



Module

MLCI



ML

CI

Module

MLCI

ML

CI

Machine **L**earning

Computational
Intelligence

Université Constantine 2 -Abdelhamid Mehri
Faculté des Nouvelles Technologies de l'Information et de la Communication
Département de l'Informatique Fondamentale et ses Applications

Module



Machine Learning and Computational Intelligence

MLCI

Unité d'enseignement: UEF3

Crédit: 5

Coefficient: 3

Cours: 1H30/semaine

TP: 1H30/semaine

Dr. Fergani

Baha.fergani@univ-constantine2.dz

Module

MLCI

Machine Learning

Deep Learning

Module

MLCI

Machine Learning

**Apprentissage par
renforcement**

Module

MLCI

Machine Learning

Logique Floue

Module

MLCI

Machine Learning

Computational Intelligence

Apprentissage automatique (Machine Learning)

Apprentissage automatique (Machine Learning)

Apprentissage

Apprentissage automatique (Machine Learning)

Apprentissage

Automatique

Apprentissage

L'apprentissage:

- Est un **comportement humain** naturel
- Qui est devenu un aspect essentiel des machines.

Apprentissage automatique

- Apprentissage automatique est un champ d'étude de l'intelligence artificielle.
- Il vise à donner aux ordinateurs la capacité d'apprendre à partir des **données** sans **assistance**.

Apprentissage automatique

- Apprentissage automatique est un champ d'étude de l'intelligence artificielle.
- Il vise à donner aux ordinateurs la capacité d'apprendre à partir des **données** sans **assistance**.

Données ???

Apprentissage automatique

- L'apprentissage automatique nécessite deux ensembles de données:
 - **Ensemble de données pour l'entraînement:** c'est les données utilisées pour entraîner l'algorithme d'apprentissage.

Pendant cette phase, les paramètres du modèle peuvent être réglés (ajustés) en fonction des performances obtenues.

Apprentissage automatique

- L'apprentissage automatique nécessite deux ensembles de données:
 - **Ensemble de données pour l'entraînement:** c'est les données utilisées pour entraîner l'algorithme d'apprentissage.

Pendant cette phase, les paramètres du modèle peuvent être réglés (ajustés) en fonction des performances obtenues.

Apprentissage automatique

- L'apprentissage automatique nécessite deux ensembles de données:
 - **Ensemble de données pour le test:** il est utilisé pour évaluer les performances du modèle sur les données non-vues.

Types d'apprentissage automatique

Types d'apprentissage automatique

Les méthodes d'apprentissage peuvent être classées en **trois** principales catégories:

**Apprentissage
supervisé**



Types d'apprentissage automatique

Les méthodes d'apprentissage peuvent être classées en **trois** principales catégories:

**Apprentissage
supervisé**



**Apprentissage non-
supervisé**



Types d'apprentissage automatique

Les méthodes d'apprentissage peuvent être classées en **trois** principales catégories:

**Apprentissage
supervisé**



**Apprentissage non-
supervisé**



**Apprentissage par
renforcement**



Types d'apprentissage automatique

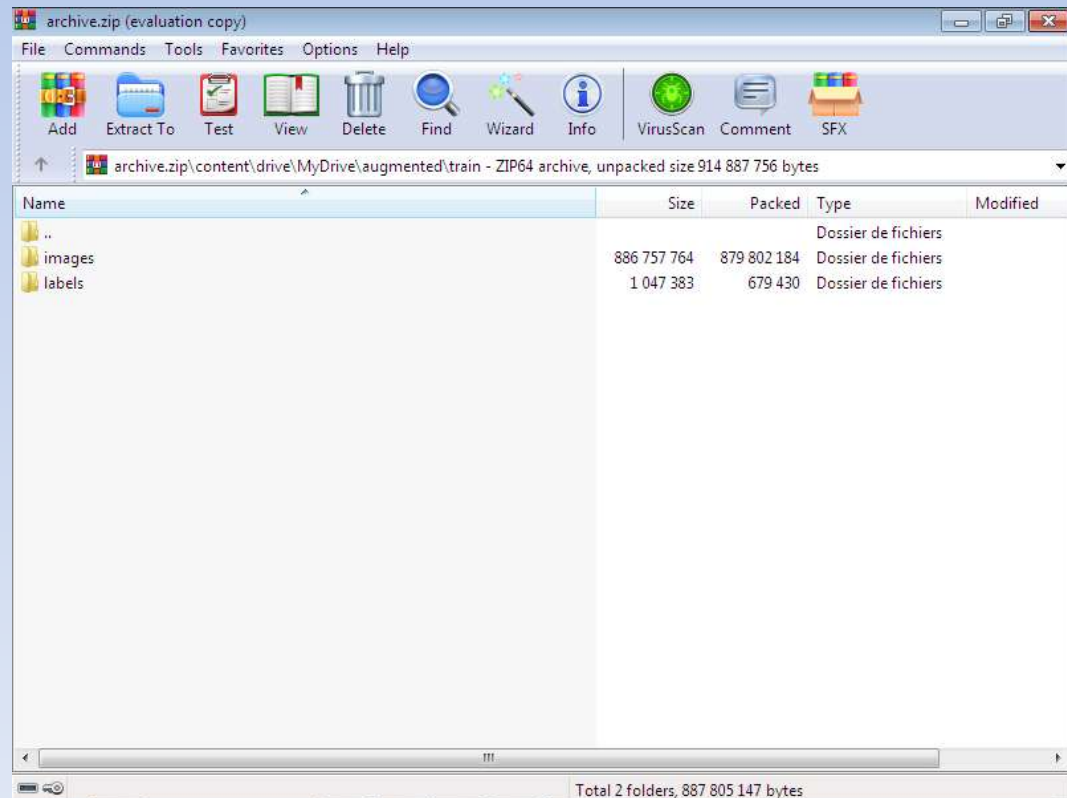
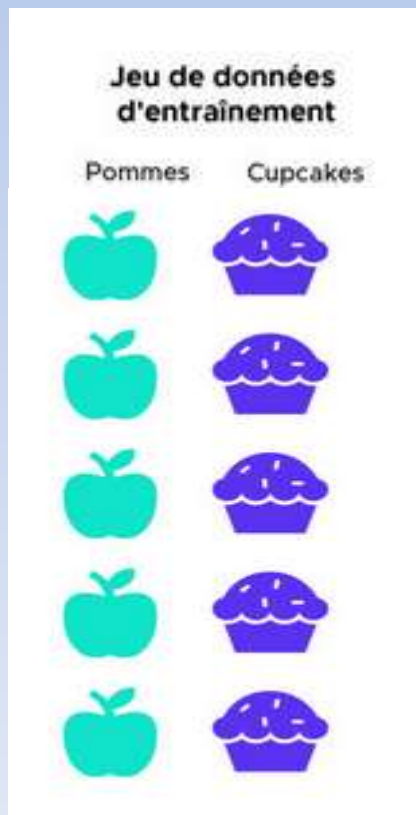
Les méthodes d'apprentissage peuvent être classées en **trois** principales catégories:

**Apprentissage
supervisé**



Apprentissage supervisé

- L'apprentissage supervisé (supervised learning) s'intéresse aux données étiquetées.



