IFT-7002 Fondements de l'apprentissage machine

SVM et méthodes à noyaux

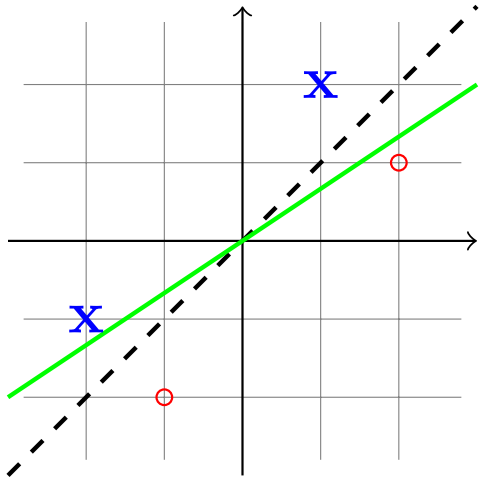
Shai Shalev-Shwartz
The Hebrew University of Jerusalem

Traduit et adapté par Mario Marchand
Université Laval

Hiver 2024

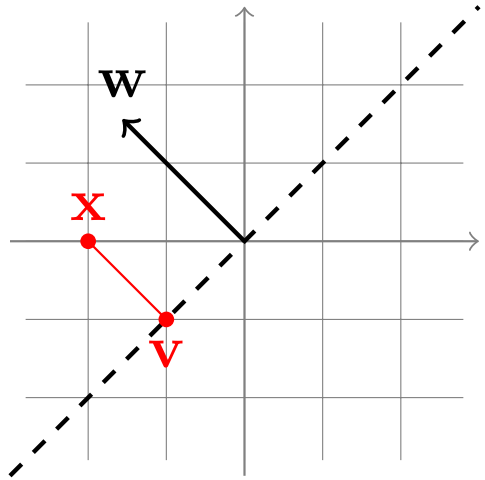
Sommaire

Quel hyperplan séparateur est le meilleur ?



- Intuitivement, le séparateur en traits interrompus est le meilleur.
- Il s'agit de l'hyperplan séparateur à marge maximale.
- Nous allons voir comment construire ce prédicteur.
- Nous allons voir aussi de quoi dépend son erreur de généralisation (*i.e.*, son risque, ou son espérance de perte).

La marge de séparation



- Soit l'hyperplan défini par $L \stackrel{\text{def}}{=} \{\mathbf{v} : \langle \mathbf{w}, \mathbf{v} \rangle + b = 0\}$ et une instance $\mathbf{x}$.
- La distance de $\mathbf{x}$ à l'hyperplan L est définie par

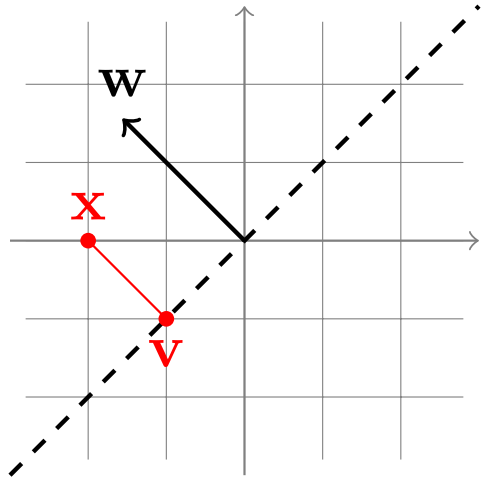
$$d(\mathbf{x}, L) \stackrel{\text{def}}{=} \min\{\|\mathbf{x} - \mathbf{v}\| : \mathbf{v} \in L\}\,.$$

- Théorème :

$$d(\mathbf{x}, L) = \frac{|\langle \mathbf{w}, \mathbf{x} \rangle + b|}{||\mathbf{w}||}$$

.

La marge de séparation



Preuve: Soit $\mathbf{v} \stackrel{\text{def}}{=} \mathbf{x} - (\langle \mathbf{w}, \mathbf{x} \rangle + b)\mathbf{w}/||\mathbf{w}||^2$. On a $\mathbf{v} \in L$ car

$$\langle \mathbf{w}, \mathbf{v} \rangle + b \;=\; \langle \mathbf{w}, \mathbf{x} \rangle + b - (\langle \mathbf{w}, \mathbf{x} \rangle + b)\langle \mathbf{w}, \mathbf{w} \rangle / ||\mathbf{w}||^2 \;=\; 0\,.$$

De plus, nous avons

$$\begin{aligned} \mathbf{x} - \mathbf{v} &= (\langle \mathbf{w}, \mathbf{x} \rangle + b)\mathbf{w}/||\mathbf{w}||^2 \\ ||\mathbf{x} - \mathbf{v}|| &= |\langle \mathbf{w}, \mathbf{x} \rangle + b|\,||\mathbf{w}||/||\mathbf{w}||^2 \;=\; |\langle \mathbf{w}, \mathbf{x} \rangle + b|/||\mathbf{w}||\,. \end{aligned}$$

La marge de séparation

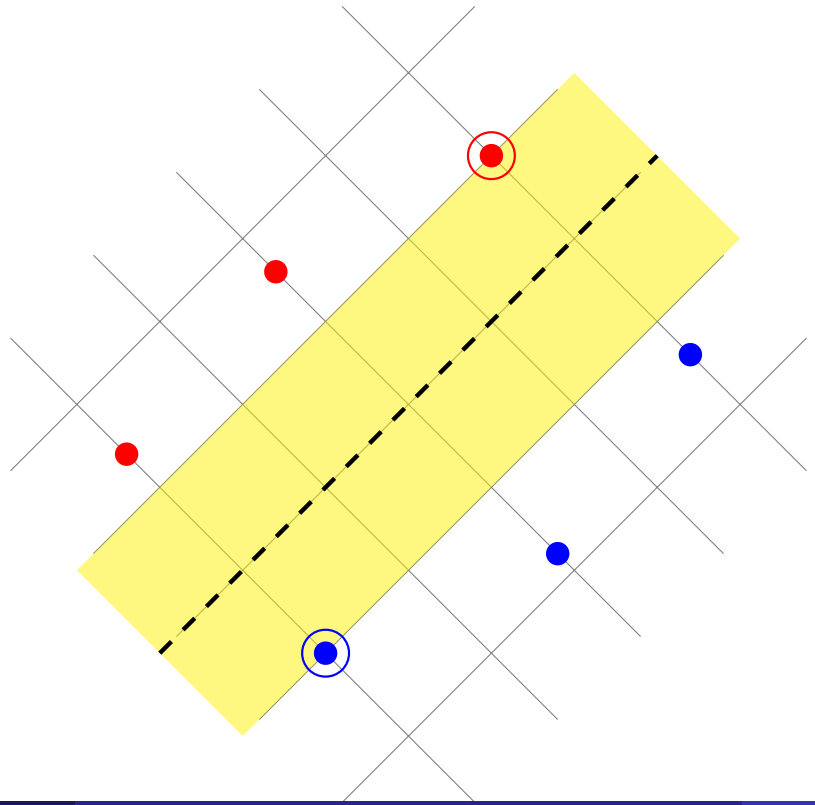
Considérons n'importe quel point $\mathbf{u}$ sur l'hyperplan, *i.e.*, $\langle \mathbf{w}, \mathbf{u} \rangle + b = 0$. Alors

$$\begin{aligned} ||\mathbf{x} - \mathbf{u}||^2 &= ||\mathbf{x} - \mathbf{v} + \mathbf{v} - \mathbf{u}||^2 \\ &= ||\mathbf{x} - \mathbf{v}||^2 + ||\mathbf{v} - \mathbf{u}||^2 + 2\langle \mathbf{x} - \mathbf{v}, \mathbf{v} - \mathbf{u} \rangle \\ &\geq ||\mathbf{x} - \mathbf{v}||^2 + 2\langle \mathbf{x} - \mathbf{v}, \mathbf{v} - \mathbf{u} \rangle \\ &= ||\mathbf{x} - \mathbf{v}||^2 + \frac{2}{||\mathbf{w}||^2}(\langle \mathbf{w}, \mathbf{x} \rangle + b)\langle \mathbf{w}, \mathbf{v} - \mathbf{u} \rangle \\ &= ||\mathbf{x} - \mathbf{v}||^2 . \end{aligned}$$

- Donc la distance entre $\mathbf{x}$ et n'importe quel autre point $\mathbf{u}$ ($\neq \mathbf{v}$) de l'hyperplan est $\geq ||\mathbf{x} - \mathbf{v}|| = |\langle \mathbf{w}, \mathbf{x} \rangle + b|/||\mathbf{w}||$.
- Donc $d(\mathbf{x}, L) = ||\mathbf{x} - \mathbf{v}|| = |\langle \mathbf{w}, \mathbf{x} \rangle + b|/||\mathbf{w}||$. ■

Marge et vecteurs de support

- Un hyperplan L, défini par $(\mathbf{w}, b)$, sépare un échantillon S lorsque pour tout $(x_i, y_i) \in S$, on a $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) > 0$. (ici, $y_i = \pm 1$)
- Définition de la marge γ de L : $\gamma \stackrel{\text{def}}{=} \min_i |\langle \mathbf{w}, \mathbf{x}_i \rangle + b|$ (ici $||\mathbf{w}|| = 1$).
- L'ensemble des exemples situés à cette distance minimale γ sont les vecteurs de support de l'hyperplan L.



