UNIVERSITÉ DE SHERBROOKE

TP3

IFT 712 Technique d'apprentissage

Eliott THOMAS — 21 164 874
thoe2303@usherbrooke.ca

Lilian FAVRE GARCIA — 21 153 421
favl2301@usherbrooke.ca

Tsiory Razafindramisa — 21 145 627
raza3902@usherbrooke.ca

Travail présenté à
Martin Vallières

Université de Sherbrooke
Département d'informatique
Date de remise : 24 mars 2022

Table des matières

1 Introduction

On s'intéresse dans ce troisième TP à des démonstrations ainsi qu'à la mise en place d'un programme informatique permettant une classification à partir de 4 algorithmes par technique de nayau. Nous utiliserons les noyaux RBFs, Polynomiaux, Linéaires et Sigmoidaux.

2 Question 1

On a comme première équation primale (6.2) :

$$
\begin{aligned}
J(\vec{w}) &= \tfrac{1}{2}\sum_{n=1}^{N}\left(t_n - w^T\phi(x_n)\right)^2 + \frac{\lambda}{2}w^T w \\
&= \sum_{n=1}^{N}\frac{1}{2}\left(w^T\phi(x_n) - t_n\right)^2 + \frac{1}{2}\lambda w^T w \\
&\Leftrightarrow 0 = \nabla_{\vec{w}} E_D(w) \\
&\Leftrightarrow 0 = \sum \phi(x_n) * \left(\phi^T(x_n)w - t_n^T\right) + \lambda w \\
&\Leftrightarrow 0 = \left(\sum \phi(x_n)\phi^T(x_n)\right)w - \left(\sum \phi(x_n)t_n^T\right) + \lambda w \\
&\Leftrightarrow 0 = \sum \left(w^T\phi(x_n) - t_n^T\right) * \phi(x_n) + \lambda w \\
&\Leftrightarrow -\lambda w = \sum \left(w^T\phi(x_n) - t_n^T\right) * \phi(x_n) \\
&\Leftrightarrow w = -\tfrac{1}{\lambda}\sum \left(w^T\phi(x_n) - t_n^T\right) * \phi(x_n) \\
&\Leftrightarrow w = \sum (a_n) * \phi(x_n) \\
&\Leftrightarrow w = \Phi^T a \qquad (A)
\end{aligned}
$$

On a donc comme notation simplifiée : $a_n = -\frac{1}{\lambda}\left[w^T\phi(x_n) - t_n^T\right]$

On remplace dans l'équation 6.2 w par $\Phi^T a$ en utilisant (A) :
On note donc J(a) car J dépend maintenant de a et non plus de w.

$$
\begin{aligned}
J(a) &= \frac{1}{2}\sum_{n=1}^{N}\left(t_n - (\Phi^T a)^T\phi(x_n)\right)^2 + \frac{\lambda}{2}(\Phi^T a)^T\Phi^T a \\
J(a) &= \frac{1}{2}\sum_{n=1}^{N}\left(t_n - (a^T\Phi)\phi(x_n)\right)^2 + \frac{\lambda}{2}a^T\Phi\Phi^T a \\
J(a) &= \frac{1}{2}\sum_{n=1}^{N}(t_n)^2 - \sum_{n=1}^{N}\left(t_n(a^T\Phi)\phi(x_n)\right) + \frac{1}{2}\sum_{n=1}^{N}\left(a^T\Phi\phi(x_n)\right)^2 + \frac{\lambda}{2}a^T\Phi\Phi^T a \\
J(a) &= \frac{1}{2}t^T t - (a^T\Phi)(t\Phi) + \frac{1}{2}\left[(a^T\Phi\Phi^T)(a^T\Phi\Phi^T)^T\right] + \frac{\lambda}{2}a^T\Phi\Phi^T a \\
J(a) &= \frac{1}{2}t^T t - (a^T\Phi\Phi^T t) + \frac{1}{2}\left[(a^T\Phi\Phi^T)\left((\Phi\Phi^T)^T a\right)\right] + \frac{\lambda}{2}a^T\Phi\Phi^T a \\
J(a) &= \frac{1}{2}t^T t - (a^T\Phi\Phi^T t) + \frac{1}{2}\left[(a^T\Phi\Phi^T)(\Phi^{T^T}\Phi^T a)\right] + \frac{\lambda}{2}a^T\Phi\Phi^T a
\end{aligned}
$$

$$J(a) = \frac{1}{2}t^T t - (a^T \Phi\Phi^T t) + \frac{1}{2}\left[(a^T \Phi\Phi^T)(\Phi\Phi^T a)\right] + \frac{\lambda}{2}a^T \Phi\Phi^T a$$

