Apprentissage semi supervisé via un SVM parcimonieux : calcul du chemin de régularisation

L_1-Norm regularization path for the sparse semi-supervised Laplacian SVM

Gilles Gasso, Karina Zapién, Stephane Canu

LITIS EA 4108 - INSA de Rouen
Avenue de l'Université, 76801,
Saint Etienne du Rouvray, France
{gilles.gasso, karina.zapien, stephane.canu}@insa-rouen.fr

Résumé

L'apprentissage semi supervisé vise à résoudre des problèmes de reconnaissance des formes comportant un petit nombre de données étiquetées pour un grand nombre de points sans labels. Ces derniers points servant à dévoiler la structure sous-jacente des données, il est pertinent d'utiliser des graphes de similarité pour représenter cette structure. Les méthodes à noyaux telles que l'algorithme du Laplacien SVM [2] offrent un cadre flexible permettant d'intégrer cette information de structure comme une contrainte dans un problème SVM sur les points étiquetés. Malheureusement, la fonction de décision obtenue s'exprime en fonction de tous les points (avec ou sans labels) et est donc peu parcimonieuse. Nous proposons ici une solution consistant à ajouter une pénalisation supplémentaire de type L_1 qui introduit la parcimonie dans les variables explicatives. En se fondant sur les travaux de Wang [19], nous présentons ici un chemin de régularisation permettant de calculer efficacement la solution. L'application sur des données simulées et réelles montre les avantages d'un modèle parcimonieux où nous obtenons les mêmes niveaux de performance que le Laplacien SVM en réduisant significativement la taille de l'ensemble des variables explicatives.

Mots-clés *: Apprentissage semi supervisé, Laplacien de graphe, SVM, parcimonie, pénalisation L_1, chemins de régularisation.*

Abstract

The goal of semi-supervised algorithms is to leverage the learning process by using unlabeled points to unravel the underlying structure of the data. A common way to represent this underlying structure is to use graphs. The flexibility of kernel methods such as the Laplacian SVM [2] offers a framework to integrate a constraint on the graph smoothness and to build a kernel machine by solving a SVM on the labeled data. A common practitioner's complaint is the algorithm's long running time for new points classification as the solution depends on all points and is not sparse. We provide an efficient way of alleviating this problem with the use of a L_1 penalization term which provides sparsity. Following [19] we derive a regularization path algorithm to efficiently compute the solution. Empirical evidence with simulated and real data shows the benefits of sparsity in semi-supervised learning demonstrating that the same level of performance as the Laplacian SVM can be achieved with an important reduction of the explicative variables set.

Key-words*: Semi-supervised learning, graph Laplacian, Laplacian SVM, sparsity constraint, regularization path.*

1 INTRODUCTION

Les algorithmes semi supervisés ont connu un regain d'intérêt ces dernières années dans la communauté de l'apprentissage statistique. A la différence de l'apprentissage supervisé, l'apprentissage semi supervisé vise les problèmes avec relativement peu de données étiquetées et une grande quantité de données non étiquetées. Ce type de situation peut se produire quand l'étiquetage des données est coûteux, comme dans le cas de la classification de pages internet. La question qui se pose est alors de savoir si la seule connaissance des points avec labels est suffisante pour construire une fonction de décision capable de prédire correctement les étiquettes des points non étiquetés. Différentes approches (voir [3, 20] pour un panorama) proposent de déduire des points non étiquetés des informations supplémentaires et de les inclure dans le problème d'apprentissage. Parmi ces travaux certains algorithmes sont basés sur l'hypothèse du regroupement (*cluster assumption* [5]) : l'idée est de rechercher une frontière de décision évitant de passer à travers des régions de grande densité de probabilité (dans l'espace des points) car on suppose que les points étiquetés et non étiquetés pouvant être groupés dans un même *cluster* ont la même étiquette. Dans la même philosophie, des algorithmes basés sur des modèles génératifs ont été proposés [1, 8, 14] pour trouver directement les *clusters*. Citons également comme techniques ayant implémenté l'hypothèse de regroupement et ayant une relation directe avec les méthodes à noyaux le *Transductive SVM* [4, 6, 17] ou les classifieurs avec des bases de connaissance de Fung et al. [9].

Un autre type d'information pouvant être extraite est la structure géométrique de la distribution marginale des données sans étiquettes. L'hypothèse

sous-jacente est que les données résident sur une sous-variété régulière de dimension réduite par rapport à la dimension intrinsèque des points. La frontière de décision recherchée doit éviter de passer à travers ces variétés régulières : c'est le *manifold assumption*. En suivant cette idée, Belkin et al. [2] ont proposé le Laplacien SVM (*Support Vector Machines*) qui consiste en un problème SVM classique sur les données étiquetées avec une contrainte supplémentaire prenant en compte la régularité des variétés. La modélisation des variétés est réalisée à travers le Laplacien du graphe de similarité des points. Cette formulation aboutit à un problème d'apprentissage semi supervisé convexe qui admet comme solution une fonction de décision s'exprimant comme la combinaison linéaire du noyau évalué en tous les points d'apprentissage (avec ou sans étiquettes). Par conséquent, la solution obtenue est peu parcimonieuse, surtout si la base d'apprentissage est de taille importante.

Pour contourner ce problème de parcimonie, nous proposons dans cet article d'intégrer une pénalisation de type L_1 sur les paramètres de la fonction de décision dans le problème d'apprentissage. La complexité du problème en est augmentée car en dehors du réglage du compromis entre le problème de classification par SVM et la régularité de la variété, il faut également déterminer le paramètre de régularisation à associer à la pénalisation L_1. Pour ce faire et dans l'objectif de rendre automatique notre algorithme pour l'utilisateur, nous proposons le calcul du chemin de régularisation qui donne l'ensemble des solutions optimales admissibles lorsque le paramètre de régularisation évoqué précédemment varie sur toute sa plage. Le calcul de ce chemin s'inspire de l'idée initiale de Hastie et al. [11] étendue par Wang et al. [18] pour la sélection de variables dans le cadre des problèmes SVM. Il est établi dans l'article qu'en parcourant le chemin de régularisation, les paramètres du modèle varient linéairement par morceaux, ce qui en fait un algorithme très efficace pouvant être couplé avec une procédure de validation croisée pour sélectionner le meilleur modèle.

Le reste de l'article est organisé de la façon suivante : dans la section 2 nous détaillons l'algorithme du Laplacien SVM puis dans la section 3, la pénalisation L_1 est introduite dans le cadre du Laplacien SVM. Le calcul du chemin de régularisation afin d'obtenir une solution parcimonieuse est alors présenté. Le potentiel de l'algorithme est illustré avec des données de simulation et des données réelles en comparaison avec l'algorithme initial du Laplacien SVM dans la section 4. Dans la section 5, nous discutons une façon (via un autre chemin de régularisation) de régler l'hyper-paramètre relatif à la régularité de la fonction de décision. Quelques conclusions et perspectives sont finalement esquissées.

Notation

Dans le reste de l'article, les notations suivantes seront adoptées. Si H est une matrice, $H_{i,\cdot}$ représente sa $i^{\text{ème}}$ ligne et $H_{\cdot,j}$ sa $j^{\text{ème}}$ colonne. De façon similaire, $H_{\mathbf{I},\mathbf{J}}$ avec $\mathbf{I}$ et $\mathbf{J}$ deux vecteurs d'entiers, représente la sous-matrice

correspondante de H. Enfin I_ℓ représente la matrice identité de dimension ℓ.

2 PRÉSENTATION DE L'ALGORITHME DU LAPLACIEN SVM

Soit $S_\mathcal{L} = \{(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}, \quad i = 1, \ldots, \ell\}$ l'ensemble des points étiquetés et soit $S_\mathcal{U} = \{x_i \in \mathcal{X}, \quad i = \ell + 1, \ldots, \ell + u\}$ celui des points sans étiquettes. L'objectif ici est de construire un classifieur SVM se basant non seulement sur les points étiquetés mais aussi sur l'information supplémentaire apportée par les points sans étiquettes.

2.1 Cadre général

Le Laplacien SVM exploite l'hypothèse de l'appartenance des points d'apprentissage à des sous-variétés c'est-à-dire qu'on considère que ces points échantillonnés selon la loi marginale $\mathcal{P}_\mathcal{X}$ (inconnue) décrivent une structure géométrique dans l'espace $\mathcal{X}$ des données. Si deux points x_i et x_j sont proches sur ces variétés alors $f(x_i) \sim f(x_j)$. Ici $f(x)$ représente la fonction de décision recherchée. L'interprétation directe de cette hypothèse est que le label des points varie de façon régulière le long des variétés. Si l'on pose un problème classique de classification par SVM des points étiquetés auquel on adjoint la pénalisation de variation régulière des labels le long des sous-variétés, on aboutit au problème d'optimisation suivant :

$$f^* = \underset{f,b}{\operatorname{argmin}} \sum_{i=1}^{l} V(y_i, f(x_i) + b) + \frac{\lambda_2}{2}\|f\|^2 + \frac{\mu}{2}\|f\|^2_\mathcal{M} \tag{1}$$

où $V(\cdot, \cdot)$ est la fonction coût et λ_2 le paramètre de régularisation classique d'un SVM (qui fait le compromis entre l'erreur de classification et la maximisation de la marge). Dans ce problème apparaît un nouveau paramètre de régularisation μ qui introduit la contrainte de régularité de la sous-variété $\mathcal{M}$ décrite par les données. Pour décider de la classe d'un point, il suffit simplement de considérer $g(x) = \text{sign}(f(x) + b)$ avec b le terme de biais.

