Apprentissage statistique
SVM

Jean-Michel Marin

Université de Montpellier
Institut Montpelliérain Alexander Grothendieck (IMAG)

HMMA303

Introduction

Support Vector Machines (SVM)

SVM attaque le problème de classification entre deux classes de façon directe

On essaie de trouver un (hyper)plan qui sépare les classes dans l'espace des prédicteurs

- on adoucit la définition de « séparer » et
- on enrichit l'espace des covariables de sorte à rendre la séparation possible

Introduction

Qu'est-ce qu'un hyperplan ?

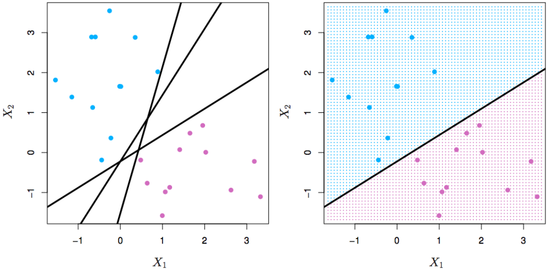
- Un hyperplan d'un espace de dimension p est un sous-espace affine de dimension $p-1$.
- En général, l'équation d'un hyperplan est de la forme

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = 0$$

- En dimension $p = 2$, un hyperplan est une droite.
- Si $\beta_0 = 0$, il passe par l'origine.

Introduction

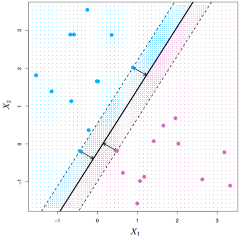
Hyperplan de séparation



- Si $f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, alors $f(X) > 0$ pour les points d'un côté de l'hyperplan et $f(X) < 0$ de l'autres côté
- Si l'on suppose que les points $Y_i = +1$ sont en bleu et $Y_i = -1$ en mauve, alors $Y_i \cdot f(X_i) > 0$ pour tout i. De plus $f(X) = 0$ définit un *hyperplan séparateur*

Classifieur de marge maximale

Parmi tout les hyperplans séparateurs, on cherche celui qui crée le plus grand écart (ou la plus grande marge) entre les deux classes



Problème d'optimisation sous contraintes qui maximise M, fonction de $\beta_0, \ldots, \beta_p$, sous les contraintes $\sum_j \beta_j^2 = 1$, et $y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geqslant M$ pour tout i

Ce problème peut se voir comme une question de programmation convexe quadratique et résolue efficacement. La fonction `svm()` du package `e1071` résout ce problème efficacement.

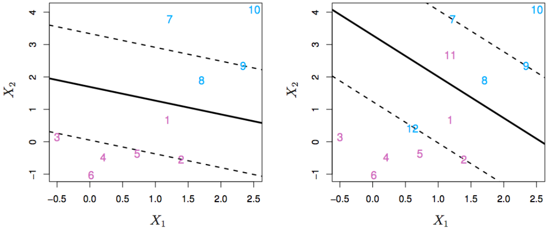
Classifieur de marge maximale

Données non séparables et données bruitées

- Parfois, les données ne sont pas séparables par un hyperplan. C'est souvent le cas, sauf si $n < p$.
- Souvent, les données sont séparables, mais bruitées. Un classifieur construit en maximisant la marge de l'hyperplan séparateur peut avoir une mauvaise erreur de test.

Pour toutes ces raisons, *le classifieur SVM* maximise une marge « douce ».

SVM classifieur



$$\text{maximiser}_{\beta_0,\ldots\beta_p} M \text{ sous les contraintes } \sum_j \beta_j^2 = 1,$$

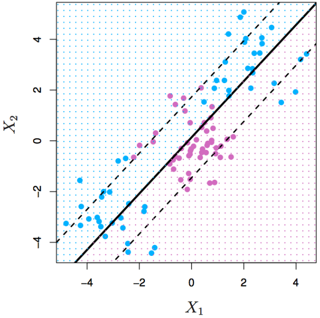
$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geqslant M(1 - \varepsilon_i)$$

$$\varepsilon_i \geqslant 0, \sum_{i=1}^n \varepsilon_i \leqslant C.$$

Le paramètre C est un paramètre de régularisation.

