

Objectif : comprendre comment l'algorithme k-NN (k nearest neighbors ou k plus proches voisins) classe un nouveau point en utilisant la proximité avec les points d'apprentissage.

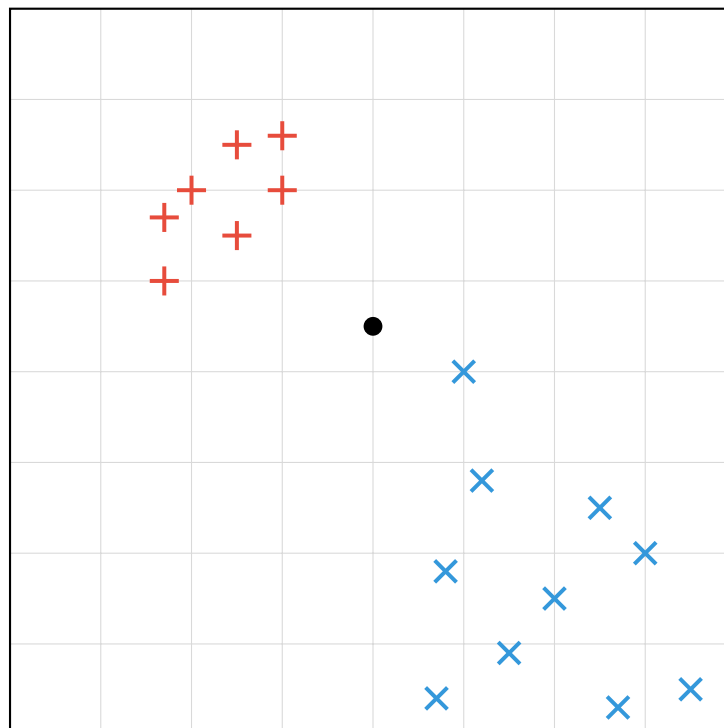
Principe

- On dispose d'un ensemble de points déjà classifiés (points d'apprentissage).
- Pour classier un nouveau point, on cherche ses k plus proches voisins.
- La classe majoritaire parmi ces k voisins devient la classe du nouveau point.

1 Avec deux classes.

Exercice 1 : classification avec $k = 3$ (3 voisins).

Voici un ensemble de points sur un plan. Les points rouges (+) sont de classe A, les points bleus (x) sont de classe B. Le point noir (●) est à classier : est-il de classe A ou de classe B ?



+ Classe A x Classe B ● À classier

1. Identifier les 3 points les plus proches du point noir ●. Mesurer les distances sur le graphique ci-dessus.

Distances : _____

2. Parmi ces 3 points, combien sont de classe A ? _____ De classe B ? _____

3. À quelle classe appartient le point noir ● avec $k = 3$? _____

 Exercice 2 : influence du paramètre k.

1 Reprendre le graphique de l'exercice précédent.

1. Avec $k = 1$ (1 seul voisin).

- Quel est le point le plus proche du point noir ● ? _____
- Classification du point noir : _____

2. Avec $k = 4$ (4 voisins).

- Entourer les 4 points les plus proches du point noir.
- Nombre de points de classe A : _____ Nombre de points de classe B : _____
- Classification du point noir : _____

3. Avec $k = 5$ (5 voisins).

- Entourer les 5 points les plus proches du point noir.
- Nombre de points de classe A : _____ Nombre de points de classe B : _____
- Classification du point noir : _____

4. Avec $k = 25$ (tous les voisins de chaque classe).

- Nombre de points de classe A : _____ Nombre de points de classe B : _____
- Classification du point noir : _____

Réflexions

1. La classification change-t-elle selon la valeur de k ? Pourquoi ?

2. Que se passe-t-il si $k = 1$?

3. Que se passe-t-il si k est supérieur au nombre total de points ?

 Exercice 3 : un cas concret.