Une approximation du dernier terme de régularisation dans l'équation (1) est obtenue via la formule [2] :

$$\|f\|^2_\mathcal{M} = f^\top L f, \tag{2}$$

avec $f = [\, f(x_1)\, f(x_2)\, \ldots\, f(x_{\ell+u})\,]^\top$ le vecteur contenant la sortie du classifieur évalué sur tous les points d'apprentissage. La matrice L représente le Laplacien du graphe de similarité des points dont les noeuds sont les $\ell + u$ points et les arcs indiquent les plus proches voisins (ce qui signifie que le point x_i est connecté à ses n plus proches voisins avec n fixé a priori ou

x_i est connecté à tous les points se trouvant dans une boule centrée en x_i et de rayon maximal ε). Pour déterminer les plus proches voisins, la métrique usuelle est la distance euclidienne. A chaque arc est associé un poids w_{ij} fonction de la distance du point x_i à x_j. Soit W la matrice d'adjacence du graphe (contenant les poids w_{ij}) et soit D une matrice diagonale dont les éléments diagonaux sont $D_{ii} = \sum_j w_{ij}$. Le Laplacien L du graphe est alors défini par :

$$L = D - W. \tag{3}$$

A partir de ces définitions, nous pouvons maintenant formuler le problème d'apprentissage semi supervisé par le Laplacien SVM.

2.2 Formulation du Laplacien SVM

On considère que la fonction de décision $f(x)$ appartient à un espace de Hilbert à noyau reproduisant $\mathcal{H}$. En utilisant le cadre du théorème de représentation des SVM [16], Belkin et al. [2] ont établi que la solution $f(x)$ du problème (1) est fonction de tous les $\ell + u$ points et s'écrit comme

$$f(x) = \sum_{j=1}^{\ell+u} \beta_j k(x, x_j) \tag{4}$$

Le problème d'optimisation posé par le Laplacien SVM se décline alors sous la forme suivante :

$$\left\{ \begin{array}{ll} \min_{\boldsymbol{\beta},b,\xi} & \sum_{i=1}^{\ell} \xi_i + \frac{\lambda_2}{2} \boldsymbol{\beta}^\top K \boldsymbol{\beta} + \frac{\mu}{2} \boldsymbol{\beta}^\top K L K \boldsymbol{\beta} \\ \text{s.c.} & y_i (f(x_i) + b) \geq 1 - \xi_i, \;\; i = 1, \ldots, \ell \\ & \xi_i \geq 0, \;\; i = 1, \ldots, \ell \end{array} \right. \tag{5}$$

où $K \in \mathbb{R}^{(\ell+u)\times(\ell+u)}$, $K = [k(x_i, x_j)]_{i,j=1,\ldots,\ell+u}$ est la matrice de Gram avec $k(\cdot,\cdot)$ la fonction noyau. Le lecteur pourra remarquer ici que la fonction coût optimisée est la fonction charnière $V(y_i, f(x_i)) = \max(0, 1 - y_i f(x_i))$.

Soit α_i et γ_i, $i = 1, \ldots, \ell$ les multiplicateurs de Lagrange associés aux contraintes inégalités du problème (5). Le lagrangien du problème est alors

$$\begin{aligned} \mathcal{J} \;\; = \;\; & \sum_{i=1}^{\ell} \xi_i + \frac{\lambda_2}{2} \boldsymbol{\beta}^\top K \boldsymbol{\beta} + \frac{\mu}{2} \boldsymbol{\beta}^\top K L K \boldsymbol{\beta} \\ & - \sum_{i=1}^{\ell} \alpha_i \left(y_i f(x_i) + y_i b - 1 + \xi_i \right) - \sum_{i=1}^{\ell} \gamma_i \xi_i \end{aligned}$$

et les conditions d'optimalité associées aux variables primales fournissent les relations suivantes

$$\nabla_b \mathcal{J} = 0 \quad : \quad \boldsymbol{\alpha}^\top \mathbf{y} = 0 \tag{6}$$

$$\begin{aligned} \nabla_{\xi_i} \mathcal{J} = 0 \quad &: \quad 1 - \alpha_i - \gamma_i = 0 \quad \Rightarrow \quad 0 \leq \alpha_i \leq 1 \\ \nabla_{\boldsymbol{\beta}} \mathcal{J} = 0 \quad &: \quad \lambda_2 K\boldsymbol{\beta} + \mu KLK\boldsymbol{\beta} - \sum_{i=1}^{\ell} \alpha_i y_i K_{i,\cdot}^\top = 0 \end{aligned} \tag{7}$$

avec $\mathbf{y} = [y_1, y_2, \ldots, y_\ell]^\top$ le vecteur contenant les labels des points étiquetés. En posant Y la matrice diagonale de $\mathbb{R}^{\ell \times \ell}$ avec $Y_{ii} = y_i, \ \forall \ i = 1, ..., \ell$, on déduit alors que le vecteur de paramètres $\boldsymbol{\beta}$ de la fonction de décision est donné par :

$$\boldsymbol{\beta} = (\lambda_2 I + \mu LK)^{-1} J^\top Y \boldsymbol{\alpha}, \tag{8}$$

avec $J = \begin{bmatrix} I_\ell & 0_{\ell \times u} \end{bmatrix}$ et $\boldsymbol{\alpha}$ le vecteur des paramètres de Lagrange α_i, pour $i = 1, \ldots, \ell$. $\boldsymbol{\alpha}$ est solution du problème dual suivant [2] qui comme le remarquera le lecteur n'implique que les points étiquetés :

$$\begin{cases} \max_{\boldsymbol{\alpha} \in \mathbb{R}^l} & -\frac{1}{2} \boldsymbol{\alpha}^\top Q \boldsymbol{\alpha} + \boldsymbol{\alpha}^\top \mathbf{1} \\ \text{s.c.} & \boldsymbol{\alpha}^\top \mathbf{y} = 0, \\ & 0 \leq \alpha_i \leq 1, \quad i = 1, \ldots, \ell \end{cases}$$

avec la définition suivante de la matrice

$$Q = YJK(\lambda_2 I_{\ell+u} + \mu LK)^{-1} J^\top Y.$$

A ce niveau, deux remarques peuvent être faites : en premier lieu, on peut constater que le vecteur de paramètres $\boldsymbol{\beta}$ issu de l'équation (8) nécessite la résolution d'un système d'équations. Cette résolution a une complexité de l'ordre de $\mathcal{O}((\ell + u)^3)$ qui peut être prohibitive si $\ell + u$ est élevé. En deuxième lieu, bien que la solution $\boldsymbol{\alpha}$ du problème dual peut être parcimonieuse, c'est-à-dire certains éléments de ce vecteur sont nuls (ceci correspond aux points étiquetés qui sont bien classés), le vecteur $\boldsymbol{\beta}$ n'est pas en général parcimonieux. Par conséquent, $f(x)$ s'exprime de façon non parcimonieuse en fonction de toutes les variables (en considérant que chaque terme $k(x, x_j)$ dans l'expression de $f(x)$ est une variable).

Une autre remarque peut être faite : en exploitant les conditions de Karush-Kuhn-Tucker (KKT) du problème (5), il est aisé d'établir (voir [16] pour plus de détails) que les points étiquetés peuvent être répartis en 3 ensembles disjoints (c'est-à-dire $S_\mathcal{L} = \mathcal{L} \cup \mathcal{R} \cup \mathcal{E}$) :

$$\begin{array}{lll} \mathcal{L} : & y_i(f(x_i) + b) < 1 & \alpha_i = 1 \\ \mathcal{R} : & y_i(f(x_i) + b) > 1 & \alpha_i = 0, \\ \mathcal{E} : & y_i(f(x_i) + b) = 1 & 0 \leq \alpha_i \leq 1, \end{array} \tag{9}$$

Les points de $\mathcal{L}$ sont des points supports (points mal classés), les points de $\mathcal{E}$ sont aussi points supports mais sur la marge alors que les points de $\mathcal{R}$ sont des points bien classés et ayant donc un paramètre de Lagrange nul. La solution (8) est entièrement déterminée par la connaissance de ces ensembles pour les paramètres de régularisation fixés.

