MI201 - Apprentissage Automatique

Résumé Théorique - kNN

12 septembre 2024

Guilherme Nunes Trofino
2022-2024

Table des matières

1 Introduction **2**

2 Apprentissage **3**
2.1 Algorithme 3
2.1.1 Avantages 4
2.1.2 Incovenients 4
2.1.3 Applications 4

1. Introduction

Repository Hello ! My name is Guilherme Nunes Trofino and this is my LaTeX notebook of MI201 - Apprentissage Automatique that can be found in my GitHub repository : https://github.com/tr0fin0/classes_ensta.

Disclaimer This notebook is made so it may help others in this subject and is not intend to be used to cheat on tests so use it by your on risk.

Suggestions If you may find something on this document that does not seam correct please reach me by e-mail : guitrofino@gmail.com.

2. Apprentissage

2.1. Algorithme

On considère que l'algorithme peut être définie par la définition suivante :

Définition 2.1. Prédire la classe d'une nouvelle donnée en annotant les k plus proches voisins du donnée à classifier.

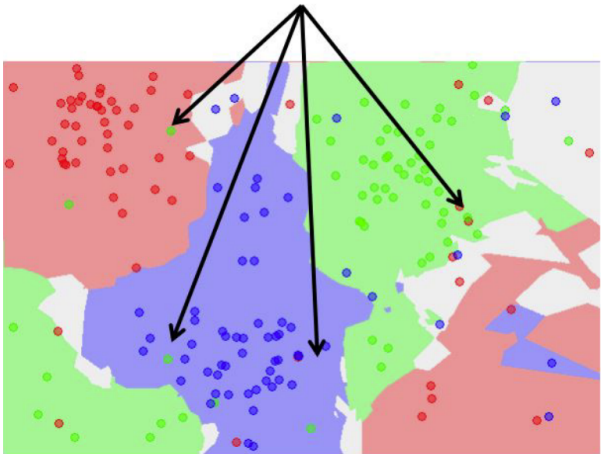

FIGURE 2.1 : Representation kNN

La décision se fait pour le vote majoritaire, s'il n'y a pas de majorité la région reste blanche.

Remarque. On nomme kNN aussi comme kPPV, k Plus Proche Voisin en français.

Remarque. kNN est très sensible aux données utilises, surtout aux bruits présentes. De cette façon-là il sera sensible à des pratiques comme la Validation Croisée.

L'algorithme va définir une région pour chaque différent caractéristique de la base de données. En plus, chaque frontière sera une polynôme de plusieurs faces droites.

La choix de k déterminera la qualité de la prédiction. Si on augmente k la qualité de la prediction améliore et le sur-apprentissage diminue. Généralement on aura le comportement suivant :

k ↑	biais ↑	variance ↓
k ↓	biais ↓	variance ↑

TABLE 2.1 : Comportement kNN

Remarque. Quand le k devient trop grand, proche de la quantité de données, les prédictions de cet algorithme échouent car, ou lieu d'analyser les données, il va juste montre la caractéristique plus présent dans la base de données.

On note qu'une variation très commun de cet algorithme est le 1NN :

Définition 2.2. Prédire la classe d'une nouvelle donnée en annotant le plus proche voisin du donnée à classifier.

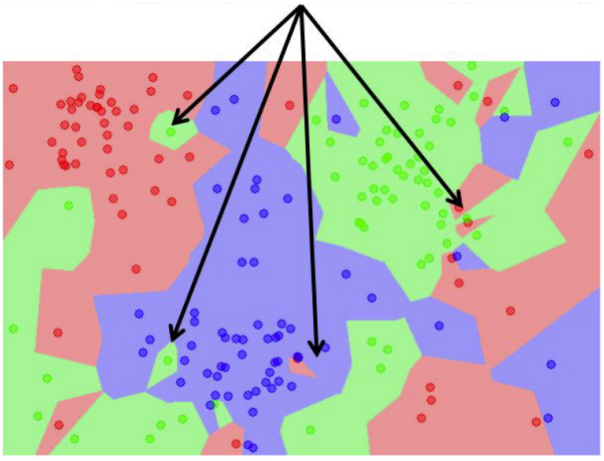

FIGURE 2.2 : Representation 1NN

Phrase. Tous les sensibilités de kNN seront encore plus grand dans 1NN.

2.1.1. Avantages

On peut citer :

1. très simple à implementer ;
2. facile à comprendre ;

2.1.2. Incovenients

On peut citer :

1. très sensible aux données ;
2. coût de prédiction croît en $O(nd)$;

2.1.3. Applications

Cet algorithme est souvent utilisé pour l'interpolation des données. Il faut remarquer que dans ce cas il n'y a pas une phase d'apprentissage, l'algorithme fait directement la prediction.