Introduction à la Régression à Vaste Marge

(Support Vector Regression — SVR)

Youssef SALMAN
youssef.salman@emse.fr

Ecole des Mines de Saint-Étienne
Majeure: Sciences de données

18 novembre 2025

Introduction à la SVR

De la classification à la régression

Dans le chapitre précédent, nous avons étudié les **SVM de classification**, dont l'objectif est de séparer deux classes à l'aide d'un hyperplan de **marge maximale**.

Cependant, dans de nombreuses applications réelles :

- la sortie n'est pas une étiquette ($+1$, -1),
- mais une **valeur réelle** : prix, température, vitesse, taux, etc.

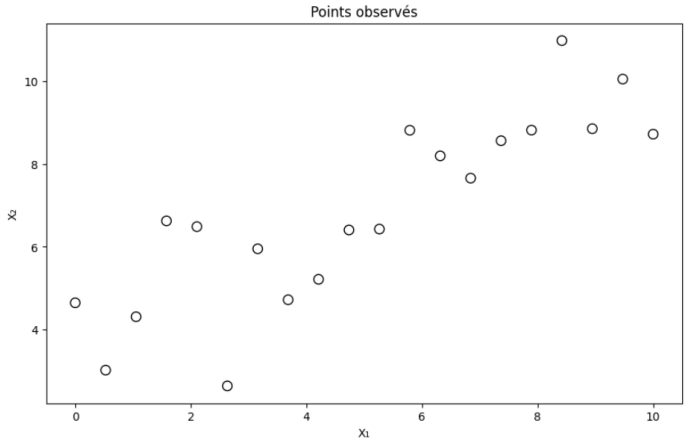
Limite du SVM classique

Le SVM ne peut pas être appliqué directement lorsque la sortie est un **nombre continu**.

Objectif du chapitre

Introduire une extension du SVM adaptée à la **régression** : la **Support Vector Regression (SVR)**.

Introduction



Introduction

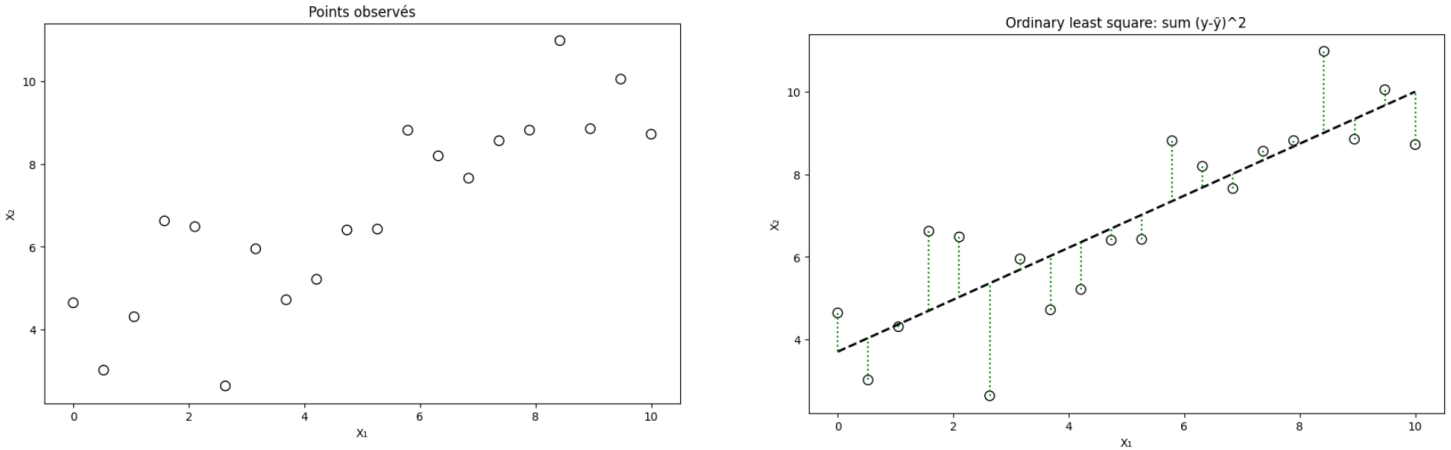


Figure – Idée de la régression linéaire simple

Introduction

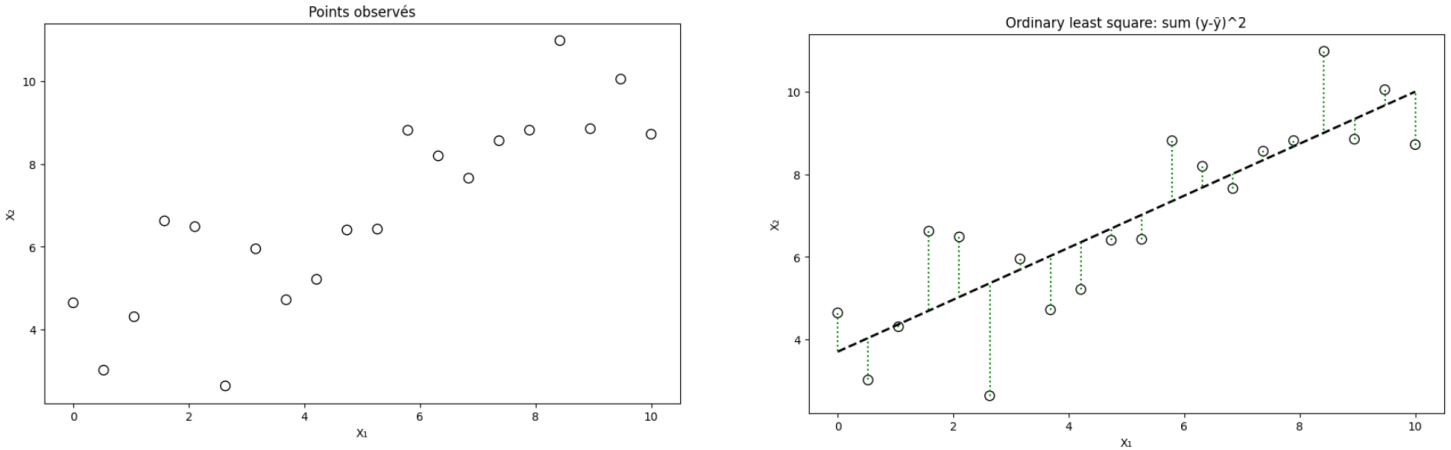
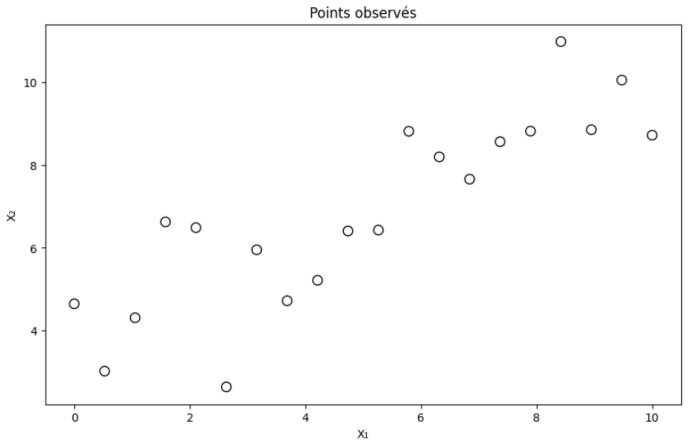