3 SVM LAPLACIEN AVEC NORME L_1

Le vecteur $\boldsymbol{\beta}$ obtenu à l'équation (8) n'étant pas parcimonieux dans le cas général, nous proposons dans cette partie d'inclure de façon explicite dans le problème d'apprentissage la contrainte de parcimonie. Ceci se fait en imposant une pénalité de type L_1 sur les paramètres $\boldsymbol{\beta}$ du modèle. Le problème à résoudre se formule alors de la manière suivante :

$$\left\{\begin{array}{ll} \min_{\boldsymbol{\beta},b,\xi} & \sum_{i=1}^{\ell}\xi_i + \frac{\lambda_2}{2}\boldsymbol{\beta}^\top K\boldsymbol{\beta} + \frac{\mu}{2}\boldsymbol{\beta}^\top KLK\boldsymbol{\beta} \\ \text{s.c.} & y_i(f(x_i)+b) \geq 1-\xi_i, \quad \forall i=1,\ldots,\ell \\ & \xi_i \geq 0, \quad \forall i=1,\ldots,\ell \\ & \sum_{j=1}^{\ell+u}|\beta_j| \leq s \end{array}\right. \tag{10}$$

La signification de cette pénalisation L_1 est la suivante : si on se réfère à l'équation (4), la fonction $f(x)$ peut être assimilée à un développement sur une base de fonctions génératrices $\mathcal{D} = \{h_1(x), h_2(x), \ldots, h_{\ell+u}(x)\}$. Une fonction de base $h_j(x)$ représente l'évaluation de la fonction noyau $k(.,.)$ au point x_j, c'est-à-dire $h_j(x) = k(x, x_j)$. Le choix d'une faible valeur pour l'hyper-paramètre s aura tendance à mettre à 0 la plupart des paramètres β_j du modèle. L'objectif de cet article est d'examiner l'évolution de la fonction de décision $f(x)$ lorsqu'on fait varier le paramètre s. Il ne s'agit pas ici de fixer a priori quelques valeurs de s, de résoudre le problème (10) puis de comparer les solutions. L'objectif visé ici est de fournir de façon automatique, en utilisant un algorithme itératif, l'ensemble des solutions admissibles $f(x)$ pour $0 \leq s \leq \infty$: c'est le calcul du chemin de régularisation. On montrera dans la sous-section suivante qu'à l'étape t, si on a la solution $f^t(x)$, les paramètres de la solution suivante se déduisent de ceux de $f^t(x)$ par une simple relation linéaire.

Pour ce faire, considérons le lagrangien associé au problème primal (10) :

$$\left\{\begin{array}{l} \mathcal{J} = \sum_{i=1}^{\ell}\xi_i + \frac{\lambda_2}{2}\boldsymbol{\beta}^\top K\boldsymbol{\beta} + \frac{\mu}{2}\boldsymbol{\beta}^\top KLK\boldsymbol{\beta} \\ \qquad - \sum_{i=1}^{\ell}\alpha_i\left(y_i f(x_i) + y_i b - 1 + \xi_i\right) - \sum_{i=1}^{\ell}\gamma_i\xi_i \\ \qquad + \lambda_1\left(\sum_{j=1}^{\ell+u}|\beta_j| - s\right) \\ \text{avec} \quad \lambda_1 \geq 0 \quad \alpha_i \geq 0, \quad \gamma_i \geq 0, \quad \forall i=1,\ldots,\ell \end{array}\right.$$

Les conditions d'optimalité par rapport aux variables primales b et ξ_i de la section précédente restent inchangées (voir les équations (6) et (7)).

La dérivation de la condition d'optimalité pour $\boldsymbol{\beta}$ est légèrement différente car ici il faut considérer uniquement les paramètres β_j non nuls. Soit $\mathcal{A} = \{j,\ \beta_j \neq 0\}$ l'ensemble de variables *actives* et $\bar{\mathcal{A}} = \{m,\ \beta_m = 0\}$ son complément sur la base $\mathcal{D}$. Il faut noter qu'à cause de la définition de $\mathcal{D}$, l'ensemble $\mathcal{A}$ correspond d'une certaine manière à l'ensemble des points utiles (étiquetés ou pas) pour décrire correctement la fonction de décision.

Considérons la matrice P telle que

$$P = \lambda_2 R \quad \text{avec} \quad R = K + \rho KLK \quad \text{et} \quad \rho = \frac{\mu}{\lambda_2} \tag{11}$$

En remarquant que $f(x_j)$ peut s'écrire sous la forme $f(x_j) = K_{j,\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}$, on déduit la condition d'optimalité par rapport aux variables actives $\boldsymbol{\beta}_{\mathcal{A}}$ comme le montre l'expression :

$$P_{\mathcal{A},\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}} - K_{\mathcal{A},\mathcal{S}_{\mathcal{L}}} Y_{S_{\mathcal{L}},S_{\mathcal{L}}}\boldsymbol{\alpha} + \lambda_1 \text{sign}\left(\boldsymbol{\beta}_{\mathcal{A}}\right) = 0 \tag{12}$$

A l'optimalité, les conditions KKT stipulent que les contraintes inégalités du problème primal deviennent des contraintes égalité. On en déduit alors les conditions suivantes :

$$\begin{aligned} y_i\left(K_{i,\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}} + b\right) &= 1, \quad \forall i \in \mathcal{E} \\ \lambda_1\left(\sum_{j\in\mathcal{A}} |\beta_j| - s\right) &= \lambda_1\left(\text{sign}\left(\boldsymbol{\beta}_{\mathcal{A}}\right)^\top \boldsymbol{\beta}_{\mathcal{A}} - s\right) = 0 \end{aligned} \tag{13}$$

En se basant sur ces équations d'optimalité, nous allons déduire dans les sous-sections à venir le calcul du chemin de régularisation lorsque le paramètre s varie. On supposera dans cette partie de l'article que les autres paramètres de régularisation λ_2 et ρ (ou de façon équivalente μ)sont fixés. Le chemin de régularisation permet de caractériser l'évolution des paramètres $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, b et λ_1 lorsque s varie.

Notons que si λ_1 (ou de façon équivalente s) et μ sont fixés, nous pouvons analyser l'évolution de $f(x)$ par rapport à λ_2, ce qui donne le chemin de régularisation associé à λ_2. Ce dernier point sera discuté dans la section 5.

3.1 Calcul du chemin de régularisation L1

Le problème (10) est convexe. Grâce à cette propriété, on peut montrer que la solution de (10) coïncide avec la solution du problème équivalent (compte tenu des définitions précédentes)

$$\min_{\boldsymbol{\beta},b} \sum_{i=1}^{\ell} \max(0, 1 - y_i(K_{i,.} \times \boldsymbol{\beta} + b)) + \frac{\lambda_2}{2}\boldsymbol{\beta}^\top R\boldsymbol{\beta} + \lambda_1\|\boldsymbol{\beta}\|_1$$

où $\|\cdot\|_1$ représente la norme L_1. Ce dernier problème fait intervenir un coût charnière qui est linéaire par morceaux, une pénalité quadratique et une

autre pénalité linéaire par morceaux (la valeur absolue). Suivant les travaux de Rosset et Zhu [15], pour λ_2 fixé, le problème admet un chemin de régularisation sur lequel la solution est linéaire par morceaux en fonction de λ_1. De façon similaire, on peut trouver le chemin de λ_2 pour λ_1 fixé. Nous référons le lecteur à [15] pour les conditions nécessaires à un chemin linéaire par morceaux.[1]

Pour pouvoir établir le chemin L_1, nous nous sommes basés sur les travaux de Wang et al. exposés dans [19]. A l'étape t, si le paramètre s^t est suffisamment petit, la contrainte sur la norme L_1 dans (10) devient active, c'est-à-dire qu'elle devient une contrainte égalité

$$\text{sign}\left(\boldsymbol{\beta}_{\mathcal{A}}\right)^{\top} \boldsymbol{\beta}_{\mathcal{A}} = s^t. \tag{14}$$

Quand s^t augmente de manière infinitésimale, la contrainte devient inactive et la solution reste inchangée jusqu'à ce qu'une valeur seuil soit atteinte. Quand s^t augmente, l'ensemble des variables actives $\mathcal{A}$, les ensembles de points étiquetés $\mathcal{E}$, $\mathcal{L}$ et $\mathcal{R}$ restent inchangés jusqu'à une valeur particulière de s. En utilisant des raisons de continuité de la solution (ce qui signifie par exemple qu'un point mal classé ne peut pas devenir bien classé sans passer sur la marge), nous pouvons exprimer les variations des paramètres $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, b et λ_1 par rapport aux variations de s en utilisant les équations (6, 12-14). On obtient alors le système d'équations linéaires suivant

$$\begin{aligned}
P_{\mathcal{A},\mathcal{A}}\Delta\boldsymbol{\beta}_{\mathcal{A}} - K_{\mathcal{A},\mathcal{E}}Y_{\mathcal{E}}\Delta\boldsymbol{\alpha}_{\mathcal{E}} + \Delta\lambda_1 \text{sign}\left(\boldsymbol{\beta}_{\mathcal{A}}\right) &= 0 \\
\mathbf{y}_{\mathcal{E}}^{\top}\Delta\boldsymbol{\alpha}_{\mathcal{E}} &= 0 \\
Y_{\mathcal{E}}\left(K_{\mathcal{E},\mathcal{A}}\Delta\boldsymbol{\beta}_{\mathcal{A}} + \mathbf{1}_{\mathcal{E}}\Delta b\right) &= 0 \\
\\
\text{sign}\left(\boldsymbol{\beta}_{\mathcal{A}}\right)^{\top}\Delta\boldsymbol{\beta}_{\mathcal{A}} &= \left(s - s^t\right)
\end{aligned}$$

où la matrice diagonale $Y_{\mathcal{E}}$ représente la matrice diagonale $Y_{\mathcal{E},\mathcal{E}}$ pour simplifier les notations. Il faut remarquer que dans ces équations, nous ne considérons que les variations $\Delta\boldsymbol{\alpha}_{\mathcal{E}}$ des multiplicateurs de Lagrange associés aux points dans $\mathcal{E}$ car les multiplicateurs des autres points étiquetés sont fixés (égaux à 0 ou 1).