Le SVM à marge rigide

- SVM-rigide : Cherche l'hyperplan séparateur de marge maximale.

$$\underset{(\mathbf{w},b):\|\mathbf{w}\|=1}{\operatorname{argmax}} \; \min_{i\in[m]} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b| \quad \text{t.q.} \quad \forall i,\ y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) > 0 \ .$$

- De manière équivalente :

$$\underset{(\mathbf{w},b):\|\mathbf{w}\|=1}{\operatorname{argmax}} \; \min_{i\in[m]} y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \quad \text{t.q.} \quad \forall i,\ y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) > 0 \ . \quad (1)$$

- Affirmation : Ce problème est équivalent à :

$$(\mathbf{w}_0, b_0) = \underset{(\mathbf{w},b)}{\operatorname{argmin}} \|\mathbf{w}\|^2 \ \text{t.q.} \ \forall i,\ y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 \ . \quad (2)$$

- C'est la forme habituel du SVM-rigide : c'est un problème d'optimisation quadratique.
- Trouver le $\mathbf{w}$ de norme 1 maximisant la marge est donc équivalent à trouver le $\mathbf{w}$ de norme minimale dont la "marge" est $= 1$.

Le SVM à marge rigide

Preuve: (une solution du problème 2 donne une solution du problème 1).

- Soit $(\mathbf{w}^*, b^*)$ une solution du problème 1.
- Soit $\gamma^* \stackrel{\text{def}}{=} \min_i y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*)$.
- On a donc $y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*) \geq \gamma^* \ \forall i$.
- Donc, $y_i(\langle \frac{\mathbf{w}^*}{\gamma^*}, \mathbf{x}_i \rangle + \frac{b^*}{\gamma^*}) \geq 1 \ \forall i$.
- Donc, $(\frac{\mathbf{w}^*}{\gamma^*}, \frac{b^*}{\gamma^*})$ satisfait les contraintes du problème 2.
- Donc, une solution $(\mathbf{w}_0, b_0)$ du problème 2 satisfait

$$||\mathbf{w}_0|| \leq ||\frac{\mathbf{w}^*}{\gamma^*}|| = 1/\gamma^* \, .$$

- Considérons $\hat{\mathbf{w}} \stackrel{\text{def}}{=} \frac{\mathbf{w}_0}{||\mathbf{w}_0||}$ et $\hat{b} \stackrel{\text{def}}{=} \frac{b_0}{||\mathbf{w}_0||}$. Pour tout i, on a

$$y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b}) = \frac{1}{||\mathbf{w}_0||} y_i(\langle \mathbf{w}_0, \mathbf{x}_i \rangle + b_0) \geq \frac{1}{||\mathbf{w}_0||} \geq \gamma^* \, .$$

- Puisque $||\hat{\mathbf{w}}|| = 1$, on a que $(\hat{\mathbf{w}}, \hat{b})$ est une solution du problème 1. ■

Le SVM à marge rigide

Preuve: (une solution du problème 1 donne une solution du problème 2).

- Soit $(\mathbf{w}^*, b^*)$ une solution du problème 1.
- Nous savons que $(\frac{\mathbf{w}^*}{\gamma^*}, \frac{b^*}{\gamma^*})$ satisfait les contraintes du problème 2.
- Pour fin de contradiction, supposons que $(\frac{\mathbf{w}^*}{\gamma^*}, \frac{b^*}{\gamma^*})$ ne solutionne pas le problème 2.
- $\exists(\mathbf{w}_0, b_0)$ t.q. $||\mathbf{w}_0|| < ||\frac{\mathbf{w}^*}{\gamma^*}|| = \frac{1}{\gamma^*}$ et $y_i(\langle \mathbf{w}_0, \mathbf{x}_i \rangle + b_0) \geq 1 \ \forall i$.
- On a alors $y_i(\langle \frac{\mathbf{w}_0}{||\mathbf{w}_0||}, \mathbf{x}_i \rangle + \frac{b_0}{||\mathbf{w}_0||}) \geq \frac{1}{||\mathbf{w}_0||} > \gamma^* \ \forall i$.
- Donc, $(\frac{\mathbf{w}_0}{||\mathbf{w}_0||}, \frac{b_0}{||\mathbf{w}_0||})$ donne une plus grande marge de séparation que celle donnée par $(\mathbf{w}^*, b^*)$.
- Ce qui implique que $(\mathbf{w}^*, b^*)$ n'est pas une solution du problème 1 (une contradiction de l'hypothèse de départ).
- Donc, $(\frac{\mathbf{w}^*}{\gamma^*}, \frac{b^*}{\gamma^*})$ doit solutionner le problème 2. ■

Le cas homogène

- Il est possible de transformer le problème de trouver un demi-espace non homogène (avec $b \neq 0$) en problème de trouver un demi-espace homogène (avec $b = 0$) en ajoutant une composante additionnelle ($= 1$) à chaque instance $\mathbf{x}$ et une composante additionnele à $\mathbf{w}$.
- Dans ce cas homogène, la classe prédite pour une instance $\mathbf{x}$ est donnée par $\text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle)$ et le problème d'optimisation à résoudre devient

$$\underset{\mathbf{w}}{\text{argmin}} \, \|\mathbf{w}\|^2 \text{ s.t. } \forall i, \; y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1 \, . \tag{3}$$

- La composante additionnelle de $\mathbf{w}$ est tenue en compte dans la fonction objectif du problème 3, mais le biais b ne fait pas parti de la fonction objectif du problème 2. (Le problème 3 régularise le biais alors que ce n'est pas le cas pour le problème 2).
- Les problèmes 2 et 3 ne sont donc pas équivalents. Mais les classificateurs retournés devraient être très similaires lorsque $\dim(\mathbf{x}) \gg 1$.

La marge doit-être spécifiée par rapport à une échelle

- La marge dépends de l'échelle choisie pour les instances :
  - Si $(\mathbf{w}, b)$ sépare $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ avec une marge γ, alors $(\mathbf{w}, 2b)$ sépare $(2\mathbf{x}_1, y_1), \ldots, (2\mathbf{x}_m, y_m)$ avec une marge de 2γ.
- La marge d'une distribution : Nous disons que $\mathcal{D}$ est séparable avec marge (γ, ρ) s'il existe $(\mathbf{w}^\star, b^\star)$ tel que $\|\mathbf{w}^\star\| = 1$ et

$$\mathcal{D}(\{(\mathbf{x}, y) : \|\mathbf{x}\| \leq \rho \wedge y(\langle \mathbf{w}^\star, \mathbf{x} \rangle + b^\star) \geq \gamma\}) = 1 \ .$$

- Théorème : (chap. 26) Si $\mathcal{D}$ est séparable avec marge (γ, ρ), alors la complexité d'échantillon du SVM-rigide satisfait

$$m(\epsilon, \delta) \ \leq \ \frac{8}{\epsilon^2} \left(2(\rho/\gamma)^2 + \log(2/\delta)\right)$$

- Ainsi, par opposition aux bornes VC, la complexité d'échantillon dépends de $(\rho/\gamma)^2$ à la place de $d + 1 = \mathrm{VCdim}$(demi-espaces non homogènes).

Le SVM à marge douce

- Le SVM-rigide ne fonctionne que si les données sont linéairement séparables.
- Le SVM-doux permet de traiter le cas non-linéairement séparable en permettant au séparateur d'effectuer des erreurs.

$$\underset{\mathbf{w},b,\boldsymbol{\xi}}{\operatorname{argmin}} \left(\lambda \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i \right)$$
$$\text{t.q. } \forall i, \quad y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i \quad \text{et} \quad \xi_i \geq 0 \,.$$

- Affirmation : Ce problème est équivalent au problème de minimisation du risque régularisé suivant :

$$\underset{\mathbf{w},b}{\operatorname{argmin}} \left(\lambda \|\mathbf{w}\|^2 + L_S^{\text{hinge}}((\mathbf{w}, b)) \right)$$

où nous avons utilisé la fonction de perte de hinge :

$$\ell^{\text{hinge}}((\mathbf{w}, b), (\mathbf{x}, y)) = \max\{0, 1 - y(\langle \mathbf{w}, \mathbf{x} \rangle + b)\} \,.$$

SVM à marge douce

Preuve:

- Considérons la minimisation sur $\boldsymbol{\xi}$ pour un $(\mathbf{w}, b)$ donné.
- Pour tout i, puisque $\xi_i \geq 0$, la meilleure valeur pour ξ_i est 0 lorsque $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$ et est $1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)$ dans le cas contraire.
- Donc, pour tout i, on a que $\xi_i = \ell^{\text{hinge}}((\mathbf{w}, b), (\mathbf{x}_i, y_i))$ qui peut être substitué dans la fonction objectif (en supprimant les contraintes). ■

Cas homogène :

- Pour simplifier l'analyse et la description des algorithmes, nous considérons uniquement le cas homogène.
- Donc $b = 0$, mais on ajoute une composante $(= 1)$ à chaque instance.
- Ceci ne change presque pas la solution lorsque $\dim(\mathbf{x}) \gg 1$.
- L'effet sur la complexité d'échantillon est faible.