Figure – Idée de la régression linéaire simple



Introduction

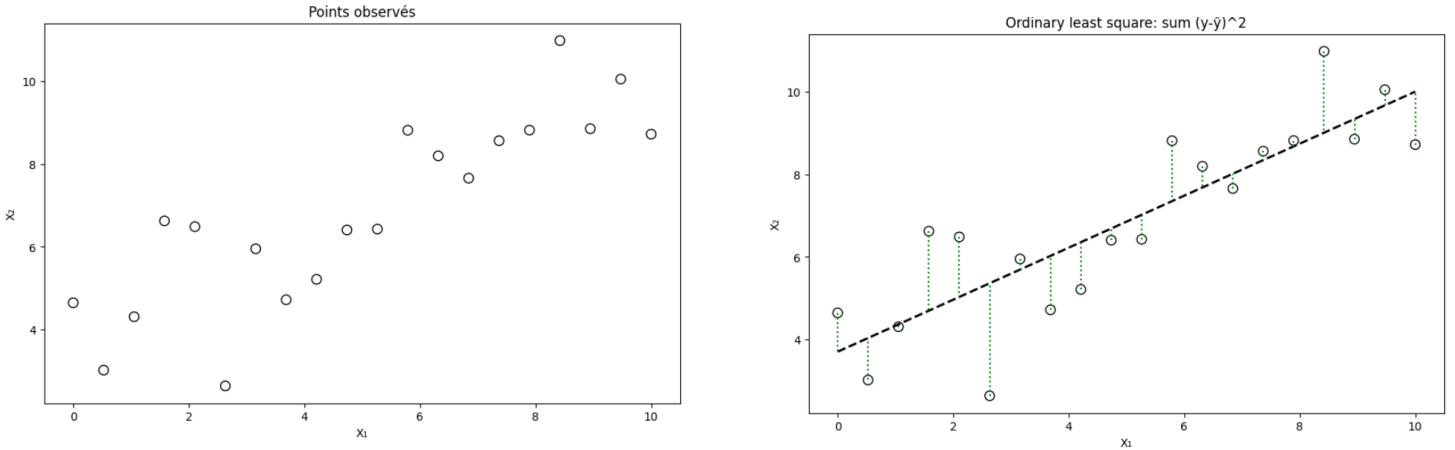


Figure – Idée de la régression linéaire simple

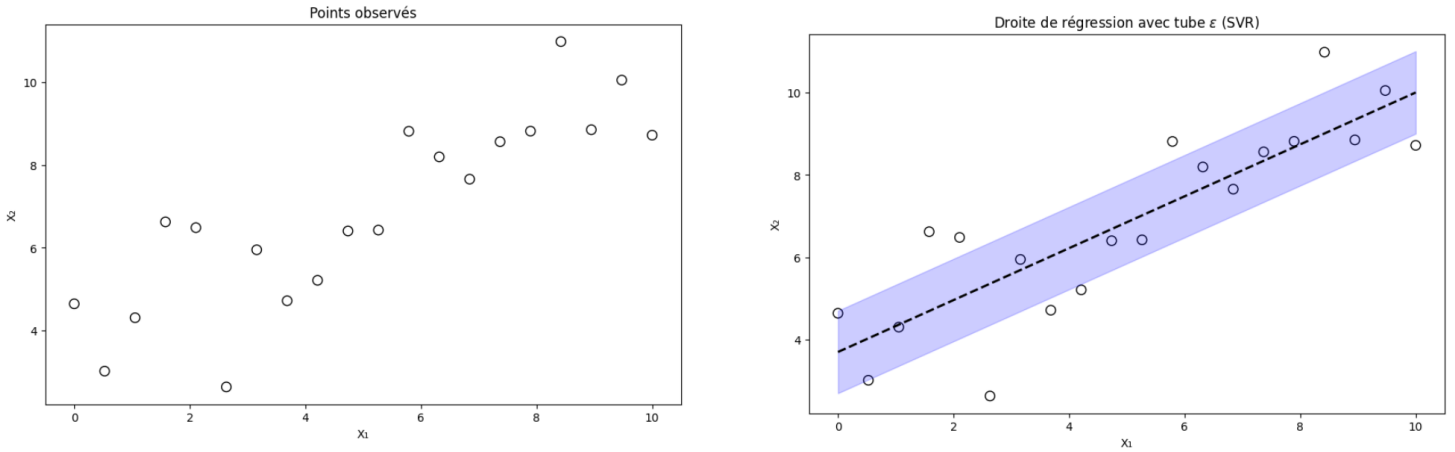
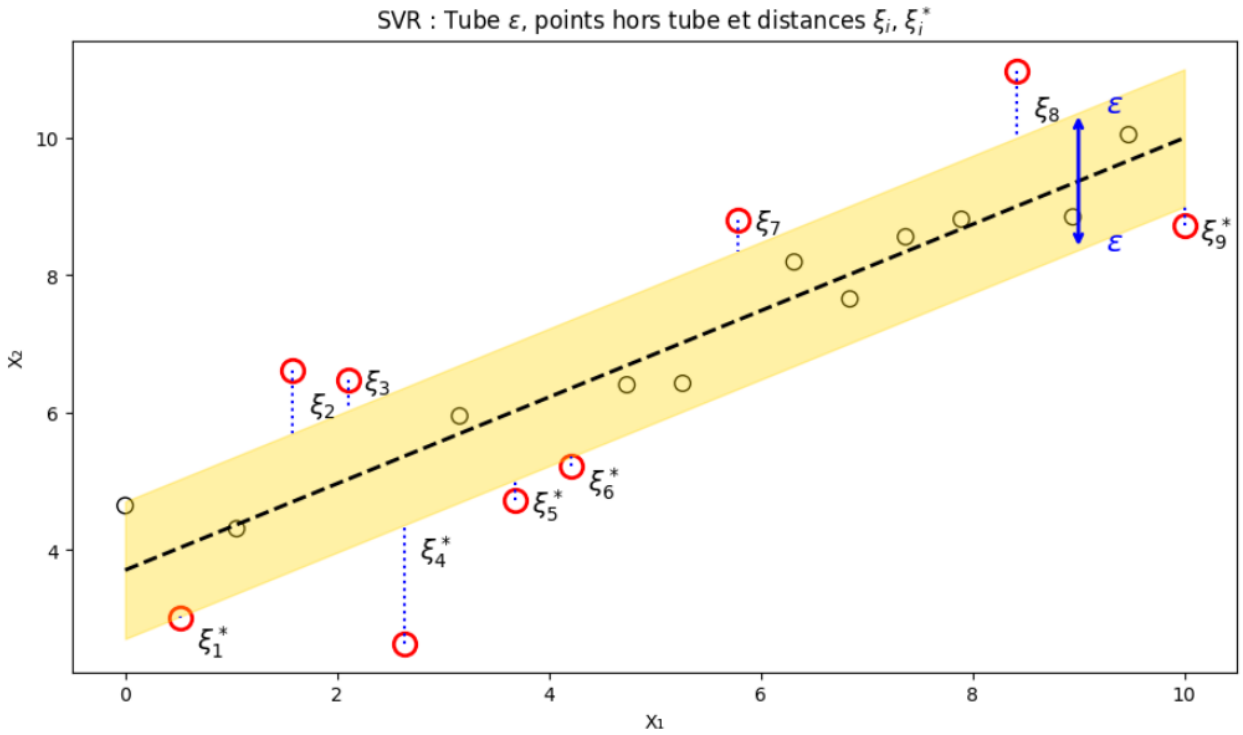


Figure – Idée de la SVR

Introduction

La SVR tolère l'existence de points **à l'intérieur du tube**, et se concentre principalement sur :

- les points **sur les parois du tube**,
- et les points **à l'extérieur du tube**.



Problème d'apprentissage supervisé

Nous nous intéressons désormais à un problème **régressif**, où la sortie est une valeur réelle.

- Chaque observation est décrite par un vecteur d'entrée :

$$\mathbf{x} \in \mathbb{R}^d.$$

- L'objectif est de prédire une quantité numérique :

$$y = f(\mathbf{x}) \in \mathbb{R}.$$

GOAL

À partir d'un ensemble d'apprentissage

$$\{(\mathbf{x}_i, y_i)\}_{i=1}^n,$$

le but est d'apprendre une fonction $\hat{f}$ capable de bien prédire y sur de nouvelles observations.

SVR : Principe et Objectif

Pourquoi la Régression à Vaste Marge ?

Les méthodes de régression classiques (ex. : régression linéaire) cherchent à **minimiser l'erreur quadratique globale** :

$$\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2.$$

Mais ce critère :

- est **sensible aux valeurs aberrantes**,
- impose des pénalités fortes aux grandes erreurs,
- ne contrôle pas la **complexité du modèle**.

Idée de la SVR

Ignorer les erreurs **petites** : on tolère une zone d'insensibilité de largeur ε autour de la prédiction.

Toutes les erreurs de taille $\leq \varepsilon$ sont considérées comme nulles.

Principe de la régression ε-insensible

On définit un tube de largeur ε autour du modèle :

$$|y - f(\mathbf{x})| \leq \varepsilon.$$

Interprétation :

- les points **dans le tube** ne sont pas pénalisés,
- seuls les points **en dehors du tube** comptent,
- la fonction doit rester **aussi plate que possible**.

Idée intuitive

On autorise une erreur contrôlée (jusqu'à ε) pour obtenir un modèle simple, robuste et peu sensible au bruit.

Que cherche-t-on à optimiser ?

L'objectif est double :

1. Minimiser la **complexité du modèle** (mesurée par $\|\beta\|^2$).
2. Limiter les **erreurs en dehors du tube** ε.

Pour cela, on introduit deux **variables de relâchement** ξ_i et $\xi_i^\star$, qui mesurent de combien un point dépasse le tube :

- ξ_i : dépassement **par au-dessus** (prédiction trop basse),
- $\xi_i^\star$: dépassement **par en-dessous** (prédiction trop haute).

Le problème primal devient :

$$\min_{\beta,\, b,\, \xi_i,\, \xi_i^\star} \quad \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^\star)$$

$$\text{s.c.} \begin{cases} y_i - \langle \beta, x_i \rangle - b \leq \varepsilon + \xi_i, \\ \langle \beta, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^\star, \\ \xi_i,\ \xi_i^\star \geq 0. \end{cases}$$

Interprétation des variables de relâchement

Dans la SVR, deux slacks sont nécessaires pour mesurer les dépassements du tube ε :

- **Dépassement par le haut** (prédiction trop grande) :

$$\hat{y}_i - y_i > \varepsilon \quad \Rightarrow \quad \xi_i^\star > 0.$$

- **Dépassement par le bas** (prédiction trop petite) :

$$y_i - \hat{y}_i > \varepsilon \quad \Rightarrow \quad \xi_i > 0.$$

Les points :

- **dans le tube** : $\xi_i = \xi_i^\star = 0$,
- **sur les parois du tube** : $\xi_i = \xi_i^\star = 0$ (vecteurs de support marginaux),
- **en dehors du tube** : $\xi_i > 0$ ou $\xi_i^\star > 0$.