On se retrouve donc avec l'équation (6.5) :

$$J(a) = \frac{1}{2}a^T \Phi\Phi^T \Phi\Phi^T a - a^T \Phi\Phi^T t + \frac{1}{2}t^T t + \frac{\lambda}{2}a^T \Phi\Phi^T a$$

On pose K, la marice de Graam telle que $K = \Phi\Phi^T$
En remplaçant dans l'équation (6.5) on obientient l'équation (6.7) :

$$J(a) = \frac{1}{2}a^T KKa - a^T KKt + \frac{1}{2}t^T t + \frac{\lambda}{2}a^T Ka$$

En dérivant cette équation par rapport à a, en en fixant le gradient à 0, et on considérant que K et λ sont des scalaires car ils ne dé

$$\nabla J(a) = \frac{K^2}{2}2a - Kt + 0 + \frac{\lambda K}{2}2a$$

$$\Leftrightarrow 0 = Ka - t + \lambda a$$

$$\Leftrightarrow t = (K + \lambda I)a$$

$$\Leftrightarrow a = (K + \lambda I)^{-1}t$$

On repart de l'équation de la régression linéaire $y = w^T\phi(x)$
On a $w = \Phi^T a$
Donc $y = (\Phi^T a)^T\phi(x) = a^T\Phi\phi(x)$
On pose $k_n(x) = k(x_n,x)$ donc $\Phi\phi(x) = k(x)$
$y = a^T k(x) = k(x)^T a = k(x)^T(K + \lambda I)^{-1}t$

3 Question 2

Un vecteur de support est un sous-ensemble de données d'une classe qui permet d'établir la marge. La marge permet de maximiser la distance entre l'hyperplan est la donnée la plus proche de l'hyperplan. C'est un vecteur parallèle à l'hyperplan séparant deux groupes de données linéairement séparables. Si les données ne sont pas linéairement séparables, alors la marge n'est pas parallèle mais elle "suis" la forme globale. On appliquera une fonction de base pour tenter au mieux de séparer ces données dans un espace de plus grande dimension. Les données qui se retrouvent sur la marge sont nommées "vecteurs de support".

Pour effectuer une prédiction, $y_w(x)$ on procède de la façon suivante :

Comme mentionné auparavant on commence par appliquer une fonction de base ϕ.

Puis on pose pour chaque donnée n : $a_n > 0$ si la donnée est sur la marge, 0 sinon. On a donc uniquement les vecteurs de support qui vont pouvoir voter dans le calcul de y_w.

Puis on a $y_w(\phi(x)) = w^T\phi(x) + w_0$

$$= (\sum_{n=1}^{N} a_n t_n \phi(x_n))^T \phi(x) + w_0$$
$$= \sum_{n=1}^{N} (a_n t_n \phi(x_n)^T \phi(x)) + w_0$$
$$= \sum_{n=1}^{N} (a_n t_n k(x_n,x)) + w_0$$

Avec a_n le coefficient des vecteurs de support et $k(x_n,x)$ le noyau. D'après l'équation (7.18) de Bishop, on calcul w_0 de la façon suivante : On pose N_S le nombre de vecteurs de supports et S l'ensemble de ces vecteurs.

$$w_0 = \frac{1}{N_S} \sum_{n \in S} \left(t_n - \sum_{m \in S} a_m t_m k(x_n,x_m) \right)$$

Le lien entre l'erreur de Hinge et la machine à vecteurs de support s'explique comme suit :

$$\arg \min_{w,w_0,\xi} \{\frac{1}{2}\|w\|^2\} + C \sum_{n=1}^{N} \xi_n \text{ tel que } t_n y_w(\phi(x_n)) \geq 1 - \xi_n \forall n, \xi_n \geq 0$$

Avec ξ_n les variables de ressort qui permettent des violations des contraintes de marge. C'est à dire qu'on peut placer des points entre la séparation et la marge ($\xi_n < 1$) voire même de l'autre côté de la séparation ($\xi_n > 1$).

Si $\xi_n > 1$ alors la donnée est mal classée. C est un hyperparamètre pondérant l'acceptation des données mal classées. Plus C est grand, moins on accepte de données mal classées.

$$\Leftrightarrow \arg \min_{w,w_0,\xi} \{\frac{1}{2}\|w\|^2\} + C \sum_{n=1}^{N} \xi_n \text{ tel que } \xi_n \geq 1 - t_n y_w(\phi(x_n)) \forall n, \xi_n \geq 0$$

$$\Leftrightarrow \arg \min_{w,w_0} \frac{1}{2}\|w\|^2 + C \sum_{n=1}^{N} max(0,1 - t_n y_w(\phi(x_n)))$$

Cette forme est similaire à celle obtenue par descente de gradient, sur la séparation linéaire :

$\arg \min_{w,w_0} \sum_{n=1}^{N} max(0,1 - t_n y_w(\phi(x_n))) + \lambda\|w\|^2$ avec $\lambda = \frac{1}{2C}$ le terme de régularisation.

