Machine Learning

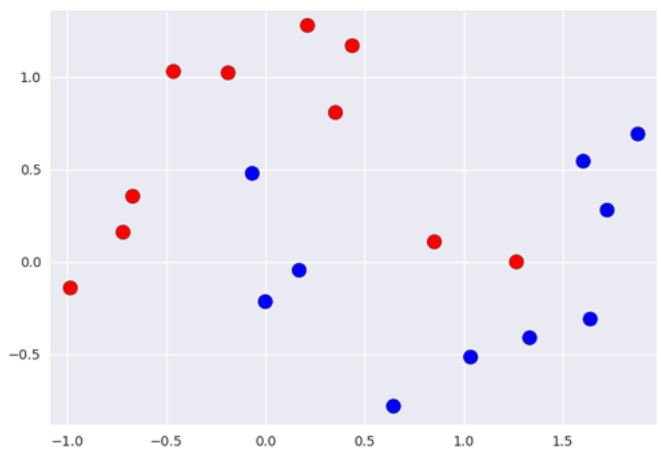
KNN: K Nearest Neighbors
(K voisins les plus proches)

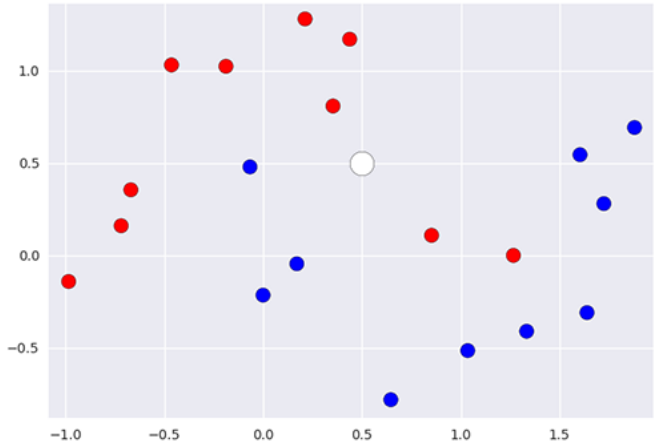
madani.a@ucd.ac.ma

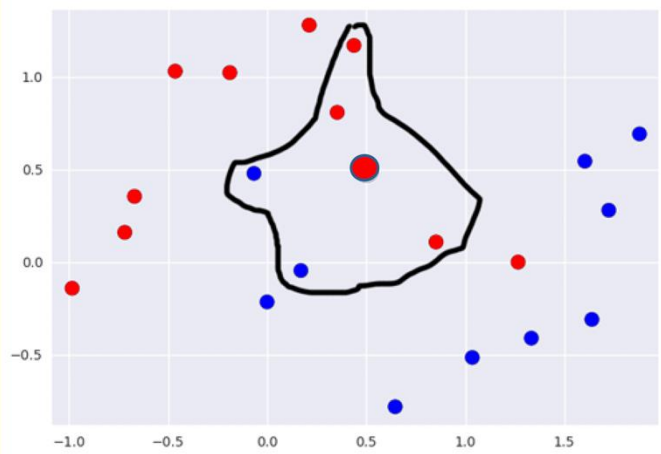
KNN c'est quoi ?

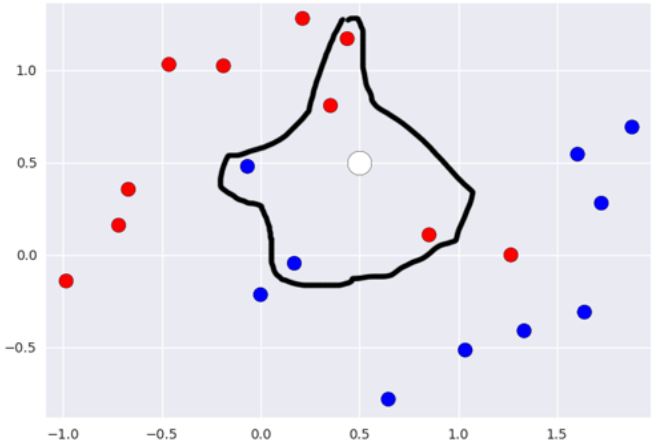
- Est un algorithme de Machine Learning supervisé, qui fonctionne sous le principe de : "Dis moi qui sont tes voisins, je te dirais qui tu es…".
- Peut être utilisé dans la classification ou dans la régression.
- Le principe de ce modèle consiste à choisir les k données les plus proches du point étudié afin d'en prédire sa valeur (sa classe).

madani.a@ucd.ac.ma









Principe de KNN

madani.a@ucd.ac.ma

Les étapes de KNN

1. Choisir la distance, la mesure de similarité entre les échantillons (instances)
2. Choisir (déterminer) la valeur de K. Cette valeur doit impair pour avoir un vote majoritaire.
3. Calculer la distance entre la nouvelle entrée et toutes les données de la base d'apprentissage.
4. Déterminer les classes des K voisins les plus proches .
5. Donner la prédiction pour la nouvelle entrée.

