

UNIVERSITÉ DE SHERBROOKE

IFT712 - TP3 Théorique

Authors

ALEXANDRE THEISSE — THEA1804
LOUIS-VINCENT CAPELLI — CAPL1101
TOM SARTORI — SART0701

November 17, 2023

1 Question 1

Montrons que l'équation de la régression linéaire à posteriori est la suivante à l'aide d'une représentation duale :

$$J(w) = \frac{1}{2} \sum_{i=1}^N (w^T \Phi(x_n) - t_n)^2 + \frac{\lambda}{2} w^T w$$

est donnée par l'équation suivante :

$$y(x) = K(x)^T (k + \lambda I)^{-1} t = a^T \Phi(x)$$

Soit la fonction $J(w)$ telle que son gradient soit égale à 0 :

$$\frac{\partial J(w)}{\partial w} = \sum_{i=1}^N (w^T \Phi(x_n) - t_n) \Phi(x_n) + \lambda w = 0$$

$$\Leftrightarrow \sum_{i=1}^N (w^T \Phi(x_n) - t_n) \Phi(x_n) = -\lambda w$$

$$\Leftrightarrow w = \frac{-1}{\lambda} \sum_{i=1}^N (w^T \Phi(x_n) \Phi(x_n) - t_n) \Phi(x_n) = - \sum_{i=1}^N \frac{(w^T \Phi(x_n) - t_n)}{\lambda} \Phi(x_n)$$

Ainsi on pose que le coefficient a_n est le suivant :

$$\forall n, a_n = \frac{(w^T \Phi(x_n) - t_n)}{\lambda} \forall n$$

On a alors pour w que :

$$w = \sum_{i=1}^N a_n \Phi(x_n) \\ \Rightarrow w = \Phi^T a$$

avec $\Phi = [\Phi(x_1), \dots, \Phi(x_n)]$ et $a = [a_1, \dots, a_n]$ On a donc $J(w)$, en utilisant la formule trouver pour w , qui est égale à :

$$J(w) = \frac{1}{2} \sum_{i=1}^N (a^T \Phi \Phi(x_n) - t_n)^2 + \frac{\lambda}{2} a^T \Phi \Phi^T a$$

$$\Leftrightarrow \frac{1}{2} \sum_{i=1}^N ((a^T \Phi \Phi(x_n))^2 - 2t_n a^T \Phi \Phi(x_n) + t_n^2) + \frac{\lambda}{2} a^T \Phi \Phi^T a$$

$$\Leftrightarrow \frac{1}{2} \sum_{i=1}^N a^T \Phi \Phi(x_n) a^T \Phi \Phi(x_n) - \sum_{i=1}^N t_n a^T \Phi \Phi(x_n) + \frac{1}{2} \sum_{i=1}^N t_n t_n + \frac{\lambda}{2} a^T \Phi \Phi^T a$$

$$\Leftrightarrow \frac{1}{2}a^T \Phi \Phi^T \Phi \Phi^T a - ta^T \Phi \Phi^T + \frac{1}{2}t^2 + \frac{\lambda}{2}a^T \Phi \Phi^T a$$

On peut alors poser $K = \Phi \Phi^T$ ce qui revient à :

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

Puis en forçant le gradient de la nouvelle formule $J(a)$ à 0 on obtient :

$$\nabla_a J(a) = 0$$

$$\Leftrightarrow K K a - K t + \lambda a^T K = 0$$

$$\Leftrightarrow K K a + \lambda a^T K = K t$$

$$\Leftrightarrow K a + \lambda a^T = t$$

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

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

Enfin pour la prédiction d'une entrée donnée on applique les fonctions à chaque données avec $x_n \Rightarrow \Phi(x_n) \forall n$ On part alors de $y_w(x) = a^T \Phi \Phi(x)$

$$\Leftrightarrow [(K + \lambda I)^{-1}t]^T \Phi \Phi(x)$$

On transpose l'ensemble de la formule car $y_w(x)^T = y_w(x)$

$$\Leftrightarrow (K + \lambda I)^{-1}t[\Phi \Phi(x)]^T$$

Il ne manque plus qu'à déterminer $[\Phi \Phi(x)]^T$:

$$[\Phi \Phi(x)]^T = \Phi(x)^T \Phi$$

et $\Phi^T = \Phi^T = (\Phi(x_1), \dots, \Phi(x_n))$ Donc on obtient :

$$\Phi(x_1)\Phi(x)^T, \dots, \Phi(x_n)\Phi(x)^T = K(x)^T$$

Soit finalement $y_w(x) = K(x)^T(K + \lambda I)^{-1}t$

2 Question 2

Afin d'expliciter le lien entre le vecteur de support et la fonction de prédiction $y_w(x)$, il nous faut donner une définition de support.

Le vecteur de support fait référence aux échantillons de données qui se trouvent le plus près de l'hyperplan de séparation (ou de la marge maximale) entre les classes. Ces échantillons sont essentiels pour déterminer l'hyperplan optimal et sont donc considérés comme des "vecteurs de support" car ils soutiennent ou définissent la position de l'hyperplan.

Ainsi, les vecteurs de supports contribue à la détermination de la frontières de decision de la fonction de prédiction et donc le calcul des prédictions et l'attribution des classes.

La Hinge loss est une fonction de pertes utilisée dans les machines à vecteurs de support. Son objectif principal est d'évaluer la performance d'un modèle de classification en mesurant la marge d'erreur pour chaque exemple d'entraînement.