RER – Support Vector Machines (SVM)

Contexte : Apprendre à utiliser SVM pour effectuer la classification et notamment dans le domaine de la régression

Problématique(s) :

- Comment utiliser le SVM (SVC et SVR) dans les problèmes d'apprentissage supervisé en machine Learning ?
- Comment choisir le noyau pour avoir le meilleur modèle ?

Mots clés :

Machine à vecteur de support (SVM) : Un ensemble de méthodes d'apprentissage supervisé utilisé pour la classification (SVC, prédiction sur des variables qualitative) et la régression (SVR, prédiction sur des variables quantitative).

Régression vectorielle de support (SVR) : Algorithme de régression (SVR, prédiction sur des variables quantitative) basé sur les vecteurs support.

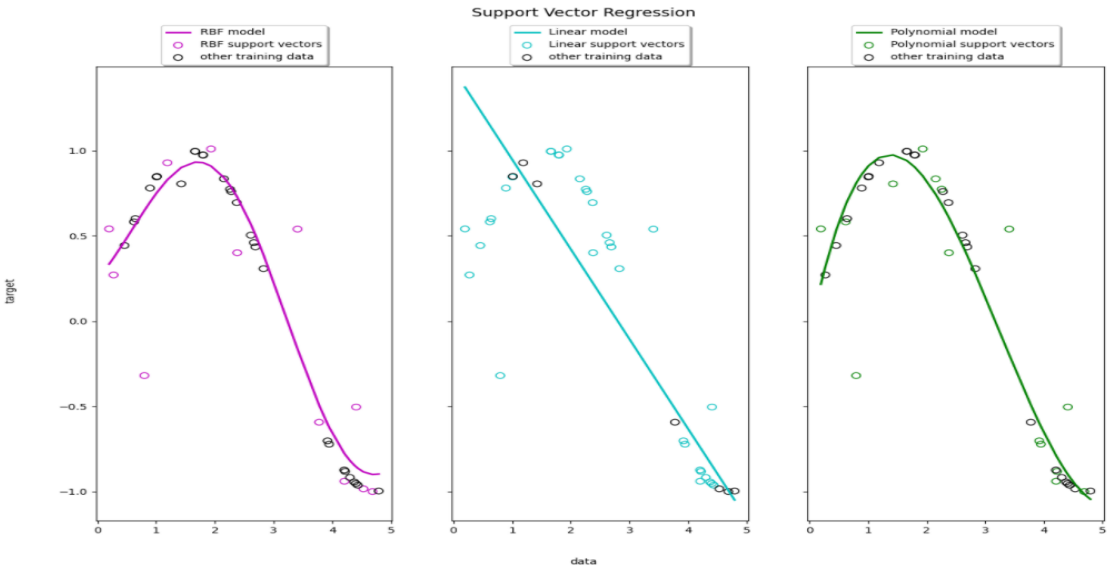
Classification vectorielle de support (SVC) : Algorithme de classification (SVC, prédiction sur des variables qualitative) basé sur les vecteurs support.

Hyperplan : Un Hyperplan est une ligne de séparation optimal entre deux classes de données dans le SVM et une ligne qui prédit la sortie continue en SVR.

Hyperplan (Maths) : Les hyperplans d'un espace vectorielle E de dimension finie n non nulle sont ces sous espaces vectorielle de dimension n-1 dans E.

Kernel (noyau) : L'application d'une transformation de données pour pouvoir les séparer linéairement. Un noyau nous aide à trouver un hyperplan dans l'espace de dimension supérieur sans augmenter le coût de calcul.

Paramètres Kernel :



- **Linéaire :** Une droite.

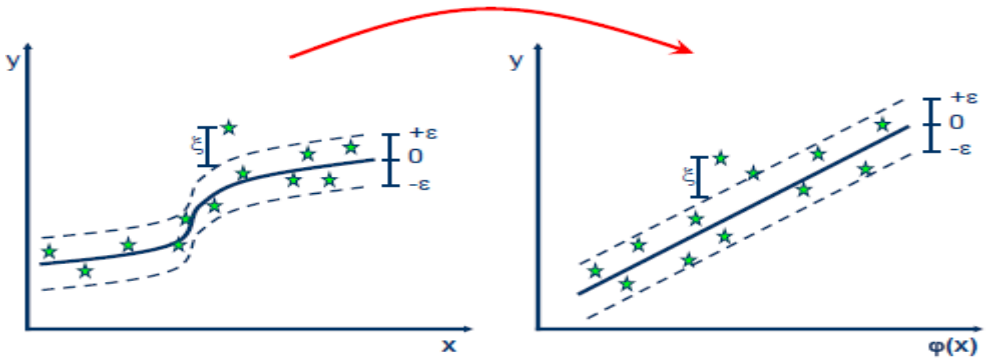
- **Polynomial :** un polynôme de degré n.

- **RBF (Radial Basis Function):** C'est un noyau gaussien.

  - **Gamma (paramètre RBF) :** C'est un hyperparamètre qui influe sur la valeur de proximité maximale entre 2 points pour être regroupés dans la même classe en considérant seulement les points proches de l'hyperplan. Un gamma donne l'allure gaussienne et contrôle la bande passante de la gaussienne.