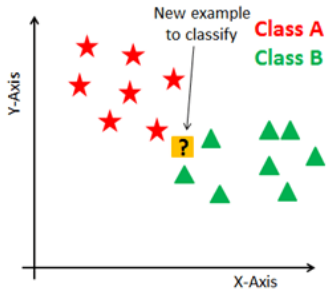
madani.a@ucd.ac.ma

Les étapes de KNN

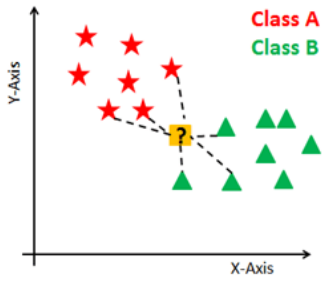
Exemples de distances

- **Distance Euclidienne :** $\mathbf{d}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- **Distance de Manhattan :** $\mathbf{d}(x, y) = \sum_{i=1}^{n}|x_i - y_i|$
- **Distance de Minkowski :** $\mathbf{d}(x, y) = \sqrt[q]{\sum_{i=1}^{n}|x_i - y_i|^q}$

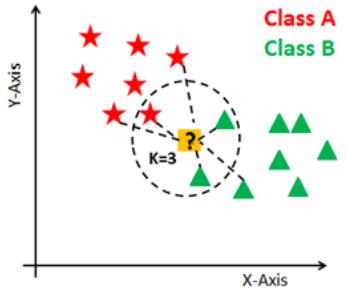
Initialisation



Calcul des distances



Définir les K voisins et voter



madani.a@ucd.ac.ma

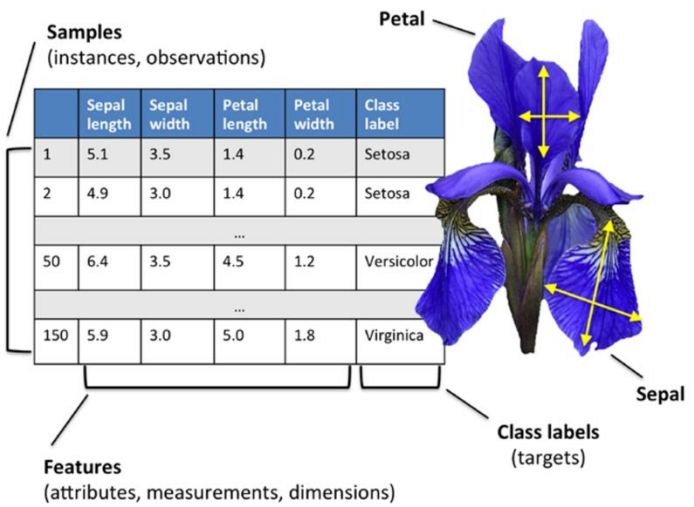
KNN : Algorithme

1. Charger les données
2. Initialiser k
3. Pour chaque exemple dans les données:
   a. Calculer la distance entre notre requête et l'observation
   b. Ajouter la distance et l'indice de l'observation concernée à une collection
4. Trier cette collection dans l'ordre croissant.
5. Sélectionner les k premières entrées de la collection de données triées
6. Obtenir les étiquettes des k entrées sélectionnées

madani.a@ucd.ac.ma

Exemple : Iris Classification

- Il s'agit dans cet exemple est de prédire la classe (Setosa, Versicolor, Virginica) d'une nouvelle fleur.
- Il s'agit d'une classification, car on fournit à l'ordinateur des exemples d'entrées et leurs sorties souhaitées
- Le dataset Iris est composé de 150 échantillons (samples, exemples) et 4 caractéristiques (features).
- On peut la représenter comme une matrice 150 x 4.



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

$$\begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & x_4^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & x_4^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{(150)} & x_2^{(150)} & x_3^{(150)} & x_4^{(150)} \end{bmatrix} \begin{bmatrix} y^{(1)} \\ \dots \\ y^{(150)} \end{bmatrix}$$

madani.a@ucd.ac.ma

Exemple : Iris Classification

```
# importer les modules nécessaires :
#pour charger les données des fleurs Iris
from sklearn.datasets import load_iris
#pour diviser les données en partie entrainement et partie test
from sklearn.model_selection import train_test_split
#pour importer le modèle (KNN dans notre cas)
from sklearn.neighbors import KNeighborsClassifier
#pour évaluer notre modèle
from sklearn.metrics import accuracy_score
```


madani.a@ucd.ac.ma

Exemple : Iris Classification

```
# sauvegarder les données (exemples et attributs) du dataset iris
iris = load_iris()
# stocker les features dans une matrice "X"
X = iris.data
# stocker les classes dans le vecteur "y"
y = iris.target
# afficher les structures (shape) de X et y
print(iris)
print(X.shape)
print(y.shape)
```


madani.a@ucd.ac.ma

Exemple : Iris Classification

```
#instancier le modèle
knn = KNeighborsClassifier()
# Diviser les données en deux sous ensembles Train et Test
X_train, X_test, y_train, y_test = train_test_split(X,
y,test_size=0.2, shuffle=True)
# Entraîner le modèle sur l'ensemble d'entraînement
knn.fit(X_train, y_train)
# evaluer le modèle sur l'ensemble de test
y_pred = clf.predict(X_test)
accuracy_score(y_test, y_pred)
```


madani.a@ucd.ac.ma

Exercice

- Quelles sont les valeurs par défaut de la distance et k dans l'implémentation sklearn de KNN
- Refaire l'exemple précédent en choisissant pour k les valeurs 3, 7 et 9
- Refaire l'exemple précédent en choisissant différentes valeurs de la distance
- Écrire un programme qui fait varier les valeurs de k entre 3 et 49 et dire quel est la valeur de k qui donne le meilleur score pour une distance donnée

madani.a@ucd.ac.ma