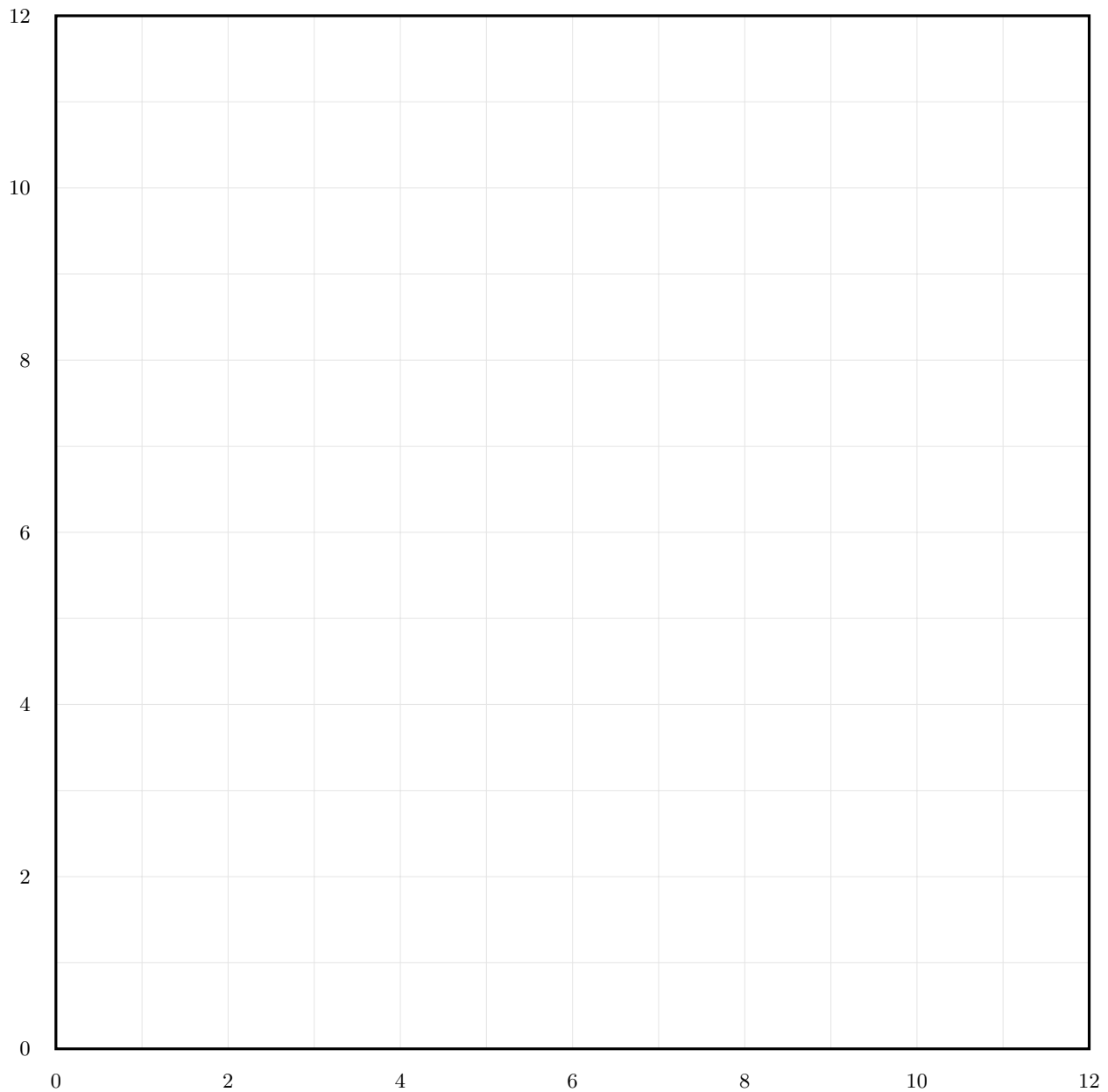
On souhaite essayer de prédire la réussite des élèves à un examen selon leur temps de révision en heure, et leur temps d'écran (scrolling sur smartphone) durant les jours précédant l'épreuve. On dispose des résultats de 12 élèves :

- Un élève a révisé pendant 1 heure et a passé 10 heures sur son smartphone, il a raté l'examen.
- Un élève a révisé pendant 2 heures et a passé 2 heures sur son smartphone, il a raté l'examen.
- Un élève a révisé pendant 3 heures et a passé 7 heures sur son smartphone, il a raté l'examen.
- Un élève a révisé pendant 4 heures et a passé 11 heures sur son smartphone, il a raté l'examen.
- Un élève a révisé pendant 6 heures et a passé 5 heures sur son smartphone, il a raté l'examen.
- Un élève a révisé pendant 10.5 heures et a passé 10 heures sur son smartphone, il a raté l'examen.
- Un élève a révisé pendant 4 heures et a passé 4 heures sur son smartphone, il a réussi l'examen.
- Un élève a révisé pendant 5 heures et a passé 0.5 heure sur son smartphone, il a réussi l'examen.
- Un élève a révisé pendant 8 heures et a passé 6 heures sur son smartphone, il a réussi l'examen.
- Un élève a révisé pendant 9 heures et a passé 3 heures sur son smartphone, il a réussi l'examen.
- Un élève a révisé pendant 10 heures et a passé 1 heures sur son smartphone, il a réussi l'examen.
- Un élève a révisé pendant 11 heures et a passé 4 heures sur son smartphone, il a réussi l'examen.

Utiliser le repère fourni ci-après et l'algorithme des k-NN pour prédire la réussite ou l'échec à l'examen de deux élèves.

