbf{x} = 5.58$ (voir les 2 points en couleur cyan sur la *Figure 2*).

La fonction de décision $G(\mathbf{x}) = sgn(g(\mathbf{x}))$ classe les points en C_1 si $g(\mathbf{x}) > 0$ et en C_2 si $g(\mathbf{x}) < 0$. Ceci est bien représenté sur notre figure, en effet les vecteurs $x_1 = 1$, $x_2 = 2$ et $x_3 = 6$ (en rouge) vérifient bien $g(\mathbf{x}) > 0$ et sont donc classés en C_1 par notre hyperplan, et $x_4 = 4$, $x_5 = 5$ (en bleu foncé) vérifient bien $g(\mathbf{x}) < 0$ et sont donc classés en C_2 par notre hyperplan.
Le vecteur x_2, x_3 et x_5 sont les plus proches de l'hyperplan, de plus x_2 est à la même distance du point $\mathbf{x} = 2.42$ que les vecteurs x_3 et x_5 sont au point $\mathbf{x} = 5.58$.
Cette distance peut être identifiée comme la marge de l'hyperplan, et notre graphique vérifie donc bien que ces vecteurs sont les vecteurs support de l'hyperplan.

Exercice 4 :

Le but de cet exercice est de mettre en oeuvre la méthode des séparateurs à vaste marge en testant l'influence du choix des paramètres.

Question 1 :

Les fonctions noyau sont de la forme suivante :

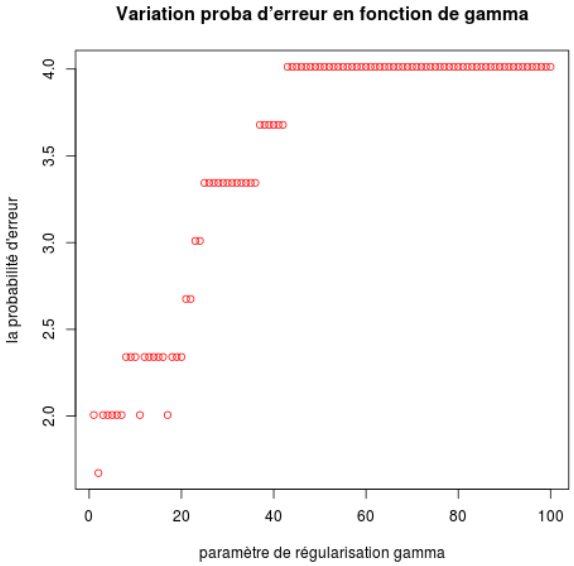
	vu en cours	fonction svm (avec $\gamma = \frac{1}{dimension}$)
noyau polynomial	$\kappa(x,y) = (x'y+1)^r, r > 0$	$\kappa(x,y) = (\gamma x'y + coef0)^r$, $r = 3$ et $coef0 = 0$ par défaut
noyau gaussien	$\kappa(x,y) = exp(-\frac{\|\|x-y\|\|^2}{2\sigma^2})$	$\kappa(x,y) = exp(-\gamma\|x-y\|^2)$

Pour la fonction noyau polynomial, nous pouvons remarquer que $coef0$ dans le cours est mis à 1 alors que par défaut dans la fonction svm, il est à 0. γ est mis à 1 dans la fonction du cours alors qu'il est par défaut à $\frac{1}{dimension}$ pour la fonction svm.

Pour la fonction noyau gaussien, en comparant le cours avec l'expression de la fonction dans svm, nous remarquons que γ correspond à l'expression $\frac{1}{2\sigma^2}$.

Question 2:

Pour cette question, nous travaillons avec le noyau gaussien.



Sur ce graphique, nous pouvons remarquer que la probabilité d'erreur a tendance à augmenter lorsque le paramètre de pénalisation γ augmente.

