#### Thème 5 : Algorithmique

15

# Algorihtme des k-plus proches voisins

Abrégé *kppv* en français. En anglais, on dit *k nearest neighbors* souvent abrégé *knn*.

L'algorithme des k plus proches voisins appartient à la famille des algorithmes d'apprentissage automatique (machine learning) qui constituent le poumon de l'intelligence artificielle actuellement.

Pour simplifier, l'apprentissage automatique part souvent de données (data) et essaye de dire quelque chose des données qui n'ont pas encore été vues : il s'agit de *généraliser*, de *prédire*.

On va utiliser l'algorithme des k plus proches voisins pour résoudre un *problème de classification* : prédire la classe d'une donnée *inconnue* à partir de la classe des données connues.

Avant de décrire cet algorithme, introduisons la situation et le problème.

## Présentation d'un problème de classification

## 1. Le jeu de données sur les Pokémons



On dispose de données sur 34 Pokémons : leur type (Psy ou Eau), leur points de vie et la valeur de leurs attaques.

Ces données sont stockées dans le fichier CSV pokemons.csv que l'on s'empresse d'importer dans un tableau de dictionnaires appelé pokemons :

```
In [1]: import csv
fichier = open('pokemons.csv', 'r', encoding = 'UTF-8')
t = csv.DictReader(fichier, delimiter=';')
pokemons = [dict(ligne) for ligne in t] # création et construction du
tableau par compréhension
fichier.close()

pokemons # pour voir les données importées
```

Out[1]: [{'Nom': 'Aligatueur', 'Type': 'Eau', 'Points de vie': '85', 'Attaque': '105'}, {'Nom': 'Bargantua', 'Type': 'Eau', 'Points de vie': '70', 'Attaque': '92'}, {'N om': 'Carabaffe', 'Type': 'Eau', 'Points de vie': '59', 'Attaque': '63'}, {'No m': 'Clamiral', 'Type': 'Eau', 'Points de vie': '95', 'Attaque': '100'}, {'Nom': 'Crefadet', 'Type': 'Psy', 'Points de vie': '75', 'Attaque': '125'}, {'Nom': 'Cr ocrodil', 'Type': 'Eau', 'Points de vie': '65', 'Attaque': '80'}, {'Nom': 'Deoxy s', 'Type': 'Psy', 'Points de vie': '50', 'Attaque': '70'}, {'Nom': 'Deoxys', 'T ype': 'Psy', 'Points de vie': '50', 'Attaque': '95'}, {'Nom': 'Deoxys', 'Type': 'Psy', 'Points de vie': '50', 'Attaque': '150'}, {'Nom': 'Deoxys', 'Type': 'Ps y', 'Points de vie': '50', 'Attaque': '180'}, {'Nom': 'Ecayon', 'Type': 'Eau', 'Points de vie': '49', 'Attaque': '49'}, {'Nom': 'Eoko', 'Type': 'Psy', 'Points de vie': '75', 'Attaque': '50'}, {'Nom': 'Gamblast', 'Type': 'Eau', 'Points de v ie': '71', 'Attaque': '73'}, {'Nom': 'Gobou', 'Type': 'Eau', 'Points de vie': '5 0', 'Attaque': '70'}, {'Nom': 'Groret', 'Type': 'Psy', 'Points de vie': '80', 'A ttaque': '45'}, {'Nom': 'Mateloutre', 'Type': 'Eau', 'Points de vie': '75', 'Att aque': '75'}, {'Nom': 'Mesmerella', 'Type': 'Psy', 'Points de vie': '60', 'Attaq ue': '45'}, {'Nom': 'Mew', 'Type': 'Psy', 'Points de vie': '100', 'Attaque': '10 0'}, {'Nom': 'Mewtwo', 'Type': 'Psy', 'Points de vie': '106', 'Attaque': '110'}, {'Nom': 'Mewtwo', 'Type': 'Psy', 'Points de vie': '106', 'Attaque': '150'}, {'No m': 'Mewtwo', 'Type': 'Psy', 'Points de vie': '106', 'Attaque': '190'}, {'Nom': 'Munna', 'Type': 'Psy', 'Points de vie': '76', 'Attaque': '25'}, {'Nom': 'Nucleo s', 'Type': 'Psy', 'Points de vie': '45', 'Attaque': '30'}, {'Nom': 'Octillery', 'Type': 'Eau', 'Points de vie': '75', 'Attaque': '105'}, {'Nom': 'Okeoke', 'Typ

```
e': 'Psy', 'Points de vie': '95', 'Attaque': '23'}, {'Nom': 'Phione', 'Type': 'E au', 'Points de vie': '80', 'Attaque': '80'}, {'Nom': 'Poissoroy', 'Type': 'Ea u', 'Points de vie': '80', 'Attaque': '92'}, {'Nom': 'Prinplouf', 'Type': 'Eau', 'Points de vie': '64', 'Attaque': '66'}, {'Nom': 'Rosabyss', 'Type': 'Eau', 'Points de vie': '55', 'Attaque': '84'}, {'Nom': 'Siderella', 'Type': 'Psy', 'Points de vie': '70', 'Attaque': '55'}, {'Nom': 'Spoink', 'Type': 'Psy', 'Points de vie': '60', 'Attaque': '25'}, {'Nom': 'Symbios', 'Type': 'Psy', 'Points de vie': '110', 'Attaque': '65'}, {'Nom': 'Tarpaud', 'Type': 'Eau', 'Points de vie': '90', 'Attaque': '75'}, {'Nom': 'Tiplouf', 'Type': 'Eau', 'Points de vie': '53', 'Attaque': '51'}]
```