Un examen du système montre que nous avons $|\mathcal{A}| + |\mathcal{E}| + 2$ équations à $|\mathcal{A}|+|\mathcal{E}|+2$ inconnues : $\Delta\boldsymbol{\beta}_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$, $\Delta\boldsymbol{\alpha}_{\mathcal{E}} \in \mathbb{R}^{|\mathcal{E}|}$, Δb et $\Delta\lambda_1$. Ce système peut être mis sous forme matricielle

$$H\ \Delta\boldsymbol{\theta} = (s - s^t)\mathbf{z}, \tag{15}$$

[1] Précisons que les algorithmes de type chemin de régularisation qui ont fait récemment leur apparition dans la communauté d'apprentissage statistique (voir [11]) sont une application de travaux anciens en économie pour la gestion des portefeuilles de risques [13]. Ces travaux sont proches de ceux d'Heller sur l'analyse de sensibilité des solutions de problème de type programmation linéaire [12] ou ceux plus généraux sur la programmation paramétrique [10].

avec

$$\boldsymbol{\delta\theta} = \begin{bmatrix} \Delta\boldsymbol{\beta}_{\mathcal{A}} \\ \Delta\boldsymbol{\alpha}_{\mathcal{E}} \\ \Delta b \\ \Delta\lambda_1 \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} \mathbf{0}_{\mathcal{A}} \\ \mathbf{0}_{\mathcal{E}} \\ 0 \\ 1 \end{bmatrix} \quad \text{et} \tag{16}$$

$$H = \begin{bmatrix} P_{\mathcal{A},\mathcal{A}} & -K_{\mathcal{A},\mathcal{E}}Y_{\mathcal{E}} & \mathbf{0}_{\mathcal{A}} & \text{sign}(\boldsymbol{\beta}_{\mathcal{A}}) \\ -Y_{\mathcal{E}}K_{\mathcal{E},\mathcal{A}} & \mathbf{0}_{\mathcal{E}} & -\mathbf{y}_{\mathcal{E}} & \mathbf{0}_{\mathcal{E}} \\ \mathbf{0}_{\mathcal{A}}^{\top} & -\mathbf{y}_{\mathcal{E}}^{\top} & 0 & 0 \\ \text{sign}(\boldsymbol{\beta}_{\mathcal{A}})^{\top} & \mathbf{0}_{\mathcal{E}}^{\top} & 0 & 0 \end{bmatrix}$$

En posant $\boldsymbol{\eta} = H^{-1}\mathbf{z}$, on constate que les paramètres impliqués dans $f(x)$ varient de façon linéaire par rapport à s car nous avons :

$$\theta = \theta^t + (s - s^t)\boldsymbol{\eta}_\theta \tag{17}$$

où θ représente de façon générique les paramètres $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, b ou λ_1 . Cette variation linéaire reste valable tant qu'aucun des ensembles de variables ou de points $\mathcal{A}$, $\mathcal{E}$, $\mathcal{L}$ ou $\mathcal{R}$ obtenus à l'étape précédente ne change. La valeur exacte de s pour laquelle la variation linéaire devient invalide correspond à un changement dans l'un des ensembles. Ce changement peut être détecté en surveillant quatre types d'évènements :

1. un point étiqueté dans $\mathcal{E}$ passe dans $\mathcal{R}$ ou $\mathcal{L}$.
2. Un point étiqueté de $\mathcal{R} \cup \mathcal{L}$ se déplace dans $\mathcal{E}$.
3. Une variable active $\beta_j \neq 0 \in \mathcal{A}$ devient inactive.
4. Une variable inactive $\beta_m = 0 \in \bar{\mathcal{A}}$ devient active.

3.1.1 Détection des évènements et détermination de la prochaine valeur de s

1. Un point étiqueté $x_i \in \mathcal{E}$ se déplace dans $\mathcal{L}$ ou $\mathcal{R}$
   En regardant la définition des ensembles de points $\mathcal{L}$ et $\mathcal{R}$, ce mouvement implique que le paramètre de Lagrange associé à ce point atteint une de ses valeurs limites c'est-à-dire respectivement $\alpha_i = 1$ ou $\alpha_i = 0$. En utilisant l'équation générale (17) de mise à jour des paramètres, on peut déduire les valeurs de l'hyper-paramètre s correspondant à ces deux évènements :

   $$\text{si } x_i \in \mathcal{E} \rightarrow \mathcal{L} \quad \Rightarrow \quad s^{t+1} = s^t + \frac{1 - \alpha_i^t}{\boldsymbol{\eta}_{\alpha_i}}, \quad \forall i \in \mathcal{E}$$

   ou

   $$\text{si } x_i \in \mathcal{E} \rightarrow \mathcal{R} \quad \Rightarrow \quad s^{t+1} = s^t + \frac{0 - \alpha_i^t}{\boldsymbol{\eta}_{\alpha_i}}, \quad \forall i \in \mathcal{E}$$

2. Un point étiqueté $x_i \in \mathcal{L} \cup \mathcal{R}$ passe dans $\mathcal{E}$
   Ces évènements correspondent au passage d'un point de $\mathcal{L}$ ou $\mathcal{R}$ sur la marge. L'apparition d'une telle situation se produit si le résidu

$$r_i = 1 - y_i(f(x_i) + b) = 1 - y_i \left(K_{i,\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}} + b\right)$$

   devient nul. La question ici est de combien doit varier le résidu r_i^t pour que le point x_i se retrouve sur la marge ? On pourra remarquer que de la définition précédente de l'expression du résidu, on peut obtenir par différentiation, la pente de variation du résidu via la formule :

$$\boldsymbol{\eta}_{r_i} = -y_i \left(\boldsymbol{\eta}_b + K_{i,\mathcal{A}}\boldsymbol{\eta}_{\beta_{\mathcal{A}}}\right),\ \forall i \in \mathcal{L} \cup \mathcal{R}$$

   Comme les paramètres $\beta_{\mathcal{A}}$ et b varient de façon linéaire par morceaux par rapport à s, on en déduit que les résidus vont également évoluer de manière similaire soit : $r_i = r_i^t + (s - s^t)\boldsymbol{\eta}_{r_i}$. Par conséquent la valeur de s correspondant à l'apparition de l'évènement $x_i \in \mathcal{L} \cup \mathcal{R} \rightarrow \mathcal{E}$ est donnée par :

$$s^{t+1} = s^t + \frac{0 - r_i^t}{\boldsymbol{\eta}_{r_i}}, \quad \forall i \in \mathcal{L} \cup \mathcal{R}$$

3. Une variable active $j \in \mathcal{A}$ devient inactive
   Cet évènement signifie que le paramètre associé à la variable β_j s'annule ce qui donne la condition suivante :

$$s^{t+1} = s^t + \frac{0 - \beta_j^t}{\boldsymbol{\eta}_{\beta_j}}, \quad \forall j \in \mathcal{A}$$

4. Une variable inactive $m \in \bar{\mathcal{A}}$ devient active
   Pour pouvoir analyser cet événement, nous considérons la condition d'optimalité liée à un paramètre β_j actif. En se basant sur l'équation (12), on déduit :

$$P_{j,\cdot}\ \boldsymbol{\beta} - K_{j,\mathcal{S}_{\mathcal{L}}} Y \boldsymbol{\alpha} \quad = \quad -\lambda_1 \text{sign}(\beta_j)\ \ \forall\, j \in \mathcal{A}$$

