



## Algorithmique III et IA DIU – EIL k-plus-proches-voisins

Christine Froidevaux

[chris@lri.fr](mailto:chris@lri.fr)

Juin 2019



Ch.Froidevaux-DIU EIL - kNN -2019

## Intelligence Artificielle et Apprentissage automatique

- Histoire de l'IA : voir Wikipedia pour des dates clés
- *IA* : naissance aux USA du terme (1956, Newell, Simon McCarthy, Minsky et Samuel) ;
- Au départ : l'IA visait à *simuler sur ordinateurs les facultés cognitives humaines* et recouvrait des approches scientifiques relevant des sciences informatiques et mathématiques, et à la frontière des sciences cognitives
- *Apprentissage automatique* : aujourd'hui on confond IA et apprentissage automatique (machine learning voire deep learning) mais l'IA ne se réduit pas à l'apprentissage
- Succès du deep learning (méthode de réseaux neuronaux à plusieurs couches) : intégration de données volumineuses (Big Data) avec des méthodes statistiques sur des machines avec une puissance de traitement accrue

Ch.Froidevaux-DIU EIL - kNN -2019

## Quelques références

### Livres :

- Cornuéjols A. et Miclet L., Apprentissage artificiel, Eyrolles, 2002
- Mitchell T.M., Machine Learning, Int. Edition, 1997 (ed. française)
- Witten I.H., Data Mining, Morgan Kaufmann Publishers, 1999 (sur la boîte à outils WEKA)

### Sites :

- <http://www.cs.waikato.ac.nz/ml/weka/> : Data Mining Software in Java
- <http://www.rdatamining.com/> : R et DataMining
- <https://scikit-learn.org/> : Machine Learning en Python commande import sklearn

Ch.Froidevaux-DIU EIL - kNN -2019

## Exemples d'activités où l'apprentissage joue un rôle important

- Professions de la relation client : conseil de produits à acheter
- Moteur de recherche sur le web
- Réseau social : proposition de nouveaux amis
- Enseignement en ligne : parcours personnalisés (analyse des actions des apprenants)
- Consommation d'énergie : apprendre les profils de consommation
- Robotique (reconfiguration automatique) : robot militaire ou soignant
- Santé : système de diagnostic, de prescriptions

Ch.Froidevaux-DIU EIL - kNN -2019

## Exemples d'activités où l'apprentissage joue un rôle important

- Véhicule autonome (reconnaissance d'images pour l'environnement)
- Reconnaissance faciale popularisée par *GoogleFace* et *Facebook*
- Traitement de la langue : traduction automatique
- Reconnaissance de l'écriture
- Chatbot
- Analyse des contributions au Grand Débat National par OpinionWay (logiciel d'IA QWAM)

Ch.Poitrenaud-DIUEIL - RNN -2019

## Ex de logiciels d'apprentissage pour les jeux

- Le système *Deep Blue* d'IBM, superordinateur spécialisé dans le jeu d'échecs développé par IBM au début des années 1990, a gagné contre le champion du monde d'échecs en 1997.
- Le système *Watson* d'IBM, programme informatique d'intelligence artificielle conçu par IBM dans le but de répondre à des questions formulées en langage naturel, a participé à trois épisodes du jeu télévisé *Jeopardy*, et a gagné (2011).
- Le système *AlphaGo* de Google DeepMind, programme informatique capable de jouer au jeu de go, a gagné contre un des meilleurs joueurs mondiaux, Lee Sedol (2016).

Ch.Poitrenaud-DIUEIL - RNN -2019

## Trois grands types d'apprentissage

### non supervisé (ou clustering)

Le **clustering** consiste à organiser une collection de données non étiquetées dans des clusters basés sur la **similarité** telle que des données regroupées dans un même cluster sont plus semblables entre elles qu'à une donnée d'un autre cluster

### supervisé

**Apprendre la définition de classes** (catégories) à partir d'exemples de ces classes (exemples dits étiquetés) dans le but de faire de la **prédiction** i.e. **classifier** de nouvelles observations

**par renforcement** : l'algorithme procède par essais et erreurs, chaque erreur l'amenant à améliorer sa performance dans la résolution du problème. Les experts définissent juste les critères de succès de l'algorithme.

