Machines à vecteurs de support (SVM)

Ahmedou Yahye Kheyri

Faculté des Sciences de Sfax

4 mai 2025

Table des matières I

Qu'est-ce qu'une SVM ?

- Méthode d'apprentissage supervisé.
- Utilisée principalement pour la classification et la régression.
- Appartient à la famille des classifieurs linéaires généralisés.
- Introduites par Vapnik et al. en 1992.

Idée Principale : Trouver le meilleur séparateur (hyperplan) entre différentes classes de données.

Pourquoi les SVM ?

- Séparateurs à vaste marge (maximisent la distance aux points les plus proches).
- Approche statistique solide pour l'approximation et l'estimation de fonctions.
- Capacité à travailler avec des données de grandes dimensions.
- Bonne capacité de généralisation (minimisation du risque structurel - SRM).
- Moins sujettes aux minima locaux que d'autres méthodes (ex : réseaux de neurones traditionnels).

Tâche de Classification : Trier des individus en fonction de leurs caractéristiques.

Données Linéairement Séparables : Il existe un hyperplan qui peut parfaitement séparer les différentes classes de données.
L'Hyperplan Séparateur :

- Représenté par l'équation $w \cdot x + b = 0$.
- w est le vecteur normal à l'hyperplan, b est le biais.
- La classification d'un point x se fait selon le signe de $w \cdot x + b$.

Problème : Pour des données linéairement séparables, il existe une infinité d'hyperplans séparateurs possibles.

Objectif : Choisir le "meilleur" hyperplan pour une meilleure généralisation.

Concept de Marge : La distance entre l'hyperplan séparateur et le point de données le plus proche de cet hyperplan.

Hyperplan Optimal : L'hyperplan qui maximise cette marge.

Pourquoi Maximiser la Marge ?

- Offre une meilleure capacité de généralisation.
- Moins sensible au bruit et aux valeurs aberrantes (outliers).
- Minimise le risque de mauvaise classification sur de nouvelles données.

Distance à l'Hyperplan : La distance d'un point x à l'hyperplan $w \cdot x + b = 0$ est donnée par :

$$\frac{|w \cdot x + b|}{||w||}$$

Marge pour un Hyperplan Canonique : Si l'hyperplan est mis à l'échelle tel que $\min_i |w \cdot x_i + b| = 1$, la marge est :

$$\frac{1}{||w||}$$

Maximiser la marge revient à minimiser $||w||$, ou de manière équivalente, minimiser $\frac{1}{2}||w||^2$.

Formulation du Problème (Cas Linéaire Séparable) - Contraintes I

Problème d'Optimisation Primal : Trouver l'hyperplan séparateur de marge maximale revient à résoudre le problème d'optimisation suivant :

$$\text{Minimiser} : \frac{1}{2}||w||^2$$
$$\text{Sous contraintes} : y_i(w \cdot x_i + b) \geq 1 \quad \forall i = 1, \ldots, l$$

où l est le nombre de points d'entraînement.

Formulation du Problème (Cas Linéaire Séparable) - Contraintes II

Explication des Contraintes :

- Assure que chaque point est du bon côté de la marge (ou sur la marge).
- Implique que l'hyperplan est "canonique".

Nature du Problème : C'est un problème d'optimisation quadratique convexe avec contraintes linéaires.

Avantage : L'existence d'un optimum global unique est garantie.

Formulation Duale et Multiplicateurs de Lagrange - Problème Dual I

Importance de la Formulation Duale : Simplifie la résolution du problème, surtout pour les cas non linéaires et l'introduction des noyaux.

Lagrangien : On utilise la méthode des multiplicateurs de Lagrange ($\alpha_i \geq 0$) pour transformer le problème primal en un problème dual.

$$L(w, b, \alpha) = \frac{1}{2} w \cdot w - \sum_{i=1}^{l} \alpha_i (y_i (w \cdot x_i + b) - 1)$$

Problème Dual :

$$\text{Maximiser : } \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

$$\text{Sous contraintes : } \sum_{i=1}^{l} \alpha_i y_i = 0$$

$$\alpha_i \geq 0 \quad \forall i = 1, \ldots, l$$

La solution de ce problème dual nous donne les valeurs α_i^*.

Définition : Les vecteurs de support sont les points d'entraînement pour lesquels les multiplicateurs de Lagrange α_i^* sont strictement positifs ($\alpha_i^* > 0$).

Propriété Clé (Conditions KKT) : Pour les vecteurs de support, la contrainte $y_i(w \cdot x_i + b) \geq 1$ est active : $y_i(w \cdot x_i + b) = 1$. Ces points se trouvent exactement sur la marge.

Importance : L'hyperplan optimal est entièrement déterminé par les vecteurs de support. Les autres points d'entraînement n'influencent pas la position de l'hyperplan.

Le vecteur de poids w peut être exprimé comme une combinaison linéaire des vecteurs de support :

$$w = \sum_{i=1}^{l} \alpha_i^* y_i x_i$$

Classification avec le Modèle Entraîné - Fonction de Décision I

Une fois les α_i^* optimaux trouvés (en résolvant le problème dual), on peut déterminer w^* et b^*.

