

KERNEL PROJECTION MACHINE: SÉLECTION DE MODÈLES POUR CE NOUVEL ALGORITHME DE CLASSIFICATION

Laurent Zwald¹ & Gilles Blanchard²

¹ *Laboratoire de Mathématique Université Paris-Sud, 91405 Orsay Cedex, France*

² *Fraunhofer First (IDA), Kékuléstr. 7, D-12489 Berlin, Allemagne.*

Abstract

This work presents a theoretical study of the Kernel Projection Machine, a classification algorithm which is an alternate to the Support Vector Machine (SVM). We propose to replace the Tikhonov regularization in the SVM by a penalized finite-dimensional projection. We show that a penalty term proportional to the dimension is appropriate under a “gap” hypothesis on the distribution of class given observation. We also show how to apply this algorithm in practice and present examples of numerical results.

Key words: Classification, SVM, Kernel Principal Component Analysis, Regularization, penalization.

Résumé

Ce travail présente une étude théorique de la Kernel Projection Machine, un algorithme de classification qui est une alternative à la Support Vector Machine (SVM). Nous proposons de remplacer la régularisation de type Tikhonov sur laquelle est fondée la SVM par une projection fini-dimensionnelle pénalisée. Nous montrons qu’une pénalité proportionnelle à la dimension est adéquate sous une hypothèse de “marge” de la loi de la classe conditionnellement aux observations. Nous montrons également comment appliquer cet algorithme en pratique et donnons des exemples de résultats numériques.

Mots clés : Classification, SVM, Analyse en composantes principales à noyau, Régularisation, Pénalisation.

1 Introduction

On s’intéresse à un problème de classification binaire: soit (X, Y) une variable aléatoire de loi conjointe \mathbb{P} à valeurs dans $\mathcal{X} \times \{-1; +1\}$. P désigne la loi marginale de X . X est parfois appelée l’observation et Y la classe ou étiquette. Notre but est de trouver une fonction de classification (fonction de \mathcal{X} à valeurs dans $\{-1; +1\}$) ayant la plus petite erreur de prédiction possible. Par définition, l’erreur de prédiction d’une fonction de classification f est $\mathbb{P}(Y \neq f(X))$. Cette quantité est minimale pour la *fonction de classification de Bayes* $f^*(x) = 2\mathbb{1}_{\{p(x) \geq \frac{1}{2}\}} - 1$ où $p(x) = \mathbb{P}(Y = 1 | X = x)$. La loi \mathbb{P} est inconnue et on ne dispose que de données i.i.d. échantillonnées suivant cette loi $(X_1, Y_1), \dots, (X_n, Y_n)$ pour estimer f^* .

La minimisation empirique de l'erreur de classification sur un modèle (ensemble fixé de fonctions de classification) est toutefois en général problématique car ce n'est pas une optimisation convexe. Pour des raisons algorithmiques, on considère donc plutôt la fonction de "perte douce" $\gamma(f, (x, y)) = (1 - yf(x))_+$ (qui est un majorant de l'erreur de classification) et le risque empirique associé $\mathcal{R}_n(f) = \frac{1}{n} \sum_{i=1}^n (1 - Y_i f(X_i))_+$. Ce changement de fonction de perte modifie la structure du problème: au lieu de chercher directement une fonction irrégulière (une fonction de classification), on cherche d'abord une fonction régulière à valeurs réelles dont on prend ensuite le signe pour avoir une fonction de classification. Ceci a notamment pour avantage d'aboutir à un problème d'optimisation convexe. De plus, cette fonction de perte est consistante au sens où $f^* = \operatorname{Argmin}_f \mathbb{E}(1 - Yf(X))_+$; enfin son risque en excès par rapport à l'optimum majore celui de la perte de classification, soit $L(g, f^*) \geq P(Yf(X) \leq 0) - P(Yf^*(X) \leq 0)$ où $L(g, f^*) = \mathbb{E}\gamma(g, (X, Y)) - \mathbb{E}\gamma(f^*, (X, Y))$.

L'algorithme bien connu de la Support Vector Machine (SVM) peut ainsi se formuler comme un problème de régularisation:

$$\hat{f} = \arg \min_{f \in \mathcal{H}_k^b} \mathcal{R}_n(f) + C \|f\|_{\mathcal{H}_k}^2, \quad (1)$$

où $\mathcal{H}_k^b = \{f(x) + b, f \in \mathcal{H}_k, b \in \mathbb{R}\}$ et \mathcal{H}_k est l'espace de Hilbert autoreproduisant associé au noyau k .

Cette formulation est très comparable aux méthodes de régularisation type Tikhonov utilisées en régression (typiquement dans les méthodes de splines, voir Wahba(1990)). Or dans ce cadre, il a pu être montré que des méthodes de projections fini-dimensionnelles pénalisées ont de meilleures propriétés d'adaptivité. L'idée de notre algorithme, appelé KPM, est de transférer ce principe à la classification: nous proposons de minimiser le risque empirique $\mathcal{R}_n(f)$ sur des espaces vectoriels de dimension finie. La régularisation se fera par une sélection adéquate de la dimension. La KPM a déjà été présentée dans Blanchard et al. (2004): le but de cette nouvelle publication est d'en étudier l'aspect sélection de modèles tant d'un point de vue théorique que pratique.

Cet article est organisé de la façon suivante: la section 2 donne un résultat de sélection de modèles qui vise à justifier la forme de la pénalité utilisée dans l'algorithme présenté dans la section 3. La section 4 est constituée de résultats numériques obtenus par la KPM et la SVM.

2 Hypothèse et Résultat de Sélection de Modèles

Soit $\{S_D\}_{D \geq 1}$ une famille emboîtée ($S_D \subset S_{D+1}$) de sous-espaces vectoriels de $L_2(P)$ tels que la dimension de S_D soit au plus D . Les estimateurs que nous proposons sont

$$\hat{f}_D = \arg \min_{f \in S_D} \mathcal{R}_n(f). \quad (2)$$

L'étape finale consiste à choisir la dimension la plus adaptée au problème de classification. Etant donné que seul le signe de ces fonctions nous intéresse, on peut, sans perte de

généralité, considérer la famille des $\tilde{f}_D = \text{clip}(\hat{f}_D)$ où

$$\text{clip}(g(x)) = \begin{cases} 1 & \text{si } g(x) \geq 1 \\ g(x) & \text{si } -1 < g(x) < 1 \\ -1 & \text{si } g(x) \leq -1 \end{cases}$$

La dimension sera ensuite choisie par pénalisation du risque empirique où la forme de la fonction de pénalisation pen est guidé par un argument théorique (Théorème 2).

$$\hat{D} = \arg \min_{D \geq 1} \left(\mathcal{R}_n(\tilde{f}_D) + \text{pen}(D) \right). \quad (3)$$

2.1 À Propos de l'Hypothèse de Marge

Il a été montré récemment (Nedelec et Massart (2003) , Tsybakov (2004)) qu'une hypothèse de "marge" $\mathbf{H}(h_0) : \forall x |p(x) - \frac{1}{2}| \geq h_0$ détermine la vitesse de convergence en classification. Par exemple, Nedelec et Massart (2003) déterminent l'ordre de grandeur du risque min-max sur une classe de Vapnik-Chervonenkis S sous l'hypothèse $\mathbf{H}(h_0) : \text{si } h_0 \geq \sqrt{V/n}$,

$$C' \frac{V}{nh_0} \leq \inf_f \sup_{P \in \Pi_{S, h_0}} (\mathbb{P}(Y \neq f(X)) - \mathbb{P}(Y \neq f^*(X))) \leq C \frac{V}{nh_0} \log \left(\frac{en h_0^2}{V} \right) \quad (4)$$

où le minimum est pris sur tous les estimateurs possibles , $\Pi_{S, h_0} = \{\mathbb{P} : f^* \in S \text{ et } |p(x) - \frac{1}{2}| \geq h_0\}$ et V est la dimension de Vapnik-Chervonenkis de S .

Dans notre cas, nous utilisons l'hypothèse de marge sous la forme suivante: une lecture attentive de la preuve du lemme 4 de Bousquet et al. (2004) donne le lemme suivant:

Lemme 1. *Soit $f : \mathcal{X} \rightarrow [-1, 1]$ et supposons que $\mathbf{H}(h_0)$ soit satisfaite. On a alors*

$$\|f - f^*\|_2^2 \leq \frac{1}{h_0} L(f, f^*)$$

2.2 Résultat Principal

En s'inspirant du Théorème 4.2 de Massart (2000), on obtient le résultat suivant.

Théorème 2. *Soit $K > 1$ et supposons que $\mathbf{H}(h_0)$ soit satisfaite. Avec les notations des paragraphes précédents, il existe des constantes universelles C_1 et C_2 telles que, si*

$$\text{pen}(D) \geq \frac{K}{nh_0} \left(C_1 D \log \left(\frac{n}{D} \right) + C_2 \log D \right),$$

alors

$$\mathbb{E}L(\tilde{f}_{\hat{D}}, f^*) \leq \frac{K}{K-1} \left(\inf_{D \geq 1} \left(\inf_{f \in S_D} L(f, f^*) + 2\text{pen}(D) \right) + \frac{C_2}{nh_0} \right)$$

L'idée de considérer une approche fini-dimensionnelle pour la classification a déjà été envisagée par Tsybakov et van de Geer (2005) sous une forme un peu différente.

