



5

Modèles SVM

5. Modèles – SVM



Les **SVM** machines à vecteurs de support sont le plus souvent utilisées sur des données avec beaucoup de variables.

L'idée des SVMs est de chercher un hyperplan (un séparateur linéaire) qui sépare au mieux les objets de chaque catégorie. En effet, il peut y avoir une infinité de séparateurs possibles comme le montre la figure ci-dessous.

La meilleure séparation entre les deux classes est celle qui maximise la marge entre les deux séparatrices. Le meilleur hyperplan selon les SVMs est celui qui maximise les marges avec les objets de chaque catégorie.

Ces objets sont d'ailleurs appelés "vecteurs de support" car ils supportent les hyper-plans parallèles.

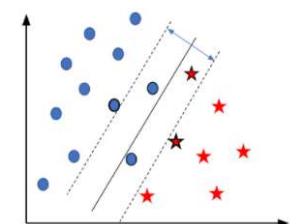
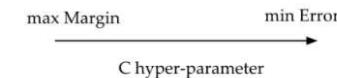
Par contre, il est fréquent qu'il n'y ait pas d'hyperplan capable de séparer parfaitement les données.

Dans ce cas, on peut permettre qu'il y ait un certain nombre d'erreurs.

L'optimisation de l'hyper-paramètre C permet justement de fixer un compromis entre la maximisation des marges et la minimisation des erreurs. Prédiction = mesure de la distance du nouveau point aux vecteurs de support.

Astuce du noyau : entraîner un modèle dans un espace de grande dimension en calculant directement la distance (produit scalaire) des points de données pour la représentation étendue des caractéristiques sans jamais calculer réellement cette extension. Ex : noyau polynomial, noyau rbf radial gaussien.

$\arg \max$ Margin
 $\arg \min$ Errors



Paramètres :

- Gamma : contrôle la largeur du noyau gaussien, détermine la portée de l'influence de certains échantillons (petite valeur : grande portée, grande : petite portée).
 - C : paramètre de régularisation (limite l'importance de chaque point).

5. Modèles – SVR



L'objectif du vecteur de support est de trouver un hyperplan dans un espace à N dimensions qui peut classer les points de données.

Les points de données situés sur la marge et les plus proches de l'hyperplan sont appelés vecteurs de support. Maintenant, lorsque la plupart des données se trouvent dans la meilleure marge vers chaque côté de l'hyperplan, alors le SVR ou la régression vectorielle de soutien peut être utilisé pour identifier et prédire les données dépendantes très clairement.

Sa formule mathématique pour les lignes marginales vers chaque côté de l'hyperplan est représentée par :

$$y_i = \langle w, x_i \rangle + b + \text{déviation}$$

$$y_i = \langle w, x_i \rangle + b - \text{déviation}$$

Étapes à suivre pour construire un modèle de régression de support :

1) Trouvez vos ensembles de données X et Y, indépendants et dépendants, pour entraîner le modèle.

2) Observez les données en un coup d'œil et essayez d'ajuster le paramètre de noyau le mieux adapté. Vous pouvez également essayer de tracer les points de données et de voir la corrélation.

Il peut être linéaire, gaussien ou polynomial, selon la complexité.

Le noyau le plus couramment utilisé est gaussien.

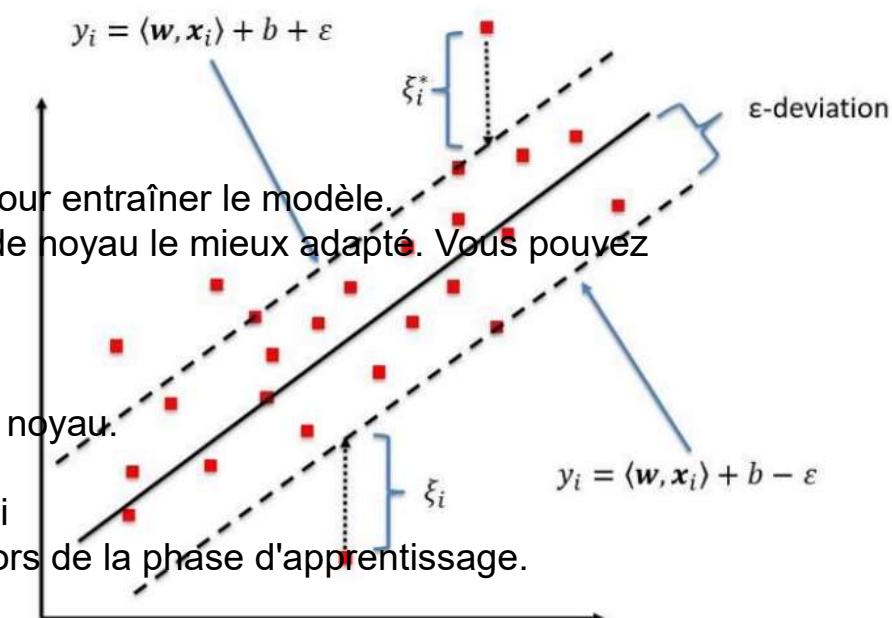
*Noyau polynomial : $K(x,y) = (x \cdot y + 1)^d$, où $d > 0$ est une constante qui définit le noyau.

Ordre.

*Noyau RBF gaussien : $K(x,y) = \exp(-|x-y|^2/2\sigma^2)$, où $\sigma > 0$ est un paramètre qui définit la largeur du noyau. Les paramètres associés d et σ sont déterminés lors de la phase d'apprentissage.

3) Tracer des graphiques

4) Prédire pour toute valeur indépendante.



5. Modèles – SVR



Avantages :

- Puissants et efficaces sur une large variété de jeux de données (peu ou grandes dimensions des features).

Inconvénients :

- Pré-traitement des données : Standardiser avec même échelle : MinMaxScaler.
- Pas pour les grandes quantités de données (temps de consommation mémoire trop grand).
- Ajustement précis des paramètres.
- Peu interprétable.