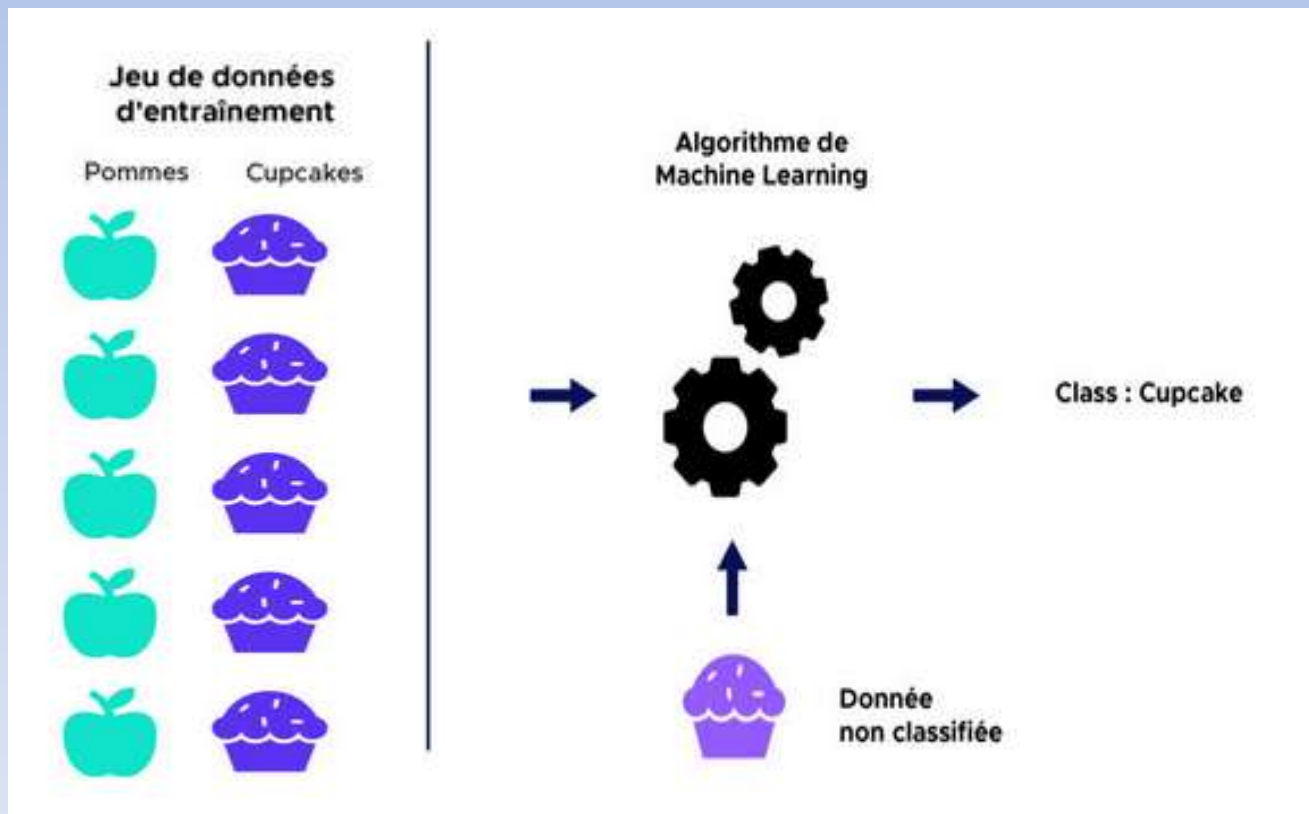
Apprentissage supervisé

Objectif:

- Est de prédire l'étiquette inconnu y
- Associée à une nouvelle observation x ,
- A partir de la connaissance fournie par les N observations étiquetées du jeu de données.

Apprentissage supervisé

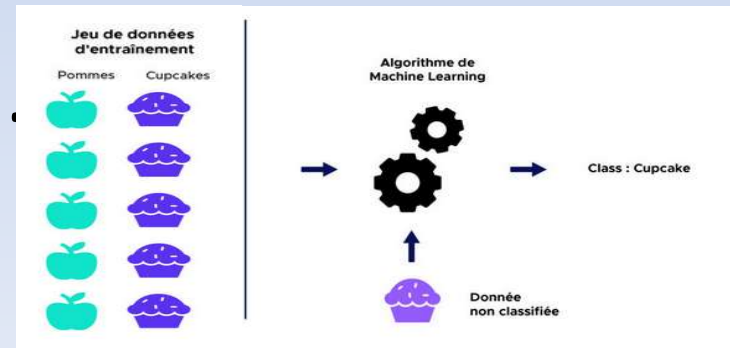
Objectif:



Apprentissage supervisé

Lors de l'apprentissage supervisé, l'algorithme reçoit:

- Un ensemble de données qui est étiqueté
- Sur lequel il va pouvoir s'entraîner **et** définir un modèle de prédiction.



Apprentissage supervisé

- Cet algorithme pourra par la suite être utilisé sur de nouvelles données.
- Afin de prédire leurs valeurs de sorties correspondantes.

Apprentissage supervisé

Exemples d'algorithmes d'apprentissage

supervisé: les algorithmes d'apprentissage

supervisé les plus utilisés sont :

- Support Vector Maching (SVM).
- L'arbre de décision.
- K-Nearest Neighbors (KNN).

Apprentissage supervisé

Exemples d'algorithmes d'apprentissage

supervisé: les algorithmes d'apprentissage

supervisé les plus utilisés sont :

- Support Vector Maching (SVM).
- L'arbre de décision.
- K-Nearest Neighbors (KNN).

Apprentissage supervisé

Exemples d'algorithmes d'apprentissage

supervisé: les algorithmes d'apprentissage

supervisé les plus utilisés sont :

- Support Vector Maching (SVM).
- L'arbre de décision.
- K-Nearest Neighbors (KNN).

Apprentissage supervisé

Exemples d'algorithmes d'apprentissage

supervisé: les algorithmes d'apprentissage

supervisé les plus utilisés sont :

- Support Vector Maching (SVM).
- L'arbre de décision.
- K-Nearest Neighbors (KNN).

Algorithme
***k* plus proches voisins**
(*k* Nearest Neighbors: KNN)

KNN

- L'algorithme KNN est utilisé pour la classification ou la régression.
- En classification, l'algorithme détermine à quelle classe appartient un échantillon en fonction de ses voisins les plus proches.

KNN

- L'algorithme KNN est utilisé pour la classification ou la régression.
- En classification, l'algorithme détermine à quelle classe appartient un échantillon en fonction de ses voisins les plus proches.

KNN

- L'algorithme KNN est utilisé pour la classification ou la régression.
- En classification, l'algorithme détermine à quelle classe appartient un échantillon en fonction de ses voisins les plus proches.
- En régression, l'algorithme calcule la moyenne des valeurs cibles des k plus proches voisins.

KNN

Principe

- L'algorithme kNN suppose que des objets similaires existent à proximité.

En d'autres termes, les éléments similaires sont proches les uns des autres.

KNN

Principe

- L'algorithme kNN suppose que des objets similaires existent à proximité.

En d'autres termes, les éléments similaires sont proches les uns des autres.

Etapes de KNN

Algorithme

KNN

Algorithme

Etape 1: Charger les données

Etape 2: Initialiser **k**: le nombre de voisins.

Etape 3: Calculer toutes les distances entre cette observation en entrée et les autres observations du jeu de données,

Etape 4: Conserver les **k** observations du jeu de données qui sont les plus « proches » de l'observation à prédire,

KNN

Algorithme

Etape 1: Charger les données

Etape 2: Initialiser **k**: le nombre de voisins.

Etape 3: Calculer toutes les **distances** entre l'**observation en entrée** et les autres **observations** du jeu de données.

Etape 4: Conserver les **k** observations du jeu de données qui sont les plus « proches » de l'observation à prédire,

KNN

Algorithme

Etape 1: Charger les données

Etape 2: Initialiser k .

Etape 3: Calculer toutes les distances entre cette observation en entrée et les autres observations du jeu de données,

Etape 4: Conserver les k observations du jeu de données qui sont les plus « **proches** » de l'observation à prédire,

KNN

Algorithme

Etape 5: Prendre les valeurs des observations retenues:

- Si on effectue une **régression**: l'algorithme calcule la moyenne (ou la médiane) des valeurs des observations retenues.
- Si on effectue une **classification**, l'algorithme assigne une étiquette (label) de la classe majoritaire à la donnée qui était inconnue.