On reconnaît la fonction de perte (erreur de Hinge) dans le premier terme de la somme.

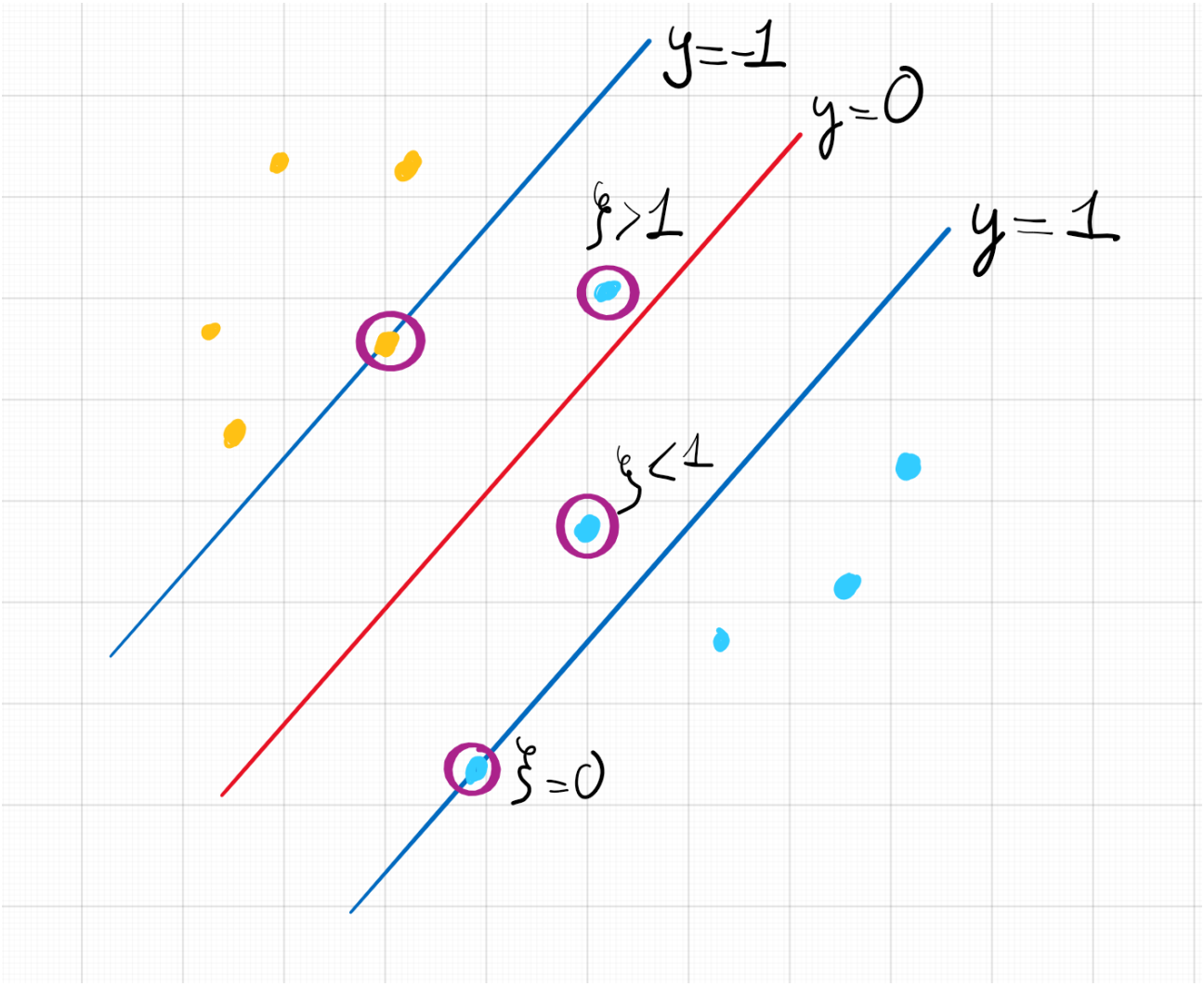

FIGURE 1 – Exemple de SVM avec des variables de ressort
En violet les vecteurs de support

4 Question 3

Voir code remis. Le lien Github est ici