Algorithmes avancés

Les *k* plus proches voisins

I L’apprentissage supervisé

L’algorithme des *k* plus proches voisins dit aussi K*NN* pour *k* Nearest Neighborgs en anglais, fait partie des algorithmes d’apprentissage supervisé.

L’apprentissage supervisé ou *supervised learning* en anglais ce fait en deux phases.

Dans la première phase, on fournit à la machine un jeu de données avec sa solution au problème. Par exemple, s'il s'agit de reconnaître des fleurs, on fournit à la machine des photos avec la bonne réponse. On dit que les données de départ sont étiquetées ou labelisées.

L'algorithme apprend de ces données, on dit alors qu'il est entrainé. Il doit maintenant pouvoir prédire le résultat avec une donnée non étiquetée.

Dans la deuxième phase, on donne un jeu de données non étiquetés et l'on compare les prédictions obtenues avec les résultats attendus. Si le taux de prédiction est bon, on peut utiliser l'algorithme, sinon on reprend la première phase avec d'autres paramètres.

Par exemple, un filtre anti-spam peut être entrainé avec de nombreux emails étiqués comme étant du spam ou ne l'étant pas. Après entrainement, il classer avec une certaine efficacité les nouveaux emails.

L’algorithme des *k* plus proches voisins fait partie des algorithmes assez simple d’apprentissage supervisé.

II L’algorithme *k*nn

"Dis-moi qui sont tes plus proches amis et je te dirai qui tu es".

Le principe de l'algorithme des k plus proches voisins est très simple : à partir d'un ensemble de données de départ déjà labélisées, la prédiction pour une nouvelle donnée est déterminée grâce aux k données labélisées qui lui sont les plus proches.

Méthode

- **Déterminer la distance :** Calculer la distance entre la nouvelle donnée et chaque donnée étiquettée.
- **Classement croissant :** Classer ces distances par ordre croissant.
- **Prédiction :** Prédire l'étiquette de la nouvelle donnée en sélectionnant la majorité des k plus proches voisins.

Remarques:

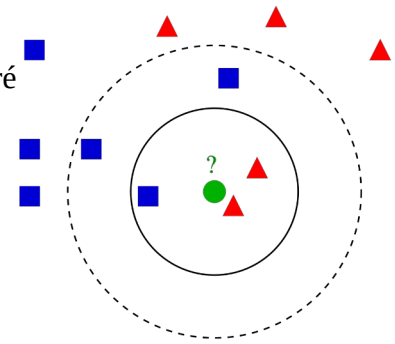
- **k impair :** Choisir un nombre impair pour k permet d'éviter les exquos lorsqu'il n'y a que deux choix et favorise l'obtention d'une majorité. En cas d'exquos, l'un des majoritaires est choisi.
- **Choix d'hyperparamètres :** Le choix de k, le choix de la distance, le choix des critères peut changer le résultat et donc influe sur la qualité des prédictions.
- **Calcul répétitif :** Calculer toutes les distances chaque fois qu'une nouvelle donnée est prévue peut être coûteux en termes de temps si le jeu d'apprentissage est grand.

Illustration

Comment classer le point vert ? Si $k = 3$ (cercle en ligne pleine) il est classé comme un triangle car il y a deux triangles et seulement un carré dans ce cercle. Si $k = 5$ (cercle en ligne pointillée) il est classé comme un carré car il y a trois carrés et uniquement deux triangles dans ce cercle.

https://nsi1.frama.io/algo/algo2/algo_1_sur_2.gif

https://nsi1.frama.io/algo/algo2/algo_2_sur_2.gif



Source: Wikpédia

Bilan

Avantages:

- L'algorithme est simple et facile à mettre en œuvre.
- Aucune hypothèse sur les données

Inconvénients:

- Le choix du nombre de voisins *k* ainsi que de la distance peut ne pas être évident
- L'étape de prédiction peut-être lente. La complexité est proportionnelle à la taille de l'échantillon. Pour chaque nouvelle prédiction, il faut recalculer la distance entre l'élément a étiqueter et les éléments du jeu de données, trier les données et en extraire l'élément majoritaire, ce qui peut être long.
-

Résumé:

L'algorithme kN-voisin (KNN) est un algorithme d'apprentissage automatique supervisé simple qui peut être utilisé pour résoudre des problèmes de classification et de régression. Il est facile à mettre en œuvre et à comprendre, mais présente l'inconvénient majeur de ralentir considérablement à mesure que la taille des données utilisées augmente.