Complexité d'échantillon du SVM à marge douce

- SVM-doux = régularisation de Tikhonov avec ℓ^{hinge}.
- La perte de hinge $\mathbf{w} \mapsto \max\{0, 1 - y\langle \mathbf{w}, \mathbf{x}\rangle\}$ est $\|\mathbf{x}\|$-Lipschitzienne.
- Soit $\mathcal{D}$, tel que $\|\mathbf{x}\| \leq \rho$ avec probabilité 1.
- Donc, puisque notre fonction de perte est convexe et ρ-Lipschitzienne, les résultats du dernier chapitre impliquent, que pour tout $\mathbf{u}$, on a

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L^{\text{hinge}}_{\mathcal{D}}(A(S))] \;\leq\; L^{\text{hinge}}_{\mathcal{D}}(\mathbf{u}) + \lambda\|\mathbf{u}\|^2 + \frac{2\rho^2}{\lambda\, m} \;.$$

- Puisque la perte de hinge borne supérieurement la perte 0-1, le terme à droite borne alors supérieurement $\mathbb{E}_{S\sim\mathcal{D}^m}[L^{0-1}_{\mathcal{D}}(A(S))]$.
- Donc pour tout $B > 0$, lorsque $\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}$, nous avons

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L^{0-1}_{\mathcal{D}}(A(S))] \;\leq\; \min_{\mathbf{w}:\|\mathbf{w}\|\leq B} L^{\text{hinge}}_{\mathcal{D}}(\mathbf{w}) + \sqrt{\frac{8\rho^2 B^2}{m}} \;.$$

Complexité d'échantillon du SVM à marge douce

Nous avons donc le théorème suivant.

Théorème (Complexité d'échantillon du SVM à marge douce)

Soit une distribution $\mathcal{D}$ sur $\mathcal{X} \times \mathcal{Y}$ telle que $||\mathbf{x}|| \leq \rho$ avec probabilité 1. Soit A l'algorithme d'apprentissage du SVM à marge douce. Alors, pour tout $m \geq 8\rho^2 B^2/\epsilon^2$, on a que

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}^{0-1}(A(S))] \;\leq\; \min_{\mathbf{w}:\|\mathbf{w}\|\leq B} L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}) + \epsilon \;.$$

- On obtient une complexité d'échantillon très similaire au SVM-rigide lorsque $\min_{\mathbf{w}:\|\mathbf{w}\|\leq B} L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}) = 0$ en remplaçant B par $1/\gamma$.
- Ceci est relié au fait que $y\langle \mathbf{w}, \mathbf{x}\rangle \geq 1$ (donc $y\langle \frac{\mathbf{w}}{||\mathbf{w}||}, \mathbf{x}\rangle \geq \frac{1}{||\mathbf{w}||} \geq 1/B$) lorsque $\ell^{\text{hinge}}(\mathbf{w}, (\mathbf{x}, y)) = 0$.

Marge/norme vs. dimensionalité

- La VCdim des demi-espaces dépend de la dimension d.
- La complexité d'échantillon augmente donc avec d.
- Par contre, la complexité d'échantillon du SVM dépends de $(\rho/\gamma)^2$, ou de manière équivalente, de $\rho^2 B^2$.
- Parfois, $d \gg \rho^2 B^2$ (comme dans le cas de la classification de texte).
- Il n'y a pas de contradiction avec le théorème fondamental puisque nous bornons l'erreur du SVM avec $L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}^\star)$.
  - Le théorème fondamental utilise $L_{\mathcal{D}}^{0-1}(\mathbf{w}^\star)$.
- Ceci constitue une connaissance a priori additionnelle, notamment que $L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}^\star)$ n'est pas beaucoup plus grand que $L_{\mathcal{D}}^{0-1}(\mathbf{w}^\star)$.

DGS pour SVM-doux

- Nous désirons solutionner

$$\min_{\mathbf{w}} \left(\lambda ||\mathbf{w}||^2 + \frac{1}{m} \sum_{i=1}^{m} \max\{0, 1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle\} \right)$$

  en utilisant la DGS pour la minimisation de la régularisation de Tikhonov présentée au chapitre précédent.
- Rappel : on tire $i \sim U(m)$ et on choisi $\mathbf{u}_i^{(t)} \in \partial \ell(\mathbf{w}^{(t)}, z_i)$.
- Pour $\ell^{\text{hinge}}(\mathbf{w}^{(t)}, (\mathbf{x}_i, y_i))$, nous pouvons choisir comme sous-gradient

$$\mathbf{u}_i^{(t)} = \begin{cases} \mathbf{0} & \text{si} \quad y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle \geq 1 \\ -y_i \mathbf{x}_i & \text{si} \quad y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle < 1 \end{cases} .$$

- Ainsi, l'algorithme de DGS pour la régularisation de Tikhonov devient celui de la page suivante.

DGS pour SVM-doux

DGS pour SVM-doux :

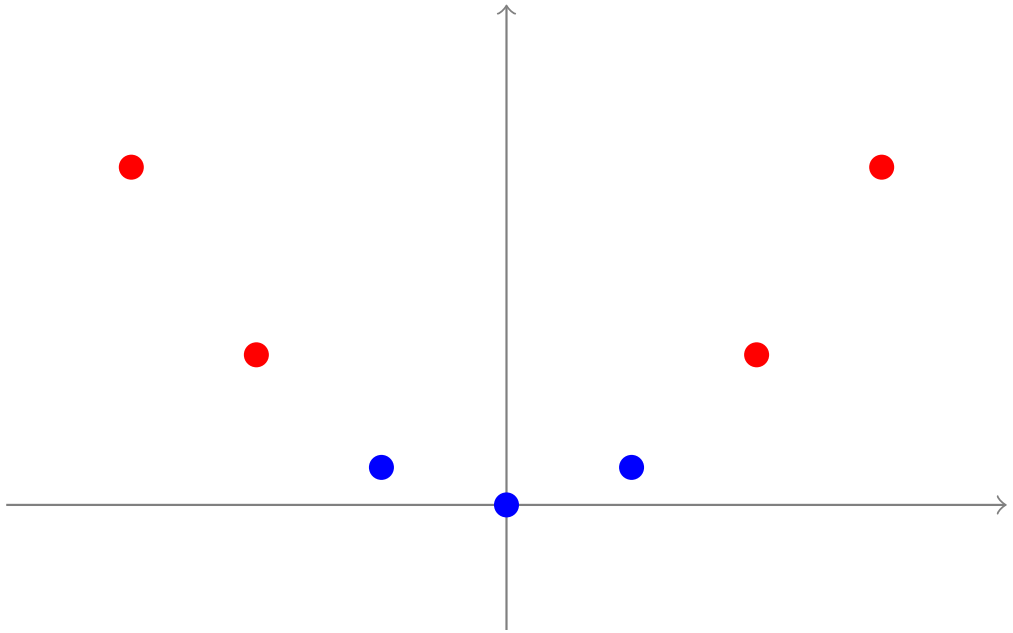
- **Objectif :** Minimiser $\lambda||\mathbf{w}||^2 + \frac{1}{m}\sum_{i=1}^{m}\max\{0, 1 - y_i\langle\mathbf{w}, \mathbf{x}_i\rangle\}$
- **initialiser :** $\mathbf{w}^{(1)} = \mathbf{0}$
- **pour** $t = 1, 2, \ldots, T$
  - tirer $i \sim U(m)$.
  - si $y_i\langle\mathbf{w}^{(t)}, \mathbf{x}_i\rangle < 1$ alors : $\mathbf{w}^{(t+1)} = \left(1 - \frac{1}{t}\right)\mathbf{w}^{(t)} + \frac{1}{2\lambda t}y_i\mathbf{x}_i$
  - si $y_i\langle\mathbf{w}^{(t)}, \mathbf{x}_i\rangle \geq 1$ alors : $\mathbf{w}^{(t+1)} = \left(1 - \frac{1}{t}\right)\mathbf{w}^{(t)}$
- **sortie :** $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}$

Sommaire

Projeter les instances dans un espace de grande dimension

- Cet échantillon n'est pas linéairement séparable dans $\mathbb{R}^1$.

- Mais si nous le transformons à l'aide de la fonction de projection $x \rightarrow (x, x^2)$, il devient linéairement séparable dans $\mathbb{R}^2$.



Projeter dans un espace de redescription

Approche générale :

- Définir une fonction de projection $\boldsymbol{\psi} : \mathcal{X} \to \mathcal{F}$.
  - $\mathcal{F}$ est un sous ensemble d'un espace vectoriel (ou de Hilbert) de dimension supérieure à $\mathcal{X}$.
  - $\mathcal{F}$ est l'espace de redescription ("feature space") des instances.
  - Chaque composante $\psi_i(\mathbf{x})$ de $\boldsymbol{\psi}(\mathbf{x})$ est une caractéristique de $\mathbf{x}$.
- Construire un demi-espace sur $\mathcal{F}$ à partir de l'échantillon

Question :

$$\hat{S} \stackrel{\text{def}}{=} ((\boldsymbol{\psi}(\mathbf{x}_1), y_1), \ldots, (\boldsymbol{\psi}(\mathbf{x}_m), y_m)) \,.$$

- Comment choisir $\boldsymbol{\psi}$?
- Si $\mathcal{F}$ est de dimension élevée, nous sommes confrontés à :
  - un défi statistique : surmontable à l'aide d'un prédicteur à large marge.
  - un défi computationnel : surmontable en utilisant des noyaux.

Choisir une bonne fonction de projection $\boldsymbol{\psi}$

- Choisir un bon $\boldsymbol{\psi}$ demande généralement de la connaissance a priori au sujet de la tâche à apprendre.
  - Une fonction de projection $\boldsymbol{\psi}$ est appropriée s'il existe un prédicteur $\mathbf{w}$ de faible norme sur $\mathcal{F}$ tel que $L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w})$ est faible.
    - Il faut aussi que $||\boldsymbol{\psi}(x)||$ soit faible avec grande probabilité.
- Néanmoins, il existe des fonctions de projection génériques qui enrichissent toujours la classe des demi-espaces et qui sont donc susceptibles d'améliorer les résultats en les utilisant.
  - *e.g.*, les fonctions de projection polynomiales.