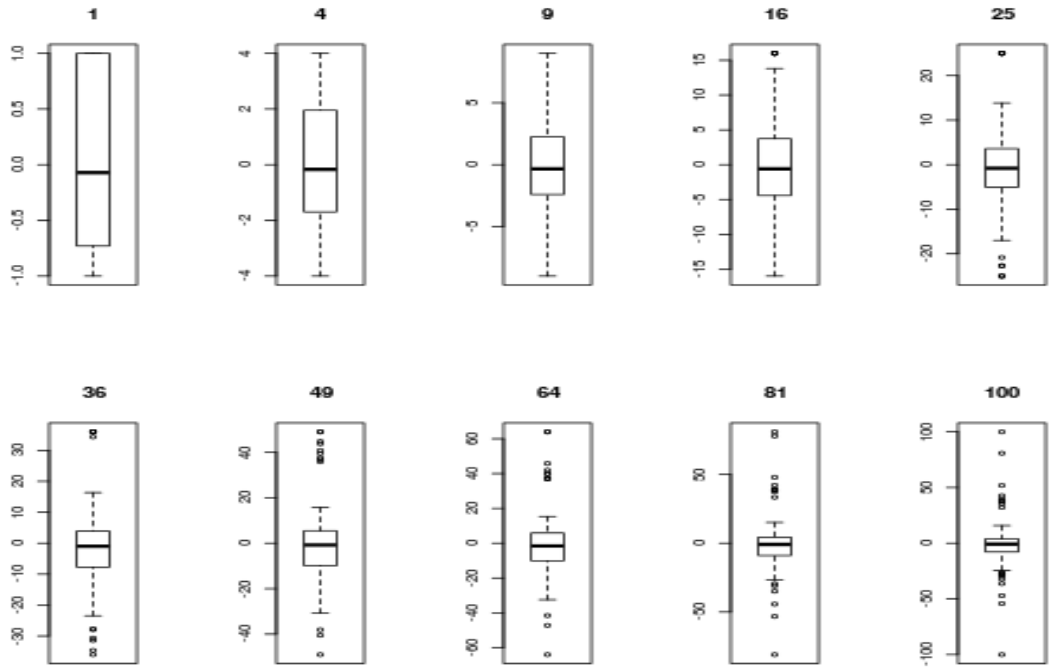


Figure 3: Boxplot sur les composantes du vecteur alpha en fonction du paramètre de pénalisation gamma

Vu que nous avons $0 < \alpha < \gamma$, ce qui nous permet de dire que plus le coût augmente, moins les α sont limités en valeur.
Nos graphiques sont cohérents en conclusion car plus α est grand, plus le vecteur de support a de l'importance ou alors ce sera un vecteur aberrant (Le modèle va donc essayer de s'y adapter si le coût est grand, ce qui nous donnera un α élevé.

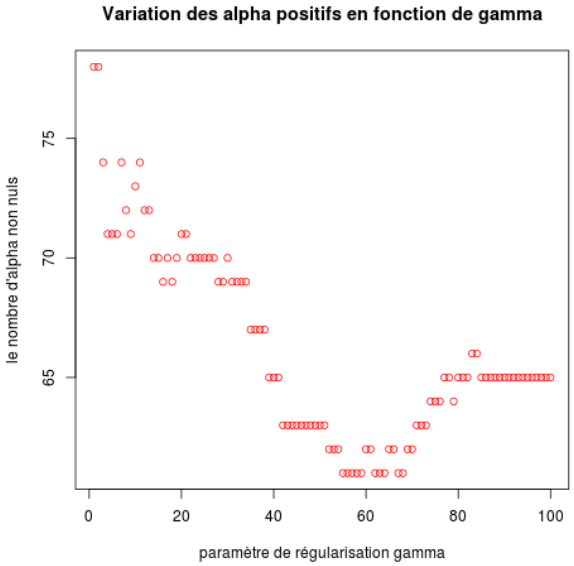
En observant les différents *boxplots*, nous avons remarqué que l'amplitude du vecteur α augmente en fonction de γ. Par exemple, nous avons pour un $\gamma = 2$, une amplitude comprise dans l'intervalle $[-2; 2]$ et pour un $\gamma = 100$, un α compris dans l'intervalle $[-100; 100]$.
Donc nous avons α compris dans l'intervalle $[-\gamma; \gamma]$.

Les valeurs d'α peuvent donc atteindre de très grandes valeurs pour un coût élevé, cependant la majorité des alphas restent dans des valeurs "centrales" ou nulles (dans ces cas de figure, le vecteur qui est lié à α n'est pas un support vector). Le nombre de valeurs extrêmes grandissant, à cause d'un coût augmentant, entraîne une augmentation de la précision de l'hyperplan, voire une trop forte adaptation de celui-ci aux données d'apprentissage quand le coût atteint ses valeurs maximum, ce qui nous donne un modèle très complexe.

De plus, nous remarquons, à travers ces graphiques, que pour un paramètre de pénalisation faible, les différentes valeurs de αs sont réparties de manière uniforme dans l'intervalle $[-1; 1]$. Alors que plus le paramètre de pénalisation augmente, plus les valeurs d'α semblent avoir tendance à se centrer sur 0 tout en ayant des valeurs aberrantes lorsqu'on dépasse l'intervalle $[-5; 5]$ en observant les *boxplots*.
Nous pouvons expliquer ce phénomère par la précision grandissante de l'hyperplan.

L'endroit où le modèle généralise le plus correspond au moment où il y a le moins d'α positifs, car un α est positif lorsque le vecteur d'apprentissage est un vecteur de support (les vecteurs de support influencent l'hyperplan).

