Apprentissage automatique (Machine Learning)

HAMID NECIR | COURS FOUILLE DE DONNÉES

Apprentissage automatique (Machine Learning)

- domaine d'étude de l'IA qui vise à donner aux machines la capacité d'apprendre.

Paradigmes d'apprentissage du Machine Learning :

1. **L'Apprentissage Supervisé** (Supervised Learning)
2. **L'Apprentissage Non-Supervisé** (Unsupervised Learning)
3. **L'Apprentissage par Renforcement** (Reinforcement Learning)

NECIR HAMID

NECIR HAMID

L'Apprentissage supervisé

objectif

Il s'agit de définir des règles permettant de classer des objets dans des classes à partir de variables qualitatives ou quantitatives caractérisant ces objets.

un algorithme d'apprentissage supervisé repose sur des données d'entrée étiquetées qui permettent de définir une fonction par apprentissage, fonction qui va fournir une réponse appropriée lorsque de nouvelles données non étiquetées sont fournies.

Exemple

Prédire le prix d'un appartement en fonction de sa surface habitable et du nombre de pièces.

Exemple

Nous voulons apprendre à un enfant à quoi ressemble une fleur.

1) Nous allons lui montrer plusieurs images différentes, en lui indiquant lesquelles sont des fleurs et lesquelles ne le sont pas !

Quand on voit une fleur, on crie «fleur !"» Quand ce n'est pas une fleur, on crie «pas fleur !».

Après cette phase d'apprentissage l'enfant doit normalement être capable si on lui montre une photo de fleur de dire "fleur !"

Domaines d'application

- **Délivrance de crédit**
- **Diagnostic médical**
- **Prédiction du cours d'une action**
- **Optimisation d'un envoi de courrier**
- **…**

Processus

Processus en deux étapes:

1. **étape d'apprentissage (entrainement)**

   Il faut disposer au départ d'un échantillon dit d'apprentissage dont le classement est connu. Cet échantillon est utilisé pour construire un modèle pour classer des données inconnues.

Processus

1. étape d'apprentissage (entrainement)

Il s'agit de tester le modèle pour s'assurer de sa capacité de généralisation.

On utilise un deuxième échantillon indépendant, dit de validation ou de test qui peut être:

1. Les données d'entrainement elles-mêmes.
2. une base de données différente appelée base de test. La base de test est un ensemble d'exemples ayant les mêmes caractéristiques que ceux de la base d'entrainement et qui sont écartés au départ de l'entrainement pour effectuer les tests.

Processus

2. Étape d'utilisation

L'étape d'utilisation dépend essentiellement du type d'information prédite.

1. Si l'attribut est catégoriel ou symbolique (appartient à un ensemble fini), il s'agit de classification.

2. si cet attribut est continu (numérique) il s'agit d'un problème de régression.

Processus

Evaluation du modèle

Il permet d'étudier la fiabilité du modèle pour l'appliquer.

a) Taux de reconnaissance (la précision du modèle).

Elle représente le rapport entre le nombre de donnée correctement classées et le nombre total des données testées.

b) Matrice de confusion

La mesure précédente donne le taux d'erreurs commises par le modèle appris mais ne donne aucune information sur la nature de ces erreurs.

Les méthodes

- Il existe de nombreuses méthodes de classification supervisée :
- les k plus proches voisins (KNN)
- analyse (factorielle) discriminante
- régression logistique
- arbres de décision
- réseaux de neuronnes
- SVM . . .

Algorithme les kNN

Algorithme KNN (base apprentissage, k, inconnu)

- **Pour chaque donnée dans l'echantillon d'apprentissage**
  - **Calculez la distance entre l'inconnu et la donnée courante.**
  - **Mémoriser cette distance et la donnée associée dans une liste de couple (distance, donnée).**
- **Triez la liste (distance, donnée) du plus petit au plus grand sur les distances.**
- **Choisissez les k premières entrées de cette liste.**
- **renvoyer la classe majoritaire.**

Exemple

NECIR HAMID

Classification d'une fleur suivant 2 espèces, suivant la tailles de ses pétales et de sa tige:

1) Echantillons de fleurs déjà étiquetée:

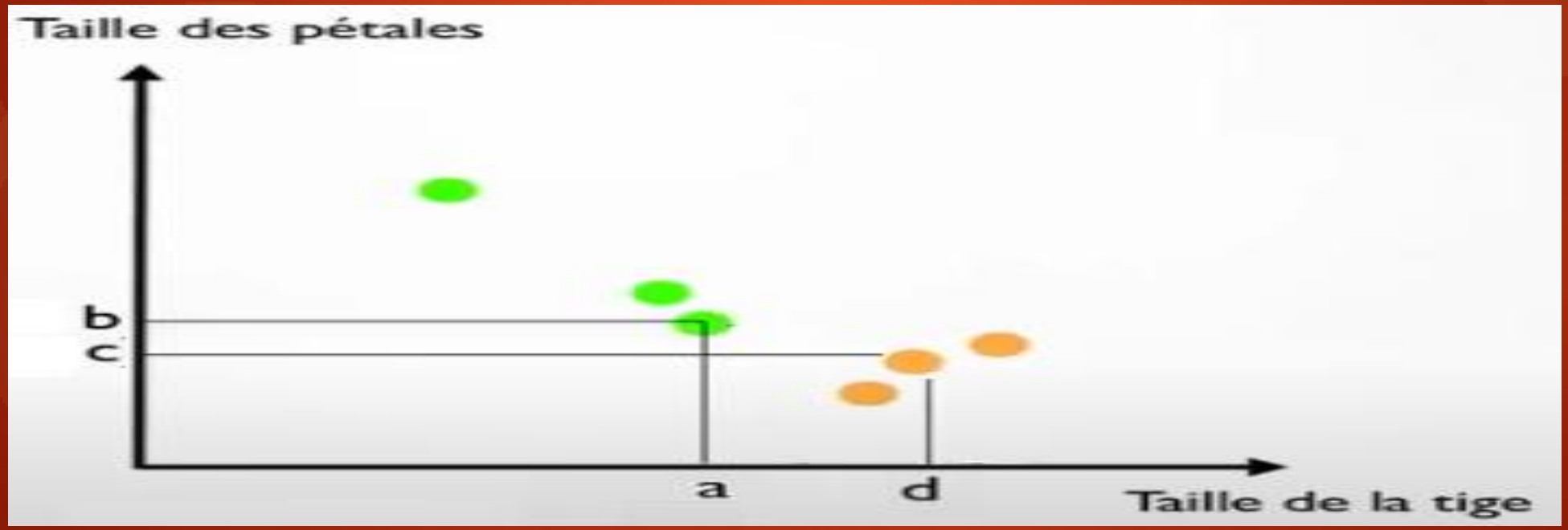
Taille des pétales	Taille de la tige	
18	43	Espèce 1
26	36	
17	44	
13	48	Espèce 2
15	57	
14	55	

Exemple

NECIR HAMID

placement des points représentant les fleurs dans un plan, suivant la tailles de leurs pétales et de leur tige:

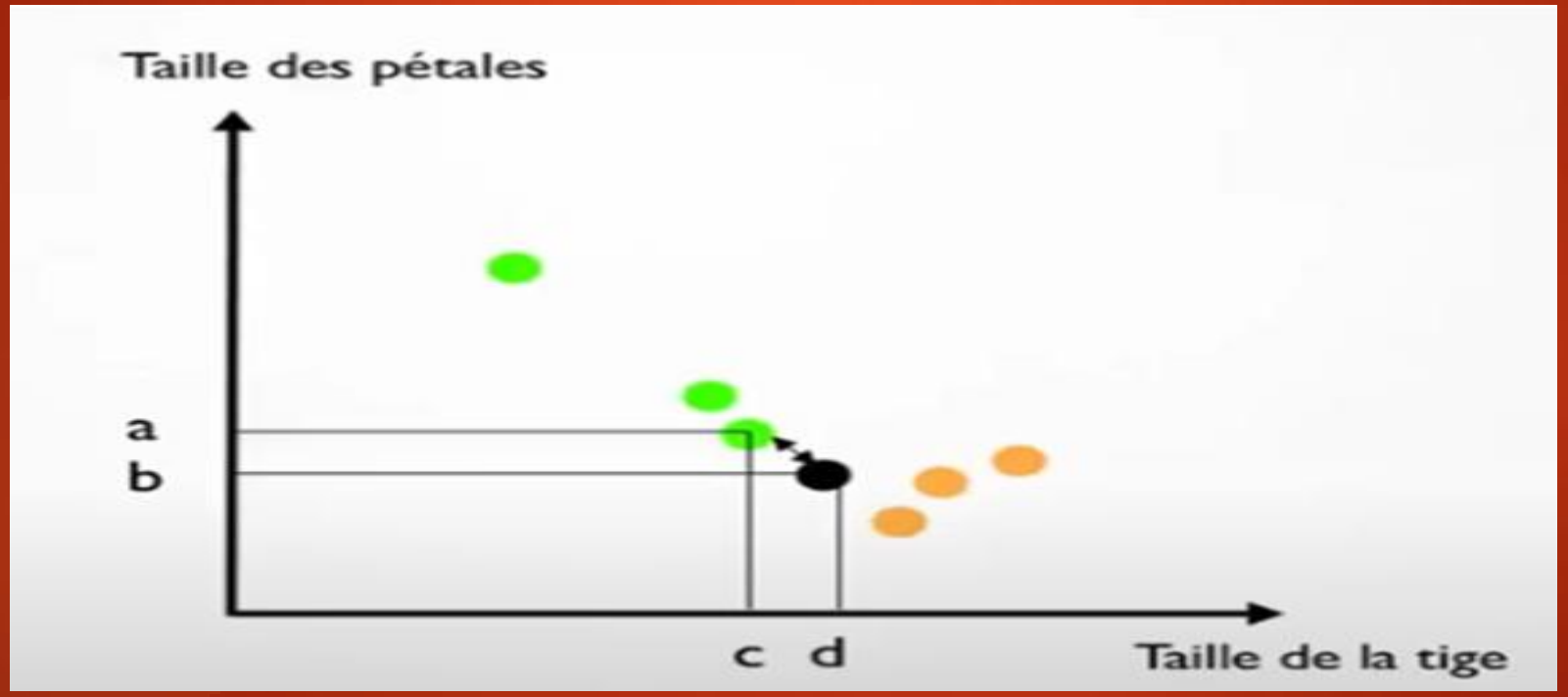
Echantillons de fleurs déjà étiquetée:



Exemple

placement la nouvelle fleur (inconnue) dans le plan.

En utilisant la notion de distance (euclidienne, City Block,..) détermination des points les plus proches d'elle.



NECIR HAMID

Exemple

F1(36,26); F2(43,18); F3(44,17) ; F4(48,13); F5(55,14); F6(57,15)

Soit F7=(46,14)

D(F4,F7)=(48-46)+(13-14)=3

D(F3,F7)=(44-46)+(17-14)=5

D(F2,F7)=(43-46)+(18-14)=7

D(F5,F7)=(55-46)+(14-14)=9

D(F6,F7)=(57-46)+(14-14)=11

D(F1,F7)=(36-46)+(26-14)=22

Pour k=3, la fleur appartient a l'espèce 1(en vert)

Pour k=5, la fleur appartient a l'espèce 2 (en jaune)

Domaine d'utilisation de KNN.

NECIR HAMID

- Comparaison de personnes possédant des caractéristiques financières similaires pour accord de prêts bancaires.
- Elaboration d'un profil pour proposer aux abonnés des films appropriés.
- Classer un électeur potentiel dans les catégories «votera» ou «ne votera pas» pour tel ou tel candidat.

Avantages

- L'algorithme est simple et facile à mettre en œuvre.
- Aucune hypothèse sur les données

Inconvénients

- L'algorithme devient beaucoup plus lent à mesure que le nombre d'exemples d'apprentissage augmente.
- Le choix de la méthode de calcul de la distance ainsi que le nombre de voisins K peut ne pas être évident
- L'étape de prédiction peut-être lente.