KNN

Algorithme

Etape 5: Prendre les valeurs des observations retenues:

- Si on effectue une **régression**: l'algorithme calcule la moyenne (ou la médiane) des valeurs des observations retenues.
- Si on effectue une **classification**, l'algorithme assigne une étiquette (label) de la classe majoritaire à la donnée qui était inconnue.

KNN

Algorithme

Etape 5: Prendre les valeurs des observations retenues:

- Si on effectue une **régression**: l'algorithme calcule la moyenne (ou la médiane) des valeurs des observations retenues.
- Si on effectue une **classification**, l'algorithme assigne une étiquette (label) de la classe majoritaire à la donnée qui était inconnue.

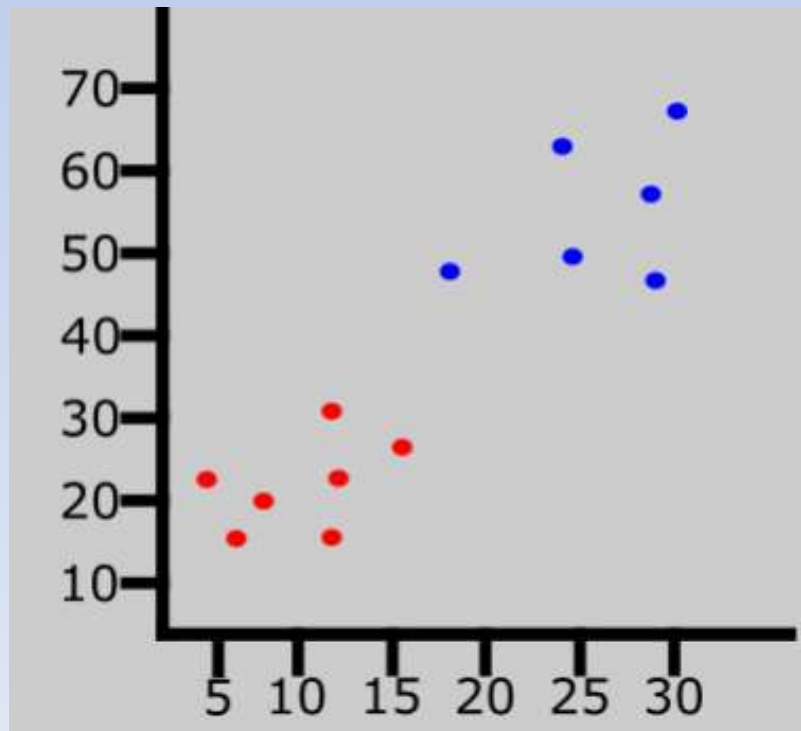
KNN

Algorithme

Etape 1: Charger les données.

Un ensemble de données regroupées en deux groupes: rouge et bleu.

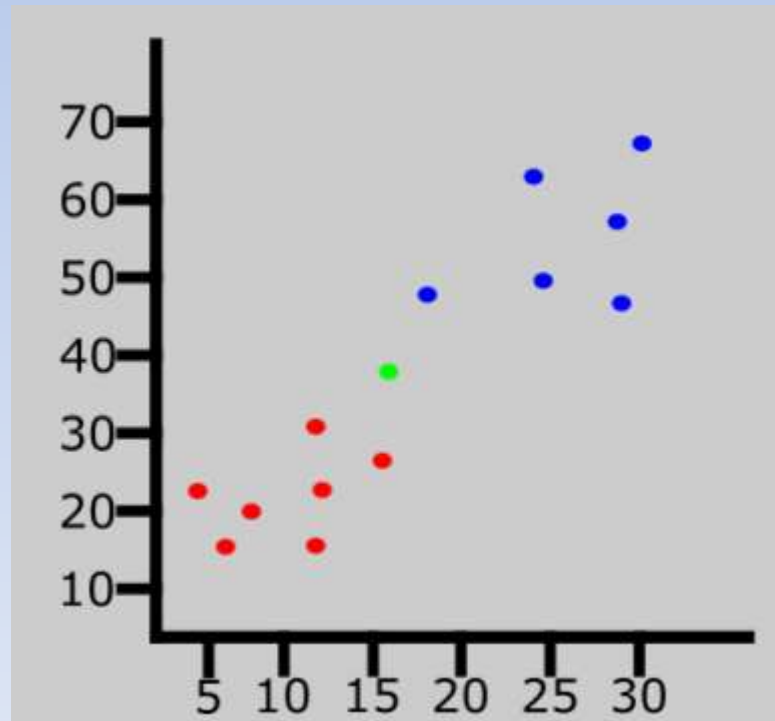
Etape 2: $k=3$.



KNN

Algorithme

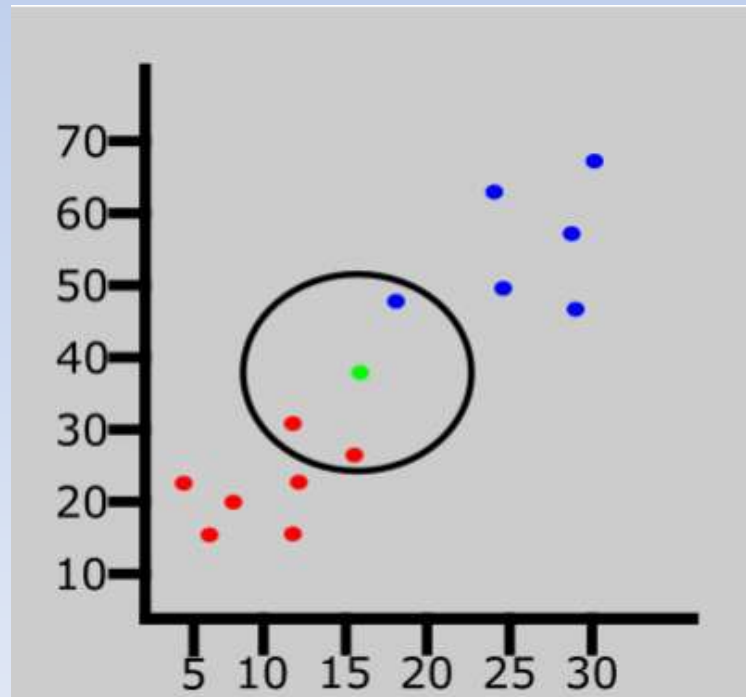
But: assigner le point **vert** à un des deux groupes.



KNN

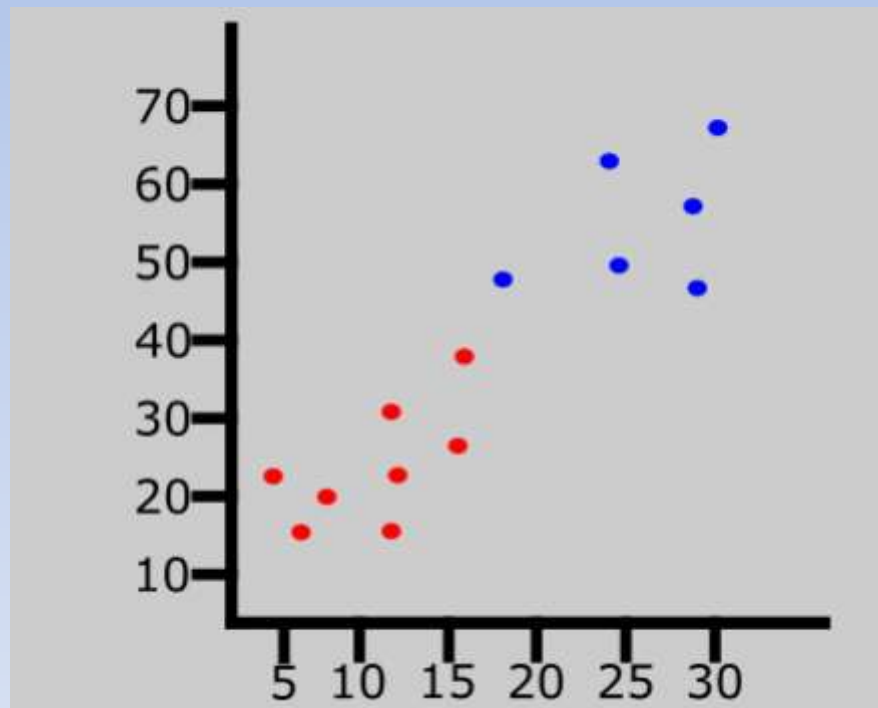
Algorithme

Etape 4: Conserver les **k=3** observations du jeu de données qui sont les plus « proches » de l'observation représentée par le point vert.