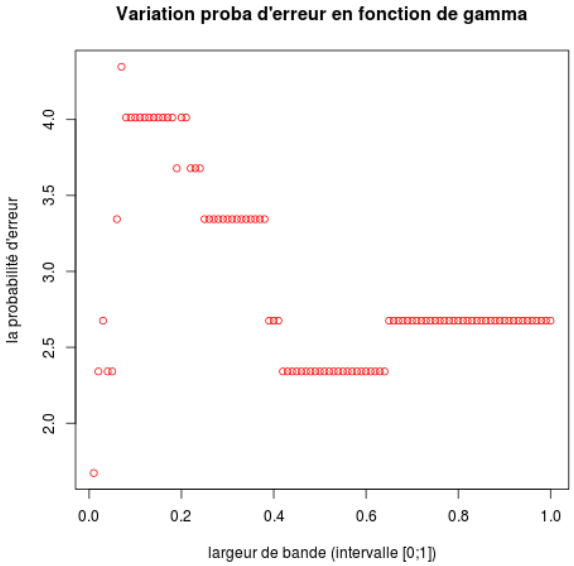
Le nombre de vecteurs de support du modèle est important car un nombre trop important implique un modèle très précis qui se généralisera mal, tandis qu'un nombre trop faible implique un modèle peu précis qui sera moins, voire peu efficace.



En observant ce graphique, nous pouvons dire que la variation des α positifs ne suit pas les variations de γ. Nous obtenons le minimum des α positifs, c'est à dire $\alpha = 61$ pour des valeurs de γ proches de 60, puis les α se stabilisent à 65 pour des γ supérieurs à 76.
Nous pouvons observer que pour de très faibles valeurs de γ, la valeur des α est très grande (elle est même à son maximum). Ces valeurs de vecteurs de supports α diminuent progressivement pour des valeurs de γ comprises entre 1 et 60 environ.

Question 3:

Nous travaillons dans cette question avec la fonction noyau gaussien.
Vu que pour des valeurs de γ proches de 100, nous obtenons des valeurs de vecteurs de support α stabilisées, nous considérons, dans notre cas, que $\gamma = \infty$ peut être assimilé à $\gamma = 100$.



Nous pouvons observer sur ce graphique l'évolution des performances de la méthode des Séparateurs à Vaste Marge suivant différentes valeurs de la largeur de bande du noyau gaussien.