On peut représenter ces données graphiquement dans un repère orthonormé, avec les valeurs d'attaque en abscisses et les points de vie en ordonnées. Les Pokémons de type "Eau" sont représentés pas les triangles bleus, et ceux de type "Psy" par les points rouges.



C'est ce que vous allez faire par la suite!

## Problème : peut-on prédire le type d'un nouveau Pokémon inconnu ?

On considère maintenant que l'on possède des données sur un nouveau Pokémon inconnu et on aimerait prédire s'il s'agit d'un Pokémon de type "Eau" ou de type "Psy".



Ce problème, qui demande à prédire à quelle catégorie, ou *classe*, appartient ce nouvel élément donné, est appelé *problème de classification*. L'algorithme des k plus proches voisins permet de trouver les k voisins les plus proches de ce nouvel élément dans le but de lui associer une *classe* plausible (Psy ou Eau, dans cet exemple). Par exemple, si k = 5 va chercher les 5 voisins les plus proches.

## Algorithme des kppv

## 1. Énoncé et spécification

À partir d'un jeu de données données (par exemple, les données sur nos 34 Pokémons) et d'une donnée cible (le nouveau Pokemon à classifier), l'algorithme de \$k\$ plus proches voisins

doit déterminer les \$k\$ données les plus proches de la cible.

Plus précisément, la spécification de l'algorithme des kppv est la suivante :

#### Entrées :

- une table donnees de taille n contenant les données et leurs classes
- une donnée cible : cible
- un nombre k
- une règle permettant de calculer la distance entre deux données
- **Sortie** : une table k\_plus\_proches\_voisins
- Rôle : trouver les k données les plus proches de cible parmi celles de la table données ("plus proches" au sens de la distance définie au départ)
- **Précondition** : \$n \geq 1\$ et \$k <= n\$.
- **Postcondition** : k\_plus\_proches\_voisins contient les k plus proches voisins de cible parmi les données de la table données

## 2. Algorithme

L'algorithme naif des kppv s'exprime de manière simple :

- 1. Créer une table distances\_voisins contenant les éléments de la table données et leurs distances avec la donnée cible.
- 2. Trier les données de la table distances\_voisins selon la distance croissante avec la donnée cible
- 3. Renvoyer les k premiers éléments de cette table triée (k\_plus\_proches\_voisins)

### Implémentation de l'algorithme

L'algorithme des plus proches voisins repose presque entièrement sur la *distance* entre deux données. Il faut donc commencer par définir une distance entre deux données.

## Étape 0 : Choix et implémentation de la distance

Dans la suite, on va choisir la distance "naturelle", c'est-à-dire celle "à vol d'oiseau". On parle de distance euclidienne.