Fonctions de projection polynomiales

- Rappel : Une fonction polynomiale $p_{\mathbf{w}}$ de degré k d'une variable x est une fonction t.q. $p_{\mathbf{w}}(x) = \sum_{j=0}^{k} w_j x^j$.
- Un séparateur polynomial est un classificateur $h_{\mathbf{w}}$ t.q. $h_{\mathbf{w}}(x) = \text{sign}(p_{\mathbf{w}}(x))$.
- Nous avons donc que $p_{\mathbf{w}}(x) = \langle \mathbf{w}, \boldsymbol{\psi}(x) \rangle$ avec $\mathbf{w} = (w_0, w_1, \ldots, w_k)$ et la fonction de projection polynomiale $\boldsymbol{\psi}(x) = (1, x, x^2, \ldots, x^k)$.
- Plus généralement, une fonction polynomiale $p_{\mathbf{w}}$ multivariée de degré k, tel que $p_{\mathbf{w}} : \mathbb{R}^n \to \mathbb{R}$, s'écrit (avec la convention $x_0 \stackrel{\text{def}}{=} 1$)

$$p_{\mathbf{w}}(\mathbf{x}) = \sum_{\mathbf{J} \in \{0,1,\ldots,n\}^k} w_{\mathbf{J}} \prod_{i=1}^{k} x_{J_i} \ .$$

- Nous avons également que $p_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \boldsymbol{\psi}(\mathbf{x}) \rangle$, mais maintenant, chaque composante de $\mathbf{w}$ et de $\boldsymbol{\psi}$ est identifiée (indexée) par un vecteur $\mathbf{J}$. Pour chaque $\mathbf{J} \in \{0, 1, \ldots, n\}^k$, on a $\psi_{\mathbf{J}}(\mathbf{x}) = \prod_{i=1}^{k} x_{J_i}$.

L'astuce du noyau

- Le noyau K associé à une fonction de projection $\boldsymbol{\psi}$ est une fonction qui implémente le produit scalaire dans l'espace de redescription. Donc, pour tout $\mathbf{x}$ et $\mathbf{x}'$ dans $\mathcal{X}$, nous avons

$$K(\mathbf{x}, \mathbf{x}') \stackrel{\text{def}}{=} \langle \boldsymbol{\psi}(\mathbf{x}), \boldsymbol{\psi}(\mathbf{x}') \rangle \,.$$

- Astuce du noyau : Nous verrons qu'il est généralement possible de calculer $K(\mathbf{x}, \mathbf{x}')$ efficacement sans utiliser $\boldsymbol{\psi}$ et que K, à lui seul, nous permet de construire un demi-espace sur $\mathcal{F}$.

Le théorème du représentant

Théorème (du représentant)

Considérons n'importe quel algorithme d'apprentissage de la forme

$$\mathbf{w}^\star \stackrel{def}{=} \operatorname*{argmin}_{\mathbf{w}} \left(f\left(\langle \mathbf{w}, \boldsymbol{\psi}(\mathbf{x}_1)\rangle, \ldots, \langle \mathbf{w}, \boldsymbol{\psi}(\mathbf{x}_m)\rangle\right) + \lambda \|\mathbf{w}\|^2 \right) \quad ,$$

tel que $f : \mathbb{R}^m \to \mathbb{R}$ *est une fonction arbitraire. Alors,* $\exists \boldsymbol{\alpha} \in \mathbb{R}^m$ *tel que* $\mathbf{w}^\star = \sum_{i=1}^m \alpha_i \boldsymbol{\psi}(\mathbf{x}_i)$.

Démonstration.

L'optimum $\mathbf{w}^\star$ s'écrit $\mathbf{w}^\star = \mathbf{v} + \mathbf{u}$, avec $\mathbf{v} \stackrel{\text{def}}{=} \sum_{i=1}^m \alpha_i \boldsymbol{\psi}(\mathbf{x}_i)$ et $\langle \mathbf{u}, \boldsymbol{\psi}(\mathbf{x}_i)\rangle = 0$, $\forall i$. Alors $\langle \mathbf{w}^\star, \boldsymbol{\psi}(\mathbf{x}_i)\rangle = \langle \mathbf{v}, \boldsymbol{\psi}(\mathbf{x}_i)\rangle \, \forall i$. Puisque $\langle \mathbf{v}, \mathbf{u}\rangle = 0$, on a $\|\mathbf{w}^\star\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{u}\|^2$. Donc, la fonction objectif en $\mathbf{w}^\star$ égale celle en $\mathbf{v}$ plus $\lambda\|\mathbf{u}\|^2$. Par optimalité de $\mathbf{w}^\star$, $\mathbf{u}$ doit être nul. □

Généralisation du théorème du représentant ?

Est-ce valide pour d'autres fonctions de régularisation ?

- C'est valide pour toute fonction de régularisation de la forme $R(||\mathbf{w}||^2)$ ssi $R : \mathbb{R} \to \mathbb{R}$ est non décroissante.
  - En effet, puisque $\|\mathbf{w}^\star\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{u}\|^2$, on a

$$R(\|\mathbf{w}^\star\|^2) = R(\|\mathbf{v}\|^2 + \|\mathbf{u}\|^2) \geq R(\|\mathbf{v}\|^2)$$

    ssi R est non décroissante.

- Le théorème n'est pas valide si la fonction de régularisation est $R(||\mathbf{w}||_1)$, avec R non décroissante et

$$||\mathbf{w}||_1 \stackrel{\text{def}}{=} \sum_{i=1}^{n} |w_i| \quad (\text{norme } L_1)\,.$$

  - En effet, il existe des exemples en deux dimensions où l'ajout d'un vecteur $\mathbf{u}$ orthogonal à $\mathbf{v}$ donne un vecteur $\mathbf{w}^\star$ de norme L_1 plus petite que la norme L_1 de $\mathbf{v}$, *i.e.*, avec $||\mathbf{w}^\star||_1 < ||\mathbf{v}||_1$.

Implications du théorème du représentant

Par le théorème du représentant, nous pouvons restreindre l'algorithme d'apprentissage à rechercher uniquement parmi les vecteurs $\mathbf{w}$ s'écrivant

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i \boldsymbol{\psi}(\mathbf{x}_i) \,.$$

Soit G la matrice t.q. $G_{i,j} = \langle \boldsymbol{\psi}(\mathbf{x}_i), \boldsymbol{\psi}(\mathbf{x}_j) \rangle$. Pour tout i, nous avons donc

$$\langle \mathbf{w}, \boldsymbol{\psi}(\mathbf{x}_i) \rangle = \langle \sum_{j=1}^{m} \alpha_j \boldsymbol{\psi}(\mathbf{x}_j), \boldsymbol{\psi}(\mathbf{x}_i) \rangle = \sum_{j=1}^{m} \alpha_j \langle \boldsymbol{\psi}(\mathbf{x}_j), \boldsymbol{\psi}(\mathbf{x}_i) \rangle = (G\boldsymbol{\alpha})_i$$

et

$$\|\mathbf{w}\|^2 = \langle \sum_{j=1}^{m} \alpha_j \boldsymbol{\psi}(\mathbf{x}_j), \sum_{j=1}^{m} \alpha_j \boldsymbol{\psi}(\mathbf{x}_j) \rangle = \sum_{i,j=1}^{m} \alpha_i \alpha_j \langle \boldsymbol{\psi}(\mathbf{x}_i), \boldsymbol{\psi}(\mathbf{x}_j) \rangle = \boldsymbol{\alpha}^\top G \boldsymbol{\alpha} \,.$$

L'optimal $\mathbf{w}^\star$ est donc donné par

$$\boldsymbol{\alpha}^\star \stackrel{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^m} \left(f\left(G\boldsymbol{\alpha}\right) + \lambda \, \boldsymbol{\alpha}^\top G \boldsymbol{\alpha} \right) .$$

L'astuce du noyau

- La matrice de Gram G dépends uniquement des produits scalaires.
- Elle peut donc être obtenu avec K sans utiliser $\boldsymbol{\psi}$.
- $\boldsymbol{\alpha}^\star$ est donc obtenu avec K sans utiliser $\boldsymbol{\psi}$.
- Ayant trouvé $\boldsymbol{\alpha}^\star$, l'étiquette prédite sur une nouvelle instance $\mathbf{x}$ est obtenue de

$$\begin{aligned}\langle \mathbf{w}^\star, \boldsymbol{\psi}(\mathbf{x})\rangle &= \langle \sum_{j=1}^{m} \alpha_j^\star \boldsymbol{\psi}(\mathbf{x}_j), \boldsymbol{\psi}(\mathbf{x})\rangle = \sum_{j=1}^{m} \alpha_j^\star \langle \boldsymbol{\psi}(\mathbf{x}_j), \boldsymbol{\psi}(\mathbf{x})\rangle \\ &= \sum_{j=1}^{m} \alpha_j^\star K(\mathbf{x}_j, \mathbf{x})\,.\end{aligned}$$

- L'apprentissage et la prédiction peuvent donc, toute les deux, se faire uniquement avec K (sans utiliser $\boldsymbol{\psi}$).