Remarque 1: La pénalité est d'ordre de grandeur:

$$\text{pen} \sim \frac{\{\text{nombre de paramètres du modèle}\}}{\{h_0\} \times \{n\}}.$$

Nous ne disposons actuellement pas de bornes inférieures montrant son optimalité mais des travaux issus de contextes un peu différents incitent à penser que c'est le bon ordre de grandeur.

- C'est le même ordre de grandeur que pour le risque min-max donné par l'équation (4) (Cependant, il s'agit là du risque de classification alors que nous considérons la "perte douce").
- En régression et dans le cadre du bruit blanc gaussien, une pénalité proportionnelle à $\frac{D}{n}$ donne une estimation minmax de la fonction cible sur une large classe d'ellipsoïdes (cf. discussion dans Blanchard et al. (2004)).

3 Présentation de la Kernel Projection Machine (KPM)

L'analyse en composantes principales à noyau (KPCA) est une version non-linéaire de l'analyse en composantes principales (ACP) usuelle: elle consiste essentiellement à faire une ACP sur les variables fonctionnelles $(k(X_1, \cdot), \dots, k(X_n, \cdot))$ où k est un noyau. Nous proposons d'utiliser les propriétés de la KPCA pour déterminer des espaces de dimension finies adaptés à nos données de la façon suivante.

Idéalement, on voudrait avoir accès aux fonctions propres de l'opérateur de covariance non-centré $C_1 = \mathbb{E}k(X, \cdot) \otimes k(X, \cdot)$: notons ϕ_i la fonction propre associée à la $i^{\text{ème}}$ plus grande valeur propre (elles sont répétées suivant leur multiplicité: $\lambda_1 \geq \lambda_2 \geq \dots$). Les espaces S_D qui nous intéressent sont les $\langle \phi_1, \dots, \phi_D \rangle$. N'ayant pas accès à la loi sous-jacente des données, nous considérons $\{\hat{\phi}_i\}_{i \geq 1}$ les fonctions propres de $C_{1,n} = \frac{1}{n} \sum_{i=1}^n k(X_i, \cdot) \otimes k(X_i, \cdot)$ ordonnées, comme les ϕ_i , suivant les valeurs propres décroissantes de $C_{1,n}$. L'algorithme de la KPM se déroule donc en deux temps:

1. Calculer $\hat{f}_D = \arg \min_{f \in \hat{S}_D} \mathcal{R}_n(f)$ pour $D = 1 \dots n$ où $\hat{S}_D = \langle \mathbb{1}, \hat{\phi}_1, \dots, \hat{\phi}_D \rangle$ et $\mathbb{1}$ est la fonction constamment égale à 1.
2. Sélectionner la meilleur fonction de classification avec la forme de pénalité suggérée par le Théorème 2: $\hat{D} = \arg \min_D (\mathcal{R}_n(\hat{f}_D) + \lambda D)$ où la constante de pénalisation λ peut être choisie par cross-validation ou grâce à l'heuristique de pente (voir Section 4).
3. La fonction de classification finale est $\text{sign}(\hat{f}_{\hat{D}})$.

L'algorithme de KPCA détaillé dans Schölkopf et Smola (2002) permet de déterminer les espaces \widehat{S}_D et la première étape consiste en la diagonalisation de la matrice noyau des données suivie de la résolution d'un problème d'optimisation linéaire. Les différentes étapes de l'algorithme sont données avec plus de détails dans Blanchard et al. (2004).

Remarque 2 : Dans l'analyse en composantes principales à noyau, la KPM sélectionne donc autant que possible une dimension optimale pour le but de la classification, réalisant un compromis entre complexité (dimension du sous-espace) et information utilisable pour la classification.

Remarque 3 : Le Théorème 2 ne justifie que partiellement la forme de pénalité utilisée pour deux raisons:

- Il ne prend pas en compte la forme particulière des modèles \widehat{S}_D et en particulier leur caractère aléatoire. Cependant, Koltchinskii (1998) et Dauxois et al. (1982) montrent que les espaces aléatoires \widehat{S}_D se stabilisent vers l'espace déterministe S_D .
- La pénalité du théorème 2 fait non-seulement intervenir des constantes universelles trop grandes pour être utiles en pratique mais, surtout, elle fait intervenir la marge qui est inconnue et reflète la complexité du problème de classification.