Dans un repère orthonormé, si A et B ont pour coordonnées respectives  $(x_A, y_A)$  et  $(x_B, y_B)$  alors la distance (euclidienne) entre ces deux points est donnée par la formule :

```
\frac{(x_B-x_A)^2 + (y_B-y_A)^2}{x_B-y_A}
```

**Important**: Sachez cependant qu'il existe d'autres distances possibles et que le choix de la distance n'est pas anodin car ce choix peut aboutir à trouver des voisins les plus proches différents et donc conduire à des prédictions différentes (voir le résumé de cours !).

Dans notre cas, chaque Pokémon possède une abscisse égale à sa valeur d'attaque et une ordonnée égale à ses points de vie. Ainsi, la formule de la distance entre deux pokémons p1 et p2 se traduit ainsi :

 $\frac{p1}{\text{valeur d'attaque de p2}-\text{valeur d'attaque de p2}-\text{valeur d'attaque de p2}-\text{valeur d'attaque de p1}}^2 + (\frac{p1}{\text{valeur d'attaque de p1}}^2).$$ 

**Question 1**: Calculez à la main la distance entre les deux Pokémons suivants.

```
p1 = {'Nom': 'Aligatueur', 'Type': 'Eau', 'Points de vie': '85', 'Attaque': '105'}
p2 = {'Nom': 'Bargantua', 'Type': 'Eau', 'Points de vie': '70', 'Attaque': '92'}
```

#### ■ Question 2 : Écrivez les instructions permettant d'accéder :

- à la valeur d'attaque de p1
- au points de vie de p2

Question 3: Complétez la fonction distance\_euclidienne(p1, p2) suivante qui renvoie la distance euclidienne entre deux Pokémons p1 et p2.

 $\triangle$  Il faut bien veiller à convertir les valeurs d'attaque et les points de vie en des nombres pour faire les calculs !\*

```
In [3]: # à vous de jouer !

from math import sqrt # pour utiliser la racine carrée

def distance_euclidienne(p1, p2):

x1 = ... # abscisse de la donnée 1

y1 = ... # ordonnée de la donnée 1

x2 = ... # abscisse de la donnée 2

y2 = ... # ordonnée de la donnée 2

return sqrt(...) # formule donnant la distance euclidienne
```

**Question 4** : Appelez cette fonction pour vérifier votre réponse à la question 1.

```
In []: # à vous de jouer!
```

## Étape 1 : Création de la table avec les distances

Maintenant que notre distance est définie, on peut passer à l'implémentation de l'algorithme. On va donc commencer par créer la table distances\_voisins contenant les distances entre toutes nos données et la donnée cible.

On choisit de reprendre les dictionnaires de la table pokemons en leur ajoutant une clé correspondant à la distance avec la cible. Par exemple, le premier élément de distances\_voisins sera un dictionnaire de la forme

```
{'Nom': 'Aligatueur', 'Type': 'Eau', 'Points de vie': '85', 'Attaque': '105',
'distance': dist}
```

dans lequel dist est la distance à calculer entre le Pokémon 'Aligatueur' et la cible.

Question 5 : Écrivez une fonction table\_avec\_distances(donnees, cible) qui renvoie la table distances\_voisins contenant les éléments de la table donnees auxquels on a ajouté la clé distance dont la valeur est leur distance avec la donnée cible.

```
Exemple: Si cible = {'Nom': 'inconnu', 'Type': 'inconnu', 'Points de vie': '92',

'Attaque': '127'} alors l'appel table_avec_distances(pokemons, cible) renvoie le tableau
dont le début est le suivant :
```

```
[{'Nom': 'Aligatueur', 'Type': 'Eau',
```

```
'Points de vie': '85',
  'Attaque': '105',
  'distance': 23.08679276123039},
 {'Nom': 'Bargantua',
  'Type': 'Eau',
  'Points de vie': '70',
  'Attaque': '92',
  'distance': 41.340053217188775},
def table_avec_distances(donnees, cible):
                                                                                       Q
              # à vous de jouer !
          # ESSAI
          cible = {'Nom': 'inconnu', 'Type': 'inconnu', 'Points de vie': '92',
          'Attaque': '127'}
          distances_voisins = table_avec_distances(pokemons, cible)
          distances_voisins
```