Théorème du représentant pour le SVM

SVM-doux :

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^m} \left(\lambda \boldsymbol{\alpha}^T G \boldsymbol{\alpha} + \frac{1}{m} \sum_{i=1}^{m} \max\{0, 1 - y_i (G\boldsymbol{\alpha})_i\} \right) .$$

SVM-rigide :

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^m} \boldsymbol{\alpha}^T G \boldsymbol{\alpha} \quad \text{t.q.} \quad \forall i,\ y_i (G\boldsymbol{\alpha})_i \geq 1 .$$

Le noyau polynomial

- Le noyau polynomial de degré k (pour $\mathbf{x}$ et $\mathbf{x}' \in \mathbb{R}^n$) est défini par

$$K(\mathbf{x}, \mathbf{x}') = (1+\langle \mathbf{x}, \mathbf{x}' \rangle)^k = \left(\sum_{i=0}^{n} x_i x_i' \right)^k \quad \text{(avec ajout de } x_0 = x_0' \stackrel{\text{def}}{=} 1) \,.$$

- Nous avons alors

$$\begin{aligned} K(\mathbf{x}, \mathbf{x}') &= \left(\sum_{i=0}^{n} x_i x_i' \right) \times \ldots \times \left(\sum_{i=0}^{n} x_i x_i' \right) \quad (k \text{ fois}) \\ &= \sum_{\mathbf{J} \in \{0,1,\ldots,n\}^k} \prod_{i=1}^{k} x_{J_i} x'_{J_i} = \sum_{\mathbf{J} \in \{0,1,\ldots,n\}^k} \left(\prod_{i=1}^{k} x_{J_i} \right) \left(\prod_{i=1}^{k} x'_{J_i} \right) \\ &= \sum_{\mathbf{J} \in \{0,1,\ldots,n\}^k} \psi_{\mathbf{J}}(\mathbf{x}) \psi_{\mathbf{J}}(\mathbf{x}') = \langle \boldsymbol{\psi}(\mathbf{x}), \boldsymbol{\psi}(\mathbf{x}') \rangle \,, \end{aligned}$$

pour $\boldsymbol{\psi} : \mathbb{R}^n \to \mathbb{R}^{(n+1)^k}$ t.q. $\psi_{\mathbf{J}}(\mathbf{x}) = \prod_{i=1}^{k} x_{J_i} \; \forall \mathbf{J} \in \{0, 1, \ldots, n\}^k$.

Le noyau polynomial

- Puisque $\boldsymbol{\psi}$ contient tout les monômes de degrés $\leq k$, un demi-espace sur l'image de $\boldsymbol{\psi}$ corresponds à un séparateur polynomial de degré k dans l'espace original $\mathcal{X}$.
  - Le paramètre k de ce noyau nous permet de contrôler le degré du séparateur polynomial.
- Notez que $K(\mathbf{x}, \mathbf{x}')$ se calcule en temps $O(n)$ malgré le fait que la dimension de $\boldsymbol{\psi}(\mathbf{x})$ est $(n+1)^k$.
  - Il s'agit du gain computationnel très substantiel que nous procure l'utilisation du noyau par rapport à l'utilisation de la fonction de projection associée.

Le noyau Gaussien (RBF)

Considérons d'abord $\mathcal{X} = \mathbb{R}$ et une fonction de projection $\boldsymbol{\psi}$ t.q. pour tout $n \in \mathbb{N}$ nous avons $\psi_n(x) = \frac{1}{\sqrt{n!}} e^{-\frac{x^2}{2}} x^n$. Alors,

$$\begin{aligned}
\langle \boldsymbol{\psi}(x), \boldsymbol{\psi}(x') \rangle &= \sum_{n=0}^{\infty} \left(\frac{1}{\sqrt{n!}} e^{-\frac{x^2}{2}} x^n \right) \left(\frac{1}{\sqrt{n!}} e^{-\frac{(x')^2}{2}} (x')^n \right) \\
&= e^{-\frac{x^2+(x')^2}{2}} \sum_{n=0}^{\infty} \left(\frac{(xx')^n}{n!} \right) \\
&= e^{-\frac{x^2+(x')^2}{2}} e^{xx'} \\
&= e^{-\frac{(x-x')^2}{2}} \stackrel{\text{def}}{=} K(x, x') \ .
\end{aligned}$$

- Notez que $\dim(\boldsymbol{\psi}) = \infty$, mais le calcul de $K(x, x')$ se fait en $O(1)$ (en précision finie).

Le noyau Gaussien (RBF)

Pour $\mathcal{X} = \mathbb{R}^d$, considérons la fonction de projection $\boldsymbol{\psi}$ t.q. pour tout $\mathbf{x} \in \mathbb{R}^d$, pour tout $n \geq 1$, pour tout $\mathbf{J} \in \{1, \ldots, d\}^n$, on a

$$\psi_{n,\mathbf{J}}(\mathbf{x}) = \frac{1}{\sqrt{n!}} e^{-\frac{1}{2}\frac{||\mathbf{x}||^2}{\sigma^2}} \prod_{i=1}^{n} \frac{x_{J_i}}{\sigma}$$

et $\psi_0(\mathbf{x}) = e^{-\frac{1}{2}\frac{||\mathbf{x}||^2}{\sigma^2}}$. Alors, nous avons

$$\begin{aligned}
\langle \boldsymbol{\psi}(\mathbf{x}), \boldsymbol{\psi}(\mathbf{x}') \rangle &= \psi_0(\mathbf{x})\psi_0(\mathbf{x}') + \sum_{n=1}^{\infty} \sum_{\mathbf{J} \in \{1,\ldots,d\}^n} \frac{1}{n!} e^{-\frac{1}{2}\frac{||\mathbf{x}||^2 + ||\mathbf{x}'||^2}{\sigma^2}} \prod_{i=1}^{n} \frac{x_{J_i}}{\sigma} \frac{x'_{J_i}}{\sigma} \\
&= e^{-\frac{1}{2}\frac{||\mathbf{x}||^2 + ||\mathbf{x}'||^2}{\sigma^2}} \left[1 + \sum_{n=1}^{\infty} \frac{1}{n!\sigma^{2n}} \sum_{\mathbf{J} \in \{1,\ldots,d\}^n} \prod_{i=1}^{n} x_{J_i} x'_{J_i} \right] \\
&= e^{-\frac{1}{2}\frac{||\mathbf{x}||^2 + ||\mathbf{x}'||^2}{\sigma^2}} \left[1 + \sum_{n=1}^{\infty} \frac{1}{n!\sigma^{2n}} \overbrace{\left(\sum_{i=1}^{d} x_i x'_i \right) \cdots \left(\sum_{i=1}^{d} x_i x'_i \right)}^{n \text{ fois}} \right]
\end{aligned}$$

Le noyau Gaussien (RBF)

Donc, nous avons

$$
\begin{aligned}
\langle \boldsymbol{\psi}(\mathbf{x}), \boldsymbol{\psi}(\mathbf{x}') \rangle &= e^{-\frac{1}{2}\frac{||\mathbf{x}||^2+||\mathbf{x}'||^2}{\sigma^2}} \left[1 + \sum_{n=1}^{\infty} \frac{1}{n!} \left(\frac{\langle \mathbf{x}, \mathbf{x}' \rangle}{\sigma^2} \right)^n \right] \\
&= e^{-\frac{1}{2}\frac{||\mathbf{x}||^2+||\mathbf{x}'||^2}{\sigma^2}} \left[\sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{\langle \mathbf{x}, \mathbf{x}' \rangle}{\sigma^2} \right)^n \right] \\
&= e^{-\frac{1}{2}\frac{||\mathbf{x}||^2+||\mathbf{x}'||^2}{\sigma^2}} e^{\frac{\langle \mathbf{x}, \mathbf{x}' \rangle}{\sigma^2}} \\
&= e^{-\frac{1}{2}\frac{||\mathbf{x}-\mathbf{x}'||^2}{\sigma^2}} \\
&\stackrel{\text{def}}{=} K(\mathbf{x}, \mathbf{x}') \, .
\end{aligned}
$$

Ce qui, par définition, est le noyau Gaussien (aussi appelé noyau RBF).

Le noyau Gaussien (RBF)

$$\psi_{n,\mathbf{J}}(\mathbf{x}) = \frac{1}{\sqrt{n!}} e^{-\frac{1}{2}\frac{||\mathbf{x}||^2}{\sigma^2}} \prod_{i=1}^{n} \frac{x_{J_i}}{\sigma} \quad \Longleftrightarrow \quad K(\mathbf{x}, \mathbf{x}') = e^{-\frac{1}{2}\frac{(||\mathbf{x}-\mathbf{x}'||^2)}{\sigma^2}} .$$

Remarques :

- Chaque composante $\psi_{n,\mathbf{J}}(\mathbf{x})$ de $\boldsymbol{\psi}(\mathbf{x})$ contient un monôme de degré n.
- $\boldsymbol{\psi}(\mathbf{x})$ contient donc les monômes de tous les degrés qu'il est possible de former avec $x_1, \ldots, x_d$.
- Notez que $K(\mathbf{x}, \mathbf{x}') \approx 1$ lorsque $||\mathbf{x} - \mathbf{x}'|| \ll \sigma$
- et que $K(\mathbf{x}, \mathbf{x}') \approx 0$ lorsque $||\mathbf{x} - \mathbf{x}'|| \gg \sigma$.
- Le paramètre σ contrôle donc l'échelle avec laquelle nous décidons si deux exemples sont près l'un de l'autre (et donc similaires).

Exemple d'un noyau pour les chaînes de caractères

- Jusqu'à maintenant les objets $\mathbf{x}$ à étiqueter étaient $\in \mathbb{R}^d$.
- Il est également possible de construire des prédicteurs sur des objets tels des chaînes de caractères.
- Considérez le problème d'apprendre à identifier si (oui ou non) un fichier x contient un virus.
- Tout fichier x est une (longue) séquence de caractères.
- Or, tout virus est constitué d'une chaine de d caractères qui l'identifie : c'est la signature du virus.
- Soit $\mathcal{X}$ = l'ensemble des fichiers possibles = l'ensemble des chaînes (finies) de caractères d'un alphabet Σ.
- Soit $\mathcal{X}_d$: l'ensemble des chaînes de d caractères.
- La classe d'hypothèses que l'on désire appendre : $\mathcal{H} = \{h_v : v \in \mathcal{X}_d\}$.
  - $h_v(x) = +1$ si v est une sous-chaîne de x et $h_v(x) = -1$ autrement.
  - $|\mathcal{H}| = |\mathcal{X}_d| = |\Sigma|^d$.