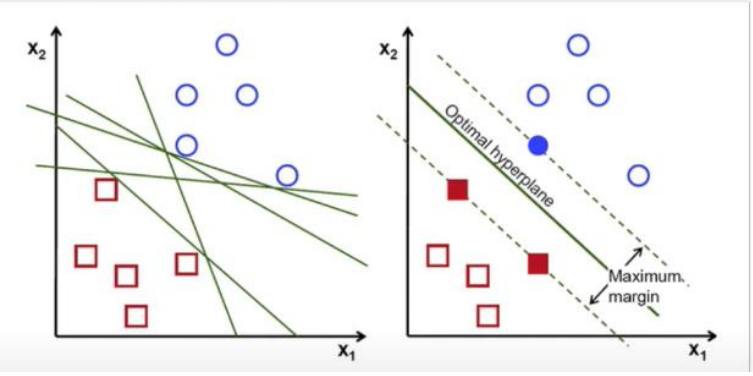
  - **C (paramètre RBF) :** C'est un hyperparamètre de Coût qui contrôle l'erreur.

Epsilon : C'est un hyperparamètre de l'algorithme SVR. Spécifie le tube entre lequel il n'y a pas de pénalité appliquée au point présent dans ce tube, sa valeur détermine le niveau de précision du modèle.



Séparateur : C'est l'hyperplan.

Marge maximale : C'est la distance entre les observations les plus proches de l'hyperplan et l'hyperplan.



Vecteur support : Les vecteurs de support sont les échantillons à l'intérieur de la marge. La droite qui passe par l'échantillon le plus proche de l'hyperplan. Ce sont les échantillons qui sont les plus proche de l'hyperplan qui influencent la position et l'orientation de l'hyperplan.

Xi (ξ) : Elles sont appelées variables ressort qui détermine la tolérance des vecteurs support.

R^2 : Cette métrique mesure la qualité de la prédiction en comparant la variance expliquée par le modèle à la variance totale de la variable cible. Plus le score R2 est proche de 1, plus le modèle explique bien la variance de la variable cible.

Adjusted R^2 : R2 c'est une mesure corrigée de la qualité de l'ajustement (précision du modèle). Il identifie le pourcentage de variance dans le variable cible qui est expliqué par l'entrée ou les entrées. Il prend ne prend que les variables indépendantes avec une certaines signification et vous pénalise pour l'ajout de fonctionnalité qui ne sont pas significatif pour prédire la variable dépendante.

MSE (Mean Squared Error): Cette métrique mesure la moyenne de l'erreur quadratique entre les valeurs prédites et les valeurs réelles. Cette métrique est souvent utilisée pour évaluer la qualité de la prédiction, car elle considère toutes les erreurs, qu'elles soient positives ou négatives, de la même manière.

RMSE (Root Mean Squared Error) : Racine carré de MSE.

MAE (Mean Absolute Error) : Cette métrique mesure la moyenne de l'erreur absolue entre les valeurs prédites et les valeurs réelles. Elle donne une idée de la magnitude moyenne des erreurs sans tenir compte de leur signe.

Hypothèses :

- SVM peut présenter très rapidement un coût important en calcul (**Etienne**) Faux
- SVM utilise nativement un processus de CV pour s'entrainer sur les données (**Etienne**) Faux
- SVM n'est efficace que dans le cas des données linéairement séparable (**Adeline**) Faux
- SVM est la méthode de référence pour les données qui nécessite l'utilisation de l'hyperplan (**Briand**) Vrai
- SVM peut ne pas se révéler optimal dans le cas de cluster superposé (**Loïc**) Vrai

- SVM peut produire rapidement un résultat peu importe la quantité de données qui lui est fourni (**Axel**) Faux
- SVM ne s'utilise uniquement pour des Dataset très grands (**Loïc**) Faux
- SVM est sujet à l'overfitting (**Adrien**) Vrai
- SVR est mieux que SVM (**Tetyana**) Faux
- SVC est plus rapide que SVR (**Osman**) Faux
- Le Kernel est un dictionnaire de clés et de valeurs (**Jean Paul**) Faux
- Minimiser l'erreur revient à minimiser la marge maximale (**Aude**) Faux
- SVM est utilisable avec tous types de données (**Adeline**) Vrai
- Le SVM ajoute des variables à ressort (slack) qui autorise une erreur (**Seydou**) Vrai
- Le Kernel est un objet en Python (**Nicolas**) Faux

Plan d'action :

- Explorer les ressources
- Définir et comprendre les mots clefs
- Répondre à la problématique
- Vérifier les hypothèses
- Faire les Workshop
- RER

Les SVM résolvent les problèmes de classification binaire en les formulant comme des problèmes d'optimisation convexe. Le problème d'optimisation consiste à trouver la marge maximale séparant l'hyperplan, tout en classant correctement autant de points d'apprentissage que possible. Les SVM représentent cet hyperplan optimal par des vecteurs de support.