Rôle du paramètre C

Interprétation

- **Petit** C : modèle très plat, grande tolérance aux erreurs.
- **Grand** C : modèle qui suit étroitement les données.

- C contrôle le compromis entre **largeur du tube** et **pénalisation des slacks** $(\xi_i, \xi_i^\star)$.
- Plus C est grand, plus on pénalise fortement les points en dehors du tube.

Formulation duale

Formulation lagrangienne

Les multiplicateurs de Lagrange sont :

$$\alpha_i, \alpha_i^\star \geq 0 \qquad \text{(contraintes du tube)}$$

$$\eta_i, \eta_i^\star \geq 0 \qquad \text{(positivité des slacks)}$$

Les contraintes sont réécrites sous la forme $g_i \leq 0$:

$$\begin{cases} y_i - \langle \beta, x_i \rangle - b - \varepsilon - \xi_i \leq 0, \\ \langle \beta, x_i \rangle + b - y_i - \varepsilon - \xi_i^\star \leq 0, \\ -\xi_i \leq 0, \quad -\xi_i^\star \leq 0. \end{cases}$$

Le Lagrangien est :

$$\mathcal{L} = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^\star) + \sum_{i=1}^{n}\alpha_i\big(y_i - \langle \beta, x_i \rangle - b - \varepsilon - \xi_i\big)$$

$$+ \sum_{i=1}^{n}\alpha_i^\star\big(\langle \beta, x_i \rangle + b - y_i - \varepsilon - \xi_i^\star\big) - \sum_{i=1}^{n}\eta_i\xi_i - \sum_{i=1}^{n}\eta_i^\star\xi_i^\star.$$

Conditions d'optimalité

En annulant les dérivées :

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = 0 \Rightarrow \boldsymbol{\beta} = \sum_{i=1}^{n} (\alpha_i - \alpha_i^\star) \boldsymbol{x}_i$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} (\alpha_i - \alpha_i^\star) = 0$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \eta_i = 0, \qquad \frac{\partial \mathcal{L}}{\partial \xi_i^\star} = C - \alpha_i^\star - \eta_i^\star = 0$$

Comme $\eta_i, \eta_i^\star \geq 0$:

$$0 \leq \alpha_i, \alpha_i^\star \leq C.$$

Conclusion : la solution dépend uniquement de $\alpha_i, \alpha_i^\star$.

Pourquoi passer au dual ?

Intérêt majeur :

- Le problème dual dépend uniquement des **produits scalaires** :

$$\langle x_i, x_j \rangle.$$

- Cela permet d'introduire directement les **noyaux** :

$$\langle x_i, x_j \rangle \rightarrow K(x_i, x_j).$$

- La solution finale s'écrit en fonction des **vecteurs de support**.

Conséquence

La SVR devient naturellement non linéaire grâce au **kernel trick**.

Problème dual de la SVR

Le dual s'écrit :

$$\max_{\alpha_i, \alpha_i^\star} W(\alpha, \alpha^\star) = -\frac{1}{2} \sum_{i,j} (\alpha_i - \alpha_i^\star)(\alpha_j - \alpha_j^\star) \langle x_i, x_j \rangle$$

$$-\varepsilon \sum_i (\alpha_i + \alpha_i^\star) + \sum_i y_i (\alpha_i - \alpha_i^\star)$$

$$\text{s.c.} \quad \sum_i (\alpha_i - \alpha_i^\star) = 0, \qquad 0 \leq \alpha_i, \alpha_i^\star \leq C.$$

Solution sous forme de vecteurs de support

La fonction estimée est :

$$\hat{f}(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^\star)\langle x_i, x\rangle + b.$$

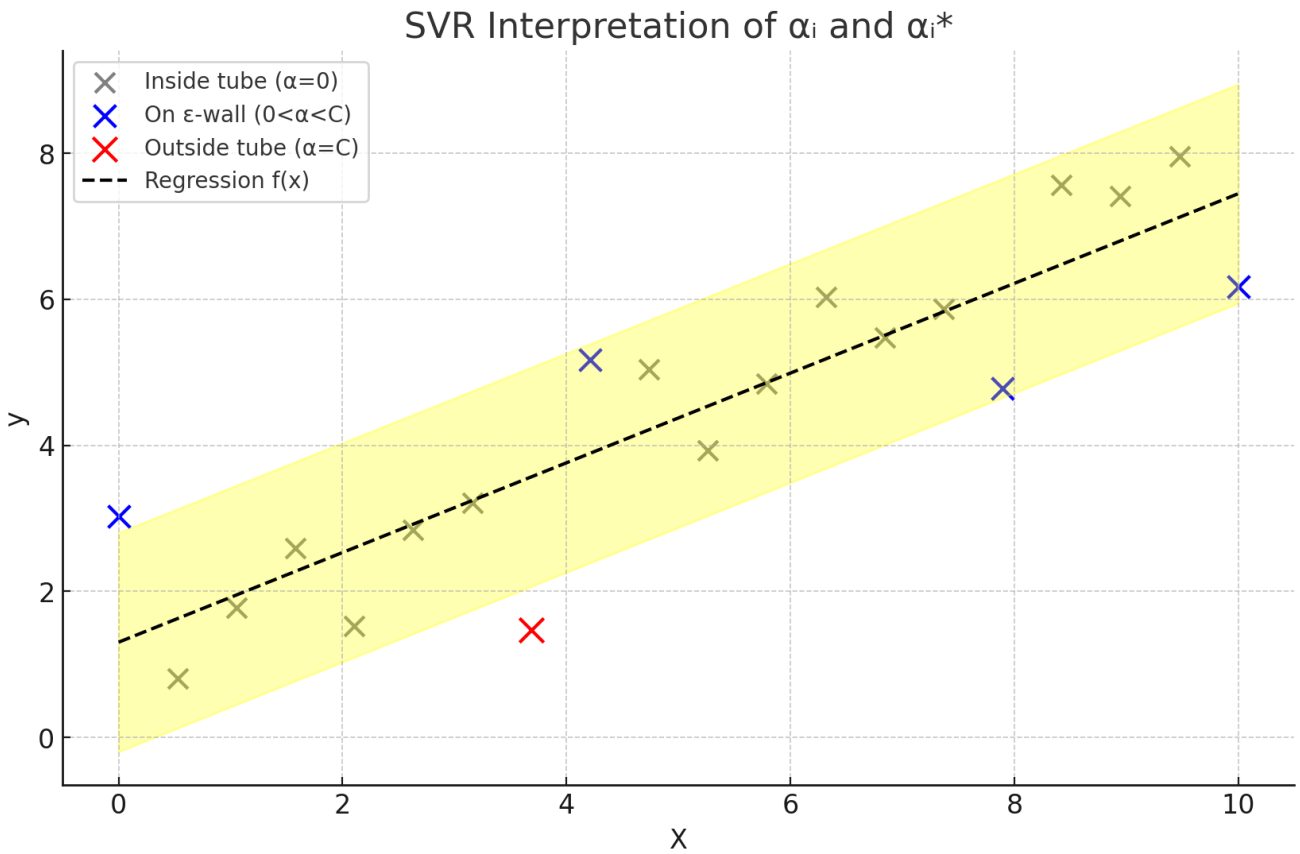
Les points utiles sont ceux pour lesquels :

$$|\alpha_i - \alpha_i^\star| > 0.$$

Interprétation :

- **Dans le tube** : $\xi_i = \xi_i^\star = 0$ et $\alpha_i = \alpha_i^\star = 0 \Rightarrow$ pas vecteur de support.
- **Sur les parois** : $0 < \alpha_i < C$ ou $0 < \alpha_i^\star < C \Rightarrow$ vecteurs de support marginaux.
- **En dehors du tube** : $\alpha_i = C$ ou $\alpha_i^\star = C \Rightarrow$ vecteurs de support avec erreur.

Solution sous forme de vecteurs de support



Astuce du noyau

Pourquoi utiliser un noyau ?

Le modèle linéaire de la SVR :

$$\hat{f}(x) = \langle \beta, x \rangle + b$$

ne permet de modéliser que des **relations linéaires**.

Dans de nombreux problèmes réels :

- la relation $x \mapsto y$ est **non linéaire**,
- augmenter la dimension manuellement est coûteux,
- la régression linéaire (ou SVR linéaire) devient insuffisante.

Idée clé

Transformer les données dans un espace de grande dimension où le problème **devient linéaire**.

But : rendre la SVR capable d'apprendre des frontières **non linéaires**.