## 3. Étape 2 : Tri de la table selon la distance croissante

On va maintenant trier notre table de la plus petite à la plus grande distance avec notre cible, soit du plus proche au plus éloigné des voisins. Pour cela, on va utiliser la fonction sorted dont on peut afficher l'aide pour rappel :

```
In []: help(sorted)
```

Question 6 : Créer une fonction tri\_distance(d) qui servira de clé à notre tri.

```
In []: # à vous de jouer!
```

Question 7 : Utilisez maintenant la fonction sorted pour créer une *nouvelle table* distances\_voisins\_triee contenant les données de distances\_voisins triés par ordre croissant de distance.

```
In []: cible = {'Nom': 'inconnu', 'Type': 'inconnu', 'Points de vie': '92', 'Attaque': '127'}
```

```
distances_voisins = table_avec_distances(pokemons, cible)
# à vous de jouer !
```

## 4. Étape 3 : Détermination des k plus proches voisins

Le plus dur est fait, il ne reste plus qu'à construire la table k\_plus\_proches\_voisins contenant les k plus proches voisins. Comme la table distances\_voisins\_triee contient les voisins du plus proches au plus éloigné, il suffit de conserver les k premiers éléments de cette table !

Question 8 : Écrivez une fonction conserve\_k\_premiers(k, table) qui renvoie une nouvelle table contenant les k premiers éléments de la table table (où k est un entier positif inférieur ou égale à len(table)).

```
In []: # à vous de jouer!
```

Question 9 : Vérifiez qu'en appelant cette fonction avec k = 3 et table = distances\_voisins\_triee vous obtenez bien les 3 premiers éléments de distances\_voisins\_triee .

```
In []: # à vous de jouer!
```

## 5. Bilan

On peut maintenant regrouper tout ce qui vient d'être fait pour chaque étape pour écrire une fonction kppv(donnees, cible, k) qui renvoie les k plus proches voisins de cible dans donnees.

Question 10 : Complétez la fonction kppv(donnees, cible, k) suivante qui renvoie les k plus proches voisins de cible dans donnees.

```
In []: # RECOPIE DES FONCTIONS UTILES:

from math import sqrt

# Calcul de la distance

def distance_euclidienne(d1, d2):

# recopiez le code écrit à la question 3

pass
```

```
# Fonction qui calcule et ajoute la distance entre la cible et chacune des
données
def table_avec_distances(donnees, cible):
    # recopiez le code écrit à la question 5
# Clé du tri
def tri_distance(d):
    # recopiez le code écrit à la question 6
def conserve_k_premiers(k, table):
    # recopiez le code écrit à la question 8
    pass
# À COMPLÉTER :
def kppv(donnees, cible, k):
    # étape 1 : création de la table avec les distances
    distances_voisins = ...
    # étape 2 : tri par distance croissante
    distances_voisins_triee = ...
    # étape 3 : récupération de k plus proches voisins
    k_plus_proches_voisins = ...
    return ...
```

Question 11: Appelez la fonction la fonction kppv pour connaître les 3 plus proches voisins de notre cible {'Nom': 'inconnu', 'Type': 'inconnu', 'Points de vie': '92', 'Attaque': '127'}.

```
In []: cible = {'Nom': 'inconnu', 'Type': 'inconnu', 'Points de vie': '92', 'Attaque': '127'}
# à vous de jouer !
```

☑ Question 12 : Même question pour les 5 plus proches voisins puis pour les 9 plus proches voisins.

```
In []: # à vous de jouer!
```

## Et notre prédiction alors ?

image sorcière

L'algorithmes des kppv en lui-même n'apporte pas la réponse à notre problème de classification puisqu'il ne fournit que les k plus proches voisins (et leurs classes) de notre donnée cible. Il reste donc une dernière étape pour prédire la classe de notre nouvel élément : pour cela, on choisit la *classe majoritaire* (la plus présente) dans les k plus proches voisins.

On est contents si k est impair car il ne peut pas y avoir d'ex-aequo!

Question 13 : Donnez la classe majoritaire parmi les plus proches voisins lorsque k = 3, 5 puis 9. Pour chaque valeur de k, quelle serait la prédiction pour notre Pokémon cible inconnu ?

