Apprentissage supervisé
Cours 2 : Support Vector Machine (SVM)

Agathe Guilloux, Geneviève Robin

Plan

Plan

Le problème de classification binaire

On a des données d'apprentissage (learning data) pour des individus $i = 1, \dots, n$.
Pour chaque individu i :

- on a un vecteur de covariables (features) $x_i \in \mathcal{X} \subset \mathbb{R}^d$
- la valeur de son label $y_i \in \{-1, 1\}$.
- on suppose que les couples (X_i, Y_i) sont des copies i.i.d. de (X, Y) de loi inconnue et que l'on observe leurs réalisations (x_i, y_i) $(i = 1, \dots, n)$.

But

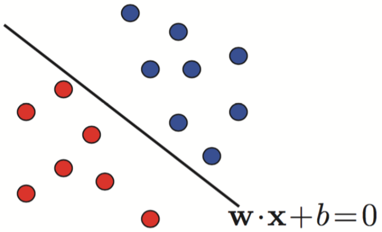
- On veut, pour un nouveau vecteur X_+ de features, prédire la valeur du label Y_+ par $\hat{Y}_+ \in \{-1, 1\}$
- Pour cela, on utilise les données d'apprentissage $\mathcal{D}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ pour construire un **classifieur** $\hat{c}$ de telle sorte que
$$\hat{Y}_+ = \hat{c}(X_+).$$
et $\hat{Y}$ est proche de Y_+ (dans un sens à préciser).

Linéairement séparable

Linéairement séparable

Un jeu de données est **linéairement séparable** si on peut trouver un hyperplan affine H tel que

- les points $x_i \in \mathbb{R}^d$ tels que $y_i = 1$ sont d'un côté de H
- les points $x_i \in \mathbb{R}^d$ tels que $y_i = -1$ sont de l'autre côté
- H ne contient aucun x_i



Rappel

Hyperplan affine

L'hyperplan affine défini par son équation normale

$$H = \{x \in \mathbb{R}^d : \langle w, x \rangle + b = 0\}$$

est une translation de b de l'ensemble de vecteur orthogonaux à w où

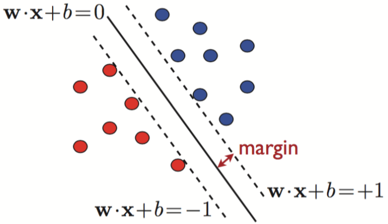
- $w \in \mathbb{R}^d$ est un vecteur non-nul normal
- $b \in \mathbb{R}$.

Par définition, H est invariant par multiplication de l'équation normale par un scalaire non-nul.

Cas linéairement séparable

Comme ici aucun x_i n'est dans H, on peut choisir w et b de telle sorte que

$$\min_{i=1,\dots,n} |\langle w, x_i \rangle + b| = 1$$



On parlera d'hyperplan **canonique**.

Point correctement classifié

x_i est correctement classifié si

$$y_i(\langle w, x_i \rangle + b) \geq 1.$$

Marge

La distance de tout point $x' \in \mathbb{R}^d$ à H est donnée par

$$\frac{|\langle w, x' \rangle + b|}{\|w\|}$$

Quand H est l'hyperplan canonique, sa **marge** est donnée par

$$\min_{i \in 1,\dots,n} \frac{|\langle w, x_i \rangle + b|}{\|w\|} = \frac{1}{\|w\|}.$$

SVM linéaire dans le cas séparable

On veut donc resoudre le problème suivant

- maximiser la marge
- en classifiant correctement les observations

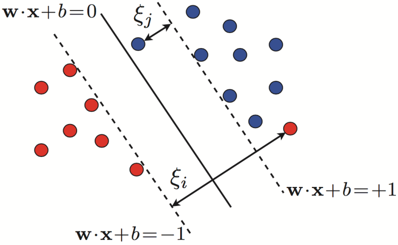
c'est équivalement à

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2} \|w\|_2^2$$
$$\text{sous contrainte} \quad y_i(\langle w, x_i \rangle + b) \geq 1 \ \text{ pour tout } \ i = 1, \ldots, n.$$

En pratique, ce n'est pas raisonnable de supposer que le jeu de données est linéairement séparable !