Ch.Poitrenaud-DIUEIL - RNN -2019

## Clustering

**Principe** : Au départ les points sont tous noirs et un algorithme de clustering va regrouper les points dans un premier cluster rouge et un deuxième cluster vert



Les flèches représentant la dissimilarité (e.g. distance) entre les données. Les flèches rouges et vertes au sein des clusters sont plus courtes que les flèches noires entre éléments de deux clusters différents

**Ex** : regrouper des gènes co-exprimés chez des patients

Ch.Poitrenaud-DIUEIL - RNN -2019

## Apprentissage supervisé

### Apprendre des classes

- En général :

- Input:**
  - Des classes (ou catégories) [par ex : classe des points rouges, des points bleus et des points verts]
  - Un ensemble d'examples étiquetés (ensemble des données pour lesquelles les classes sont définies) [par ex, des points du plan avec chacun l'une des trois couleurs]
  - Output : une description des classes OU un ensemble de règles permettant de discriminer les classes entre elles OU une fonction numérique f apprise (cf régression) OU une méthode pour prédire la classe d'un objet non étiqueté...
- Principe : apprendre à un enfant à reconnaître un chien à partir d'une base d'images

ChP-Aideaux-DU-Et-kNN-2019

## Principe des K plus proches voisins

(k-Nearest Neighbors, kNN, années 1970, en statistiques et pattern recognition)

2 classes : cercle ● et triangle ▲

« Dis-moi qui tu fréquentes je te dirai qui tu es »

Suppose une certaine régularité dans les données

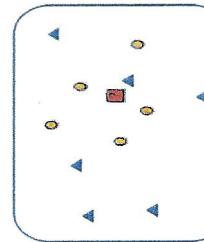
- Utilisé pour classer des exemples en utilisant les  $k$  données étiquetées « les plus proches » dans l'espace des caractéristiques → notion de **distance** / **dissimilarité** entre données
- Exemple : deux classes « triangle » et « cercle »

Quelle classe pour le nouvel élément rouge ?  
 $K = 4$  : classe cercle ●

ChP-Aideaux-DU-Et-kNN-2019

## Principe des K plus proches voisins

2 classes : ● et ▲

- 
- Un des meilleurs algorithmes de fouille de données et parmi les plus simples pour la classification
  - Classifieur paresseux : le calcul est reporté au moment où la classification d'un nouvel élément doit se faire (on n'apprend pas un classifieur / une fonction cible)

- Instance-based learning

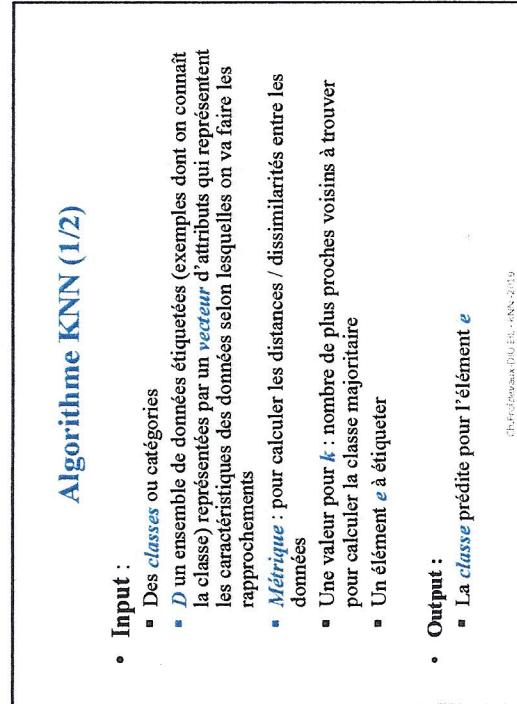
Quelle classe pour l'instance rouge ?  
 $K = 4$  : classe ●

ChP-Aideaux-DU-Et-kNN-2019

## Algorithme KNN (1/2)

- Input :**
  - Des classes ou catégories
  - D'un ensemble de données étiquetées (exemples dont on connaît la classe) représentées par un **vecteur** d'attributs qui représentent les caractéristiques des données selon lesquelles on va faire les rapprochements
  - Métrique : pour calculer les distances / dissimilarités entre les données
  - Une valeur pour  $k$  : nombre de plus proches voisins à trouver pour calculer la classe majoritaire
  - Un élément  $e$  à étiqueter
- Output :**
  - La **classe** prédictée pour l'élément  $e$

ChP-Aideaux-DU-Et-kNN-2019



$D = \{ [e_1, \text{dist}(e, e_1), \text{classe } e_1], [e_2, \text{dist}(e, e_2), \text{classe } e_2], \dots, [e_n, \text{dist}(e, e_n), \text{classe } e_n] \}$   
 $D = \{e_1, e_2, \dots, e_n\}$