KNN

Algorithmme



KNN

Algorithme

Remarques

KNN

Algorithme

Pour cet algorithme,

- Le choix du nombre **k**.
- et
- Le choix de la **fonction de similarité**

Sont des étapes qui peuvent conduire à une forte variabilité des résultats.

KNN

Algorithme

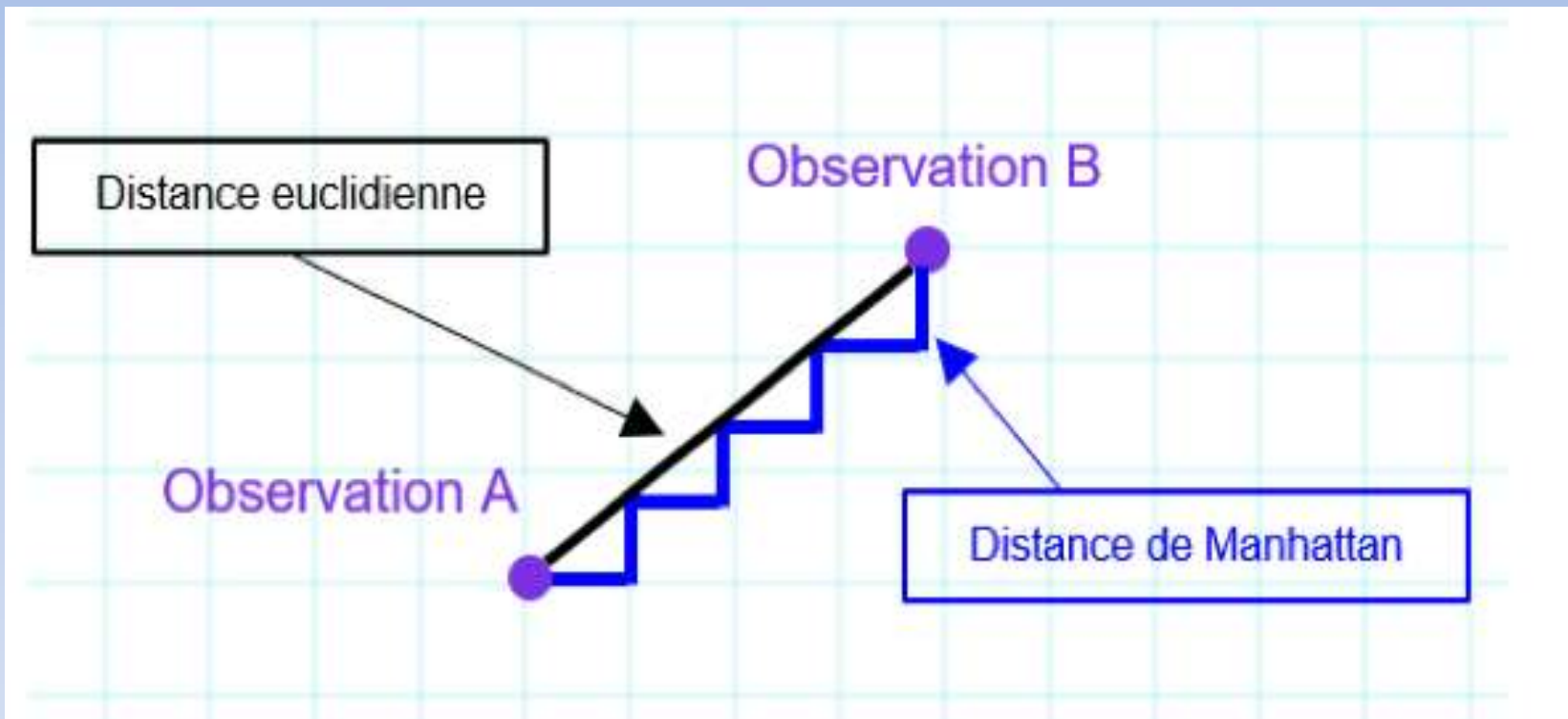
Pour cet algorithme,

- Le choix du nombre **k**.
- et
- Le choix de la **fonction de similarité**

Sont des étapes qui peuvent conduire à une forte variabilité des résultats.

KNN

Distance



$$\sum_{i=1}^n |x_i - y_i|$$

Manhattan

$$\sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Euclidienne

KNN

Avantages

KNN

Avantages

- L'algorithme est simple et facile à mettre en œuvre.
- Il n'est pas nécessaire de construire un modèle, d'ajuster plusieurs paramètres ou de faire des hypothèses supplémentaires.
- L'algorithme est polyvalent. Il peut être utilisé pour la classification, la régression et la recherche d'informations.

KNN

Avantages

- L'algorithme est simple et facile à mettre en œuvre.
- Il n'est pas nécessaire de construire un modèle, d'ajuster plusieurs paramètres ou de faire des hypothèses supplémentaires.
- L'algorithme est polyvalent. Il peut être utilisé pour la classification, la régression et la recherche d'informations.

KNN

Avantages

- L'algorithme est simple et facile à mettre en œuvre.
- Il n'est pas nécessaire de construire un modèle, d'ajuster plusieurs paramètres ou de faire des hypothèses supplémentaires.
- L'algorithme est **polyvalent**. Il peut être utilisé pour la **classification**, la **régression** et la recherche d'informations.

KNN

Inconvénients

KNN

Inconvénients

L'algorithme ralentit si:

- Le nombre d'observations et/ou de variables augmente.
- Parce que, **KNN** parcourt l'ensemble des observations pour calculer chaque distance.

KNN

Inconvénients

L'algorithme ralentit si:

- Le nombre d'observations et/ou de variables augmente.
- Parce que, **KNN** parcourt l'ensemble des observations pour calculer chaque distance.

KNN

Inconvénients

L'algorithme ralentit si:

- Le nombre d'observations et/ou de variables augmente.
- Parce que, **KNN** parcourt l'ensemble des observations pour calculer chaque distance.

Types d'apprentissage automatique

Types d'apprentissage automatique

Les méthodes d'apprentissage peuvent être classées en trois principales catégories:

**Apprentissage
supervisé**



**Apprentissage non-
supervisé**



Apprentissage non supervisé

- Aucun expert n'est disponible.
- L'algorithme doit découvrir par lui-même la structure des données.
- Le clustering est un algorithme d'apprentissage non supervisés.

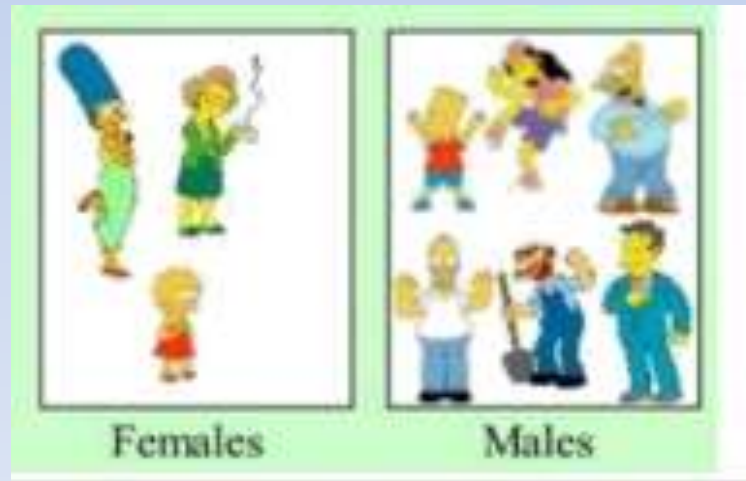
Apprentissage non supervisé

- Aucun expert n'est disponible.
- L'algorithme doit découvrir par lui-même la structure des données.
- Le clustering est un exemple d'application des algorithmes d'apprentissage non supervisés.