Plan

SVM linéaire dans le cas non-séparable (1)



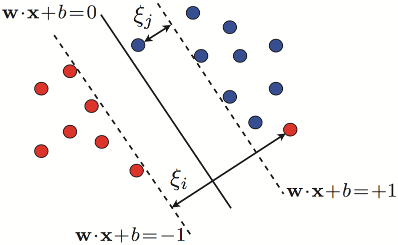
On va remplacer les contraintes

$$y_i(\langle w, x_i \rangle + b) \geq 1 \quad \text{pour tout} \quad i = 1, \ldots, n,$$

par des contraintes plus souples (relaxées).

Marge souple (soft margin)

$$y_i(\langle w, x_i \rangle + b) \geq 1 - s_i \quad \text{pour tout} \quad i = 1, \ldots, n, \quad \text{avec } s_1, \ldots, s_n \geq 0$$



- Le “slack” $s_i \geq 0$ mesure la distance par laquelle x_i viole l'inégalité
- Si $s_i = 0$ alors i est correctement classifié.
- $s_i \in]0, 1]$ alors i est correctement classifié mais est un outlier.
- Si $s_i > 1$ alors i n'est pas correctement classifié.
- Si on enlève les i pour lesquels $s_i > 0$, on se ramène au problème séparable.

SVM linéaire dans le cas non-séparable (2)

On remplace donc le problème de minimisation par

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}, s\in\mathbb{R}^n} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^n s_i$$

$$\text{sous contrainte} \quad y_i(\langle w, x_i\rangle + b) \geq 1 - s_i \;\; \text{et} \;\; s_i \geq 0 \;\; \text{pour tout} \;\; i = 1, \ldots, n$$

où $C > 0$ est le "goodness-of-fit strength".

- Ce problème admet une solution unique.
- La constante C doit être choisie par cross-validation.

Plan

Écriture lagrangienne

$$L(w, b, s, \alpha, \beta) = \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} s_i$$
$$+ \sum_{i=1}^{n} \alpha_i \big(1 - s_i - y_i(\langle w, x_i\rangle + b)\big) - \sum_{i=1}^{n} \beta_i s_i$$

A l'optimum, on va écrire les conditions KKT et la condition complémentaire.

Optimum

Conditions KKT

$$\nabla_w L(w, b, s, \alpha, \beta) = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \quad \text{i.e.} \quad w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\nabla_b L(w, b, s, \alpha, \beta) = -\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{i.e.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\nabla_s L(w, b, s, \alpha, \beta) = C - \alpha_i - \beta_i = 0 \quad \text{i.e.} \quad \alpha_i + \beta_i = C$$

Condition complémentaire

$$\alpha_i \big(1 - s_i - y_i(\langle w, x_i \rangle + b)\big) = 0 \text{ i.e. } \alpha_i = 0 \;\text{ ou }\; y_i(\langle w, x_i \rangle + b) = 1 - s_i$$

$$\beta_i s_i = 0 \quad \text{i.e.} \quad \beta_i = 0 \;\text{ ou }\; s_i = 0$$

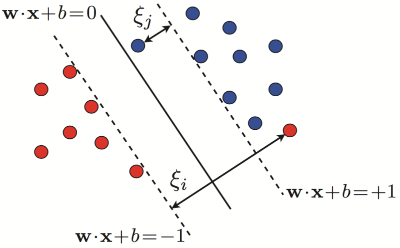
pour tout $i = 1, \ldots, n$

Optimum

On obtient

- $w = \sum_{i=1}^{n} \alpha_i y_i x_i$
- Si $\alpha_i \neq 0$, on dit que x_i est un vecteur de support ("support vector") et $y_i(\langle w, x_i \rangle + b) = 1 - s_i$
  - Si $s_i = 0$ alors x_i appartient à l'hyperplan marginal
  - Si $s_i \neq 0$ alors x_i est un outlier et $\beta_i = 0$ et donc $\alpha_i = C$

Les "support vectors" appartiennent soit à l'hyperplan marginal, ou sont des outliers avec $\alpha_i = C$.



Plan

Classifieur

La règle de classification s'exprime alors comme

$$x \mapsto \mathrm{signe}(\langle w, x\rangle + b) = \mathrm{signe}\Big(\sum_{i=1}^{n} \alpha_i y_i \langle x, x_i\rangle + b\Big)$$

L'intercept b peut s'exprimer pour un "support vector" x_i tel que $0 < \alpha_i < C$ comme

$$b = y_i - \sum_{j=1}^{n} \alpha_j y_j \langle x_i, x_j\rangle.$$

Lien avec le “hinge loss”

On peut récrire le problème

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}, s \in \mathbb{R}^n} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} s_i$$

$$\text{s.c.} \quad y_i(\langle w, x_i \rangle + b) \geq 1 - s_i \;\; \text{et} \;\; s_i \geq 0 \;\; \text{pour tout} \;\; i = 1, \ldots, n$$

comme

$$\operatorname*{argmin}_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \max\Big(0, 1 - y_i(\langle w, x_i \rangle + b)\Big).$$

Plan

Problème dual (1)

- Si on remplace w par $\sum_{i=1}^n \alpha_i y_i x_i$ dans $L(w, b, s, \alpha, \beta)$, on obtient

$$\sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

- avec les contraintes

$$\alpha_i \geq 0, \quad \beta_i \geq 0, \quad \sum_{i=1}^n \alpha_i y_i = 0, \alpha_i + \beta_i = C$$

  ce qui se réécrit

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^n \alpha_i y_i = 0$$

  pour tout $i = 1, \ldots, n$.

Problème dual (2)

On obtient alors le problème dual

$$\max_{\alpha \in \mathbb{R}^n} \qquad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

$$\text{sous contrainte} \quad 0 \leq \alpha_i \leq C \;\text{ et }\; \sum_{i=1}^{n} \alpha_i y_i = 0 \;\text{ pour tout }\; i = 1, \ldots, n$$

Une remarque très importante

Le problème dual

$$\max_{\alpha \in \mathbb{R}^n} \qquad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

$$\text{sous la contrainte} \quad 0 \leq \alpha_i \leq C \;\; \text{et} \;\; \sum_{i=1}^{n} \alpha_i y_i = 0 \;\; \text{pour tout} \;\; i = 1, \ldots, n$$

et le classifieur

$$x \mapsto \text{signe}(\langle w, x \rangle + b) = \text{signe}(\sum_{i=1}^{n} \alpha_i y_i \langle x_i, x \rangle + b)$$

ne dépendent que des features x_i via les produits scalaire $\langle x_i, x_j \rangle$!

Plan

Feature engineering

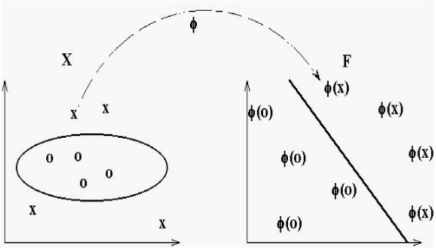
- A partir des $x_1, \ldots, x_n \in \mathbb{R}^d$, on peut construire de **nouvelles** features
- Par exemple, en considérant des polynômes d'ordre 2

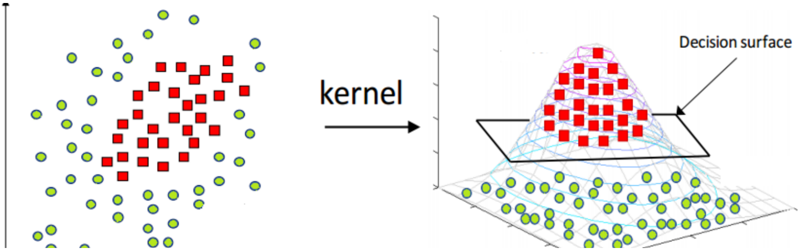
$$x_{i,j}^2, \quad x_{i,j} \times x_{i,k} \qquad \text{pour tout} \quad 1 \leq j, k \leq d$$

- Cela grandit la dimension du problème (dimension de w).