Interprétation géométrique

Idée : Appliquer une transformation :

$$x \longmapsto \varphi(x)$$

vers un espace de dimension élevée où les données deviennent plus facilement séparables.

Exemple (idée) :

$$x = (x_1, x_2) \quad \mapsto \quad \varphi(x) = (x_1, x_2, x_1^2, x_1 x_2, x_2^2).$$

Dans cet espace :

- le tube SVR devient un **tube linéaire**,
- le modèle devient une simple SVR linéaire,
- mais dans l'espace original, la fonction est **non linéaire**.

Pourquoi le noyau est-il si puissant ?

Calculer explicitement $\varphi(x)$ peut être :

- très coûteux (dimension énorme),
- voire infini (noyau RBF $\Rightarrow$ espace de dimension infinie).

Astuce du noyau

On ne calcule **jamais** $\varphi(x)$. On calcule directement :

$$K(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle.$$

Conséquence :

- complexité très faible,
- expressivité très élevée,
- capacité à apprendre des relations très **non linéaires**.

Astuce du noyau (Kernel Trick)

Dans le dual, les données apparaissent uniquement via :

$$\langle x_i, x_j \rangle.$$

Définition d'un noyau

Un noyau calcule un **produit scalaire dans un espace transformé** :

$$K(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle,$$

sans jamais calculer explicitement la transformation φ.

On remplace :

$$\langle x_i, x_j \rangle \quad \longrightarrow \quad K(x_i, x_j)$$

où K peut être linéaire, polynomial, RBF, sigmoïdal...

Modèle final avec noyau

$$\hat{f}(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^\star)\, K(x_i, x) + b.$$

Conclusion

Grâce au noyau, la SVR peut apprendre des fonctions **non linéaires** tout en restant un problème **linéaire dans l'espace transformé**.

Résumé : SVR avec noyau en 4 étapes

1. Choisir un noyau $K(x_i, x_j)$ (linéaire, polynomial, RBF, sigmoïdal. . .).
2. Résoudre le problème dual :

$$\max_{\alpha_i, \alpha_i^\star} W(\alpha, \alpha^\star)$$

sous

$$\sum_i (\alpha_i - \alpha_i^\star) = 0, \quad 0 \leq \alpha_i, \alpha_i^\star \leq C.$$

3. Calculer la fonction :

$$\hat{f}(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^\star) K(x_i, x) + b.$$

4. Les seuls points utiles sont les **vecteurs de support** :

$$|\alpha_i - \alpha_i^\star| > 0.$$

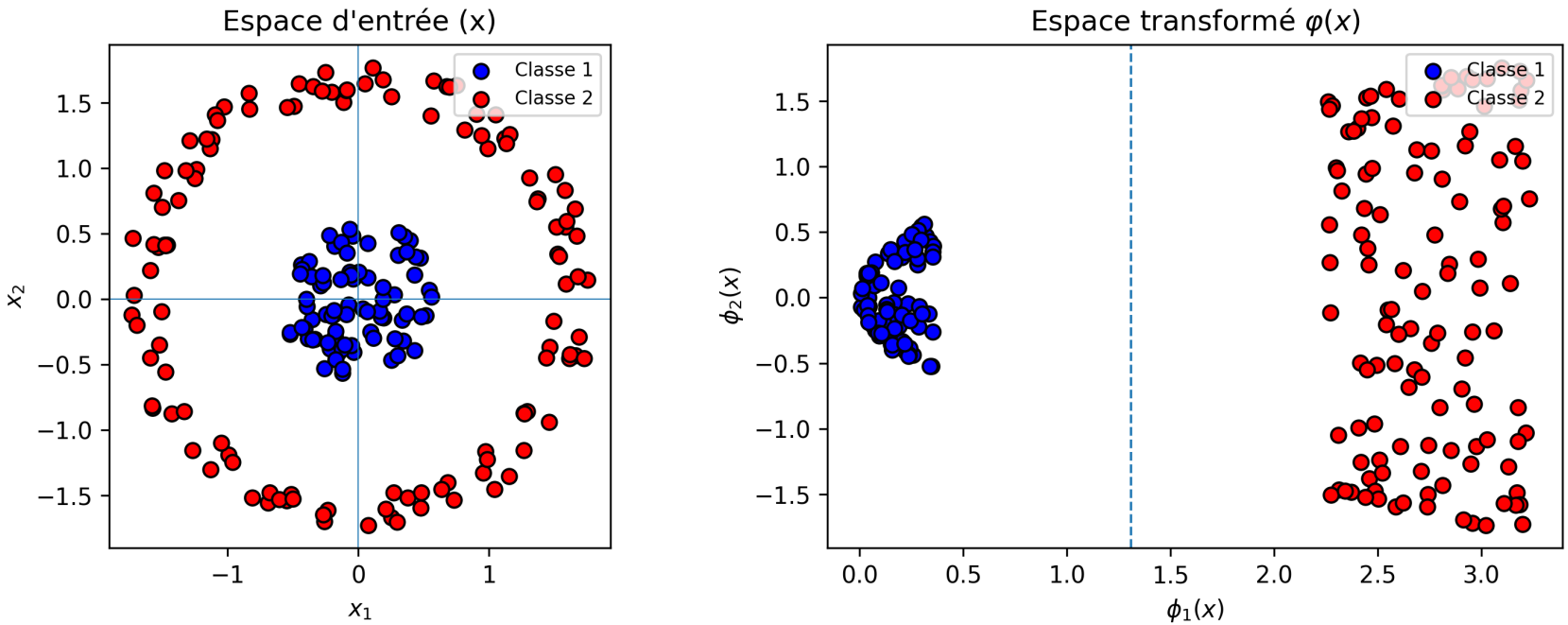
Quels noyaux utiliser en pratique ?