- $w^* = \sum_{i=1}^{l} \alpha_i^* y_i x_i$
- b^* peut être calculé en utilisant n'importe quel vecteur de support x_k (pour lequel $\alpha_k^* > 0$) et l'égalité $y_k(w^* \cdot x_k + b^*) = 1$.

Classification avec le Modèle Entraîné - Fonction de Décision II

Fonction de Décision (pour un nouveau point x) :

$$f(x) = w^* \cdot x + b^*$$

En utilisant l'expression de w^* :

$$f(x) = \sum_{i=1}^{l} \alpha_i^* y_i K(x_i, x) + b^*$$

Classification : Le signe de $f(x)$ détermine la classe du point x.

$$\text{Classe}(x) = \text{signe}(f(x))$$

Problèmes Non Linéairement Séparables I

- **Réalité :** Dans de nombreux cas pratiques, les données ne sont pas linéairement séparables dans l'espace d'entrée d'origine.
- **Limitation des SVM Linéaires :** Les SVM linéaires ne peuvent pas trouver d'hyperplan séparateur si les données ne sont pas linéairement séparables.

Idée Principale : Transformer les données de l'espace d'entrée ($\mathbb{R}^n$) vers un espace de redescription (feature space) de dimension potentiellement plus élevée ($\mathbb{R}^r$) où elles deviennent linéairement séparables.

- Transformation $\Phi : \mathbb{R}^n \to \mathbb{R}^r$, $x \mapsto \Phi(x)$.

Problème : Appliquer explicitement Φ et calculer les produits scalaires $\Phi(x_i) \cdot \Phi(x_j)$ dans l'espace de redescription peut être très coûteux computationnellement si r est grand.

L'Astuce du Noyau : Utiliser une fonction noyau $K(x_i, x_j)$ qui calcule directement le produit scalaire dans l'espace de redescription, sans avoir besoin de calculer explicitement $\Phi(x_i)$ et $\Phi(x_j)$.

$$K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$$

Avantage Clé : Permet de travailler dans l'espace de redescription (où les données sont séparables) tout en effectuant les calculs dans l'espace d'entrée d'origine (moins coûteux).

La formulation duale du problème d'optimisation et la fonction de décision ne dépendent des données d'entraînement que par le biais de produits scalaires $(x_i \cdot x_j)$.

En remplaçant simplement $x_i \cdot x_j$ par $K(x_i, x_j)$ dans la formulation duale et la fonction de décision, on peut utiliser les SVM pour les problèmes non linéaires.

Problème Dual (avec noyau) :

$$\text{Maximiser : } \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$\text{Sous contraintes : } \sum_{i=1}^{l} \alpha_i y_i = 0$$

$$\alpha_i \geq 0 \quad \forall i = 1, \ldots, l$$

Fonction de Décision (avec noyau) :

$$f(x) = \sum_{i=1}^{l} \alpha_i^* y_i K(x_i, x) + b^*$$

b^* est calculé en utilisant un vecteur de support x_k (pour lequel $\alpha_k^* > 0$) et l'égalité $y_k(\sum_{i=1}^{l} \alpha_i^* y_i K(x_i, x_k) + b^*) = 1$.

Le choix du noyau est crucial et dépend du problème.

Quelques Noyaux Populaires :

- **Noyau Linéaire :** $K(x_i, x_j) = x_i \cdot x_j$ (Revient à une SVM linéaire).
- **Noyau Polynomial :** $K(x_i, x_j) = (x_i \cdot x_j + c)^d$ (où d est le degré, $c \geq 0$).
- **Noyau Gaussien (RBF - Radial Basis Function) :**

$$K(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right)$$

  (où σ^2 est un paramètre). Très utilisé et donne souvent de bons résultats.
- **Noyau Sigmoïde :** $K(x_i, x_j) = \tanh(\kappa x_i \cdot x_j + \theta)$.

Marge Souple (Soft Margin SVM) I

Problème avec la Marge Stricte : Les données réelles contiennent souvent du bruit ou des points aberrants, rendant une séparation parfaite difficile ou conduisant au sur-apprentissage.

Idée de la Marge Souple : Permettre à certains points d'être mal classés ou de se trouver à l'intérieur de la marge, en échange d'une frontière de décision plus générale et robuste.

Variables d'Écart (Slack Variables ξ_i) : Introduire une variable $\xi_i \geq 0$ pour chaque point x_i qui mesure à quel point le point viole la contrainte de marge.

- Contrainte modifiée : $y_i(w \cdot x_i + b) \geq 1 - \xi_i$.
- $\xi_i = 0$: point bien classé et en dehors de la marge.
- $0 < \xi_i < 1$: point bien classé mais à l'intérieur de la marge.
- $\xi_i \geq 1$: point mal classé.

Objectif : Minimiser $\frac{1}{2}||w||^2$ (maximiser la marge) ET minimiser la somme des écarts $\sum \xi_i$ (minimiser les erreurs de classification et les violations de marge).