21/06/2019

## Algorithme KNN (2/2)

- Algorithme pour classer un nouvel élément  $e$ , étant donné un ensemble de données étiquetées  $D$  :

1. Calculer la distance de  $e$  à tous les autres éléments de  $D$
2. Identifier les  $k$  plus proches voisins de  $e$
3. Utiliser la classe des  $k$  plus proches voisins pour déterminer la classe du nouvel élément par vote majoritaire

→ Cela suppose qu'on a choisi :

- Les descripteurs pour les données [ici les coordonnées des points]
- Une distance entre éléments (euclidienne, Manhattan, Hamming)

Ch.Froidveaux-DIU EIL - KNN - 2019

adopté le tri  
 Selection pour quel  
 ne classe que les  
 k  
 élément s dans la  
 tableau  
 Puis l'compt par classe.  
 rechuds le max des compteurs

dist de tuples IN

## Classification

- Déterminer la classe à partir de celles des  $k$  plus proches voisins

➤ Vote majoritaire des étiquettes des  $k$  plus proches voisins

➤ Mais si une classe est plus représentée qu'une autre on trouvera plus de voisins dans cette classe et la prédiction sera biaisée vers cette classe

→ facteur de pondération possible :

chaque proche voisin intervient plus ou moins selon sa distance à l'élément à classer (par exemple, poids  $1/d$  si le voisin est à la distance  $d$ )

Ch.Froidveaux-DIU EIL - KNN - 2019

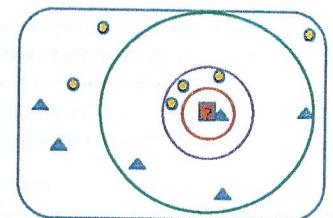
## Régression

- On peut aussi utiliser le principe des  $k$  plus proches voisins dans le cadre de la **régression** : on cherche alors à calculer la valeur d'une fonction cible à **valeurs continues**

→ l'output est alors la moyenne des valeurs des fonctions cibles des  $k$  plus proches voisins

Ch.Froidveaux-DIU EIL - KNN - 2019

## Choix de $k$



K = 1 : classe ?
K = 4 : classe ?
K = 7 : classe ?

On a un nouvel élément (en rouge sur le dessin) et on se demande quelle forme il a : triangle ou cercle ? On regarde ses  $k$  voisins.

Ch.Froidveaux-DIU EIL - KNN - 2019

exemple :  $D$  : eniris constitué de Versicolor (45°) et Setosa 60  
 on partage en 2  $E = 2/3 D$  et  $V = 1/3 D$  avec une proportion de Vert et Seto.  
 21/06/2019

$E = 30$  Vertich.  
 $V = 15$  vertic.

50 Setosa  
 20 Setosa

on tente sur V  
 on prend un élément connu

### Choix de k

**Trop loin de  $x$  ?**

K = 1 : classe	▲
K = 4 : classe	●
K = 7 : classe	—

Que conclure ?  
 Éliminer  $k = 7$  car les points sont trop loin de  $x$ ? Se limiter à  $k = 1$  car point très proche de  $x$ ?

- On a rajouté un élément rouge et on se demande quelle forme il a : selon la valeur de  $k$ , on n'a pas la même réponse ...

Ch.Froidveaux-DIU EIL - kNN - 2019

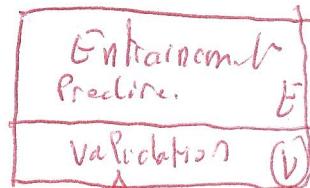
### Choix du paramètre k

- Si  $k$  est **trop petit**, sensible aux données bruitées
  - Si  $k$  est **trop grand**, le voisinage peut inclure des éléments d'autres classes
- ⇒ Différentes **heuristiques** sont possibles pour le choix de  $k$

On peut utiliser un **ensemble de validation**. On sépare l'ensemble  $D$  des données étiquetées en un ensemble de **données d'entraînement**  $E$  et un ensemble de **données de validation**  $V$  (plus petit). On fait ensuite tourner l'algorithme kNN sur  $E$  avec différentes valeurs de  $k$  pour chaque élément de  $V$ , et on calcule le taux d'erreurs à chaque fois, en confrontant la classe prédictive par les  $k$  plus proches voisins dans  $E$  et son étiquette. On choisit  $k$  qui minimise le taux d'erreur.

Rem : on peut aussi utiliser la validation croisée (cross-validation)  
 ...