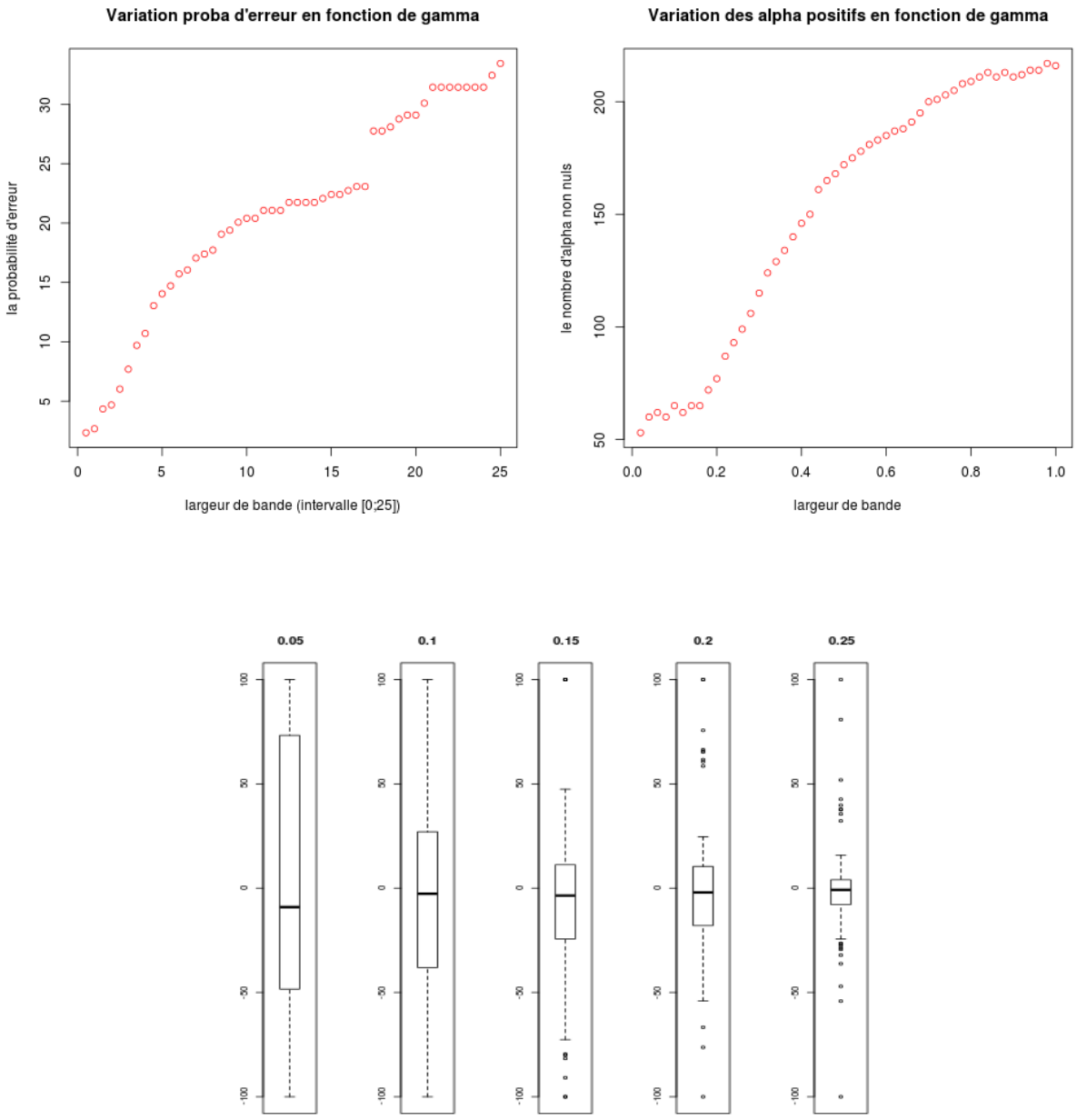


Figure 4: Boxplot sur les composantes du vecteur alpha en fonction du paramètre de pénalisation gamma

Le paramètre γ correspond à l'étendue de l'influence de chaque point. Plus γ est faible, plus l'influence est forte; plus γ est élevé, plus l'influence est faible.

Les valeurs intéressantes se situent pour $\gamma < 1$ car comme nous pouvons l'observer sur le graphique avec une largeur de bande développée sur un intervalle de $[0; 25]$, lorsque $\gamma > 1$, le modèle perd rapidement en efficacité.
Alors que pour un $\gamma < 1$, la courbe correspondant au pourcentage d'erreur n'est pas linéaire.

À l'inverse de ce que l'on a pu observer pour le paramètre de régularisation dans la **question 2**, le nombre d'α_i non nuls varie ici presque linéairement en fonction de la largeur de bande.

Pour un γ très petit, il y a peu de vecteurs de support et les valeurs des α_i sont distribuées uniformément. Cela signifie que la frontière est lisse et qu'elle n'est pas très précise par rapport aux données d'apprentissage.
Lorsque le γ augmente, le nombre de vecteurs de support augmente linéairement : l'hyperplan est donc plus précis et adapté à l'ensemble d'apprentissage.

Question 4:

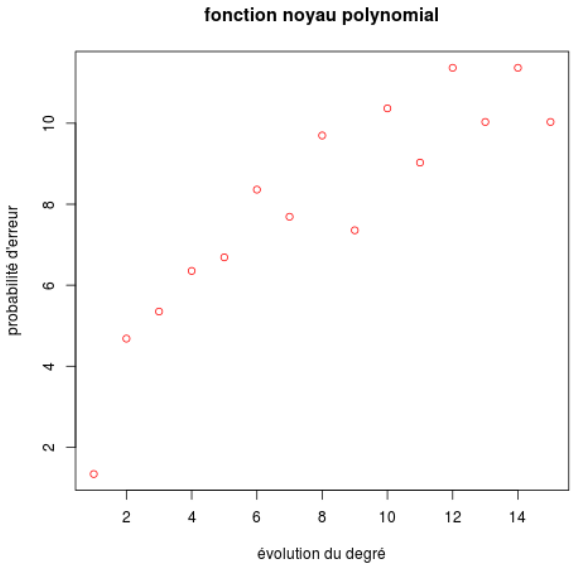


Figure 5: Les probabilités d'erreur sur la fonction noyau polynomiale suivant différentes valeurs de degré.

Le graphique obtenu montre le pourcentage d'erreur obtenu avec les paramètres optimaux pour le kernel polynomial.
Pour ce kernel polynomial, nous avons fait varier le *degré* de 1 à 10, et le meilleur résultat, observé sur le graphique, est obtenu pour un degré égal à 1.
Il semble que le kernel polynomial avec un $degré = 1$ est efficace.

L'autre méthode très efficace est le kernel gaussien étudié tout au long de ce chapitre, avec un coût égal à 2 (obtenu grâce à la fonction tune()).

Conclusion

Pour conclure, tout au long de ce tp, nous avons pu observer l'efficacité des Séparateurs à Vaste Marge dans le cas linéaire et dans le cas non linéaire.
Nous avons pu ensuite juger de l'importance des paramètres dans l'utilisation des noyaux polynomial et gaussien pour obtenir un modèle efficace pour résoudre des problème de discrimination en 2 classes.