#### Réponse :

**Moralité** : La valeur du paramètre k est importante car elle a une influence forte sur la prédiction. En pratique, elle doit être judicieusement choisie : voir dans le cours comment trouver expérimentalement une bonne valeur de k.

## Visualisation graphique

On peut utiliser le module matplotlib pour représenter graphiquement nos Pokémons, et observer graphiquement les plus proches voisins selon la valeur de k choisies.

Pour créer une nuage de points, il suffit d'utiliser la fonction plot qui prend en paramètres deux tableaux contenant respectivement la liste des abscisses et la liste des ordonnées des points à construire, ainsi qu'un paramètre permettant de définir le type de points et leur couleur :

```
In []: import matplotlib.pyplot as plt

liste_abscisses = [0, 2, 5, 9, 7] # tableau avec les abscisses
liste_ordonnees = [5, 7, 12, 3, 8] # tableau avec les ordonnées

plt.plot(liste_abscisses, liste_ordonnees,'ro') # r pour red, o pour un cercle

plt.xlim(-2, 12) # pour définir la fenêtre en abscisse
plt.ylim(0, 14) # pour définir la fenêtre en ordonnée
plt.gca().set_aspect('equal', adjustable='box') # pour que le repère soit
orthonormé (même unité sur chaque axe)

plt.show() # affichage du graphique
```

Pour représenter nos Pokémons sous forme de nuages de points (valeur d'attaque en abscisse et points de vie en ordonnée), il suffit alors de créer un tableau contenant les abscisses (valeur

d'attaque) et un tableau contenant les ordonnées (points de vie) puis d'utiliser la fonction plot pour construire les points.

Pour bien visualiser les deux classes, on va créer un nuage de points pour chaque classe : les Pokémons de type "Eau" seront représentés par des points bleus et les Pokémons "Psy" par des points rouges. On crée aussi le point correspondant à notre Pokémon cible en cyan :

```
In [ ]:
         import matplotlib.pyplot as plt
                                                                                      Q
         # construction des Pokémons de type 'Eau'
         valeur_attaque_eau = [int(p['Attaque']) for p in pokemons if p['Type'] ==
         'Eau'] # tableau des abscisses
         points_de_vie_eau = [int(p['Points de vie']) for p in pokemons if p['Type']
         == 'Eau'] # tableau des orodonnées
         plt.plot(valeur_attaque_eau, points_de_vie_eau, 'bv') # construction du
         graphique avec des triangles bleus
         # construction des Pokémons de type 'Psy'
         valeur_attaque_psy = [int(p['Attaque']) for p in pokemons if p['Type'] ==
         points_de_vie_psy = [int(p['Points de vie']) for p in pokemons if p['Type']
         == 'Psy']
         plt.plot(valeur_attaque_psy, points_de_vie_psy, 'ro') # construction du
         graphique avec des cercles rouges
         # construction du Pokémon cible
         cible = {'Nom': 'inconnu', 'Type': 'inconnu', 'Points de vie': '92',
         'Attaque': '127'}
         valeur_attaque_cible = int(cible['Attaque'])
         points_de_vie_cible = int(cible['Points de vie'])
         plt.plot(valeur_attaque_cible, points_de_vie_cible, 'cD')
         # Pour définir le titre, noms de axes et la légende
         plt.title("Pokémons")
         plt.xlabel("Valeur d'attaque")
         plt.ylabel("Points de vie")
         plt.legend(["Type Eau", "Type Psy"], loc='upper center', bbox_to_anchor=(1.2,
         0.8))
         plt.xlim(0, 200)
         plt.ylim(40, 120)
         plt.gca().set_aspect('equal', adjustable='box')
         plt.show()
```

Question 14 : Complétez ci-dessous le programme précédent pour construire également le nuage de points correspondant aux k plus plus proches voisins déterminés par notre algorithme.