Feature map / transformation de feature

- Considérons une tranformation $\varphi : \mathbb{R}^d \to \mathbb{F}$
- $\mathbb{F}$ est un espace de Hilbert (qui peut être de dimension infinie), muni du produit scalaire $\langle \cdot, \cdot \rangle_{\mathbb{F}}$, qu'on appelle **feature space**.
- La frontière de décision $\{x : \langle w, \varphi(x) \rangle + b = 0\}$ n'est plus un hyperplan (mais $\{\varphi(x) :\to \langle w, \varphi(x) \rangle + b = 0\}$ l'est)



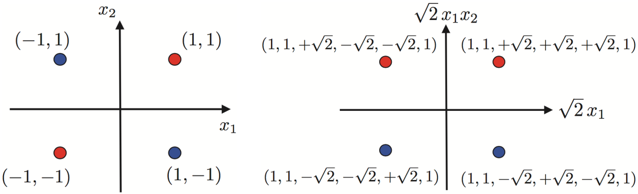

kernel
Decision surface

Transformation polynomiale d'ordre 2 (1)

La transformation polynomiale $\varphi : \mathbb{R}^2 \to \mathbb{R}^6$ pour $x = (x_1, x_2) \in \mathbb{R}^2$

$$\varphi(x) = (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1)$$

résoud le problème de classification XOR (Exclusive OR).



XOR : y_i est bleu ssi une des coordonnées de x_i vaut 1.

Transformation polynomiale d'ordre 2 (2)

Il faut remarquer que pour $x, x' \in \mathbb{R}^2$ nous avons

$$
\begin{aligned}
\langle \varphi(x), \varphi(x') \rangle &= \left\langle \begin{bmatrix} x_1^2 \\ x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ 1 \end{bmatrix}, \begin{bmatrix} x_1^2 \\ x_1'^2 \\ x_2'^2 \\ \sqrt{2}x_1'x_2' \\ \sqrt{2}x_1' \\ \sqrt{2}x_2' \\ 1 \end{bmatrix} \right\rangle \\
&= (x_1x_1' + x_2x_2' + 1)^2 \\
&= (\langle x, x' \rangle + 1)^2
\end{aligned}
$$

Noyau polynomial

Cela motive la définition de

$$K(x, x') = \langle \varphi(x), \varphi(x') \rangle = (\langle x, x' \rangle + c)^q$$

où $q \in \mathbb{N}^\star$ et $c > 0$. K est alors appelé **noyau polynomial** de degré q.

Noyau

Soit un espace de feature $\mathcal{X}$ (sous $\mathcal{X} = \mathbb{R}^d$), une fonction

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

est appelée un **noyau** sur $\mathcal{X}$.

Noyau symétrique

On dit que le noyau K est **symétrique** quand

$$K(x, x') = K(x', x)$$

pour tout $x, x' \in \mathcal{X}$

Plan

PDS kernel

PDS kernel / noyau symétrique défini positif

On dit qu'un noyau est PDS ssi

- il est symétrique
- pour tout $N \in \mathbb{N}$ et tout$\{x_1, \dots x_N\} \subset \mathcal{X}$ on a

$$\mathbb{K} = [K(x_i, x_j)]_{1 \leq i,j \leq N} \succeq 0$$

$\mathbb{K}$ est une matrice symétrique définie positive, ou de façon équivalente que

$$u^\top \mathbb{K} u = \sum_{1 \leq i,j \leq N} u_i u_j K(x_i, x_j) \geq 0$$

pour $u \in \mathbb{R}^N$, toutes les valeurs propres de $\mathbb{K}$ sont strictement positive.

Intérêt des noyaux PDS

- L'intérêt des noyaux positifs c'est qu'il est possible de leur associer un produit scalaire dans un espace $\mathcal{F}$ de features.
- Pour un échantillon $x_1, \ldots, x_n$ on nomme $\mathbb{K} = [K(x_i, x_j)]_{1 \leq i,j \leq n}$ la **matrice de Gram** ou la **matrice de similarité.** On peut imaginer définir des similarités dans des cas où les x_i de départ sont des données plus complexes (pas dans $\mathbb{R}^d$) : images, séquences d'ADN, graphes, etc
- Tout cela est associé à la théorie des espaces de Hilbert a noyau reproduisant (RKHS : Reproducing kernel Hilbert space).

Propriété d'un noyau PDS

Produit d'Hadamard

$\mathbb{A} \odot \mathbb{B}$ entre les matrices $\mathbb{A}$ et $\mathbb{B}$ de même dimension est donné par

$$(\mathbb{A} \odot \mathbb{B})_{i,j} = \mathbb{A}_{i,j} \odot \mathbb{B}_{i,j}$$

Théorème

La somme, le produit, la composition par une série de puissance $\sum_{n\geq 0} a_n x^n$ avec $a_n \geq 0$ pour tout $n \geq 0$ préserve la propriété PDS.

