NSI

ALGORITHME DES K VOISINS

SOMMAIRE

- LA NOTION D'ALGORITHME D'APPRENTISSAGE
- LA NOTION DE CLUSTER / CLUSTERING
- L'OBJECTIF DE NOTRE ALGORITHME
- COMMENT IL MARCHE
- DANS QUELS CAS IL PEUT S'AVÉRER UTILE
- EXEMPLE D'UN ALGORITHME DES K VOISINS

UN ALGORITHME D'APPRENTISSAGE C'EST :

UN ALGORITHME QUI PREND UN ENSEMBLE CONNU DE DONNÉES D'ENTRÉE (C'EST POUR ÇA QUE L'ON UTILISE LE TERME APPRENTISSAGE PARCE QUE C'EST L'ENSEMBLE D'APPRENTISSAGE) ET DES RÉPONSES CONNUE(SORTIE) ET FORME UN MODÈLE POUR GÉNÉRER DES PRÉVISIONS RAISONNABLES POUR LA RÉPONSE AUX NOUVELLES DONNÉES D'ENTRÉE

LE CLUSTER C'EST :

UNE MÉTHODE D'APPRENTISSAGE NON SUPERVISÉ. AINSI ON N'ESSAIE PAS D'APPRENDRE UNE OBSERVATION ET UNE VALEUR A PRÉDIRE.

UN ALGORITHME DES K VOISINS C'EST :

D'UTILISER UN GRAND NOMBRE DE DONNÉES AFIN D'INSTRUIRE LA MACHINE À RÉSOUDRE UN CERTAIN TYPE DE PROBLÈME. ELLE SERT PLUS PRÉCISÉMENT À CLASSER NOS K PLUS PROCHES VOISINS COMME L'INDIQUE LE NOM.

L'ALGORITHME MARCHE DE CETTE FAÇON :

- FIXER LE NOMBRE DE VOISINS K
- ON DÉTECTE LES K VOISINS LES PLUS PROCHES DES NOUVELLES DONNÉES D'ENTRÉE QUE L'ON VEUT CLASSER
- ATTRIBUER LES CLASSES CORRESPONDANTES PAR VOTE MAJORITAIRE
- N'OUBLIONS PAS QUE L'ON PEUT FAIRE VARIER K ET QUE POUR CHAQUE VALEUR DE K, ON CALCULE LE TAUX D'ERREUR DE L'ENSEMBLE DE TEST (ON VÉRIFIE). PUIS POURS FINIR ON GARDE LE PARAMÈTRE K QUI MINIMISE CE TAUX D'ERREUR

IL PEUT S'AVÉRER UTILE DANS CES CAS-LÀ

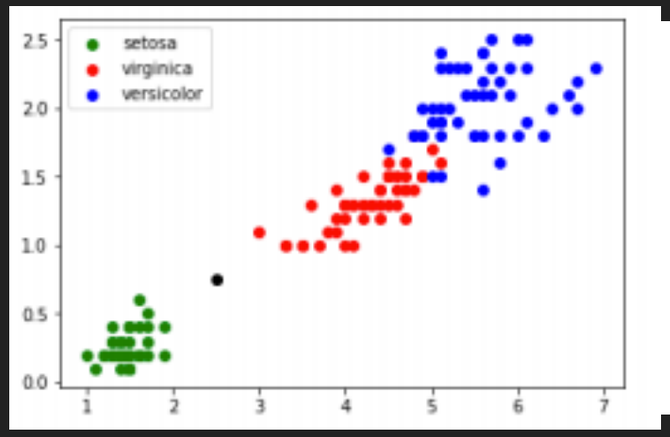
- LA CLASSIFICATION DES DOCUMENTS INDEXÉS DNS UN MOTEUR DE RECHERCHE
- DANS LE GÉO-MARKETING : PLACER DIFFÉRENTS POINTS DE VENTE POUR ANTICIPER SON CHIFFRE D'AFFAIRES
- CONSTRUIRE UN MOTEUR DE RECOMMANDATION SUR UN SITE INTERNET C'EST-À-DIRE QUE L'ACHAT DES AUTRES VISITEURS DU SITE PERMETTENT DE DÉFINIR MES SUGGESTIONS À D'AUTRES POTENTIELS ACHETEURS SOUS FORME DE RECOMMANDATION D'ACHATS
- DÉTERMINER QUEL CLIENT D'UNE ENSEIGNE EST APPÉTANT À UNE CARTE DE FIDÉLITÉ OU UN CRÉDIT : ON PEUT REGARDER QUI DANS LA BASE DE DONNÉES À DISPOSITION LUI RESSEMBLE LE MIEUX