- **Noyau linéaire** Simple, rapide, adapté si les données sont presque linéaires.
- **Noyau polynomial** Apprend des relations du type interactions, courbes douces.
- **Noyau RBF (Gaussian)** Le plus utilisé. Permet d'apprendre des relations très **non linéaires**. Ne nécessite pas de choisir un degré.
- **Noyau sigmoïdal** Proche des réseaux neuronaux. Rarement utilisé aujourd'hui.

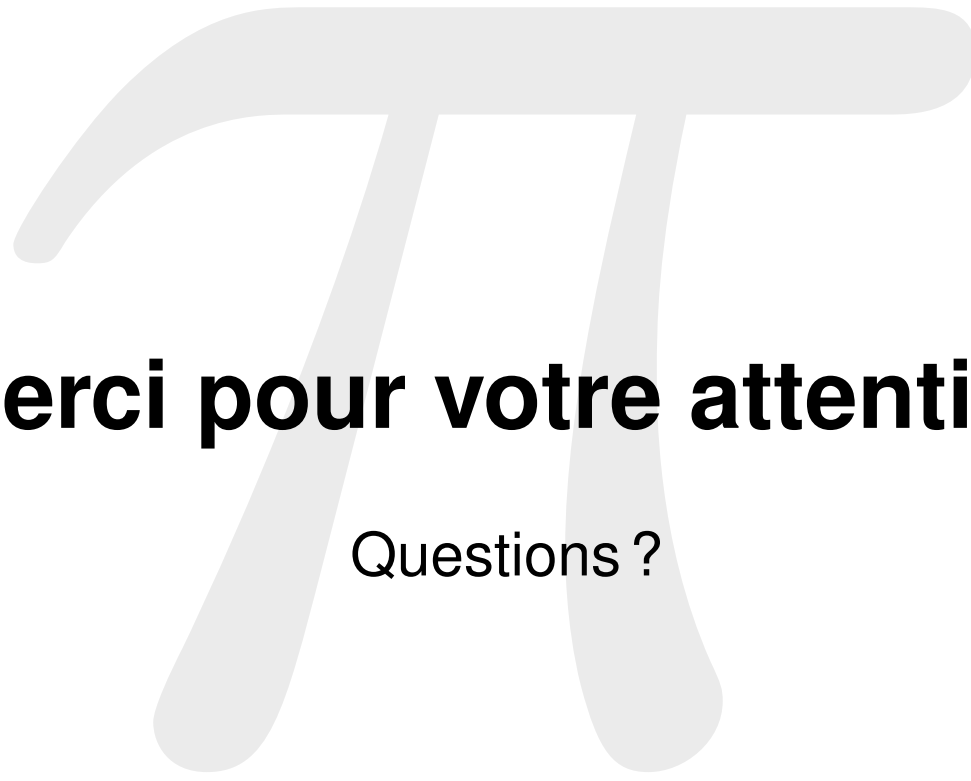
Recommandation

En pratique : **RBF** est presque toujours le meilleur choix par défaut.

Vision schématique du kernel trick



Dans l'espace d'origine : données non linéaires $\Rightarrow$ **SVR linéaire impossible**.
Après transformation implicite $\varphi(x)$: hyperplan + tube linéaire $\Rightarrow$ **SVR linéaire possible**.



Merci pour votre attention

Questions ?