4 Résultats Numériques

Afin de tester les performances de la KPM, elle a été utilisée sur les données 'Banana', 'Breast Cancer', 'Diabetis', 'Flare Solar', 'German' et 'Heart' disponibles sur <http://ida.first.gmd.de/~raetsch/data/benchmarks.htm>. Elles sont organisées de la façon suivante: 100 échantillons d'entraînement ont été créés par tirage sans remise dans un même jeu de données. A chaque fois, le reste de ces données a été regroupé dans l'échantillon test correspondant. Dans les expériences décrites ci-dessous, la fonction de classification a été calculée sur chacun des 100 échantillons d'entraînement indépendamment les uns des autres. Les erreurs sont calculées sur l'échantillon test correspondant: les résultats sont reportés sous la forme {moyenne des 100 erreurs de test} \pm {écart-type}. Le noyau gaussien a été utilisé pour toutes les expériences: $k(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ où le paramètre σ est celui donné sur le site précédemment cité.

La première colonne du tableau suivant répertorie les résultats obtenus par les SVM où la constante de régularisation C est choisie par 5-cross-validation. La deuxième colonne donne les résultats de la KPM où la constante de pénalisation λ est aussi choisie par 5-cross-validation. Pour la dernière colonne, la constante de pénalisation λ est choisie grâce à l'heuristique de pente. Cette heuristique est basée sur l'observation des différents régimes de décroissance du risque empirique $D \rightarrow \mathcal{R}_n(\widehat{f}_D)$: empiriquement, on constate que si on considère la suite $\widehat{D}_1 \leq \widehat{D}_2 \leq \dots$ des dimensions sélectionnées par le critère (3)

lorsqu'on fait décroître λ de $+\infty$ vers 0, on observe à un moment un saut de dimension plus important que tous les autres. Il correspond à un changement de régime de décroissance du risque empirique et aussi, approximativement, à la dimension minimisant le nombre d'erreurs de test.

	SVM	KPM/CV	KPM/pente
Banana ($\sigma = 0.7071$)	10.69 ± 0.67	10.91 ± 0.57	11.19 ± 0.74
Breast Cancer ($\sigma = 5$)	26.68 ± 5.23	28.73 ± 4.42	27.53 ± 4.77
Diabetis ($\sigma = 3.1623$)	23.79 ± 2.01	23.77 ± 1.69	23.97 ± 1.82
Flare Solar ($\sigma = 3.8730$)	32.62 ± 1.86	32.52 ± 1.78	33.02 ± 1.81
German ($\sigma = 5.2440$)	23.79 ± 2.12	24.09 ± 2.38	23.96 ± 2.37
Heart ($\sigma = 7.7460$)	16.23 ± 3.18	17.35 ± 3.54	17.73 ± 3.46

En conclusion, notre algorithme est compétitif avec la SVM. De plus, l'heuristique de pente permet une implémentation plus rapide que la cross-validation avec des résultats similaires.

Remerciements

Les auteurs remercient P. Massart pour l'idée originale de l'algorithme et ses précieux commentaires ainsi que E. Lebarbier et R. Vert sans qui la partie expérimentale n'aurait pas existé.

Bibliographie

- [1] Blanchard, G., Massart, P., Vert, R. et Zwald, L. (2004) *Kernel projection Machine: a New Tool for Pattern Recognition*, proceedings of NIPS 2004.
- [2] Blanchard, G., Bousquet, O. et Massart, P. (2004) *Statistical performance of Support Vector Machines*, preprint, Université Paris-Sud.
- [3] Nedelec, E. et Massart, P. (2003) *Risk bounds for statistical learning*, preprint, Université Paris-sud.
- [4] Tsybakov, A. (2004) *Optimal aggregation of classifiers in statistical learning*, Annals of Statistics, 32(1).
- [5] Tsybakov, A.B. et van de Geer, S.A. (2005) *Adaptivity of Support Vector Machines with ℓ_1 Penalty*, The Annals of Statistics.
- [6] Schölkopf, B et Smola, A.J. (2002) *Learning with Kernels*, MIT Press.
- [7] Massart, P. (2000) *Some applications of concentration inequalities to statistics*, Annales de la Faculté des Sciences de Toulouse, IX(245–303).
- [8] Dauxois, P., Pousse, A. et Romain, Y. (1982) *Asymptotic Theory for the Principal Component Analysis of a Vector Random Function: Some Applications to Statistical Inference*, Journal of multivariate analysis.
- [9] Wahba, G. (1990) *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics 59, SIAM.