

Support Vector Machines (SVM)

Théorie complète : Classification et Régression

1 Cadre général

On considère un ensemble de données supervisées :

$$\{(x_i, y_i)\}_{i=1}^m$$

avec :

$$x_i \in \mathbb{R}^d$$

- **Classification** : $y_i \in \{-1, +1\}$
- **Régression** : $y_i \in \mathbb{R}$

Le modèle général SVM est :

$$f(x) = w^\top \phi(x) + b$$

où :

- ϕ est une transformation vers un espace de caractéristiques,
- w est le vecteur des poids,
- b est le biais.

Le principe fondamental des SVM est de **contrôler la complexité du modèle** via la norme $\|w\|$ tout en assurant un bon ajustement aux données.

2 SVM pour la classification

2.1 Hyperplan et distance

Un hyperplan dans \mathbb{R}^d est défini par :

$$H = \{x \mid w^\top x + b = 0\}$$

La distance signée d'un point x à l'hyperplan est :

$$d(x, H) = \frac{w^\top x + b}{\|w\|}$$

2.2 Marge fonctionnelle et géométrique

La marge fonctionnelle d'un point (x_i, y_i) est :

$$\gamma_i = y_i(w^\top x_i + b)$$

La marge géométrique est :

$$\hat{\gamma}_i = \frac{y_i(w^\top x_i + b)}{\|w\|}$$

La marge globale est :

$$\hat{\gamma} = \min_i \hat{\gamma}_i$$

2.3 Normalisation canonique

On impose :

$$\min_i y_i(w^\top x_i + b) = 1$$

Les hyperplans de marge deviennent :

$$w^\top x + b = \pm 1$$

La largeur de la marge est :

$$\frac{2}{\|w\|}$$

Maximiser la marge revient donc à minimiser $\|w\|^2$.

2.4 SVM linéaire à marge dure

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

sous contraintes :

$$y_i(w^\top x_i + b) \geq 1 \quad \forall i$$

2.5 SVM à marge souple

On introduit des variables de relâchement $\xi_i \geq 0$:

$$y_i(w^\top x_i + b) \geq 1 - \xi_i$$

Le problème primal devient :

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i$$

2.6 Hinge loss

Le problème précédent est équivalent à :

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \max(0, 1 - y_i f(x_i))$$

2.7 Dualité (classification)

Le Lagrangien est :

$$\mathcal{L} = \frac{1}{2} \|w\|^2 - \sum_i \alpha_i [y_i(w^\top x_i + b) - 1 + \xi_i] - \sum_i \mu_i \xi_i$$

Conditions stationnaires :

$$w = \sum_i \alpha_i y_i x_i$$

$$\sum_i \alpha_i y_i = 0$$

—

2.8 Problème dual

$$\max_{\alpha} \quad \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

Sous contraintes :

$$0 \leq \alpha_i \leq C, \quad \sum_i \alpha_i y_i = 0$$

—

2.9 Fonction de décision

$$f(x) = \sum_i \alpha_i y_i \langle x_i, x \rangle + b$$

Les points tels que $\alpha_i > 0$ sont appelés **support vectors**.

—

3 Noyaux (kernels)

3.1 Définition

Un noyau est une fonction :

$$K(x, z) = \langle \phi(x), \phi(z) \rangle$$

Il permet de travailler dans un espace de dimension élevée sans calculer explicitement ϕ .

—

3.2 Condition de Mercer

Un noyau valide vérifie :

$$\sum_{i,j} c_i c_j K(x_i, x_j) \geq 0$$

pour tout ensemble fini $\{x_i\}$.

La matrice de Gram est donc semi-définie positive.

—

3.3 Exemples de noyaux

- Linéaire : $K(x, z) = x^\top z$
 - Polynômial : $K(x, z) = (\gamma x^\top z + r)^p$
 - RBF : $K(x, z) = \exp(-\gamma \|x - z\|^2)$
 - Sigmoïde : $K(x, z) = \tanh(\gamma x^\top z + r)$
-

3.4 Kernel trick

Dans le dual :

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

—

4 Support Vector Regression (SVR)

4.1 Principe

On cherche une fonction $f(x)$ telle que :

$$|y_i - f(x_i)| \leq \varepsilon$$

Les erreurs inférieures à ε ne sont pas pénalisées.

—

4.2 Primal SVR

On introduit $\xi_i, \xi_i^* \geq 0$:

$$\min_{w, b, \xi, \xi^*} \frac{1}{2} \|w\|^2 + C \sum_i (\xi_i + \xi_i^*)$$

Sous contraintes :

$$\begin{cases} y_i - w^\top x_i - b \leq \varepsilon + \xi_i \\ w^\top x_i + b - y_i \leq \varepsilon + \xi_i^* \end{cases}$$

—

4.3 Perte ε -insensible

$$\ell_\varepsilon(y, f) = \max(0, |y - f| - \varepsilon)$$

—

4.4 Dual SVR

$$w = \sum_i (\alpha_i - \alpha_i^*) x_i$$

$$\sum_i (\alpha_i - \alpha_i^*) = 0$$

—

4.5 Problème dual (SVR)

$$\max_{\alpha, \alpha^*} -\frac{1}{2} \sum_{i,j} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(x_i, x_j) - \varepsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i (\alpha_i - \alpha_i^*)$$

Sous :

$$0 \leq \alpha_i, \alpha_i^* \leq C$$

4.6 Fonction de régression

$$f(x) = \sum_i (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

Les support vectors sont les points tels que $\alpha_i > 0$ ou $\alpha_i^* > 0$.

5 Synthèse classification / régression

$$f(x) = \sum_i \beta_i K(x_i, x) + b$$

- Classification : $\beta_i = \alpha_i y_i$
- Régression : $\beta_i = \alpha_i - \alpha_i^*$

Les SVM reposent sur :

- une optimisation convexe,
- une régularisation explicite,
- un petit nombre de support vectors,
- le kernel trick pour la non-linéarité.