Apprentissage supervisé - Regression, DecisionTreeRegressor, RandomForestRegressor

Rappel sur les arbres de décisions

Les arbres de classification et de régression (parfois aussi appelés arbres de segmentation ou de décision) sont des méthodes qui permettent d'obtenir des modèles à la fois explicatifs et prédictifs. Parmi leurs avantages on notera d'une part leur simplicité du fait de la visualisation sous forme d'arbres, d'autre part la possibilité d'obtenir des règles en langage naturel.

On distingue notamment deux cas d'utilisation de ces modèles :

- on utilise les arbres de classification pour expliquer et/ou prédire l'appartenance d'objets (observations, individus) à une classe (ou modalité ou catégorie) d'une variable qualitative, sur la base de variables explicatives quantitatives et/ou qualitatives.
- on utilise les arbres de régression pour expliquer et/ou prédire les valeurs prise par une variable dépendante quantitative, en fonction de variables explicatives quantitatives et/ou qualitatives.

RandomForestRegressor

Le random forest regressor est ici utilisé pour palier a une problématique du decision Tree.

Les arbres de décisions ont une architecture qui leur permet de coller parfaitement aux données d'entrainement. C'est ce que l'on demande à un algorithme d'apprentissage supervisé. Néanmoins, à partir d'un certain point, le modèle devient beaucoup trop influencé par les données d'entraînements, ce qui peut engendrer des biais.
L'utilisation de plusieurs arbres offre une meilleure flexibilité et permet de réduire ce problème.

L'idée du random forest est de créer un grand nombre d'arbres de décisions de façon aléatoires, à partir de différents sous-ensembles de données de l'ensemble de données initial.Le fait de considérer différents sous-ensembles est important : cela réduit les risques d'erreur, puisque nos arbres seront peu corrélés.
On évite alors également le problème du surapprentissage, qui intervient lorsque l'arbre construit s'est trop adapté à l'échantillon considéré.

Il aura considéré tous les tests possibles, chaque feuille ne représentant qu'un unique candidat. Cet arbre sera donc fiable à 100% pour l'échantillon ayant permis sa construction (puisqu'il prend en compte tous les tests possibles), mais ne sera pas généralisable à d'autres échantillons.

Le candidat est alors testé sur chacun de ces arbres (qui peuvent être plus d'une centaine). L'intérêt de la forêt est de procéder par vote majoritaire quant aux résultats obtenus. On réduit ainsi la marge d'erreur que peut avoir un arbre seul. Plus l'on dispose d'arbres, plus la forêt sera fiable.