   Cette dernière équation montre que pour toutes les variables actives, le terme de droite est égale à $-\lambda_1$ au signe du paramètre près. Cette expression peut être vue alors comme la corrélation généralisée de la variable $k(x, x_j)$ (par analogie à l'algorithme du LARS [7]). Par conséquent, pour toute variable m (active ou non), on définit la corrélation généralisée sous la forme $c_m = P_{m,\cdot}\ \boldsymbol{\beta} - K_{m,\mathcal{S}_{\mathcal{L}}} Y \boldsymbol{\alpha}$ qui dépend des paramètres du modèle. En utilisant les mêmes arguments sur la variation linéaire des paramètres, on établit que cette corrélation évolue aussi de façon linéaire. Sa variation unitaire est donnée par

$$\boldsymbol{\eta}_{c_m} = P_{m,\mathcal{A}}\ \boldsymbol{\eta}_{\beta_{\mathcal{A}}} - K_{m,\mathcal{E}} Y_{\mathcal{E}}\ \boldsymbol{\eta}_{\alpha_{\mathcal{E}}}$$

Si une variable inactive m rejoint l'ensemble des variables actives, sa corrélation généralisée doit être égale à la valeur courante λ_1 au signe près. Ceci signifie que c_m doit obligatoirement vérifier la contrainte

$$|c_m^{t+1}| = |\lambda_1^{t+1}|$$

où

$$c_m^{t+1} = c_m^t + \boldsymbol{\eta}_{c_m}(s^{t+1} - s^t)$$

et

$$\lambda_1^{t+1} = \lambda_1^t + \boldsymbol{\eta}_{\lambda_1}(s^{t+1} - s^t).$$

A partir de cette contrainte, on déduit alors que la valeur de s qui rend possible l'activation de la variable m est :

$$s^{t+1} = s^t + \min\left(\frac{\lambda_1^t - c_m^t}{\boldsymbol{\eta}_{c_m} - \boldsymbol{\eta}_{\lambda_1}}, \frac{-\lambda_1^t - c_m^t}{\boldsymbol{\eta}_{c_m} + \boldsymbol{\eta}_{\lambda_1}}\right)_+ \quad m \in \bar{\mathcal{A}}$$

avec $(x)_+ = \max(0, x)$

5. Le dernier évènement susceptible de se produire est l'annulation de la valeur de la corrélation généralisée des variables actives c'est-à-dire l'annulation de l'hyper-paramètre λ_1 associé à la contrainte L_1. Le paramètre λ_1 devient nul (considéré ici comme une critère d'arrêt) pour

$$s^{t+1} = s^t + \frac{0 - \lambda_1^t}{\boldsymbol{\eta}_{\lambda_1}}.$$

Vu qu'on parcourt le chemin de régularisation en augmentant progressivement la valeur de s, la prochaine valeur s^{t+1} à retenir correspond à la plus petite variation positive de s c'est-à-dire s^{t+1} est la plus petite valeur immédiatement supérieure à s^t. Les ensembles de points et de variables sont mis à jour en tenant compte de l'évènement qui a été détecté.

3.1.2 Initialisation de l'algorithme

Nous décrivons brièvement la procédure d'initialisation. Deux cas peuvent être considérés en fonction de la répartition des points étiquetés : le cas équilibré (la classe « positive » I_+ a le même effectif que la classe « négative » I_- c'est-à-dire $\ell_+ = \ell_-$) et le cas déséquilibré.

Cas équilibré ($\ell_+ = \ell_-$)

Pour $s^t = 0$, tous les paramètres $\boldsymbol{\beta}$ sont nuls ($\mathcal{A}$ est vide) et la frontière de décision est une droite d'équation $y = b$. La solution b n'est pas unique ; elle peut être déterminée de telle manière que tous les points étiquetés appartiennent à $\mathcal{L}$. La conséquence directe est que $\boldsymbol{\alpha} = \mathbf{1}_\ell$. La procédure d'initialisation consiste alors à identifier les variables présentant en valeur absolue la corrélation généralisée maximale $|c_m| = |K_{m,\mathcal{S}_\mathcal{L}} Y \boldsymbol{\alpha}|$. Ces variables

sont candidates pour intégrer l'ensemble actif $\mathcal{A}$ avec un signe sign(c_m). Les signes étant connus, on peut déduire $\Delta\boldsymbol{\beta}_{\mathcal{A}}$ et $\Delta\lambda_1$ en utilisant l'équation (15) de laquelle on élimine les termes relatifs à $\Delta\boldsymbol{\alpha}_{\mathcal{E}}$ et Δb. En combinant toutes ces informations, on obtient alors

$$1 - (b + (s - s^t)K_{i,\mathcal{A}}\Delta\boldsymbol{\beta}_{\mathcal{A}}) \geq 0, \quad \forall i \in I_+$$

et

$$1 + (b + (s - s^t)K_{i,\mathcal{A}}\Delta\boldsymbol{\beta}_{\mathcal{A}}) \geq 0, \quad \forall i \in I_-$$

puisque tous les points sont dans $\mathcal{L}$ et

$$f(x_i) = (s - s^t)K_{i,\mathcal{A}}\Delta\boldsymbol{\beta}_{\mathcal{A}}.$$

Les deux inégalités font intervenir deux inconnues b et s. Ces derniers paramètres sont alors déterminés de sorte qu'au moins un point de chaque classe I_+ et I_- intègre la marge $\mathcal{E}$. Le lecteur remarquera que la configuration initiale équivaut à mettre tous les points étiquetés soit à l'intérieur de la marge, soit sur la marge. L'ensemble $\mathcal{R}$ est par conséquent vide à l'initialisation.

Cas déséquilibré ($\ell_+ \neq \ell_-$)

L'initialisation dans ce cas est plus ardue. Pour déterminer les ensembles de points et de variables actives, un problème de programmation linéaire est résolu avec une valeur prédéfinie de s (généralement petite). Nous renvoyons le lecteur à l'article de Wang et al. [19] pour de plus amples détails. Dans la suite de l'article, nous nous placerons dans le contexte de la première initialisation.

3.1.3 L'algorithme du chemin de régularisation L_1

La démarche de construction du chemin de régularisation peut être résumée par les étapes de l'algorithme 1.

L'arrêt de l'algorithme peut être jugé sur différents critères : le paramètre de régularisation λ_1 s'annule, l'erreur de validation croisée des différents modèles obtenus le long du chemin atteint un minimum ou simplement le nombre maximal d'itérations a été atteint.

3.2 Complexité numérique

Au delà du coût de calcul du graphe de similarité (qui vaut $\mathcal{O}(\ell + u)^2$), du Laplacien du graphe L (voir équation (3)) et de la matrice de Gram K, l'algorithme du chemin L_1 induit d'autres coûts qui sont énumérés ci-après.

La résolution du système d'équations linéaires (15) nécessite un coût de calcul pouvant être évalué à $\mathcal{O}((|\mathcal{A}| + |\mathcal{E}| + 2)^3)$. Ce coût peut être réduit à $\mathcal{O}((|\mathcal{A}| + |\mathcal{E}| + 2)^2)$ si l'on utilise la technique de Sherman-Morrison pour la mise à jour de l'inverse de la matrice H car d'une itération à l'autre de

Algorithm 1 Algorithme du chemin L_1

Initialiser s et les ensembles $\mathcal{A}$, $\mathcal{E}$, $\mathcal{L}$ et $\mathcal{R}$. En déduire une initialisation des paramètres $\boldsymbol{\beta}_{\mathcal{A}}$, $\boldsymbol{\alpha}$, b, λ_1 et b.

Répeter

Calculer la direction de mise à jour des paramètres à savoir $\boldsymbol{\eta}_{\beta_{\mathcal{A}}}$, $\boldsymbol{\eta}_{\alpha_{\mathcal{E}}}$, $\boldsymbol{\eta}_b$ et $\boldsymbol{\eta}_{\lambda_1}$ en résolvant l'équation $H\boldsymbol{\eta} = \mathbf{z}$.

Calculer la variation unitaire des résidus $\boldsymbol{\eta}_{r_i}, i \in \mathcal{L} \cup \mathcal{R}$ et des corrélations généralisées $\boldsymbol{\eta}_{c_m}, m \in \bar{\mathcal{A}}$.

Calculer la valeur suivante de s en détectant les évènements appropriés.

Mettre à jour les paramètres du modèle en utilisant l'équation (17).

Mettre à jour les ensembles $\mathcal{A}$, $\mathcal{E}$, $\mathcal{L}$ et $\mathcal{R}$ en fonction de l'évènement détecté à l'étape 3.

Jusqu'à satisfaction d'un critère d'arrêt

l'algorithme, les systèmes d'équations ne diffèrent les uns des autres que d'une ligne ou d'un colonne à cause des mouvements des points ou des variables. La détermination de $\boldsymbol{\eta}_{r_i}$ et $\boldsymbol{\eta}_{c_m}$ implique des coûts respectifs $\mathcal{O}(|\mathcal{A}|)$ et $\mathcal{O}(|\mathcal{A}| + |\mathcal{E}|)$. Par conséquent le coût de calcul de toutes les variations de résidus et de corrélation généralisée se ramène respectivement à $\mathcal{O}(|\mathcal{A}| \times \ell)$ et $\mathcal{O}((\ell + u) \times (|\mathcal{A}| + |\mathcal{E}|))$.