Noyau polynomial

Noyau polynomial

Pour $c > 0$ et $q \in \mathbb{N} - \{0\}$ on définit le noyau polynomial

$$K(x, x') = (\langle x, x' \rangle + c)^q.$$

C'est un noyau PDS.

Preuve. C'est une puissance du noyau PDS $(x, x') \mapsto \langle x, x' \rangle + b$.

Nous avons déjà calculé la transformation associée $\varphi(x)$

Noyaux RBF (Radial basis function) et tanh

Noyau RBF

Pour $\gamma > 0$ il est donné par

$$K(x, x') = \exp(-\gamma \|x - x'\|_2^2)$$

C'est un noyau PDS.

Noyau tanh

Il est aussi appelé le noyau sigmoïde et est défini par

$$K'(x, x') = \tanh(a\langle x, x'\rangle + c) = \frac{e^{a\langle x, x'\rangle + c} - e^{a\langle x, x'\rangle + c}}{e^{a\langle x, x'\rangle + c} + e^{a\langle x, x'\rangle + c}}$$

pour $a, c > 0$.

C'est aussi un noyau PDS.

Plan

Rappel de la remarque avec les features brutes

Le problème dual

$$\max_{\alpha \in \mathbb{R}^n} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

$$\text{sous la contrainte} \quad 0 \leq \alpha_i \leq C \;\; \text{et} \;\; \sum_{i=1}^{n} \alpha_i y_i = 0 \;\; \text{pour tout} \;\; i = 1, \ldots, n$$

et le classifieur

$$x \mapsto \text{signe}(\langle w, x \rangle + b) = \text{signe}(\sum_{i=1}^{n} \alpha_i y_i \langle x_i, x \rangle + b)$$

ne dépendent que des features x_i via les produits scalaire $\langle x_i, x_j \rangle$!

Remarque dans l'espace des features transformées $\mathbb{F}$

- Le problème dual est donné par

$$\max_{\alpha \in \mathbb{R}^n} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \varphi(x_i), \varphi(x_j) \rangle$$

$$= \max_{\alpha \in \mathbb{R}^n} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$\text{sous la contrainte} \quad 0 \leq \alpha_i \leq C \ \text{ et } \ \sum_{i=1}^{n} \alpha_i y_i = 0 \ \text{ pour tout } \ i = 1, \ldots, n$$

- Le classifieur s'écrit

$$x \mapsto \mathrm{signe}(\langle w, \varphi(x) \rangle + b)$$

$$= \mathrm{signe}(\sum_{i=1}^{n} \alpha_i y_i \langle \varphi(x_i), \varphi(x) \rangle + b) = \mathrm{signe}(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b)$$

Ils ne dépendent que des features $\varphi(x_i)$ via le noyau K.

Kernel trick

Pour entrainer un SVM à noyau, on n'a pas besoin de calculer les $\varphi(x_i)$.

Exemple avec un noyau gaussien (1)

On reprend le problème

$$\max_{\alpha \in \mathbb{R}^n} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$\text{sous la contrainte} \quad 0 \leq \alpha_i \leq C \ \text{ et } \ \sum_{i=1}^{n} \alpha_i y_i = 0 \ \text{ pour tout } \ i = 1, \ldots, n$$

et la prédiction

$$x \mapsto \text{signe}\left(\sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b \right)$$

avec l'intercept

$$b = y_i - \sum_{j=1}^{n} \alpha_j y_j K(x_j, x_i)$$

pour tout i tel que $0 < \alpha_i < C$.

Exemple avec un noyau gaussien (2)

Le classifieur est donc donné par

$$\hat{c}(x) = \text{signe}\Big(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\Big),$$

c'est une combinaison des $K(x_i, \cdot)$ où x_i sont les vecteurs de support

Pour le noyau gaussien, la fonction de décision est donnée par

$$x \mapsto \sum_{i:\alpha_i \neq 0} \alpha_i y_i \exp\big(-\gamma \|x - x_i\|_2^2\big) + b$$

c'est un mélange de "densités" gaussiennes.

$$x \mapsto \sum_{i:\alpha_i \neq 0} \alpha_i y_i \exp\left(-\gamma \|x - x_i\|_2^2 \right) + b$$