SVM classifieur

Une frontière linéaire n'est pas toujours une bonne idée



Parfois, il est simplement abscons d'utiliser une frontière linéaire, quelle que soit la valeur de C que l'on utilise.

SVM classifieur

Extension de l'espace des covariables

- Agrandir les espaces des covariables en ajoutant des transformations non linéaires des X_j (polynomiales, etc.) On passe alors d'un espace de dimension p à un espace de dimension M.
- Ajuster un classifieur SVM dans l'espace agrandi
- Cela fournit une frontière de décision non-linéaire dans l'espace original.

Exemple Si on utilise $(X_1, X_2, X_1^2, X_2^2, X_1X_2)$ au lieu de (X_1, X_2), alors la frontière entre les deux décisions est de la forme

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 = 0$$

c'est-à-dire une conique.

SVM classifieur

Non-linéarités et noyaux

- Les polynômes (en particulier en grande dimension) devient rapidement incontrôlables.
- Il y a une façon plus élégante et mesurée d'introduire des non-linéarités dans les classifiers SVM, via l'utilisation de *noyaux*
- Avant de présenter ce point, nous devons comprendre le rôle du *produit scalaire* dans les SVM.

SVM classifieur

Produit scalaire et SVM

$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}$ est le produit scalaire entre les vecteurs x_i et $x_{i'}$

- Le classifieur construit par SVM peut s'écrire

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle$$

  avec n paramètres
- Pour estimer ces paramètres α_i et β_0, tout ce dont on a besoin est l'ensemble des $n(n-1)/2$ produits scalaires $\langle x_i, x_{i'} \rangle$ entre paires d'observations de l'ensemble d'entrainement

SVM classifieur

Il s'avère que la plupart des $\hat{\alpha}_i$ peuvent être nuls :

$$\hat{f}(x) = \beta_0 + \sum_{i \in \mathscr{S}} \hat{\alpha}_i \langle x, x_i \rangle$$

$\mathscr{S}$ est l'ensemble support des indices i tels que $\hat{\alpha}_i > 0$.

Noyaux et SVM

- Si l'on peut calculer les produits scalaires entre observations, on peut ajuster un classifier SVM. Celui ci peut être très abstrait.
- Quelques fonctions noyaux très particulières le font pour nous. Par exemple

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d$$

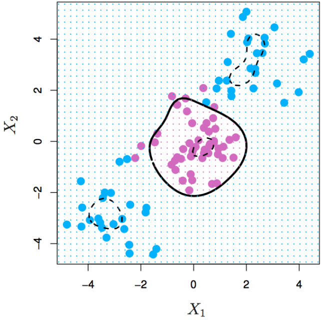
  calcule les produits scalaires nécessaires pour des polynômes de degré d.
- La solution est de la forme

$$\hat{f}(x) = \beta_0 + \sum_{i \in \mathscr{S}} \hat{\alpha}_i \langle x, x_i \rangle$$

Noyaux et SVM

Noyaux radiaux

$$K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2\right)$$



$\hat{f}(x) = \beta_0 + \sum_{i \in \mathscr{S}} \hat{\alpha}_i \langle x, x_i \rangle$

L'espace des covariables complétées est implicite, mais de très grande dimension.

Contrôle la variance en annulant la plupart des dimensions sévérement.

SVM et régression logistique

Avec $f(X) = \beta_0 + \sum_j \beta_j X_j$, on peut ré-écrire l'optimisation au coeur de SVM comme

$$\text{minimize}_{(\beta_j, j=1\ldots p)} \left\{ \sum_{i:1}^{n} \max\left[0, 1 - y_i\, f(x_i)\right] + \lambda \sum_j \beta_j^2. \right\}$$

- Classes bien séparées : SVM souvent meilleur que la régression logistique.
- Lorsque ce n'est pas le cas, la régression logistique (avec pénalité ridge) et SVM sont similaires.