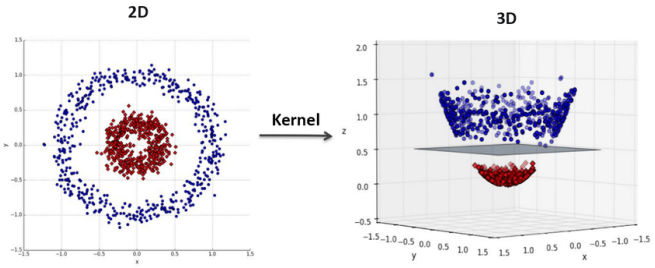
SVM Non Linéaire avec Noyaux

Introduction aux SVM Non Linéaires

Les SVM linéaires fonctionnent bien si les données sont linéairement séparables. Grâce au SVM non linéaires, on peut utiliser des fonctions noyaux pour traiter des problèmes où la séparation linéaire n'est pas possible dans l'espace d'origine.

Le Kernel Trick

Le noyau $k(x, z) = \langle \Phi(x), \Phi(z) \rangle_H$ permet de calculer implicitement un produit scalaire dans un espace transformé H sans construire $\Phi(x)$ explicitement. Cela permet de traiter des données non linéairement séparables.



Propriétés des Noyaux

- **Symétrie** : $k(x, z) = k(z, x)$.
- **Définie positive** :

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0.$$

- Les noyaux définis positifs assurent l'existence d'une solution au problème dual du SVM.

Matrice de Gram

La matrice de Gram G est définie par $G_{ij} = k(x_i, x_j)$. Elle est symétrique et semi-définie positive pour un noyau valide. Utilisée dans le problème dual, elle représente les relations entre les données dans l'espace transformé. Elle est de dimension $n \times n$ (où n est le nombre d'observations).

Exemples de Noyaux

Les noyaux permettent de modéliser des relations complexes entre les données. Voici les noyaux les plus courants et leurs paramètres principaux :

Noyau Linéaire

$$k(x, z) = \langle x, z \rangle$$

Noyau Polynomial

$$k(x, z) = (\langle x, z \rangle + c)^d$$

- d : Degré du polynôme ($d \geq 2$).
- c : Constante d'ajustement (souvent $c \geq 0$).
- **Utilisation** : Modélise des interactions quadratiques, cubiques ou de degré supérieur.

Hugo Tondenier

Noyau Gaussien (RBF)

$$k(x, z) = \exp\left(-\gamma \|x - z\|^2\right)$$

où γ est un paramètre positif défini par la relation :

$$\gamma = \frac{1}{2\sigma^2}$$

- γ : Contrôle la portée d'influence du noyau.
  - γ grand (σ petit) : Modèle complexe, influence locale.
  - γ petit (σ grand) : Modèle simple, influence globale.
- σ : Écart-type qui définit la largeur du noyau gaussien.
- **Utilisation** : Très flexible, adapté à des frontières de décision complexes.

Noyau Sigmoïde

$$k(x, z) = \tanh(\kappa \langle x, z \rangle + c)$$

- κ : Contrôle la pente de la fonction sigmoïde.
- c : Décalage (*offset*).
- Fonctionne bien pour des problèmes où les données peuvent être séparées par une frontière douce, mais non linéaire. (rarement utile)

Formulation pour la Classification

Problème Primal

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \xi_i \\ \text{sous contraintes} \quad & y_i \left(\langle \mathbf{w}, \Phi(x_i) \rangle + b \right) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned}$$

Problème Dual

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \\ \text{sous contraintes} \quad & 0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n} \alpha_i y_i = 0. \end{aligned}$$

Fonction de Décision

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i k(x_i, x) + b \right).$$

Seuls les observations x_i avec $\alpha_i > 0$ (vecteurs de support) sont utilisés, car pour les autres, $\alpha_i = 0$, ce qui rend leur contribution nulle. Cela optimise le calcul.

Formulation pour la Régression SVR

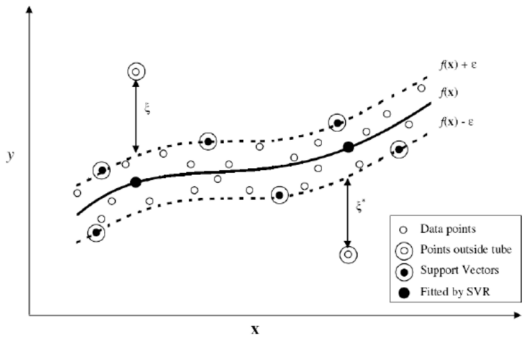
La SVR (Support Vector Regression) est une adaptation des SVM pour les problèmes de régression. Elle utilise une marge ε qui définit une zone de tolérance autour de la fonction de régression.

