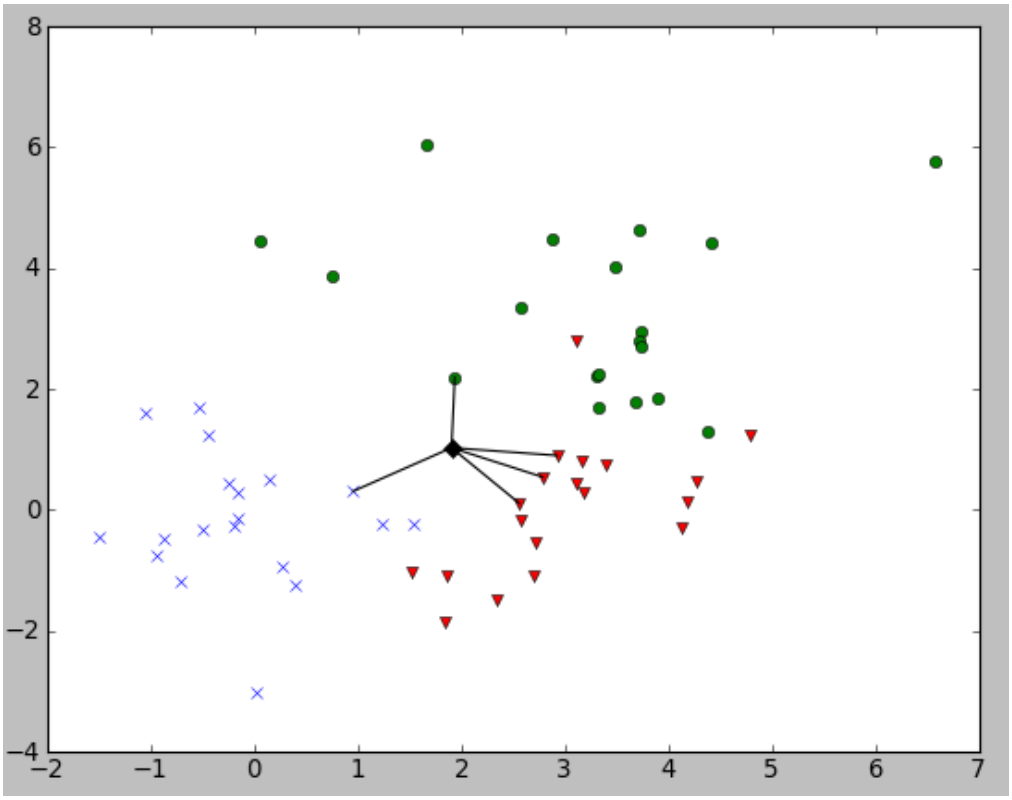
TP4 M1 IAA

Classification par K Plus Proches Voisins

On désire écrire en Python (avec Pylab) un programme qui réalise la classification par K-plus proches voisins sur les données qui se trouvent dans l'embryon de programme kppv.py. Il faut implémenter la partie qui prend une décision. Vous disposez de 20 échantillons représentatifs (en dimension 2) pour chaque classe et d'un échantillon de chaque classe pour tester.

La méthode des K-plus proches voisins consiste à affecter, à un point à classer, la classe majoritaire de ses K plus proches voisins.



(la classe triangle rouge est affectée au losange noir)

Algorithme des K-PPV

Soit X un point à classer parmi des classes définies. Chaque classe est représentée par une liste de taille fixe de points représentatifs.

Soit K l'entier représentant le nombre de voisins à considérer.

(1) Calculer la distance euclidienne entre X et chacun des points représentatifs de chaque classe. Stocker cette distance, ainsi que la classe d'appartenance, dans un tableau.

(2) Trier ce tableau distance/classe d'appartenance par ordre croissant suivant les distances.

(3) Extraire le nombre d'occurences des classes associées aux K plus petites distances.

(4) Affecter à X la classe ayant obtenu le plus grand nombre d'occurences (vote majoritaire).

1. Implémentation

Écrivez le programme qui réalise la décision par K-plus proches voisins. Ne prenez pas de décision s'il y a égalité pour chaque classe du nombre de points les plus proches.

2. Traitement des cas d'égalité

Implémentez des heuristique pour gérer les cas d'égalité sur le nombre de voisins. Il existe de nombreuses solutions, essayer d'en comparer quelques unes (avantages/inconvénients).

3. Affichage

Réalisez un affichage des données (vous pourrez représenter les K plus proches voisins en affichant les segments entre les K points les plus proches et le point à classer).

4. Choix des paramètres

Faut-il prendre un K élevé pour prendre une bonne décision ?

Faut-il prendre un nombre de points représentatifs élevés comme ensemble ■ d'apprentissage ■ ?

D'après-vous, qu'est ce qui va vous guider pour arriver à obtenir un choix idéal sur ces deux paramètres ?

5. Frontières de décision

Représentez les frontières de décision graphiquement : parcourez tous les points qui composent votre affichage et représentez tous les points qui présentent des indécisions.

Quelles forme prennent ces frontières quand K varie ?

6. Généralisation

Vérifiez que votre programme puisse réaliser la classification d'un nombre quelconque de classes et également que les observations puissent être de dimension quelconque.