Apprentissage non supervisé

- Aucun expert n'est disponible.
- L'algorithme doit découvrir par lui-même la structure des données.
- Le clustering est un exemple d'application des algorithmes d'apprentissage non supervisés.

Exemple de Clustering



Exemple de Clustering

- Exemple de clustering



Apprentissage supervisé / non-supervisé

	Apprentissage supervisé	Apprentissage non-supervisé
Données d'entrée	Données connues en entrée	Données inconnues en entrée
Complexité informatique	Complexe	Moins complexe
Domaines d'activités	Classification et régression	Clustering
Précision	Produit des résultats précis	Génère des résultats modérés

Apprentissage supervisé / non-supervisé

	Apprentissage supervisé	Apprentissage non-supervisé
Données d'entrée	Données connues en entrée	Données inconnues en entrée
Complexité informatique	Complexe	Moins complexe
Domaines d'activités	Classification et régression	Clustering
Précision	Produit des résultats précis	Génère des résultats modérés

Apprentissage supervisé / non-supervisé

	Apprentissage supervisé	Apprentissage non-supervisé
Données d'entrée	Données connues en entrée	Données inconnues en entrée
Complexité informatique	Complexe	Moins complexe
Domaines d'activités	Classification et régression	Clustering
Précision	Produit des résultats précis	Génère des résultats modérés

Apprentissage supervisé / non-supervisé

	Apprentissage supervisé	Apprentissage non-supervisé
Données d'entrée	Données connues en entrée	Données inconnues en entrée
Complexité informatique	Complexe	Moins complexe
Domaines d'activités	Classification et régression	Clustering
Précision	Produit des résultats précis	Génère des résultats modérés

Examples

Exemples

Exemple 1:

Supposons que l'on dispose d'une collection d'articles de journaux.

Comment identifier des groupes d'articles portant sur un même sujet?

Exemples

- ***Exemple 1:*** discussion

On cherche à regrouper les articles portant sur un même sujet, sans disposer d'exemples d'articles dont on sait a priori qu'ils portent sur ce sujet, et sans connaître à l'avance les sujets à identifier.

On parlera donc de problème d'apprentissage non-supervisé.

Exemples

Exemple 2:

- Supposons que l'on dispose d'un certain nombre d'images représentant des chiens, et d'autres représentant des chats.
- Comment classer automatiquement une nouvelle image dans une des catégories « chien » ou « chat » ?

Exemples

- ***Exemple 3:***

Supposons que l'on dispose d'une base de données regroupant les caractéristiques de logements dans une ville :

superficie, quartier, étage, prix, année de construction, nombre d'occupants, montant des frais de chauffage.

Exemples

- *Exemple 3:*

Comment prédire la facture de chauffage à partir des autres caractéristiques pour un logement qui n'appartiendrait pas à cette base ?

Exemples

- **Exemple 2 et exemple 3:**

Dans les exemples 2 et 3, on cherche à prédire une caractéristique qui est soit une catégorie (exemple 2), soit un montant de facture (exemple 3), à partir d'exemples pour lesquels on connaît la valeur de cette caractéristique. Il s'agit de problèmes d'apprentissage supervisé.

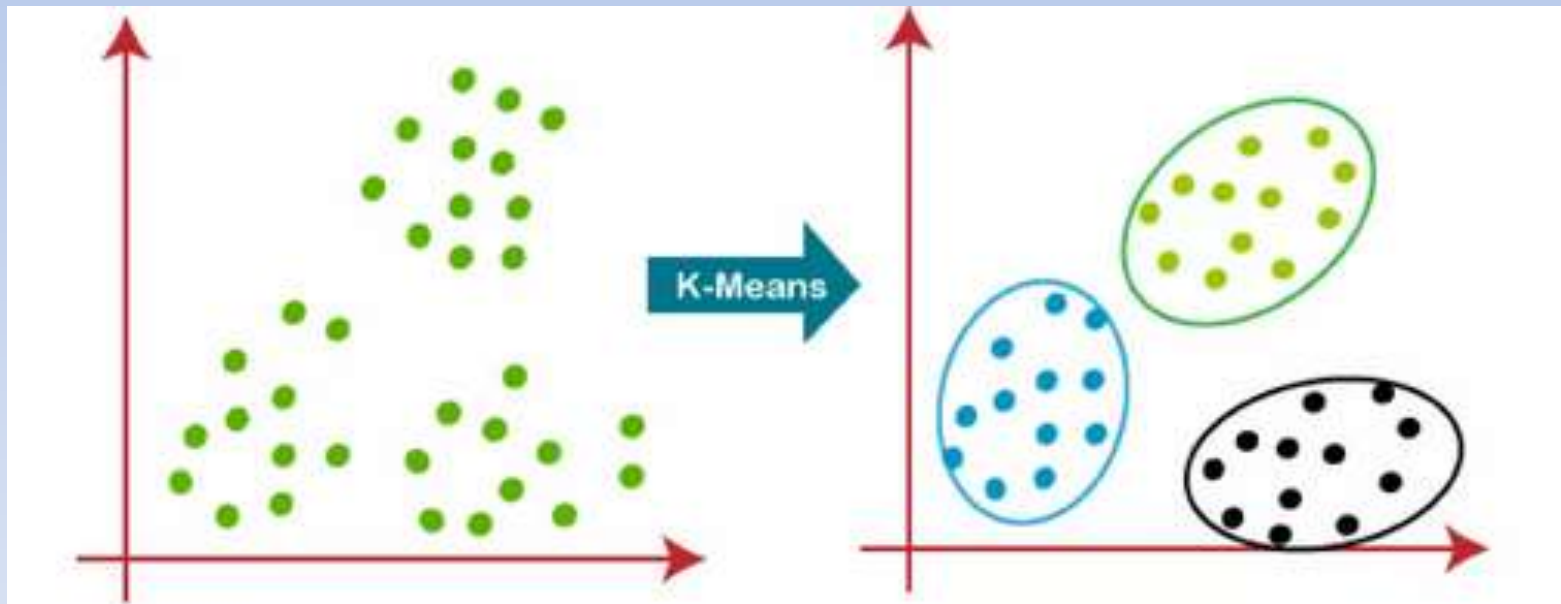
Algorithme de k-moyennes

Algorithme des centres
mobiles

K-means

K-means

- **But:** assigner les éléments aux groupes



Avant k-means

Après K-means

Etapes

K-means

K-means

Etape 1: initialisation des centroïdes

Choisir aléatoirement **K** points.