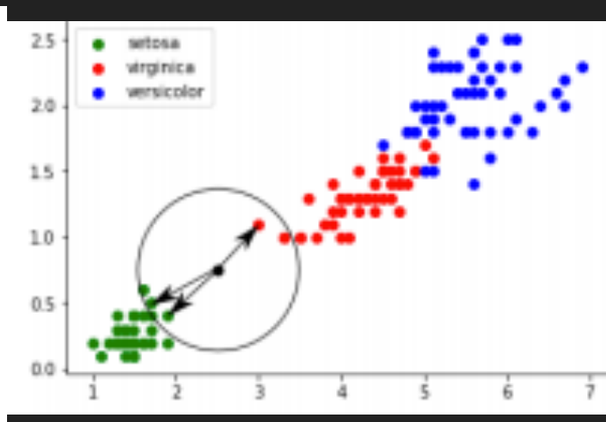
ON PEUT EN CONSTATER QUE CET ALGORITHME EST EXTRÊMEMENT EFFICACE DANS LE DOMAINE DU COMMERCE OU ENCORE CELUI DE LA FINANCE

EXEMPLE SIMPLE ET ASSEZ CONNU :

- ON PREND 3 IRIS DIFFÉRENTES. DANS UN DOSSIER NOMMÉ FLEUR (par ex) ON NOTE CHACUNE DE LEURS CARACTÉRISTIQUE: LA LONGUEUR ET LA LARGEUR DES SÉPALES. QUAND TOUT CELA EST FAIT ON LES ENREGISTRE DANS UN FICHIER CSV ET LES METTRE DANS UN TABLEUR.
- NOUS ALLONS UTILISER 3 BIBLIOTHÈQUES PYTHON : -PANDAS (IMPORTER LES DONNÉES DU FICHIERS CSV), -MAT PLOT LIB (VISUALISER LES DONNÉES) ET ENSUITE POUR FINIR SCIKIT-LEARN (PROPOSE UNE IMPLÉMENTATION DE L'ALGO DES K PLUS PROCHES VOISINS)
- APRÈS AVOIR MODIFIER LE FICHIER CSV, ON VA ÉCRIRE UN PROGRAMME PERMETTANT DE VISUALISER LES DONNÉES SOUS FORME DE GRAPHIQUE (VOIR 1ERE IMAGE)
- ENSUITE NOUS AURONS UNE REPRÉSENTATION GRAPHIQUE DES DONNÉES (2ÈME IMAGE) DANS LAQUELLE NOUS AVONS PLACER NOTRE POINT K
- PUIS POUR FINIR ON REPÈRE BIEN ÉVIDEMMENT NOS POINTS LES PLUS PROCHES DE K (2ÈME IMAGE)

```
import pandas
import matplotlib.pyplot as plt
iris=pandas.read_csv("iris.csv")
x=iris.loc[:,"petal_length"]
y=iris.loc[:,"petal_width"]
lab=iris.loc[:,"species"]
plt.scatter(x[lab == 0], y[lab == 0], color='g', label='setosa')
plt.scatter(x[lab == 1], y[lab == 1], color='r', label='virginica')
plt.scatter(x[lab == 2], y[lab == 2], color='b', label='versicolor')
plt.legend()
plt.show()
```





MERCI D'AVOIR LU