1. Un élève a révisé pendant 6 heures et a passé 2 heures sur son smartphone.

2. Un élève a révisé pendant 6 heures et a passé 8 heures sur son smartphone.



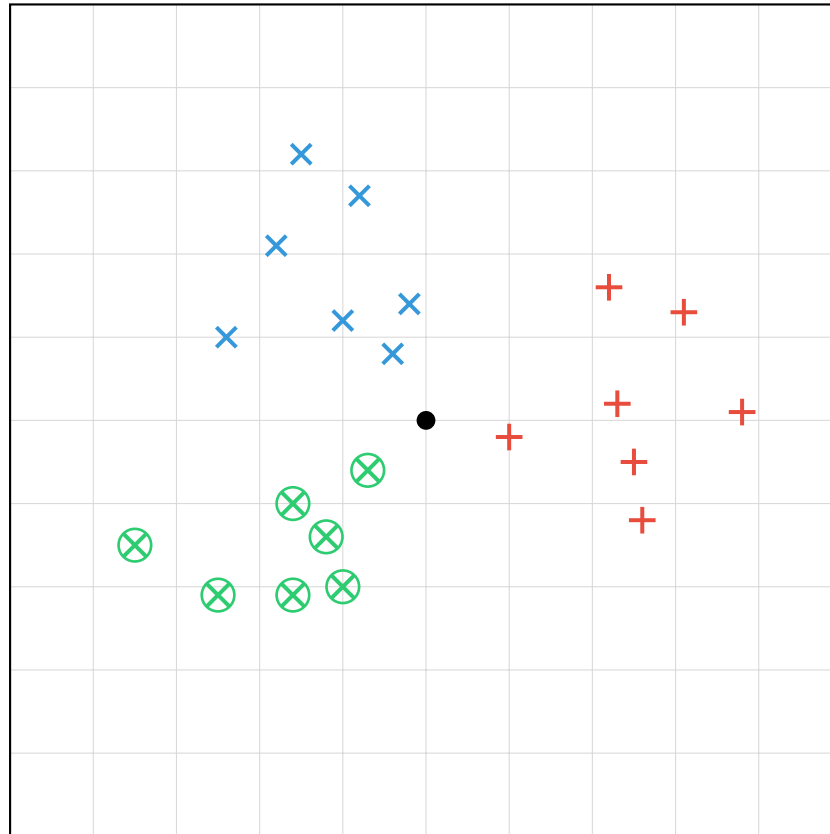
Exercice 4 : illustration.

Aller sur http://http://ia.dellasantina.corsica/knn_corse.html et déterminer si un point de la carte de la Corse est en Haute-Corse ou en Corse-du-Sud d'après l'algorithme k-NN. Ne pas hésiter à zoomer pour bien voir la frontière réelle si nécessaire.

2 Avec trois classes.

§ Exercice 5 : classification avec $k = 3$ (3 voisins).

Voici un ensemble de points sur un plan. Les points rouges (+) sont de classe A, les points bleus (x) sont de classe B et les points verts (x) sont de classe C. Le point noir (●) est à classer : est-il de classe A, de classe B ou de classe C ?



+ Classe A x Classe B x Classe C ● À classer

1. Identifier et entourer les 3 points les plus proches du point noir ●.
2. À quelle classe appartient le point noir ● avec $k = 3$? _____

§ Exercice 6 : influence du paramètre k.

Reprendre le graphique de l'exercice précédent.

1. Avec $k = 5$ (5 voisins).
 - Entourer les 5 points les plus proches du point noir.
 - Nombre de points de classe A : _____ Nombre de points de classe B : _____
Nombre de points de classe C : _____
 - Classification du point noir : _____

2. Avec $k = 7$ (7 voisins).

- Entourer les 7 points les plus proches du point noir.
- Nombre de points de classe A : _____ Nombre de points de classe B : _____
Nombre de points de classe C : _____
- Classification du point noir : _____

3. Avec $k = 9$ (9 voisins).

- Entourer les 9 points les plus proches du point noir.
- Nombre de points de classe A : _____ Nombre de points de classe B : _____
Nombre de points de classe C : _____
- Classification du point noir : _____

Réflexions

1. Qu'est-ce qui change par rapport à 2 classes seulement ?

2. Peut-il y avoir des cas d'égalité ? Donner un exemple.

3. Comment résoudre un cas d'égalité ?

Exercice 7 : illustration.

Aller sur http://http://ia.dellasantina.corsica/knn_microregions.html et déterminer dans quelle micro-région se situe un point de la carte de la Corse d'après l'algorithme k-NN.

3 Applications réelles de k-NN.

- Reconnaissance d'écriture manuscrite.
- Reconnaissance de chiffres.
- Systèmes de recommandation (films, musique, produits).
- Diagnostic médical (classification de maladies).
- Détection de spam dans les emails.
- Reconnaissance vocale.
- Reconnaissance faciale.
- Prévisions météorologiques.
- Zonage géographique (micro-régions, quartiers, zones climatiques).
- Classification de fleurs (différentes espèces).

4 Défi.

Défi

Peut-on trouver un ensemble de données telles que l'algorithme des k-NN ne fonctionne jamais quelle que soit la valeur de k ?