Principe de la SVR

La SVR cherche à trouver une fonction $f(x)$ telle que :

- Les points s'écartent au maximum de ε de leurs vraies valeurs y_i
- La fonction est aussi "plate" que possible (minimisation de $\|\mathbf{w}\|$)
- Les déviations supérieures à ε sont permises mais pénalisées (variables ξ_i, ξ_i^*)

Hugo Tondenier



Variables d'écart dans la SVR

Dans la SVR, nous avons deux types de déviations possibles par rapport au tube de tolérance ε :

- ξ_i : mesure la déviation au-dessus du tube ($f(x_i) > y_i + \varepsilon$)
- ξ_i^* : mesure la déviation en-dessous du tube ($f(x_i) < y_i - \varepsilon$)

Cette distinction est nécessaire car :

- Contrairement à la classification où l'erreur est unidirectionnelle (être du mauvais côté de la marge), en régression l'erreur peut être positive ou négative
- Le tube de tolérance ε crée deux frontières :
  - Une frontière supérieure : $y_i + \varepsilon$
  - Une frontière inférieure : $y_i - \varepsilon$
- Pour chaque point x_i, au maximum une seule variable d'écart est non nulle : $\xi_i \cdot \xi_i^* = 0$

Problème Primal

$$\begin{aligned} \min_{\mathbf{w},b,\xi,\xi^*} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ \text{sous contraintes} \quad & y_i - \langle \mathbf{w}, \Phi(x_i)\rangle - b \le \varepsilon + \xi_i, \\ & \langle \mathbf{w}, \Phi(x_i)\rangle + b - y_i \le \varepsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \ge 0. \end{aligned}$$

Où :

- ε définit la largeur du tube de tolérance
- ξ_i, ξ_i^* sont les variables d'écart (slack variables)
- C est le paramètre de régularisation qui contrôle le compromis entre la platitude de f et la tolérance aux déviations

Problème Dual

$$\begin{aligned} \max_{\alpha,\alpha^*} \quad & -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(x_i, x_j) \\ & -\varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) \\ \text{sous contraintes} \quad & 0 \le \alpha_i, \alpha_i^* \le C, \quad \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0. \end{aligned}$$

Où :

- α_i, α_i^* sont les multiplicateurs de Lagrange
- Pour chaque point, au moins un des multiplicateurs est nul : $\alpha_i \cdot \alpha_i^* = 0$
- Les points avec $\alpha_i \neq 0$ ou $\alpha_i^* \neq 0$ sont les vecteurs de support

Hugo Tondenier

Fonction de Décision

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) k(x_i, x) + b$$

Propriétés importantes

- **Éparsité** : Seuls les points hors du tube ε deviennent des vecteurs de support
- **Robustesse** : Les points à l'intérieur du tube n'influencent pas la solution
- **Non-linéarité** : L'utilisation de noyaux permet de modéliser des relations non-linéaires
- **Convexité** : Le problème d'optimisation est convexe, garantissant une solution optimale unique

Formulation Matricielle de la classification

Problème Primal

$$\min_{\mathbf{w}, b, \boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\mathbf{1}^\top \boldsymbol{\xi}$$

sous les contraintes :

$$\mathbf{y} \odot (\Phi(\mathbf{X})\mathbf{w} + b\mathbf{1}) \geq \mathbf{1} - \boldsymbol{\xi}, \quad \boldsymbol{\xi} \geq 0.$$

$\odot$ est le produit d'Hadamard (terme à terme)

Lagrangien

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\mathbf{1}^\top \boldsymbol{\xi} - \boldsymbol{\alpha}^\top \big(\mathbf{y} \odot (\Phi(\mathbf{X})\mathbf{w} + b\mathbf{1}) - \mathbf{1} + \boldsymbol{\xi}\big) - \boldsymbol{\mu}^\top \boldsymbol{\xi},$$

où :

- $\boldsymbol{\alpha} \in \mathbb{R}^n$ $(\alpha_i \geq 0)$: multiplicateurs pour $y_i(\langle \mathbf{w}, \Phi(x_i) \rangle + b) \geq 1 - \xi_i$,
- $\boldsymbol{\mu} \in \mathbb{R}^n$ $(\mu_i \geq 0)$: multiplicateurs pour $\xi_i \geq 0$.

Conditions KKT

1. Gradient par rapport à $\mathbf{w}$:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} - \Phi(\mathbf{X})^\top (\mathbf{y} \odot \boldsymbol{\alpha}) = 0$$
$$\implies \mathbf{w} = \Phi(\mathbf{X})^\top (\mathbf{y} \odot \boldsymbol{\alpha}).$$

2. Gradient par rapport à b :

$$\frac{\partial \mathcal{L}}{\partial b} = -\mathbf{1}^\top (\mathbf{y} \odot \boldsymbol{\alpha}) = 0$$
$$\implies \mathbf{1}^\top (\mathbf{y} \odot \boldsymbol{\alpha}) = 0.$$

3. Gradient par rapport à $\boldsymbol{\xi}$:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\xi}} = C\mathbf{1} - \boldsymbol{\alpha} - \boldsymbol{\mu} = 0$$
$$\implies \boldsymbol{\alpha} + \boldsymbol{\mu} = C\mathbf{1}.$$

Formulation Duale

En substituant $\mathbf{w}$ et $\boldsymbol{\xi}$ dans le lagrangien, le problème dual devient :

$$\max_{\boldsymbol{\alpha}} \quad \mathbf{1}^\top \boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^\top (\mathbf{Y}G\mathbf{Y})\boldsymbol{\alpha}$$

sous les contraintes :

$$0 \leq \alpha_i \leq C, \quad \mathbf{1}^\top (\mathbf{y} \odot \boldsymbol{\alpha}) = 0,$$

où :

- $\mathbf{Y} = \mathrm{diag}(\mathbf{y})$ est une matrice diagonale contenant les étiquettes,
- $G = \Phi(\mathbf{X})\Phi(\mathbf{X})^\top$ est la matrice de Gram $(G_{ij} = k(x_i, x_j))$.

Hugo Tondenier