Ch.Froidveaux-DIU EIL - kNN - 2019



cherche dans cette zone permet de fixer K

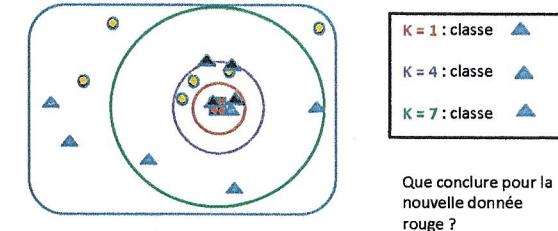
### Choix des données d'apprentissage : on a rajouté dans D des triangles (en noir)

**Que conclure pour la nouvelle donnée rouge ?**

K = 1 : classe	▲
K = 4 : classe	●
K = 7 : classe	—

Ch.Froidveaux-DIU EIL - kNN - 2019

### Choix des données d'apprentissage : on a rajouté dans D des triangles (en noir)



Ch.Froidveaux-DIU EIL - kNN - 2019

## Importance des données D

### ◆ Importance du choix des données :

S'il y a beaucoup plus de cercles que de triangles dans les données étiquetées utilisées, on va plus souvent prédire un cercle qu'un triangle :

**!! attention aux BIAIS des données choisies pour classifier !!**

Plus généralement, en apprentissage supervisé, **bien choisir les données d'entraînement**

### ◆ Choisir des descripteurs des données pertinents

Ch.Froiddevaux-DIU EEL - KNN -2019

## Biais des données d'apprentissage ?

### Algorithmes d'apprentissage sexistes, racistes ?

- Projet interne de **recrutement du personnel** chez Amazon, arrêté en 2014 : détection de tendances sexistes.

Pourquoi ? La base de données qui a été utilisée pour entraîner l'algorithme était constituée majoritairement de profils masculins. Par conséquent, l'algorithme écartait systématiquement les femmes des postes techniques.

- Logiciel Compas, utilisé pour **prédir les risques de récidive des détenus**, accusé de surévaluer ces risques pour les afro-américains

**!! Ce n'est pas l'algo qui est en cause mais la personne qui a entraîné l'algo sur des données biaisées : il faut choisir des données d'entraînement de qualité et représentatives de l'ensemble des données qu'on étudie !!**

Ch.Froiddevaux-DIU EEL - KNN -2019

## Forces et faiblesses (1/2)

### • Forces :

- Simple à comprendre
- Facile d'expliquer la prédiction
- Simple à implémenter et utiliser
- Robuste aux données bruitées par la moyenne calculée sur plusieurs voisins*
- Très performants pour certaines applications de bioinformatique

Ch.Froiddevaux-DIU EEL - KNN -2019

## Forces et faiblesses (2/2)

### • Faiblesses

- Classification relativement **coûteuse** à cause du calcul des distances du nouvel élément à tous les éléments étiquetés
- La précision peut être dégradée par trop de **données bruitées** ou des **attributs mal choisis** :
  - Certains attributs sont plus importants que d'autres : leur donner plus de **poids** dans le calcul de la distance
  - Certains attributs peuvent jouer à tort un rôle prépondérant si le domaine de leurs valeurs est trop grand dans le calcul de la distance  
➔ il faut les **normaliser**
  - Certains attributs ne sont pas pertinents : ils ne devraient pas être considérés dans la distance  
➔ **sélection d'attributs**
  - Ne pas choisir un trop grand nombre d'attributs  
➔ techniques de **réduction de dimension**

Ch.Froiddevaux-DIU EEL - KNN -2019

**Exercice :** Quelle est la classe qu'on obtient avec KNN pour la fleur e9 décrite par (petit, long, grand, moyen) ?

Exemples étiquetés	Longueur_sépale	Largeur_sépale	Longueur_pétale	Largeur_pétale	Classe : setosa ou versicolor
e1	grand	long	petit	long	setosa
e2	grand	moyen	grand	moyen	versicolor
e3	grand	moyen	petit	long	setosa
e4	petit	court	petit	court	versicolor
e5	grand	court	petit	long	setosa
e6	grand	moyen	petit	moyen	setosa
e7	petit	long	grand	court	versicolor
e8	grand	long	petit	court	versicolor

Ch.Froidevaux-DIU EIL - kNN - 2019

**Prédiction :** Quelle est la classe prédicta par KNN pour e9 = (petit, long, grand, moyen) ?

On utilisera la **distance de Hamming** pour  $k = 1$ , puis  $k = 3$  ? En cas d'égalité dans la distance de plusieurs exemples étiquetés à e9, on considérera que différer sur l'attribut « longueur\_sépale » est pénalisant.