Exemple d'un noyau pour les chaînes de caractères (suite)

- Pour apprendre $\mathcal{H}$: construisons un demi-espace sur un espace de redescription approprié.
  - Imaginons alors une fonction de projection $\boldsymbol{\psi}$ appropriée.
  - Trouvons ensuite le noyau K associé à $\boldsymbol{\psi}$.
- Soit $\boldsymbol{\psi} : \mathcal{X} \to \{0,1\}^{|\mathcal{X}_d|+1}$ t.q. $\psi_0(x) = 1 \, \forall x$ et : $\psi_v(x) = 1$ si v est une sous-chaîne de x et $\psi_v(x) = 0$ autrement.
  - Donc $\dim(\boldsymbol{\psi}(x)) = |\Sigma|^d + 1$.
  - Donc, hors de question d'utiliser $\boldsymbol{\psi}$ explicitement.
- Chaque h_v de $\mathcal{H}$ peut être réalisé avec $\text{sign}(\langle \mathbf{w}, \boldsymbol{\psi}(x) \rangle)$ t.q. $w_0 = -1$, $w_v = 2$ et $w_u = 0 \, \forall u \neq v$ car, dans ce cas

$$\langle \mathbf{w}, \boldsymbol{\psi}(x) \rangle \;=\; w_0 \psi_0(x) + w_v \psi_v(x) \;=\; -1 + 2\psi_v(x) \;=\; h_v(x)\,,$$

- et on a alors (trivialement) que $\text{sign}(\langle \mathbf{w}, \boldsymbol{\psi}(x) \rangle) \;=\; h_v(x)$.

Exemple d'un noyau pour les chaînes de caractères (suite)

- Notez que pour ce $\mathbf{w}$, on a $||\mathbf{w}|| = \sqrt{w_0^2 + w_v^2} = \sqrt{1 + 2^2} = \sqrt{5}$.
- Nous avons aussi

$$[1 - y\langle \mathbf{w}, \boldsymbol{\psi}(x)\rangle]_+ \;=\; [1 - y h_v(x)]_+ \;=\; 0 \text{ lorsque } y = h_v(x)\,.$$

- Donc, s'il existe une signature de d caractères, il existe $\mathbf{w}$: $||\mathbf{w}|| = \sqrt{5}$ et tel que $[1 - y\langle \mathbf{w}, \boldsymbol{\psi}(x)\rangle]_+ = 0 \,\forall (x, y)$, et donc $L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}) = 0$.
- Soit $|x|$ le nombre de caractères dans (le fichier) x.
- Le nombre de sous-chaînes de taille d dans x est au plus $|x| - d + 1$.
- Donc $||\boldsymbol{\psi}(x)|| \leq \sqrt{1 + (|x| - d + 1)} \leq \sqrt{|x|}$ lorsque $d \geq 2$.
- Complexité d'échantillon du SVM-doux : il suffit d'avoir $m \geq 8\rho^2 B^2/\epsilon^2$ avec $\rho = \max_x ||\boldsymbol{\psi}(x)||$ (donc $\rho^2 = \max_x |x|$) et avec $B^2 = 5$.
- Donc, il suffit d'avoir $m \geq 40 \max(|x|)/\epsilon^2$ exemples.

Exemple d'un noyau pour les chaînes de caractères (suite)

- On ne peut pas utiliser $\boldsymbol{\psi}$ explicitement car $\dim(\boldsymbol{\psi}(x)) = |\Sigma|^d + 1$.
- Le noyau $K_d(x, x') = \langle \boldsymbol{\psi}(x), \boldsymbol{\psi}(x') \rangle$ donne $1+$ le nombre de sous-chaînes distinctes de taille d qui sont à la fois dans x et x'.
- Pour calculer $K_d(x, x')$ il suffit, pour chaque position i de x et j de x', d'examiner si il y a une sous-chaîne commune de d caractères.
  - Cela nécessite au plus d comparaisons par paire (i, j).
  - Si une sous-chaîne commune a été trouvée, on ajoute 1 à $K_d(x, x')$ ssi la sous-chaîne n'a pas été trouvée avant pour (x, x').
  - Cette vérification s'effectue en temps $O(d)$ si on utilise une table de hachage suffisamment grande pour stocker les sous-chaines communes trouvées.
- Donc, $K_d(x, x')$ se calcule en $O(|x|\ |x'|\ d)$ comparaisons de caractères.
- Notez que le calcul explicite de $\langle \boldsymbol{\psi}(x), \boldsymbol{\psi}(x') \rangle$ nécessiterait $|\Sigma|^d + 1$ multiplications.

Autres noyaux pour les chaînes de caractères

- Pour certaines applications, il serait plus intéressant d'avoir $\psi_v(x) =$ au nombre de fois que la sous-chaîne v est présente dans x.
- On a encore $\dim(\boldsymbol{\psi}(x)) = |\Sigma|^d + 1$.
- Mais $||\boldsymbol{\psi}(x)|| \leq \sqrt{1 + |x|^2}$; réalisé pour le cas *extrême* où x contient le même caractère $|x|$ fois. Mais, typiquement $||\boldsymbol{\psi}(x)|| \approx \sqrt{|x|}$.
- Ici $K(x, x') = \langle \boldsymbol{\psi}(x), \boldsymbol{\psi}(x') \rangle = 1 + \sum_{u \in \mathcal{X}_d} \psi_u(x)\psi_u(x')$ est appelé le noyau "d-gram" ou "spectrum".
- Pour calculer $K(x, x')$ en temps $O(|x|\ |x'|\ d)$, on initialise $K(x, x')$ à 1 et pour chaque position i de x et j de x', on ajoute $+1$ à $K(x, x')$ lorsque il y a une sous-chaîne commune de d caractères.
- On généralise au "blended spectrum" lorsque l'on ajoute $+1$ à $K(x, x')$ dès qu'une sous-chaîne commune *d'au plus* d caractères à été trouvée en (i, j).
  - C'est le noyau associé à la fonction de projection $\boldsymbol{\psi}$ constituée de toutes les composantes ψ_v tel que v est un mot *d'au plus* d caractères.

Caractérisation de noyaux

- Jusqu'ici notre approche a été de concevoir une fonction de projection $\boldsymbol{\psi}$ appropriée à une tâche d'apprentissage donnée.
- Ensuite, ayant $\boldsymbol{\psi}$, nous avons trouvé le noyau $K(\mathbf{x}, \mathbf{x}')$ associé ayant la propriété de se calculer beaucoup plus efficacement que le produit scalaire explicite $\langle \boldsymbol{\psi}(\mathbf{x}), \boldsymbol{\psi}(\mathbf{x}') \rangle$.
- Cependant, il est parfois plus naturel de procéder de manière inverse : on propose d'abord une fonction de similarité $K(\mathbf{x}, \mathbf{x}')$ (conforme à notre connaissance a priori) et, ensuite, nous démontrons qu'il existe une fonction de projection $\boldsymbol{\psi}$ telle que $K(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\psi}(\mathbf{x}), \boldsymbol{\psi}(\mathbf{x}') \rangle$.
- Nous allons présenter un théorème nous permettant de déterminer si oui ou non une fonction $K(\mathbf{x}, \mathbf{x}')$ est un noyau (et donc une implémentation d'un produit scalaire dans un espace de redescription).

Matrice (semi-) définie positive

- Mais il faut d'abord présenter cette définition importante.

Définition

Une matrice symétrique M est *définie positive* lorsque pour tout vecteur non-nul $\mathbf{v}$ on a

$$\mathbf{v}^\top M \mathbf{v} > 0 \,.$$

Une matrice symétrique M est *semi-définie positive* lorsque pour tout vecteur non-nul $\mathbf{v}$ on a

$$\mathbf{v}^\top M \mathbf{v} \geq 0 \,.$$

Propriétés des matrices (semi-) définies positives

- Considérez les vecteurs propres $\mathbf{v}_1, \ldots, \mathbf{v}_m$ et les valeurs propres $\lambda_1, \ldots, \lambda_m$ d'une matrice $m \times m$ semi-définie positive M, *i.e.*, pour tout $i \in [m]$, on a
$$M\mathbf{v}_i = \lambda_i \mathbf{v}_i \,.$$
- Selon le théorème Hilbert-Schmidt, ces vecteurs propres forment une base orthogonale de l'espace dans lequel agit M, et l'on peut écrire
$$M = \sum_{i=1}^{m} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top \,.$$
- Donc pour tout $\mathbf{u} \in \mathbb{R}^m$, on a
$$\mathbf{u}^\top M \mathbf{u} = \sum_{i=1}^{m} \lambda_i \langle \mathbf{v}_i, \mathbf{u} \rangle^2 \,.$$
- Puisque $\mathbf{u}$ est arbitraire, cette équation implique que toutes les valeurs propres λ_i d'une matrice semi-définie positive sont ≥ 0.

Caractérisation complète des noyaux

Théorème (Caractérisation des noyaux)

Une fonction symétrique $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ implémente un produit scalaire dans un espace d'Hilbert (appelé l'espace des redescriptions) ssi pour tout échantillon $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \stackrel{def}{=} S$ d'instances, la matrice de Gram $G_{i,j} \stackrel{def}{=} K(\mathbf{x}_i, \mathbf{x}_j)$ est une matrice symétrique semi-définie positive.

Preuve:

- Si K implémente un produit scalaire, alors il existe $\boldsymbol{\psi}$ t.q. $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \boldsymbol{\psi}(\mathbf{x}_i), \boldsymbol{\psi}(\mathbf{x}_j) \rangle$. Alors pour tout $\mathbf{v}$ et pour tout S on a

$$\mathbf{v}^\top G \mathbf{v} = \sum_{i=1}^{m} \sum_{j=1}^{m} v_i \langle \boldsymbol{\psi}(\mathbf{x}_i), \boldsymbol{\psi}(\mathbf{x}_j) \rangle v_j$$

$$= \langle \sum_{i=1}^{m} v_i \boldsymbol{\psi}(\mathbf{x}_i), \sum_{j=1}^{m} v_j \boldsymbol{\psi}(\mathbf{x}_j) \rangle \geq 0 \,,$$

  et donc G est semi-définie positive.

Caractérisation complète des noyaux (suite)

- Maintenant, supposons que G est symétrique semi-définie positive **pour tout** $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$.
- On doit démontrer que $K(\mathbf{x}, \mathbf{x}')$ implémente un produit scalaire.
- Soit $\phi : \mathbf{x} \mapsto K(\mathbf{x}, \cdot)\ \forall \mathbf{x} \in \mathcal{X}$.
  - Donc $\phi(\mathbf{x})$ est la fonction $K(\mathbf{x}, \cdot)$ telle que $\phi(\mathbf{x})(\mathbf{x}') = K(\mathbf{x}, \mathbf{x}')$.
- Soit l'espace vectoriel $\mathcal{V}$ des fonctions de $\mathcal{X}$ vers $\mathbb{R}$ formé par toutes les combinaisons linéaires **finies** de la forme $\sum_i \alpha_i K(\mathbf{x}_i, \cdot)$.
- Soit $f_\alpha, f_\beta \in \mathcal{V}$. Définissons l'opérateur $\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ par

$$\langle f_\alpha, f_\beta \rangle \;=\; \langle \sum_i \alpha_i K(\mathbf{x}_i, \cdot), \sum_j \beta_j K(\mathbf{x}'_j, \cdot) \rangle \;\stackrel{\text{def}}{=}\; \sum_{i,j} \alpha_i \beta_j K(\mathbf{x}_i, \mathbf{x}'_j)\,.$$

- Si $\langle \cdot, \cdot \rangle$ est un produit scalaire valide, on a pour tout $\mathbf{x}$ et $\mathbf{x}'$

$$\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle \;=\; \langle K(\mathbf{x}, \cdot), K(\mathbf{x}', \cdot) \rangle \;=\; K(\mathbf{x}, \mathbf{x}')\,,$$

ce qui implique que $K(\mathbf{x}, \mathbf{x}')$ implémente un produit scalaire dans $\mathcal{V}$.

Caractérisation complète des noyaux (suite)

- Or, $\langle \cdot, \cdot \rangle$ est un produit scalaire valide ssi cet opérateur est :
  1. symétrique (évident)
  2. linéaire (évident)
  3. défini positif (pas du tout évident).
- Or, $\langle \cdot, \cdot \rangle$ est défini positif ssi $\langle f, f \rangle \geq 0 \; \forall f \in \mathcal{V}$ (ce qui est le cas car G est semi-définie positive) **et** $\langle f, f \rangle = 0$ seulement lorsque f est la fonction nulle, *i.e.*, $f(\mathbf{x}) = 0 \; \forall \mathbf{x} \in \mathcal{X}$.
- Pour démontrer que $\langle f, f \rangle = 0$ seulement lorsque $f(\mathbf{x}) = 0 \; \forall \mathbf{x} \in \mathcal{X}$, notez que pour tout $f_\alpha \in \mathcal{V}$, nous avons

$$\langle f_\alpha, K(\mathbf{x}, \cdot) \rangle = \sum_i \alpha_i \langle K(\mathbf{x}_i, \cdot), K(\mathbf{x}, \cdot) \rangle = \sum_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) = f_\alpha(\mathbf{x}) \, .$$

- Si $\langle \cdot, \cdot \rangle$ satisfait l'inégalité de Cauchy-Schwarz, nous avons alors

$$f_\alpha^2(\mathbf{x}) \; \leq \; \langle f_\alpha, f_\alpha \rangle K(\mathbf{x}, \mathbf{x}) \, .$$

Donc $\langle f_\alpha, f_\alpha \rangle = 0$ implique que l'on a $f_\alpha^2(\mathbf{x}) = 0 \; \forall \mathbf{x}$.

Caractérisation complète des noyaux (suite)

- Pour compléter la preuve, démontrons que $\langle \cdot, \cdot \rangle$ satisfait l'inégalité de Cauchy-Schwarz.
- Considérons alors

$$Q \stackrel{\text{def}}{=} \begin{pmatrix} K(\mathbf{x}, \mathbf{x}) & K(\mathbf{x}, \mathbf{x}') \\ K(\mathbf{x}', \mathbf{x}) & K(\mathbf{x}', \mathbf{x}') \end{pmatrix} .$$

- Puisque, par hypothèse, Q est semi-définie positive, le produit de ses valeurs propres $(= \det(Q))$ doit être ≥ 0. Ce qui implique que

$$K(\mathbf{x}, \mathbf{x})K(\mathbf{x}', \mathbf{x}') - K(\mathbf{x}, \mathbf{x}')K(\mathbf{x}', \mathbf{x}) \geq 0 ,$$

- Alors

$$\langle K(\mathbf{x}, \cdot), K(\mathbf{x}', \cdot) \rangle^2 \leq \langle K(\mathbf{x}, \cdot), K(\mathbf{x}, \cdot) \rangle \langle K(\mathbf{x}', \cdot), K(\mathbf{x}', \cdot) \rangle .$$

- Donc $\langle \cdot, \cdot \rangle$ satisfait l'inégalité de Cauchy-Schwarz. ■

Espace de Hilbert à noyau reproduisant (RKHS)

- L'espace vectoriel $\mathcal{V}$ de fonctions de la preuve du théorème précédent, munie de son produit scalaire $\langle \cdot, \cdot \rangle$, devient un espace de Hilbert $\mathcal{H}$ lorsqu'on le complète en lui ajoutant toutes les fonctions constituées de sommes infinies convergentes $\sum_{i=1}^{\infty} \alpha_i K(\mathbf{x}_i, \cdot)$.
- Le noyau K possède la propriété dite de **reproduction** car pour tout $f_\alpha \in \mathcal{H}$, on peut vérifier que $\langle f_\alpha, K(\mathbf{x}, \cdot) \rangle = f_\alpha(\mathbf{x})$. En effet, pour tout $f_\alpha \in \mathcal{H}$, nous avons

$$\langle f_\alpha, K(\mathbf{x}, \cdot) \rangle = \sum_i \alpha_i \langle K(\mathbf{x}_i, \cdot), K(\mathbf{x}, \cdot) \rangle = \sum_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) = f_\alpha(\mathbf{x}) \,.$$

- Le produit scalaire $\langle \cdot, \cdot \rangle$ de $\mathcal{H}$ étant donné par K, fait en sorte que l'on dénote $\mathcal{H}$ comme étant un espace de Hilbert à noyau reproduisant (RKHS).

SVM-doux avec noyaux

De retour au problème d'optimisation SVM-doux avec un noyau :

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^m} \left(\lambda \boldsymbol{\alpha}^T G \boldsymbol{\alpha} + \frac{1}{m} \sum_{i=1}^{m} \max\{0, 1 - y_i (G\boldsymbol{\alpha})_i\} \right) .$$

- Au lieu de résoudre directement ce problème, retournons à notre DGS permettant de résoudre le problème original

$$\min_{\mathbf{w}} \left(\lambda ||\mathbf{w}||^2 + \frac{1}{m} \sum_{i=1}^{m} \max\{0, 1 - y_i \langle \mathbf{w}, \boldsymbol{\psi}(\mathbf{x}_i) \rangle\} \right) ,$$

et modifions l'algorithme de manière à ce que toutes les opérations impliquent uniquement le calcul de $K(\mathbf{x}_i, \mathbf{x}_j)$ sans utiliser la fonction de projection associée $\boldsymbol{\psi}$.

SVM-doux avec noyaux

Ré-écrivons alors notre DGS en utilisant la fonction de projection $\boldsymbol{\psi}$:

DGS pour SVM-doux avec fonction de projection :

- **Objectif :** Minimiser $\lambda||\mathbf{w}||^2 + \frac{1}{m}\sum_{i=1}^{m}\max\{0, 1 - y_i\langle \mathbf{w}, \boldsymbol{\psi}(\mathbf{x}_i)\rangle\}$
- **initialiser :** $\mathbf{w}^{(1)} = \mathbf{0}$
- **pour** $t = 1, 2, \ldots, T$
  - tirer $i \sim U(m)$
  - si $y_i\langle \mathbf{w}^{(t)}, \boldsymbol{\psi}(\mathbf{x}_i)\rangle < 1$ alors : $\mathbf{w}^{(t+1)} = \left(1 - \frac{1}{t}\right)\mathbf{w}^{(t)} + \frac{1}{2\lambda t} y_i \boldsymbol{\psi}(\mathbf{x}_i)$
  - si $y_i\langle \mathbf{w}^{(t)}, \boldsymbol{\psi}(\mathbf{x}_i)\rangle \geq 1$ alors : $\mathbf{w}^{(t+1)} = \left(1 - \frac{1}{t}\right)\mathbf{w}^{(t)}$
- **sortie :** $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}$

Utilisons le fait que $\mathbf{w}^{(t)} = \sum_{i=1}^{m}\alpha_i^{(t)}\boldsymbol{\psi}(\mathbf{x}_i)$ à chaque itération t et ré-écrivons le code en utilisant $\boldsymbol{\alpha}^{(t)} \stackrel{\text{def}}{=} (\alpha_1^{(t)}, \ldots, \alpha_m^{(t)})$ au lieu de $\mathbf{w}^{(t)}$.

SVM-doux avec noyaux

- Dans cet algorithme, le test du produit scalaire peut s'écrire

$$y_i\langle \mathbf{w}^{(t)}, \boldsymbol{\psi}(\mathbf{x}_i)\rangle = \sum_{j=1}^{m} y_i \alpha_j^{(t)} \langle \boldsymbol{\psi}(\mathbf{x}_j), \boldsymbol{\psi}(\mathbf{x}_i)\rangle = \sum_{j=1}^{m} y_i \langle \boldsymbol{\psi}(\mathbf{x}_i), \boldsymbol{\psi}(\mathbf{x}_j)\rangle \alpha_j^{(t)}$$
$$= \sum_{j=1}^{m} y_i G_{i,j} \alpha_j^{(t)} \;=\; y_i (G\boldsymbol{\alpha}^{(t)})_i$$

- La mise à jour $\mathbf{w}^{(t+1)} = (1 - \frac{1}{t})\mathbf{w}^{(t)}$ peut s'effectuer à l'aide de $\boldsymbol{\alpha}^{(t+1)} = (1 - \frac{1}{t})\boldsymbol{\alpha}^{(t)}$ car, dans ce cas, on a

$$\mathbf{w}^{(t+1)} \;=\; \sum_{i=1}^{m} \alpha_i^{(t+1)} \boldsymbol{\psi}(\mathbf{x}_i)$$
$$= \; (1 - \frac{1}{t}) \sum_{i=1}^{m} \alpha_i^{(t)} \boldsymbol{\psi}(\mathbf{x}_i) \;=\; (1 - \frac{1}{t})\mathbf{w}^{(t)} \;.$$

SVM-doux avec noyaux

- La mise à jour $\mathbf{w}^{(t+1)} = (1 - \frac{1}{t})\mathbf{w}^{(t)} + \frac{1}{2\lambda t} y_i \boldsymbol{\psi}(\mathbf{x}_i)$ peut s'effectuer à l'aide de

$$\alpha_j^{(t+1)} = \begin{cases} (1 - \frac{1}{t})\alpha_j^{(t)} + \frac{1}{2\lambda t} y_j & \text{si} \quad j = i \\ (1 - \frac{1}{t})\alpha_j^{(t)} & \text{si} \quad j \neq i \end{cases}$$

- car, dans ce cas, on a

$$\begin{aligned} \mathbf{w}^{(t+1)} &= \sum_{j=1}^{m} \alpha_j^{(t+1)} \boldsymbol{\psi}(\mathbf{x}_j) \\ &= (1 - \frac{1}{t}) \sum_{j \neq i} \alpha_j^{(t)} \boldsymbol{\psi}(\mathbf{x}_j) + \left[(1 - \frac{1}{t})\alpha_i^{(t)} + \frac{1}{2\lambda t} y_i \right] \boldsymbol{\psi}(\mathbf{x}_i) \\ &= (1 - \frac{1}{t}) \sum_{j=1}^{m} \alpha_j^{(t)} \boldsymbol{\psi}(\mathbf{x}_j) + \frac{1}{2\lambda t} y_i \boldsymbol{\psi}(\mathbf{x}_i) = (1 - \frac{1}{t})\mathbf{w}^{(t)} + \frac{1}{2\lambda t} y_i \boldsymbol{\psi}(\mathbf{x}_i) \end{aligned}$$

SVM-doux avec noyaux

Nous obtenons donc la DGS suivante :

DGS pour SVM-doux utilisant un noyau :

- **Objectif :** Minimiser $\left(\lambda \boldsymbol{\alpha}^T G \boldsymbol{\alpha} + \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i (G\boldsymbol{\alpha})_i\}\right)$
- **initialiser :** $\boldsymbol{\alpha}^{(1)} = \mathbf{0}$
- **pour** $t = 1, 2, \ldots, T$
  - $\boldsymbol{\alpha}^{(t+1)} = (1 - \frac{1}{t})\boldsymbol{\alpha}^{(t)}$ (pré-calcul temporaire de $\boldsymbol{\alpha}^{(t+1)}$)
  - tirer $i \sim U(m)$
  - si $y_i (G\boldsymbol{\alpha})_i < 1$ alors : $\alpha_i^{(t+1)} = \alpha_i^{(t+1)} + \frac{1}{2\lambda t} y_i$
- **sortie :** $\bar{\boldsymbol{\alpha}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\alpha}^{(t)}$

- Chaque itération requiert un temps $\in \Theta(m)$ si G est pré-calculé.
- Notez que $\bar{\mathbf{w}} = \sum_{i=1}^{m} \bar{\alpha}_i \boldsymbol{\psi}(\mathbf{x}_i)$.
- L'étiquette prédite pour $\mathbf{x}$: $\text{sign}[\langle \bar{\mathbf{w}}, \boldsymbol{\psi}(\mathbf{x}) \rangle] = \text{sign}[\sum_{i=1}^{m} \bar{\alpha}_i K(\mathbf{x}_i, \mathbf{x})]$.

Régression de ridge avec noyaux

- Le méthodes à noyaux ne s'appliquent pas uniquement au SVM.
- Examinons le cas de la régression de ridge, où l'on tente de résoudre

$$\operatorname*{argmin}_{\mathbf{w}} \left(\lambda \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{i=1}^{m} (\langle \mathbf{w}, \boldsymbol{\psi}(\mathbf{x}_i) \rangle - y_i)^2 \right) .$$

- Au chapitre précédent, nous avons que que la solution doit satisfaire

$$\left(m\lambda I + XX^\top \right) \mathbf{w} = X\mathbf{y} ,$$

où, cette fois-ci, X est la matrice $d \times m$ formée des vecteurs colonnes $\boldsymbol{\psi}(\mathbf{x}_1), \ldots, \boldsymbol{\psi}(\mathbf{x}_m)$ et $d = \dim(\boldsymbol{\psi}(\mathbf{x})) = \dim(\mathbf{w})$.

- En utilisant $\mathbf{w} = \sum_{j=1}^{m} \alpha_j \boldsymbol{\psi}(\mathbf{x}_j)$ et $G_{i,j} \overset{\text{def}}{=} \langle \boldsymbol{\psi}(\mathbf{x}_i), \boldsymbol{\psi}(\mathbf{x}_j) \rangle$ et le fait que $XX^\top = \sum_i \boldsymbol{\psi}(\mathbf{x}_i) \boldsymbol{\psi}(\mathbf{x}_i)^\top$, cette équation devient

$$m\lambda \sum_{j=1}^{m} \alpha_j \boldsymbol{\psi}(\mathbf{x}_j) + \sum_{j=1}^{m} (G\boldsymbol{\alpha})_j \boldsymbol{\psi}(\mathbf{x}_j) = \sum_{j=1}^{m} y_j \boldsymbol{\psi}(\mathbf{x}_j) .$$

Régression de ridge avec noyaux

- Ce qui se ré-écrit comme suit

$$\sum_{j=1}^{m} [m\lambda\alpha_j + (G\boldsymbol{\alpha})_j - y_j]\,\boldsymbol{\psi}(\mathbf{x}_j) \;=\; \mathbf{0}\,.$$

- Une condition suffisante (et nécessaire lorsque les $\boldsymbol{\psi}(\mathbf{x}_j)$ sont linéairement indépendants) pour satisfaire cette équation est d'avoir $m\lambda\alpha_j + (G\boldsymbol{\alpha})_j - y_j = 0$ pour tout j.
- Cela implique qu'il suffit pour $\boldsymbol{\alpha}$ de satisfaire

$$(m\lambda I + G)\boldsymbol{\alpha} \;= \mathbf{y}\,.$$

- Pour $\lambda > 0$, $(m\lambda I + G)$ est une matrice défini positive (car G est semi-définie positive) et elle est donc inversible. La solution pour la régression de ridge avec noyau est donc unique et est donnée par

$$\boldsymbol{\alpha} \;= (m\lambda I + G)^{-1}\mathbf{y}\,.$$

Résumé

- Utilisation d'une marge de séparation comme connaissance à priori.
  - S'il existe une marge γ séparant les exemples ayant $||\mathbf{x}|| \leq \rho$, le SVM peut donner de bons classificateurs car sa complexité d'échantillon dépends uniquement de $\rho^2/(\gamma^2\epsilon^2)$ (sans dépendre de la dimension).
  - SVM-doux : utilisation de la fonction de perte de hinge dans le cas non-linéairement séparable : complexité d'échantillon dépends de $B^2\rho^2/\epsilon^2$ par rapport au meilleur prédicteur $\mathbf{w}$ avec $||\mathbf{w}|| \leq B$.
- L'astuce du noyau : permet de construire des séparateurs non-linéaires sans utiliser explicitement la fonction de projection.
- Les noyaux sont utilisables pour le SVM et la régression de ridge (et autres cas de la régularisation de Tikhonov).
- Utilisation de la DGS pour construire un SVM avec ou sans noyau.