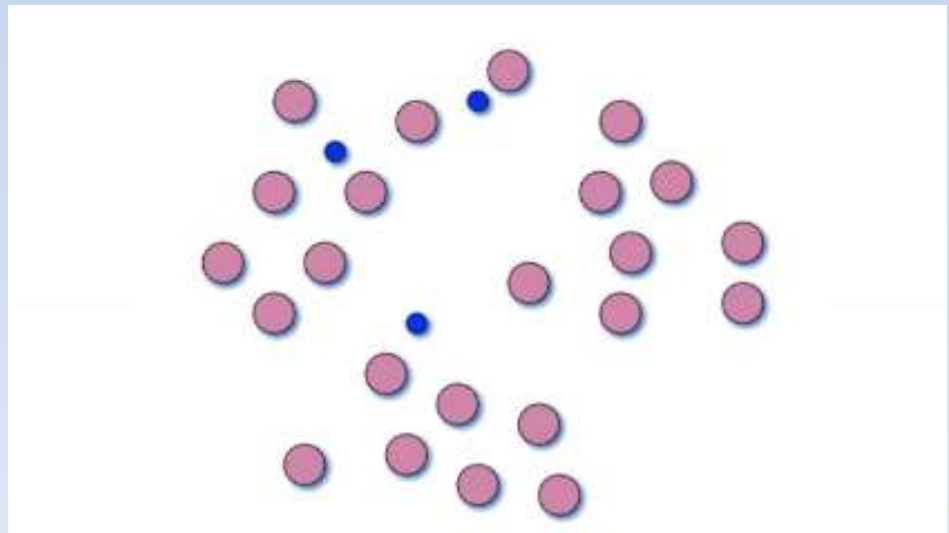
Ces points sont les centres des clusters initiaux (nommé centroïde).

K-means

Etape 1: initialisation des centroïdes

Choisir aléatoirement **K** points.

Ces points sont les centres des clusters initiaux (nommé centroïde).



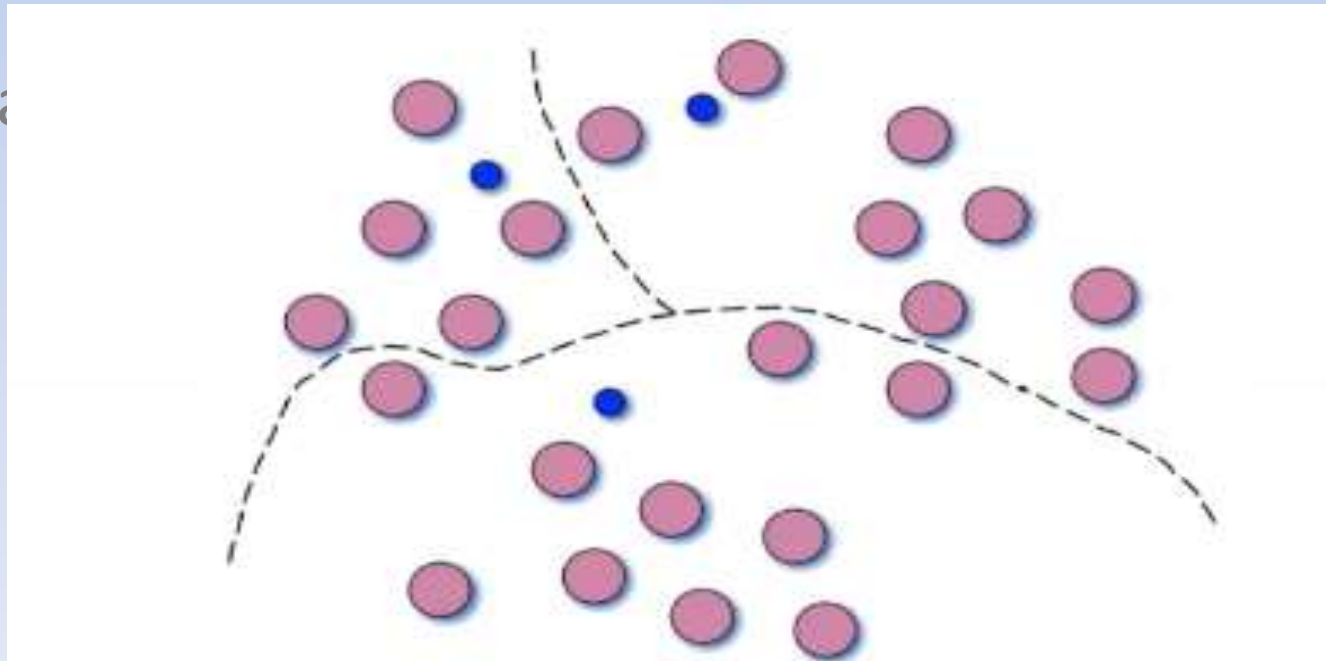
K-means

REPETER

- Affecter chaque point au cluster du centroïde le plus proche.

– Recalculer les centroïdes de chaque cluster.

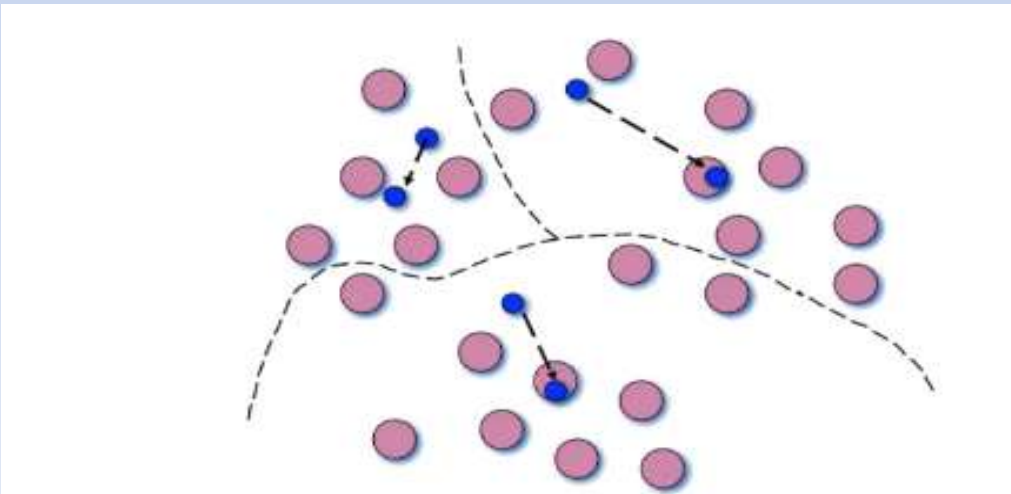
JUSQU'À



K-means

REPETER

- Affecter chaque point au cluster du centroïde le plus proche.
- Recalculer le centre de chaque cluster.

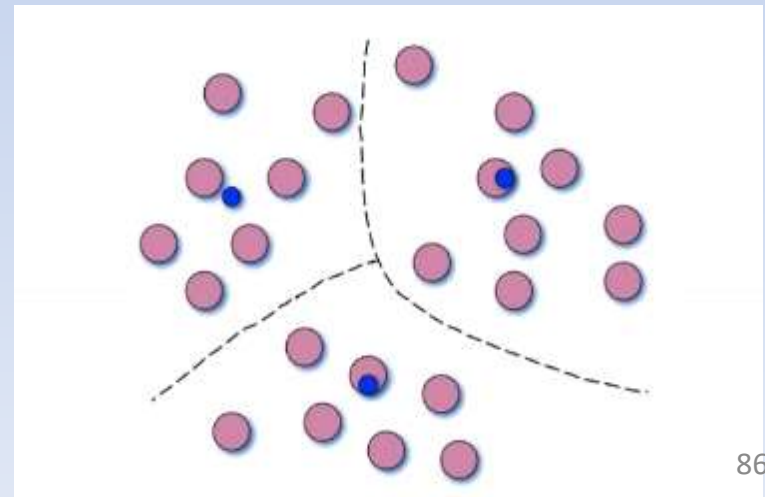


K-means

REPETER

- Affecter chaque point au cluster du centroïde le plus proche.
- Recalculer le centre de chaque cluster.

JUSQU'À CONVERGENCE



K-means

Remarque:

La convergence correspond au fait que les centroïdes ne changent pas après une mise à jour.

Attention : La convergence des centroïdes n'est pas garantie dans cet algorithme. Il faut en tenir compte dans lors de l'implémentation et ajouter une autre condition de sortie pour la boucle principale.