Ch.Froidevaux-DIU EIL - kNN - 2019

### Correction des algorithmes d'apprentissage ?

- L'algorithme kNN est « correct » au sens de la preuve de programme par rapport à sa spécification mais cela ne veut pas dire qu'il fera de bonnes prédictions !
- On évalue la **performance d'un classifieur** en termes de taux d'erreur (*error rate*). On considère les erreurs faites par le classifieur sur un ensemble de données étiquetées :

Taux d'erreur = Nb d'erreurs / Nb total de données considérées

⇒ **Nécessité d'un ensemble de données de test, indépendant des données d'entraînement.** On fait l'hypothèse que ces deux ensembles sont des échantillons représentatifs du problème sous-jacent et on évalue le taux d'erreur sur le test set

❖ *Les données sur lesquelles on teste ne doivent pas avoir été utilisées pour apprendre le classifieur*

Ch.Froidevaux-DIU EIL - kNN - 2019

### Apprendre, valider et tester

- On va séparer les données étiquetées dont on dispose en trois ensembles disjoints :
  - **Données d'apprentissage** (training data) : utilisées par un ou plusieurs algorithmes d'apprentissage pour obtenir des classificateurs
  - **Données de validation** (validation data) : utilisées pour régler (optimiser) les paramètres de ces classificateurs ou pour en choisir un en particulier (ici choix de k pour l'algorithme kNN)
  - **Données de test** (test data) : utilisées pour calculer le taux d'erreur du classificateur final optimisé

Ch.Froidevaux-DIU EIL - kNN - 2019

## Evaluation de la qualité de la classification

- But** : distinguer les éléments absents dans une classe des éléments présents à tort dans une classe
- Matrice de confusion** : matrice carrée qui indique le nombre d'éléments de classe i qui ont été prédits (classés) en classe j sur un ensemble test.

⇒ Cas de la classification binaire :

Classe prédictive \ Classe réelle	oui	non
oui	TP	FN
non	FP	TN

TP : vrai positif (true positive)  
TN : vrai négatif (true negative)  
FP : faux positif (false positive)  
FN : faux négatif (false negative)

## Rappel, précision

**Matrice de confusion** : la prédiction est bonne sur un ensemble test si le plus grand nombre d'éléments se retrouve sur la **diagonale**

**Taux d'erreur** : rapport entre le nbre d'éléments mal classés dans l'ens test et le nbre d'éléments de l'ens test

Mais selon les applications les **faux positifs** et les **faux négatifs** n'ont pas le même coût (ex: analyse transcriptome, appendicite)

⇒ notions de rappel, précision :

- Rappel** = 1 : on a appris tous les exs > 0 (complétude)
- Précision** = 1 : tous les exs prédits > 0 sont bien > 0 ; on n'a pas trouvé de faux > 0 (correction)

Ch.Froiddevaux-DIU EL - KNN -2019

## A retenir

- Importance du **choix** et de la **qualité** des **données d'entraînement** (pas de données **bruitées**, pas de **biais** dans les données, cad, pas de sur-représentation d'une classe dans les données)
- Importance du choix des **descripteurs** des données et du poids qu'on donne à chacun
- Choisir un **grand ensemble de données** (big data)
- Pas simple d'utiliser des algorithmes d'une boîte à outils qui restent à **paramétrier** (e.g. trouver k)
- Les prédictions sont **incertaines** : nécessité de valider les classificateurs appris.
- On évalue les classificateurs sur des ensembles de **données de test différentes** des ensembles de **données d'entraînement**.
- Toutes les erreurs n'ont pas la même importance (**rappel** vs **précision**).
- Problèmes **d'explicabilité** ; algorithmes d'IA boîtes noires ?

⇒ éthiques par conception

Ch.Froiddevaux-DIU EL - KNN -2019

## Activités

- Sur les **iris** : prendre le jeu de données « iris » (ou Fisher's Iris data ou encore Anderson's Iris data set) qui a été introduit par le statisticien et biologiste britannique Ronald Fisher en 1936 (article "The use of multiple measurements in taxonomic problems"). On peut récupérer un fichier .csv
- Voir aussi :  
<https://gist.github.com/curnan/a08a1080b88344b0c8a7>
- Sur un exemple de reconnaissance de la langue dans un texte à partir de l'analyse des fréquences des lettres utilisées :  
<https://www.isnbreizh.fr/lnsi/activity/algoRefKnn/index.html>

Ch.Froiddevaux-DIU EL - KNN -2019