FACULTÉ DES SCIENCES
DÉPARTEMENT D'INFORMATIQUE

IFT712 - Techniques d'apprentissage
TP3

préparé par
Agathe Le Bouler (leba3207)
Philippe Spino (spip2401)
Gabriel Gibeau Sanchez (gibg2501)

présenté à
Pierre-Marc Jodoin

13 mars 2020

1 Question 1

Dans me cas d'une régression, avec l'utilisation d'une optimisation de type Maximisation A Priori, on dispose de la fonction de coût E_D à minimiser telle que :

$$E_D(\overrightarrow{w}) = \frac{\beta}{2}\sum_{n=1}^{N}\left(y(\overrightarrow{x_n}, w) - t_n\right)^2 + \frac{\alpha}{2}\overrightarrow{w}^T\overrightarrow{w}$$

Si l'on écrit : $\lambda = \frac{\alpha}{\beta}$ et $y(\overrightarrow{x_n}, w) = \overrightarrow{w}^T\overrightarrow{\phi}(\overrightarrow{x_n})$

$$\iff \frac{1}{2}\sum_{n=1}^{N}\left(\overrightarrow{w}^T\overrightarrow{\phi}(\overrightarrow{x_n}) - t_n\right)^2 + \frac{\lambda}{2}\overrightarrow{w}^T\overrightarrow{w}$$

Cette expression est la représentation primale que l'on note $J(\overrightarrow{w})$.

$$\begin{aligned} J(\overrightarrow{w}) &= \frac{1}{2}\sum_{n=1}^{N}\left(\overrightarrow{w}^T\overrightarrow{\phi}(\overrightarrow{x_n}) - t_n\right)^2 + \frac{\lambda}{2}\overrightarrow{w}^T\overrightarrow{w} \\ &= \frac{1}{2}\sum_{n=1}^{N}\left(\overrightarrow{w}^T\overrightarrow{\phi}(\overrightarrow{x_n})^2\overrightarrow{w} - 2\overrightarrow{w}^T\overrightarrow{\phi}(\overrightarrow{x_n})t_n + t_n^2\right) + \frac{\lambda}{2}\overrightarrow{w}^T\overrightarrow{w} \end{aligned} \tag{1}$$

Afin de déterminer sa représentation duale dans un second temps, on va déterminer $\overrightarrow{w}$ par le calcul du gradient de E_D par rapport à $\overrightarrow{w}$ et le forcer à 0.

$$\begin{aligned} \nabla_{\overrightarrow{w}} E_D(\overrightarrow{w}) = \frac{\partial}{\partial\overrightarrow{w}}&\left(\frac{1}{2}\sum_{n=1}^{N}\left(\overrightarrow{w}^T\overrightarrow{\phi}(\overrightarrow{x_n})^2\overrightarrow{w} - 2\overrightarrow{w}^T\overrightarrow{\phi}(\overrightarrow{x_n})t_n + t_n^2\right) + \frac{\lambda}{2}\overrightarrow{w}^T\overrightarrow{w}\right) = 0 \\ \iff &\frac{1}{\cancel{2}}\sum_{n=1}^{N}\left(\cancel{2}\overrightarrow{w}^T\overrightarrow{\phi}(\overrightarrow{x_n})^2 - \cancel{2}\overrightarrow{\phi}(\overrightarrow{x_n})t_n\right) + \lambda\overrightarrow{w} = 0 \\ \iff &\sum_{n=1}^{N}\left(\overrightarrow{w}^T\overrightarrow{\phi}(\overrightarrow{x_n})^2 - \overrightarrow{\phi}(\overrightarrow{x_n})t_n\right) + \lambda\overrightarrow{w} = 0 \\ \iff &\sum_{n=1}^{N}\left(\overrightarrow{w}^T\overrightarrow{\phi}(\overrightarrow{x_n})^2 - \overrightarrow{\phi}(\overrightarrow{x_n})t_n\right) = -\lambda\overrightarrow{w} \\ \iff &\overrightarrow{w} = -\frac{1}{\lambda}\sum_{n=1}^{N}\left(\overrightarrow{w}^T\overrightarrow{\phi}(\overrightarrow{x_n}) - t_n\right)\overrightarrow{\phi}(\overrightarrow{x_n}) \\ \iff &\overrightarrow{w} = \sum_{n=1}^{N}\left(-\frac{\overrightarrow{w}^T\overrightarrow{\phi}(\overrightarrow{x_n}) - t_n}{\lambda}\overrightarrow{\phi}(\overrightarrow{x_n})\right) \end{aligned} \tag{2}$$

On peut simplifier cette expression par :

$$\iff \overrightarrow{w} = \sum_{n=1}^{N} a_n \overrightarrow{\phi}(\overrightarrow{x_n}) \tag{3}$$

en considérant a_n tel que : $a_n = -\frac{\overrightarrow{w}^T \overrightarrow{\phi}(\overrightarrow{x_n}) - t_n}{\lambda}$
Aussi, on peut exprimer $\sum_{n=1}^{N} \overrightarrow{\phi}(\overrightarrow{x_n}) = \Phi^T$, matrice de design où la n^{ieme} ligne est donnée par $\overrightarrow{\phi}(\overrightarrow{x_n})^T$. Et a tel que $\overrightarrow{a} = (a_1, ..., a_N)^T$. On écrit alors : $\overrightarrow{w} = \Phi^T \overrightarrow{a}$.
Maintenant, revenons à notre représentation primale (sous forme développée) de $J(\overrightarrow{w})$ et exprimons $\overrightarrow{w}$ tel que $\overrightarrow{w} = \Phi^T \overrightarrow{a}$. J se devient être exprimé en fonction de $\overrightarrow{a}$. On inscrit : $\Phi^T = \sum_{n=1}^{N} \overrightarrow{\phi}(\overrightarrow{x_n})$ et $\overrightarrow{t} = \sum_{n=1}^{N} t_n$.

$$J(\overrightarrow{a}) = \frac{1}{2}\left[(\Phi \overrightarrow{a}^T)(\Phi^T \Phi)(\Phi^T \overrightarrow{a}) - 2(\Phi \overrightarrow{a}^T)\Phi^T t + t^T t\right] + \frac{\lambda}{2}(\Phi \overrightarrow{a}^T)(\Phi^T \overrightarrow{a}) \tag{4}$$

que l'on peut réécrire tel que :

$$\iff \frac{1}{2}\overrightarrow{a}^T \Phi\Phi^T\Phi\Phi^T \overrightarrow{a} - \overrightarrow{a}^T \Phi\Phi^T t + \frac{1}{2}t^T t + \lambda 2 \overrightarrow{a}^T \Phi\Phi^T \overrightarrow{a} \tag{5}$$

qui est la représentation duale de $J(\overrightarrow{w})$.
On définit la matrice de Gram K telle qu'une matrice symétrique $N \times N$ où chaque élément est une fonction noyau telle que : $K_{nm} = \overrightarrow{\phi}(\overrightarrow{x_n})^T \overrightarrow{\phi}(\overrightarrow{x_m}) = k(\overrightarrow{x_n}, \overrightarrow{x_m})$.
On exprime donc K tel que : $K = \Phi\Phi^T$
On peut alors réécrire la représentation duale de $J(\overrightarrow{w})$ telle que :

$$J(\overrightarrow{a}) = \frac{1}{2}\overrightarrow{a}^T K K \overrightarrow{a} - \overrightarrow{a}^T K t + \frac{1}{2}t^T t + \frac{\lambda}{2}\overrightarrow{a}^T K \overrightarrow{a} \tag{6}$$

On détermine le gradient de $J(\overrightarrow{a})$ par rapport à $\overrightarrow{a}$ et on le force à 0 pour déterminer le meilleur vecteur tel que les données sont les mieux classifiées.

$$
\begin{aligned}
\nabla_{\vec{a}} J(\vec{a}) &= 0 \\
\iff \frac{\partial}{\partial \vec{a}} \left(\frac{1}{2} \vec{a}^T K K \vec{a} - \vec{a}^T K \vec{t} + \frac{1}{2} \vec{t}^T \vec{t} + \frac{\lambda}{2} \vec{a}^T K \vec{a} \right) &= 0 \\
\iff \vec{a} K K - K \vec{t} + \lambda \vec{a} K &= 0 \\
\iff \vec{a}(KK + \lambda K) &= K \vec{t} \\
\iff \vec{a} &= \frac{K \vec{t}}{KK + \lambda K} \\
\iff \vec{a} &= \frac{\cancel{K} \vec{t}}{\cancel{K}(K + \lambda I_N)} \\
\iff \vec{a} &= (K + \lambda I_N)^{-1} \vec{t} \qquad (7)
\end{aligned}
$$

Reprenons maintenant notre modèle de régression linéaire tel que : $y(\vec{x}) = \vec{w}^T \vec{\phi}(\vec{x})$, où nous pouvons exprimer $\vec{w}^T$ tel que : $\vec{w}^T = \Phi \vec{a}^T$.

$$
y(\vec{x}) = \vec{w}^T \vec{\phi}(\vec{x}) = \Phi \vec{a}^T \vec{\phi}(\vec{x}) \qquad (8)
$$

Nous pouvons remplacer $\vec{a}^T$ par l'expresion trouvée via le calcul du gradient.

$$
\iff y(\vec{x}) = \Phi \left((K + \lambda I_N)^{-1} \vec{t} \right)^T \vec{\phi}(\vec{x}) \qquad (9)
$$

On sait que : $\Phi = \sum_{n=1}^{N} \vec{\phi}(\vec{x_n})$

$$
\iff y(\vec{x}) = \left((K + \lambda I_N)^{-1} \vec{t} \right)^T \sum_{n=1}^{N} \vec{\phi}(\vec{x_n})^T \vec{\phi}(\vec{x}) \qquad (10)
$$

Or on sait que la fonction noyau est définie telle que : $k(\vec{x_n}, \vec{x_m}) = \vec{\phi}(\vec{x_n})^T \vec{\phi}(\vec{x})$

$$
\iff y(\vec{x}) = \left((K + \lambda I_N)^{-1} \vec{t} \right)^T k(\vec{x_n}, \vec{x}) \qquad (11)
$$

On définit maintenant $k(\vec{x})$ comme la forme vectorisée des éléments de fonctions noyaux $k_n(\vec{x}) = k(\vec{x_n}, \vec{x})$

$$
\begin{aligned}
\iff y(\vec{x}) &= \left((K + \lambda I_N)^{-1} \vec{t} \right)^T k(\vec{x}) \\
\iff y(\vec{x}) &= k(\vec{x})^T (K + \lambda I_N)^{-1} \vec{t} \qquad (12)
\end{aligned}
$$

2 Question 2

Un vecteur de support est un exemple de l'ensemble des données d'entraînement qui est situé sur la marge d'un modèle Machine à Vecteur de Support. On définit la marge comme la plus petite distance entre la surface de séparation du modèle entre les classes et l'ensemble des données d'entraînement.
Pour un modèle, on exprime ses prédictions telles que : $y_{\overrightarrow{w}}(\overrightarrow{\phi}(\overrightarrow{x_n})) = \overrightarrow{w}^T\overrightarrow{\phi}(\overrightarrow{x_n})$, où $\overrightarrow{\phi}(\overrightarrow{x_n})$ peut se révéler être un vecteur de support.
On se place dans un cas où les données sont non linéairement séparables et on envisage une utilisation d'une représentation primal. Aussi, on accepte qu'un nombre restreint de données soient en-deçà de la marge afin d'éviter au maximum le sur-apprentissage. On dispose alors des contraintes telles que :

$$\begin{cases} t_n\left(\overrightarrow{w}^T\overrightarrow{\phi}(\overrightarrow{x_n}) + w_0\right) = 1 \\ C\sum_{n=1}^{N}\xi_n = 1 \end{cases} \tag{13}$$

ξ_n représente la variable de ressort permettant d'accepter des données mal classées pour notre modèle et C est un hyper-paramètre.
On a alors :

$$argmin_{\overrightarrow{w},w_0}\left(\frac{1}{2}\left\|\overrightarrow{w}\right\|^2 + C\sum_{n=1}^{N}\xi_n\right) \tag{14}$$

tel que : $t_n\left(y_{\overrightarrow{w}}(\overrightarrow{\phi}(\overrightarrow{x_n}))\right) \geqslant 1-\xi_n$ et $\forall n, \xi_n \geqslant 0$
La contrainte $t_n\left(y_{\overrightarrow{w}}(\overrightarrow{\phi}(\overrightarrow{x_n}))\right) \geqslant 1-\xi_n$ peut être réécrite telle que :

$$\xi_n \geqslant 1 - t_n\left(y_{\overrightarrow{w}}(\overrightarrow{\phi}(\overrightarrow{x_n}))\right)$$

Si on remplace ξ_n dans l'expression de la représentation primale (équation 7.21 dans Bishop)

$$C\sum_{n=1}^{N}\xi_n + \frac{1}{2}||w||^2 \tag{15}$$

on a alors :

$$\iff argmin_{\overrightarrow{w},w_0}\left(\frac{1}{2}\left\|\overrightarrow{w}\right\|^2\right) + C\sum_{n=1}^{N}1 - t_n\left(y_{\overrightarrow{w}}(\overrightarrow{\phi}(\overrightarrow{x_n}))\right) \tag{16}$$

considérant la contrainte $1 - t_n \left(y_{\overrightarrow{w}}(\overrightarrow{\phi}(\overrightarrow{x_n})) \right) \geqslant 0$

$$\begin{aligned} &\iff argmin_{\overrightarrow{w},w_0} \left(\frac{1}{2} \left\| \overrightarrow{w} \right\|^2 \right) + C \sum_{n=1}^{N} max \left(0, 1 - t_n \left(y_{\overrightarrow{w}}(\overrightarrow{\phi}(\overrightarrow{x_n})) \right) \right) \\ &\iff argmin_{\overrightarrow{w},w_0} \sum_{n=1}^{N} max \left(0, 1 - t_n \left(y_{\overrightarrow{w}}(\overrightarrow{\phi}(\overrightarrow{x_n})) \right) \right) + \lambda \left\| \overrightarrow{w} \right\|^2 \end{aligned} \tag{17}$$

avec $\lambda = \frac{C}{2}$

On reconnaît ici l'expression de la fonction d'erreur de Hinge.