La détection d'un évènement fait appel à des calculs numériques de l'ordre de $\mathcal{O}(\ell+u)$. A partir de ces éléments, on établit que la complexité numérique d'une étape de l'algorithme est approximativement de l'ordre de $\mathcal{O}((|\mathcal{A}| + |\mathcal{E}| + 2)^2 + (\ell + u) \times (|\mathcal{A}| + |\mathcal{E}|)$. Ce qui est difficile par contre, c'est la prédiction du nombre total d'étapes de l'algorithme afin de parcourir tout le chemin de régularisation. Néanmoins, ce nombre peut être estimé à un multiple peu élevé de $\ell + u$. La justification est qu'on fait appel à au moins $\ell + u$ mouvements de variables afin d'examiner les possibilités de sélection de ces variables dans le modèle. Néanmoins, si la mise en oeuvre du chemin de régularisation est couplée avec une procédure de validation de modèle (validation croisée ou autre technique), il n'est pas nécessaire de parcourir entièrement le chemin de régularisation. Un arrêt prématuré sur le chemin peut être fait afin d'obtenir une solution parcimonieuse.

4 APPLICATIONS DE L'ALGORITHME L_1

Pour évaluer les performances de l'algorithme, nous nous sommes servis de deux exemples : d'une part, des données simulées et de l'autre des données réelles issues de la base de données UCI.

4.1 Test sur des données jouets

L'algorithme du chemin de régularisation a été testé sur des données synthétiques représentant deux demi-lunes comme le montre la figure 1. Chaque demi-lune représente une classe et on constate que le problème est non-linéairement séparable. La base d'apprentissage comprend 200 points. Pour évaluer l'algorithme, la procédure expérimentale suivante a été mise en oeuvre : un nombre prédéterminé ℓ de points étiquetés est tiré de manière aléatoire et l'algorithme est ensuite testé sur ces points et les 200-ℓ points sans étiquettes. Un noyau gaussien de largeur de bande σ a été utilisé. Le graphe de similarité (ou de proximité) des points a été construit en utilisant la méthode des n-ppv (n plus proches voisins) avec $n = 7$.

L'application de l'algorithme lorsqu'on considère un point étiqueté par classe est illustrée sur la figure 1. On constate qu'on arrive à obtenir une erreur de classification nulle. Cette performance est obtenue pour 31 variables dans le modèle, ce qui correspond à un degré de parcimonie égale à 31/200=0.16.

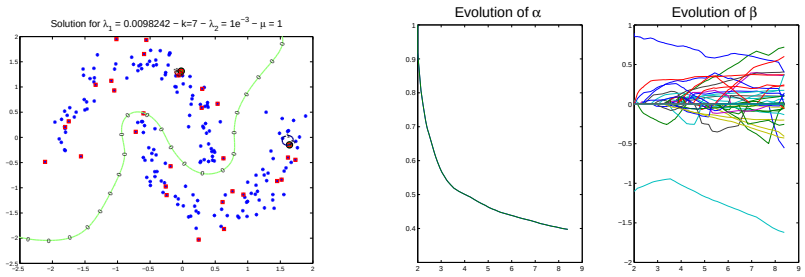


FIG. 1 – Illustration de l'algorithme sur le données demi-lunes lorsqu'on considère 1 point étiquetté par classe. Les autres graphiques montrent l'évolution des paramètres α_i et β_j au fil du chemin de régularisation.

En effet, à l'initialisation de l'algorithme, nous avons 200 variables $k(x, x_i)$ candidates au modèle. Après le parcours total du chemin de régularisation, nous sélectionnons la fonction de décision la plus parcimonieuse (c'est-à-dire celle comprenant le plus faible nombre de variables) et ayant une erreur de classification nulle aussi bien sur les points étiquetés que sur les points non étiquetés.

Nous avons répété cette procédure 10 fois pour différentes valeurs de ℓ à savoir $\ell = 4, 8, 16, 32, 64$. Les résultats obtenus sont résumés dans le tableau 1 ainsi que les paramètres utilisés pour l'algorithme.

On constate aisément que le nombre de variables sélectionnées est important pour de petites valeurs de ℓ (nombre de points étiquetés). La frontière de décision est en grande partie portée par les variables $k(x, x_i)$ liées aux points

ℓ	4	8	16	32	64
σ	0.5	0.5	0.5	0.5	0.5
λ_2	0.001	0.01	0.01	0.01	0.01
μ	1	1	1	1	1
$\lvert\mathcal{A}\rvert$	48.87 (12.58)	32.2 (11.69)	28.4 (12.81)	23.8 (6.54)	22.5 (2.27)

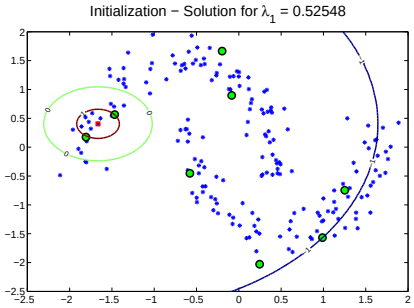
TAB. 1 – Illustration de l'algorithme sur le données demi-lunes. Récapitulation du nombre de variables sélectionnées en fonction du nombre de points étiquetés. Sont indiqués dans le tableau, les nombres moyens de variables sélectionnées avec les écart-types correspondants. Les résultats ont été moyennés sur 10 expériences.

non étiquetés x_i comme le montre la figure 2. Cette figure illustre l'évolution de la frontière de décision sur quelques étapes de l'algorithme. On constate que la solution à l'étape initiale est mauvaise en termes de bonne classification ainsi que la solution intermédiaire après une dizaine d'itérations. Par contre la solution finale montre une fonction de décision réalisant une classification parfaite.

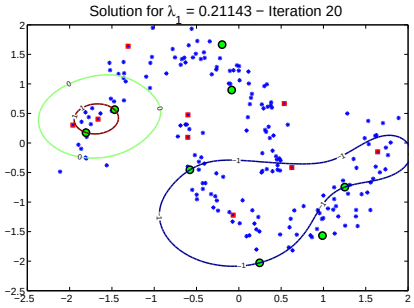
Dans le tableau 1, la variabilité relativement importante du nombre de variables sélectionnées $\mathcal{A}$ peut s'expliquer par la position des points étiquetés sélectionnés sur la variété. Si ces points couvrent de façon appropriée la variété, il apparaît qu'un faible nombre de variables est nécessaire pour construire une fonction de décision performante. On remarquera pour $\ell = 4$, une importance accrue est donnée à la pénalisation sur la régularité de la variété puisque λ_2 a été fixé à 0.001 dans ce cas alors que ce paramètre a été fixé à 0.01 dans les autres cas.

4.2 Test sur des données réelles (données USPS)

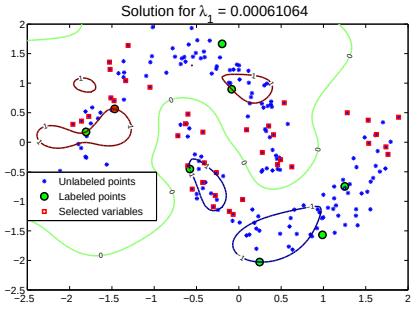
L'algorithme a été également testé sur des images représentant des chiffres manuscrits. Ces données sont issues de la base UCI bien connue en apprentissage statistique. Pour évaluer les performances de l'algorithme, nous avons testé un problème de classification semi supervisé consistant à classer les chiffres '2' et '5'. La base d'apprentissage contient 847 points et la base de test 1274 points. Un noyau gaussien de paramètre $\sigma = 10$ a été utilisé. Les autres paramètres de régularisation ont été respectivement fixés aux valeurs suivantes : $\lambda_2 = 0.001$ et $\mu = 0.1$. Comme dans l'exemple précédent, nous sélectionnons de manière aléatoire dans la base d'apprentissage un certain de points à étiqueter par classe. L'algorithme est ensuite lancé dans cette configuration et il est stoppé quand ses performances en test équivalent les performances de l'algorithme initial du Laplacien SVM (qui comprend 847 variables et dont les résultats sont préalablement calculés). Le tableau 2 donne un récapitulatif des résultats obtenus qui sont moyennés sur 7 tests de l'algorithme pour chaque valeur de ℓ. Comme on peut le constater, l'al-



a) Initialisation



b) Evolution de la frontière de décision après quelques itérations



c) Solution finale parcimonieuse

FIG. 2 – Illustration de l'algorithme sur le données demi-lunes. La solution parcimonieuse fournit une erreur de classification nulle pour seulement 40 variables contre 200 variables que nécessite l'algorithme initial du Laplacien SVM. Les points étiquetés sont des cercles alors que les points correspondant aux variables sélectionnées sont des carrés rouges. Ces derniers points ont tendance à couvrir les variétés.

gorithme atteint des niveaux de performance similaires au Laplacien SVM avec un nombre réduit de variables.