Formulation du Problème (Marge Souple) - Formulation Duale I

Problème Primal :

$$\text{Minimiser : } \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\xi_i$$

$$\text{Sous contraintes : } y_i(w \cdot x_i + b) \geq 1 - \xi_i \quad \forall i = 1, \ldots, l$$

$$\xi_i \geq 0 \quad \forall i = 1, \ldots, l$$

Paramètre C : Une constante positive qui contrôle le compromis entre la maximisation de la marge et la minimisation des erreurs d'entraînement.

- Grand C : Pénalise fortement les erreurs, frontière plus complexe, risque de sur-apprentissage.
- Petit C : Tolère plus d'erreurs, frontière plus simple, risque de sous-apprentissage.
- Le choix de C se fait généralement par validation croisée.

Formulation du Problème (Marge Souple) - Formulation Duale II

Formulation Duale : La formulation duale de la marge souple est très similaire à celle de la marge stricte, avec une contrainte supplémentaire sur α_i : $0 \leq \alpha_i \leq C$.

$$\begin{aligned}
\text{Maximiser :} \quad & \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\
\text{Sous contraintes :} \quad & \sum_{i=1}^{l} \alpha_i y_i = 0 \\
& 0 \leq \alpha_i \leq C \quad \forall i = 1, \ldots, l
\end{aligned}$$

Dans le cas de la marge souple, les vecteurs de support sont les points pour lesquels $\alpha_i^* > 0$.

Types de Vecteurs de Support : Ces points peuvent se trouver :

- Exactement sur la marge ($y_i(w \cdot x_i + b) = 1$, $\xi_i = 0$, $0 < \alpha_i^* < C$). Appelés vecteurs de support libres/non bornés.
- À l'intérieur de la marge ($y_i(w \cdot x_i + b) < 1$, $\xi_i > 0$, $\alpha_i^* = C$). Appelés vecteurs de support bornés. Ces points sont soit mal classés, soit bien classés mais violent la marge.

Comme précédemment, l'hyperplan optimal est déterminé uniquement par les vecteurs de support ($\alpha_i^* > 0$).

Domaines d'Application des SVM :

- **Reconnaissance de Formes (Pattern Recognition) :**
  - Classification d'images.
  - Reconnaissance d'écriture manuscrite.
  - Détection de visages.
- **Classification de Texte :**
  - Filtrage de spam.
  - Classification de documents par catégorie.

- **Bioinformatique :**
  - Classification de séquences d'ADN/protéines.
  - Analyse de données d'expression génique.
- **Finance :**
  - Prédiction de marché.
  - Détection de fraude.
- **Santé :**
  - Diagnostic médical.
- **Autres Domaines :** Les SVM sont utilisées dans de nombreux autres domaines où la classification est nécessaire.

Forces des SVM I

Forces :

- Bonne performance empirique et théorique.
- Gestion efficace des espaces de grande dimension.
- Moins de risque de tomber dans des minima locaux (problème d'optimisation convexe).
- Contrôle explicite du compromis complexité/erreur (via la marge et le paramètre C).
- Détermination unique de la solution optimale pour des paramètres donnés.

Faiblesses :

- Le choix du noyau et des hyperparamètres (C, σ pour RBF, etc.) peut être délicat et nécessite souvent une validation croisée.
- Peut être coûteux en calcul pour de très grands ensembles de données d'entraînement (même si des méthodes rapides existent).
- L'interprétabilité du modèle peut être moins directe que d'autres algorithmes (comme les arbres de décision).

Scénario : Classifier des emails en "spam" (+1) et "non-spam" (-1) basés sur deux caractéristiques (par exemple, fréquence du mot "promotion", nombre de points d'exclamation).

Visualisation 2D : Placer les emails dans un graphique 2D selon leurs caractéristiques, en marquant les spams (+) et les non-spams (-).
SVM Linéaire (si possible) : Trouver la ligne (hyperplan en 2D) qui sépare le mieux les spams des non-spams, avec la plus grande marge. Identifier les emails qui sont les vecteurs de support (ceux sur la marge).
SVM Non Linéaire (si nécessaire) : Si les points ne sont pas séparables linéairement, utiliser un noyau (par exemple RBF) pour trouver une frontière de décision non linéaire dans l'espace 2D d'origine.

Marge Souple : Si certains emails sont difficiles à classer ou s'il y a du bruit, utiliser la marge souple pour permettre quelques erreurs en échange d'une meilleure généralisation.

Conclusion I

- Les SVM sont des algorithmes de classification puissants basés sur la maximisation de la marge.
- L'utilisation de l'astuce du noyau permet d'étendre les SVM aux problèmes non linéairement séparables.
- La marge souple gère efficacement le bruit et les données non parfaitement séparables.
- Les vecteurs de support sont les points clés qui définissent le modèle.
- Les SVM ont démontré leur efficacité dans de nombreux domaines d'application.
- Le choix des hyperparamètres (C, noyau et ses paramètres) est crucial pour la performance.

Questions

Merci de votre attention !

Avez-vous des questions ?