Vérifiez ensuite graphiquement pour différentes valeurs de k que les plus proches voisins sont les bons.

```
In [ ]:
         import matplotlib.pyplot as plt
                                                                                      Q
         # construction des Pokémons de type 'Eau'
         valeur_attaque_eau = [int(p['Attaque']) for p in pokemons if p['Type'] ==
         'Eau'] # tableau des abscisses
         points_de_vie_eau = [int(p['Points de vie']) for p in pokemons if p['Type']
         == 'Eau'] # tableau des orodonnées
         plt.plot(valeur_attaque_eau, points_de_vie_eau, 'bv') # construction du
         graphique avec des triangles bleus
         # construction des Pokémons de type 'Psy'
         valeur_attaque_psy = [int(p['Attaque']) for p in pokemons if p['Type'] ==
         points_de_vie_psy = [int(p['Points de vie']) for p in pokemons if p['Type']
         == 'Psy']
         plt.plot(valeur_attaque_psy, points_de_vie_psy, 'ro') # construction du
         graphique avec des cercles rouges
         # construction du Pokémon cible
         cible = {'Nom': 'inconnu', 'Type': 'inconnu', 'Points de vie': '92',
         'Attaque': '127'}
         valeur_attaque_cible = int(cible['Attaque'])
         points_de_vie_cible = int(cible['Points de vie'])
         plt.plot(valeur_attaque_cible, points_de_vie_cible, 'cD')
         # construction des KPPV
         # À COMPLÉTER :
         # Pour définir le titre, noms de axes et la légende
         plt.title("Pokémons")
         plt.xlabel("Valeur d'attaque")
         plt.ylabel("Points de vie")
         plt.legend(["Type Eau", "Type Psy"], loc='upper center', bbox_to_anchor=(1.2,
         0.8))
         plt.xlim(0, 200)
         plt.ylim(40, 120)
         plt.gca().set_aspect('equal', adjustable='box')
         plt.show()
```

## Et si on change de distance?

On a utilisé la distance euclidienne pour mesurer la distance entre nos données et la cible. C'est la distance "naturelle" mais il existe d'autres distances. Par exemple, la **distance de Manhattan** entre deux données  $d_1(x_1, y_1)$  et  $d_2(x_2, y_2)$  est définie ainsi  $\frac{d_1}{d_1} = |x_2-x_1| + |y_2-y_1|$ .

```
Article Wikipédia : Distance de Manhattan
```

On peut modifier notre programme simplement en définissant et en utilisant cette nouvelle distance .

```
In []: # définition de la distance utilisée
                                                                                      Q
         def distance_manhattan(d1, d2):
             x1 = float(d1['Attaque'])
             y1 = float(d1['Points de vie'])
             x2 = float(d2['Attaque'])
             y2 = float(d2['Points de vie'])
             return abs(x2-x1) + abs(y2-y1) # abs est la fonction valeur absolue
         def table_avec_distances(donnees, cible):
             distances_voisins = [d for d in donnees]
             for d in distances_voisins:
                 distance = distance_manhattan(d, cible) # CHANGEMENT DE DISTANCE
                 d['distance'] = distance
             return distances_voisins
         def tri_distance(d):
             return d['distance']
         def kppv(donnees, cible, k):
             # étape 1 : création de la table avec les distances
             distances_voisins = table_avec_distances(donnees, cible)
             # étape 2 : tri par distance croissante
             distances_voisins_triee = sorted(distances_voisins, key=tri_distance)
             # étape 3 : récupération de k plus proches voisins
             k_plus_proches_voisins = conserve_k_premiers(k, distances_voisins_triee)
             return k_plus_proches_voisins
```

En déterminant les 3 plus proches voisins on obtient :

Dans ce cas, pour k = 3, on prédirait que notre Pokémon inconnu est de type "Eau" (2 "Eau" contre 1 "Psy") et la prédiction serait l'inverse de celle utilisant la distance euclidienne !

**Moralité**: La distance utilisée a également toute son importance puisque son choix peut aboutir à des prédictions différentes. De plus, nous n'avons parlé ici que de distances *géométriques* s'appliquant à des données chiffrées. Toutes les données ne sont pas adaptées à ce type de distance : si on veut comparer la distance entre deux chaînes de caractères (dans le but de prédire la langue d'origine de certains mots par exemple) d'autres types de distances sont à considérer : *la distance de Hamming* ou la *distance d'édition* qui seront abordées en Terminale.

#### SUIVANT : TP iris

In []: