NSI · PREMIÈRE / COURS

Les k plus proches voisins

Astrolab Cours · astrolabcours.fr

01 Le problème

On connaît la taille et le poids de 200 chiens, avec leur race. Un nouveau chien arrive : 45 cm, 14 kg. De quelle race est-il probablement ? On n'a pas de règle écrite, seulement des exemples.

01 L'idée : dis-moi qui sont tes voisins

Pour deviner le quartier préféré d'un nouvel élève, regarde où vont ses trois amis les plus proches : il ira probablement au même endroit. L'algorithme des **k plus proches voisins** classe une donnée nouvelle d'après les données connues qui lui ressemblent le plus.

01 L'algorithme

Pour classer une nouvelle donnée :

1. calculer sa **distance** à chacune des données connues ;
2. garder les **k** plus proches ;
3. lui attribuer la **classe majoritaire** parmi ces k voisins.

01 La distance

Avec deux caractéristiques (x_1, y_1) et (x_2, y_2), on prend en général la distance euclidienne :

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

En Python

```
from math import sqrt

def distance(a, b):
    return sqrt((a[0] - b[0]) ** 2 + (a[1] - b[1]) ** 2)

def knn(donnees, nouveau, k):
    """donnees : liste de (point, classe)."""
    tri = sorted(donnees, key=lambda d: distance(d[0], nouveau))
    votes = {}
    for point, classe in tri[:k]:
```

© Astrolab Cours · astrolabcours.fr/nsi/premiere/k-plus-proches-voisins/ · Tous droits réservés : usage personnel uniquement, republication et usage commercial interdits

```
        votes[classe] = votes.get(classe, 0) + 1
    return max(votes, key=votes.get)
```

Le choix de k

- $k = 1$: très sensible à une donnée isolée ou à une erreur de mesure.
- k trop grand : on finit par donner la classe la plus fréquente, quelle que soit la donnée.
- On prend souvent un k **impair** (3, 5, 7) pour éviter les égalités entre deux classes.

Cet algorithme ne construit aucune règle : il compare simplement aux exemples. C'est une méthode d'**apprentissage automatique** : plus les exemples sont nombreux et fiables, meilleures sont les prédictions.

Les pièges

- Caractéristiques d'échelles très différentes (centimètres et kilogrammes) : la plus grande écrase l'autre. On les ramène à des échelles comparables.
- Le résultat est une prédiction probable, pas une certitude.
- Calculer toutes les distances coûte cher quand il y a des millions d'exemples.

© Astrolab Cours · astrolabcours.fr/nsi/premiere/k-plus-proches-voisins/ · Tous droits réservés : usage personnel uniquement, republication et usage commercial interdits