ℓ	16	32	64
$\lvert\mathcal{A}\rvert$	132	126	105.87
$\widehat{\sigma}_{\lvert\mathcal{A}\rvert}$	104.17	154.18	114
Erreur de test	19 (8.73)	13 (2.78)	9 (3.33)

TAB. 2 – Résultats de l'algorithme sur les images des chiffres manuscrits de la base USPS. Apprentissage de la classe '2' contre la classe '5'. L'erreur de classification fournie est une moyenne des points mal classés. Les écart-types sont indiqués entre parenthèses.

La même expérience a été réalisée pour la classification des chiffres 1 contre 7. Les résultats sont compilés dans le tableau 3. Dans cet exemple, on a considéré $\ell + u = 954$ points d'apprentissage. On constate ici aussi que la réduction du nombre de variables est significative.

TAB. 3 – Résultats de la classification des chiffres 1 contre 7.

ℓ	16	32	64
$\lvert\mathcal{A}\rvert$	348 (118.7)	317 (37.42)	408 (190.40)
Test error	36.8 (23.63)	23.25 (6.19)	13.75 (2.63)

5 DICUSSIONS

Pour déterminer le chemin de régularisation, nous avons supposé les paramètres λ_2 et μ (ou ρ, voir équation 11) sont connus. Si on s'intéresse à l'analyse de l'influence de l'hyper-paramètre λ_2 sur la solution $f(x)$, on peut établir aussi un chemin de régularisation que nous nommerons chemin L_2. Nous donnons ci-dessous les détails d'élaboration de ce chemin.

5.1 Principe du chemin L_2

Pour ce faire, fixons [2] ρ et λ_1 et reprenons les équations (6) et (12-13). En exprimant de façon explicite la dépendance de $P_{\mathcal{A}}$ par rapport à λ_2 dans la relation (12), on aboutit à l'expression suivante

$$\lambda_2 R_{\mathcal{A},\mathcal{A}} \boldsymbol{\beta}_{\mathcal{A}} - K_{\mathcal{A},\mathcal{S}_{\mathcal{L}}} Y \boldsymbol{\alpha} + \lambda_1 \text{sign}\,(\boldsymbol{\beta}_{\mathcal{A}}) \quad = \quad 0$$

où la matrice R a été définie en (11). Avec les changements de variables $\tilde{\lambda}_2 = \frac{1}{\lambda_2}$ et $\tilde{\alpha}_i = \tilde{\lambda}_2 \alpha_i$, on peut reécrire l'équation précédente sous la forme

$$R_{\mathcal{A},\mathcal{A}} \boldsymbol{\beta}_{\mathcal{A}} - K_{\mathcal{A},\mathcal{S}_{\mathcal{L}}} Y \tilde{\boldsymbol{\alpha}} \quad = \quad -\tilde{\lambda}_2 \ \lambda_1 \text{sign}\,(\boldsymbol{\beta}_{\mathcal{A}}) \tag{18}$$

[2] Fixer λ_1 correspond à fixer s dans le problème d'optimisation (10). Une grande valeur de λ_1 correspond à s petit et inversement

Ici le raisonnement sera légèrement différent par rapport au chemin L_1 car certains paramètres vont varier linéairement par rapport à λ_2 et d'autres en raison inverse de λ_2. En effet, supposons que pour une valeur λ_2^t (correspondant à $\tilde{\lambda}_2^t$) à l'itération t, les solutions $\boldsymbol{\beta}^t$, $\boldsymbol{\alpha}^t$, b^t et les ensembles $\mathcal{E}^t, \mathcal{R}^t, \mathcal{L}^t, \mathcal{A}^t$ et $\bar{\mathcal{A}}^t$ sont connus. Sous l'hypothèse que ces ensembles ne sont pas modifiés lorsqu'on passe de $\tilde{\lambda}_2^t$ à $\tilde{\lambda}_2$, on peut établir par différentiation respective des équations (18), (13) et (6) les relations

$$R_{\mathcal{A},\mathcal{A}} \frac{\Delta \boldsymbol{\beta}_{\mathcal{A}}}{\Delta \tilde{\lambda}_2} - K_{\mathcal{A},\mathcal{S}_{\mathcal{L}}} Y \frac{\Delta \tilde{\boldsymbol{\alpha}}}{\Delta \tilde{\lambda}_2} = -\lambda_1 \text{sign}\left(\boldsymbol{\beta}_{\mathcal{A}}\right) \quad (19)$$

$$Y_{\mathcal{E}} \left(K_{\mathcal{E},A} \frac{\Delta \boldsymbol{\beta}_{\mathcal{A}}}{\Delta \tilde{\lambda}_2} + \mathbf{1}_{\mathcal{E}} \frac{\Delta b}{\Delta \tilde{\lambda}_2} \right) = 0 \quad (20)$$

$$\mathbf{y}_{\mathcal{E}}^{\top} \frac{\Delta \boldsymbol{\alpha}_{\mathcal{E}}}{\Delta \lambda_2} = 0 \quad (21)$$

avec $\Delta\tilde{\lambda}_2 = \tilde{\lambda}_2 - \tilde{\lambda}_2^t$. On remarquera que les variations des β_j, $\tilde{\alpha}_i$ et b sont pris par rapport à $\tilde{\lambda}_2$ contrairement à α_i. On montre (voir annexe) facilement que

$$\frac{\Delta \tilde{\alpha}_i}{\Delta \tilde{\lambda}_2} = \alpha_i^t - \lambda_2^t \frac{\Delta \alpha_i}{\Delta \lambda_2}$$

En injectant cette relation dans l'équation (19), on établit finalement

$$R_{\mathcal{A},\mathcal{A}} \frac{\Delta \boldsymbol{\beta}_{\mathcal{A}}}{\Delta \tilde{\lambda}_2} + \lambda_2^t K_{\mathcal{A},\mathcal{E}} Y_{\mathcal{E}} \frac{\Delta \boldsymbol{\alpha}_{\mathcal{E}}}{\Delta \lambda_2} = \lambda_2 R_{\mathcal{A},\mathcal{A}} \boldsymbol{\beta}_{\mathcal{A}}^t \quad (22)$$

En combinant ensemble les équations (20), (21) et (22), on s'aperçoit qu'on peut calculer les variations unitaires des paramètres à partir d'un système d'équations linéaires à $|\mathcal{E}|+|\mathcal{A}|+1$ équations et $|\mathcal{E}|+|\mathcal{A}|+1$ inconnues que sont $\frac{\Delta \boldsymbol{\beta}_{\mathcal{A}}}{\Delta \tilde{\lambda}_2}$, $\frac{\Delta \boldsymbol{\alpha}_{\mathcal{E}}}{\Delta \lambda_2}$ et $\frac{\Delta b}{\Delta \tilde{\lambda}_2}$. Notons $\boldsymbol{\gamma}$ le vecteur contenant ces variations unitaires. On en déduit alors que lorsque l'hyper-paramètre λ_2 varie, les paramètres $\boldsymbol{\alpha}$ varient linéairement par rapport à λ_2 alors que les paramètres $\boldsymbol{\beta}$ et b varient linéairement en fonction de $\tilde{\lambda}_2$

A partir de cette variation linéaire, il est alors possible de mener le même type de raisonnement que pour le chemin L_1 et d'en déduire les conditions de réalisation des quatre évènements précédemment analysés.

5.2 Détection des évènements pour le chemin L_2

– Un point étiquetté x_i passe de $\mathcal{E}$ à $\mathcal{L}$ ou $\mathcal{R}$
Les valeurs du paramètre de régularisation correspondant à ces évènements sont données par

$$\lambda_2 = \lambda_2^t + \frac{1-\alpha_i^t}{\gamma_{\alpha_i}} \quad \text{ou} \quad \lambda_2 = \lambda_2^t + \frac{0-\alpha_i^t}{\gamma_{\alpha_i}}$$

– Un point $x_i \in \mathcal{L} \cup \mathcal{R} \rightarrow \mathcal{E}$
Ce mouvement correspond à l'annulation du résidu $r_i = 1 - y_i(f(x_i) + b)$ pour ces points. On en déduit

$$\tilde{\lambda}_2 = \tilde{\lambda}_2^t + \frac{0 - r_i^t}{\gamma_{r_i}} \quad \text{avec} \quad \gamma_{r_i} = -y_i(K_{i,\mathcal{A}}\gamma_{\beta_\mathcal{A}} + \gamma_b).$$

– $\beta_j \in \mathcal{A} \rightarrow \bar{\mathcal{A}}$
L'évènement se produit pour $\tilde{\lambda}_2 = \tilde{\lambda}_2^t + \frac{0-\beta_i^t}{\gamma_{\beta_i}}$
– $\beta_j \in \bar{\mathcal{A}} \rightarrow \mathcal{A}$
On sait que cet évènement est lié à la corrélation généralisée des variables inactives qui s'exprime pour la variable m comme

$$c_m = R_{m,\mathcal{A}}\boldsymbol{\beta}_\mathcal{A} - K_{m,\mathcal{S}_\mathcal{L}}\ Y\ \tilde{\boldsymbol{\alpha}}$$

en vertu de (18). La variation unitaire est donc

$$\gamma_{c_m} = R_{m,\mathcal{A}}\ \gamma_{\beta_\mathcal{A}} - K_{m,\mathcal{S}_\mathcal{L}}\ Y\ \boldsymbol{\gamma}_{\tilde{\alpha}}$$