K-means

Remarque:

La convergence correspond au fait que les centroïdes ne changent pas après une mise à jour.

Attention : La convergence des centroïdes n'est pas garantie dans cet algorithme. Il faut en tenir compte dans lors de l'implémentation et ajouter une autre condition de sortie pour la boucle principale.

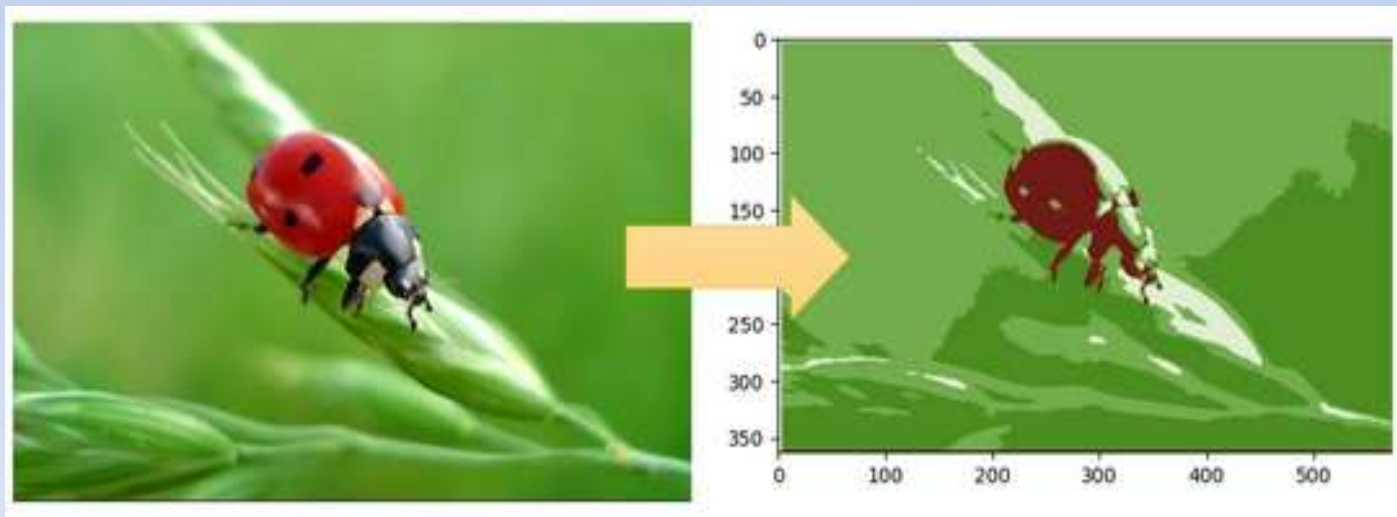
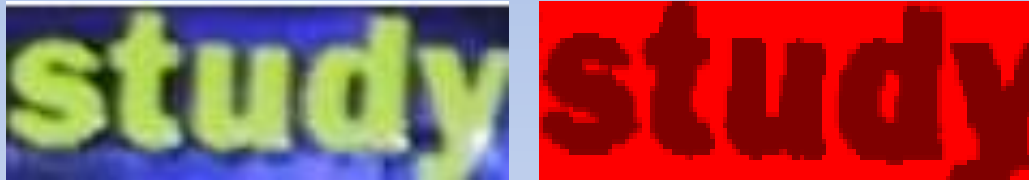
K-means

- K-means pour la segmentation d'images



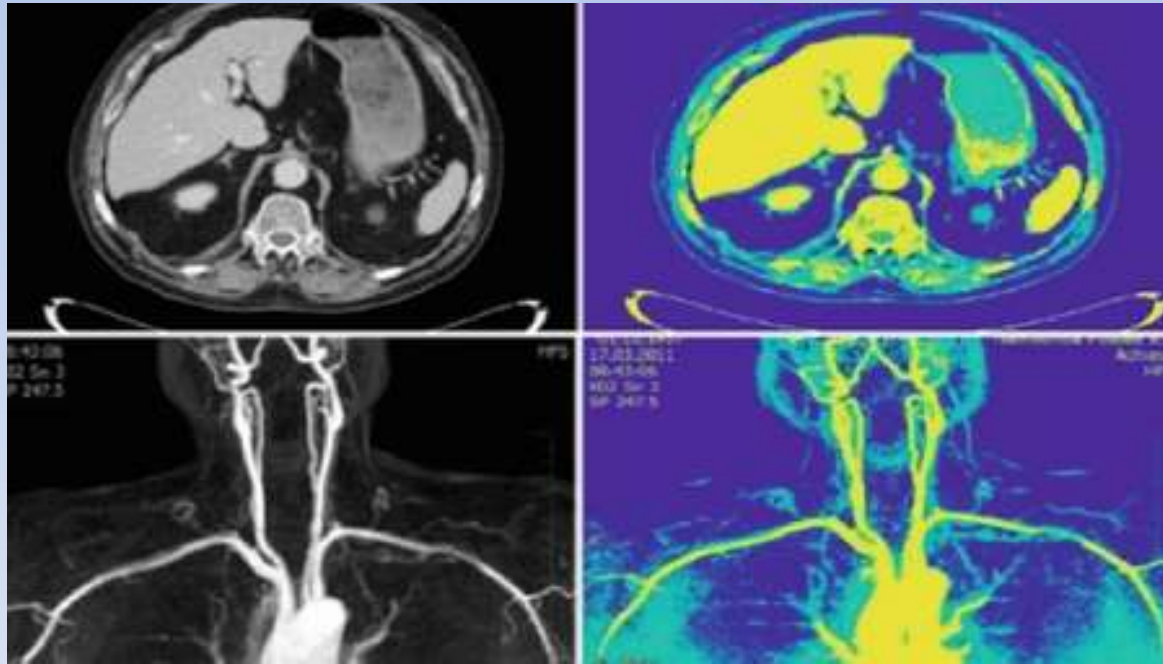
K-means

- K-means pour la segmentation d'images



K-means

- K-means pour la segmentation d'images médicales.



Avantages du k-means

Avantages du k-means

L'algorithme de k-means est:

- 1) Très facile à comprendre et à mettre en œuvre.
- 2) Simple et rapide.
- 3) Applicable à des données de grandes tailles, et aussi à tout type de données (même textuelles), en choisissant une bonne notion de distance.

Avantages du k-means

L'algorithme de k-means est:

- 1) Très facile à comprendre et à mettre en œuvre.
- 2) Simple et rapide.
- 3) Applicable à des données de grandes tailles, et aussi à tout type de données (mêmes textuelles), en choisissant une bonne notion de distance.

Avantages du k-means

L'algorithme de k-means est:

- 1) Très facile à comprendre et à mettre en œuvre.
- 2) Simple et rapide.
- 3) Applicable à des données de grandes tailles, et aussi à tout type de données (mêmes textuelles), en choisissant une bonne notion de distance.

Inconvénients du k-means

Inconvénients du k-means

- 1) Le nombre de cluster doit être fixé au départ.
- 2) Le résultat dépend de l'initialisation des centres des classes.
- 3) Les clusters sont construits par rapports à des objets inexistantes (les milieux)

Inconvénients du k-means

- 1) Le nombre de classe doit être fixé au départ.
- 2) Le résultat dépend de l'initialisation des centres des classes.
- 3) Les clusters sont construits par rapports à des objets inexistantes (les milieux)

Inconvénients du k-means

- 1) Le nombre de classe doit être fixé au départ.
- 2) Le résultat dépend de l'initialisation des centres des classes.
- 3) Les clusters sont construits par rapports à des objets inexistantes (les milieux)

Types d'apprentissage automatique

Les méthodes d'apprentissage peuvent être classées en trois principales catégories:

**Apprentissage
supervisé**



**Apprentissage non-
supervisé**



**Apprentissage par
renforcement**



Apprentissage par renforcement

L'apprentissage par renforcement
(reinforcement learning) est:

- Un processus
- Dans lequel un agent (robot,...)
- Apprend à prendre des décisions
- A partir d'expérimentations et d'erreurs.