L'expression de $\boldsymbol{\gamma}_{\tilde{\alpha}}$ est établie dans les annexes. D'après (18), la variable devient active si sa corrélation généralisée vérifie $|c_m^{t+1}| = \lambda_1\tilde{\lambda}_2$ où $c_m^{t+1} = c_m^t + \gamma_{c_m}(\tilde{\lambda}_2 - \tilde{\lambda}_2^t)$. Il en découle la valeur de l'hyper-paramètre correspondant à cet évènement

$$\tilde{\lambda}_2 = \min\left(\frac{\gamma_{c_m}\tilde{\lambda}_2^t - c_m^t}{\gamma_{c_m} - \lambda_1}, \frac{\gamma_{c_m}\tilde{\lambda}_2^t - c_m^t}{\gamma_{c_m} + \lambda_1}\right)_+, \quad \forall m \in \bar{\mathcal{A}}$$

On aura noté que certains évènements fournissent directement λ_2 et d'autres $\tilde{\lambda}_2$. A l'itération suivante, on retiendra la valeur de λ_2 immédiatement supérieure ou inférieure à λ_2^t selon le sens dans lequel on désire parcourir le chemin de régularisation.

L'algorithme du chemin L_2 est similaire à celui du chemin L_1. Sa complexité numérique est également très proche car le système linéaire à résoudre comporte une inconnue et une équation en moins.

Notons pour clore cette partie que le passage du chemin L_1 au chemin L_2 est immédiat. En effet, à chaque étape, sur l'un ou l'autre chemin, si on dispose de tous les ensembles et des paramètres correspondants, on peut analyser soit l'évolution de ces paramètres par rapport à s, soit par rapport à λ_2. Dans cette configuration, le chemin de régularisation est en fait une surface qu'il n'est pas toujours aisé d'explorer.

6 CONCLUSION

Dans cet article nous avons décrit une approche de solution de problème semi supervisé qui combine l'approche SVM et la prise en compte de la

forme géométrique décrite par les données. Cet algorithme appelé Laplacien SVM a pour inconvénient d'exprimer la solution du problème en fonction de tous les points d'apprentissage. Pour obtenir la parcimonie de la solution, nous avons proposé d'adjoindre au problème initial une pénalisation de type L_1 sur les paramètres de la fonction de décision. Le réglage du compromis entre la parcimonie et l'erreur de classification n'étant pas aisé, nous avons proposé une méthode efficace de calcul de l'ensemble des fonctions de décision parcimonieuse lorsqu'on relâche progressivement la contrainte de parcimonie : c'est le chemin de régularisation. Pour montrer l'intérêt de notre approche, des tests de l'algorithme ont été réalisés sur des données simulées et des données réelles. Ces tests montrent sans conteste qu'il est possible d'obtenir des niveaux de performances similaires à l'algorithme initial du Laplacien SVM mais avec un nombre réduit de variables.

Pour parcourir le chemin de régularisation correspondant à la parcimonie, les paramètres de régularisation λ_2 et μ ont été pris de manière ad hoc. L'examen de l'évolution de la fonction de décision lorsque varient ces paramètres peut également être fait en se basant sur un autre chemin de régularisation dont nous avons précisé les différents ingrédients. La combinaison du calcul de ces chemins conduit à une surface de régularisation dont nous étudions actuellement la meilleure stratégie pour son exploration efficiente.

RÉFÉRENCES

[1] M.-R. Amini et P. Gallinari. Semi-supervised learning with explicit misclassification modeling. In *Proc of 18th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 555–560, 2003.

[2] M. Belkin, P. Niyogi et V. Sindhwani. Manifold regularization : a geometric framework for learning from label and unlabeled examples. *Journal of Machine Learning Research*, 1 :1–48, 2006.

[3] O. Chapelle, B. Schölkopf et A. Zien. *Semi-Supervised Learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass., USA, 09 2006.

[4] O. Chapelle, V. Sindhwani et S. Keerthi. Optimization techniques for semi-supervised support vector machines. *Journal of Machine Learning Research*, 9 :203–233, 2008.

[5] O. Chapelle et A. Zien. Semi-supervised classification by low density separation. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 2005.

[6] R. Collobert, F. Sinz, J. Weston et L. Bottou. Large scale transductive svms. *Journal of Machine Learning Research*, 7 :1687–1712, 2006.

[7] B. Efron, T. Hastie, T. Johnstone et R. Tibshirani. Least angle regression. *Annals of Statistics*, 32 :407–499, 2004.

[8] A. Fujino, N. Ueda et K. Saito. A hybrid generative/discriminative approach to semi-supervised classifier design. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, 2005.

[9] Glenn Fung, Olvi Mangasarian et Jude Shavli. Knowledge-based nonlinear kernel classifiers. In Manfred Warmuth et Bernhard Schölkopf, éditeurs, *Conference On Learning Theory (COLT 03) and Workshop on Kernel Machines*, pages 102–113. Springer Verlag, Berlin, August 2003.

[10] T. Gal. *Postoptimal analyses, parametric programming and related topics*. McGraw-Hill, New-York, USA, 1979.

[11] Trevor Hastie, Saharon Rosset, Robert Tibshirani et Ji Zhu. The entire regularization path for the support vector machine. *JMLR*, 5 :1391–1415, 2004.

[12] I. Heller. Sensitivity analysis in linear programming. Technical report, George Washington University, 1954.

[13] H. Markowitz. The optimization of a quadratic function subject to linear constraints. *Naval Research Logistics Quarterly*, 3 :111–133, 1956.

[14] K. Nigam, A. mcCallum, A. Thurn et T. Mitchell. Text classification from labeled and unlabeled documents using em. *machine Learning*, 39(2/3) :127–163, 2000.

[15] Saharon Rosset et Ji Zhu. Piecewise linear regularized solution paths. *Annals of Statistics*, 35(3) :1012–1030, 2007.

[16] Bernhard Schölkopf et Alexander J. Smola. *Learning with Kernels*. MIT Press, 2002.

[17] V. Vapnik. *Statistical learning theory*. John Wiley & Sons, 1998.

[18] G. Wang, T.C. Yeund et F. H. Lochovsky. Solution path of semi-supervised classification with manifold regularization. In *Proceedings of 6th International Conference on Data Mining*, pages 1124–1129, 2006.

[19] L. Wang, J. Zhu et H. Zou. The doubly regularized support vector machine. *Statistica Sinica*, 16 :589–615, 2006.

[20] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005. http ://www.cs.wisc.edu/∼jerryzhu/pub/ssl_survey.pdf.

ANNEXES

Lien entre $\Delta\tilde{\lambda}_2$ et $\Delta\lambda_2$

On a la relation évidente

$$\begin{aligned} \Delta\tilde{\lambda}_2 &= \frac{1}{\lambda_2} - \frac{1}{\lambda_2^t} \\ \Delta\tilde{\lambda}_2 &= -\frac{\Delta\,\lambda_2}{\lambda_2\,\lambda_2^t} \quad \text{avec} \quad \Delta\lambda_2 = \lambda_2 - \lambda_2^t \end{aligned} \tag{23}$$

Lien entre $\frac{\Delta\tilde{\alpha_i}}{\Delta\tilde{\lambda}_2}$ et $\frac{\Delta\alpha_i}{\Delta\lambda_2}$

La variation de $\Delta\tilde{\alpha}_i$ est fournie par

$$\begin{aligned} \Delta\tilde{\alpha}_i &= \frac{\alpha_i}{\lambda_2} - \frac{\alpha_i^t}{\lambda_2^t} \\ &= \frac{\lambda_2^t(\alpha_i - \alpha_i^t) - (\lambda_2 - \lambda_2^t)\alpha_i^t}{\lambda_2\,\lambda_2^t} \\ \Delta\tilde{\alpha}_i &= \frac{\lambda_2^t\Delta\alpha_i - \Delta\lambda_2\,\alpha_i^t}{\lambda_2\,\lambda_2^t} \end{aligned}$$

En se basant sur l'équation (23), on peut alors établir que

$$\frac{\Delta\tilde{\alpha_i}}{\Delta\tilde{\lambda}_2} = \alpha_i^t - \lambda_2^t\frac{\Delta\alpha_i}{\Delta\lambda_2}$$

On en déduit alors la relation entre les variations unitaires

$$\gamma_{\tilde{\alpha}_i} = \alpha_i^t - \lambda_2^t\gamma_{\alpha_i}$$

Le lecteur notera que γ_{α_i} pour tous les points $x_i \in \mathcal{L} \cup \mathcal{R}$. En revanche $\gamma_{\tilde{\alpha}_i}$ n'est nul que pour les points $x_i \in \mathcal{R}$.