Apprentissage par renforcement

L'apprentissage par renforcement
(reinforcement learning) est:

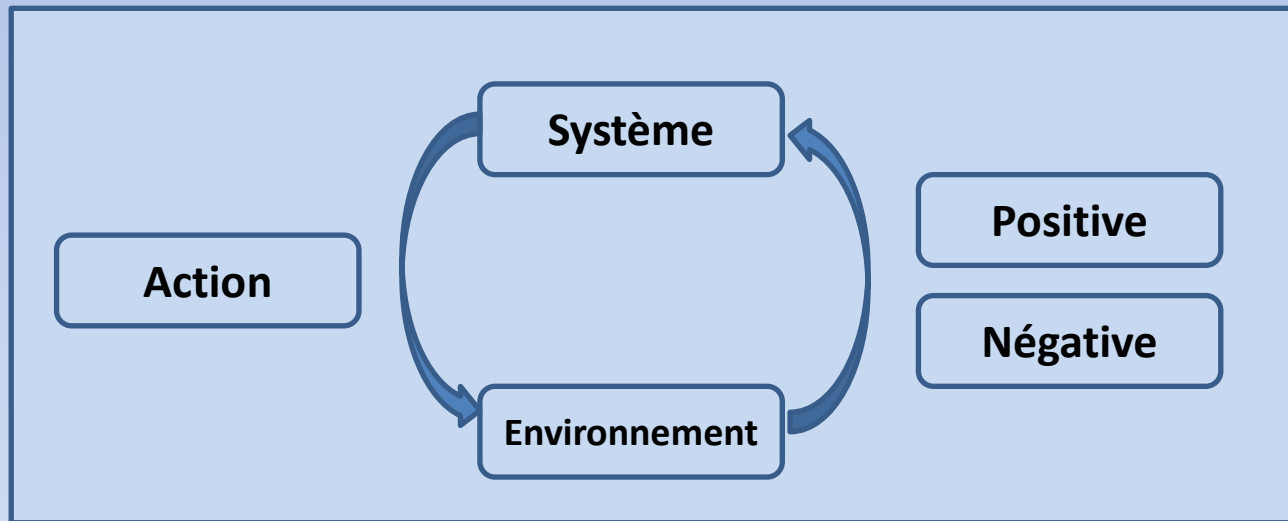
- Un processus
- Dans lequel un agent (robot,...)
- Apprend à prendre des décisions
- A partir d'expérimentations et d'erreurs.

Apprentissage par renforcement

L'apprentissage par renforcement est:

- Une technique de machine learning (ML) qui entraîne les logiciels à **prendre des décisions** en vue d'obtenir les **meilleurs résultats**.

Apprentissage par renforcement



Comparaison AS, AN-S, AR

	Apprentissage supervisé	Apprentissage non-supervisé	Apprentissage par renforcement
Définition	L'algorithme apprend à partir de données étiquetées	L'algorithme est entraîné à partir de données inconnues en entrée	L'algorithme interagit avec son environnement en apprenant de ses erreurs et succès
Types de problèmes	Classification et régression	Association et Clustering	Basé sur un système de récompense
Type de données	Données étiquetées	Données non étiquetées	Pas de données fournies au préalable
Approche	Etudie les relations sous-jacentes qui lient les données en entrée aux labels	Découvre les motifs communs au sein de données d'entrée	Apprend une stratégie de comportement en fonction d'expériences passées et des récompenses perçues.

Processus du ML/DL

Processus du ML

1. Collection des données: Il s'agit de regrouper les données d'un problème à résoudre.

=> La construction du Dataset.

Processus du ML

2. Prétraitement des données: Afin de rendre la Dataset utilisable à l'apprentissage, il faut le nettoyer:

- La suppression de données inutiles.
- La suppression de données répétées.
- La suppression de données incomplètes et manquantes.
- L'enrichissement par d'autres données, décomposition des données.

Processus du ML

2. Prétraitement des données: Afin de rendre la Dataset utilisable à l'apprentissage, il faut le nettoyer:

- La suppression de données inutiles.
- La suppression de données répétées.
- La suppression de données incomplètes et manquantes.
- L'enrichissement par d'autres données, décomposition des données.

Processus du ML

2. Prétraitement des données: Afin de rendre la Dataset utilisable à l'apprentissage, il faut le nettoyer:

- La suppression de données inutiles.
- La suppression de données répétées.
- La suppression de données incomplètes et manquantes.
- L'enrichissement par d'autres données, décomposition des données.

Processus du ML

2. Prétraitement des données: Afin de rendre la Dataset utilisable à l'apprentissage, il faut le nettoyer:

- La suppression de données inutiles.
- La suppression de données répétées.
- La suppression de données incomplètes et manquantes.
- L'enrichissement par d'autres données, décomposition des données.

Processus du ML

3. Choix du modèle:

Selon le problème traité, on peut choisir:

- La régression: s'il s'agit d'un problème de prédiction.
- Le clustering: pour les problèmes tels que la détection d'anomalies, la segmentation d'images, etc.
- Naïve bayes: s'il s'agit d'un problème de classification.

Processus du ML

3. Choix du modèle:

Selon le problème traité, on peut choisir:

- La régression: s'il s'agit d'un problème de prédiction.
- Le clustering: pour les problèmes tels que la détection d'anomalies, la segmentation d'images, etc.
- Naïve bayes: s'il s'agit d'un problème de classification.

Processus du ML

3. Choix du modèle:

Selon le problème traité, on peut choisir:

- La régression: s'il s'agit d'un problème de prédiction.
- Le clustering: pour les problèmes tels que la détection d'anomalies, la segmentation d'images, etc.
- Naïve bayes: s'il s'agit d'un problème de classification.

Processus du ML

4. Entrainement:

Les données de Dataset sont séparées en:

- 80% pour entraîner l'algorithme choisi.
- 20% pour tester et vérifier la performance du résultat.

Processus du ML

5. Evaluation: (l'étude des valeurs prédictives)

Elle permet:

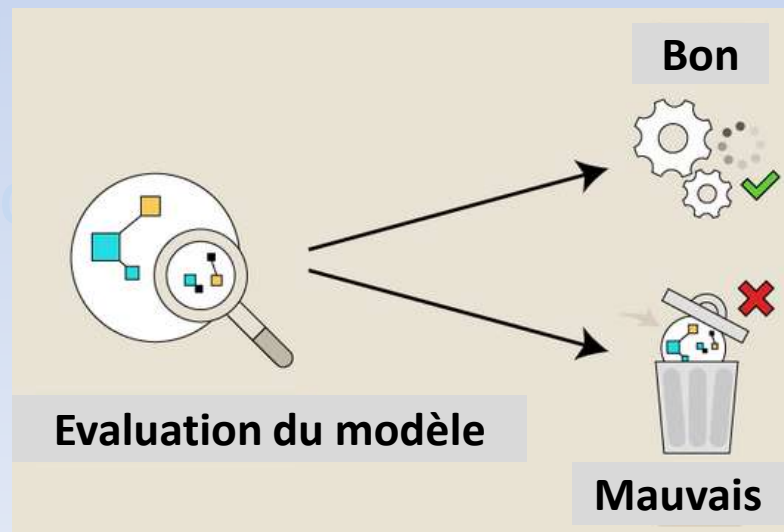
- De définir si le modèle du Machine Learning est fiable.
- Dans quels cas il commet des erreurs.
- Dans quelle mesure.

Métriques d'évaluation

Métriques d'évaluation

Une métrique est: une valeur numérique.

- Elle permet de quantifier la qualité des prédictions d'un modèle.



Métriques d'évaluation

Une métrique est: une valeur numérique.

- Elle permet de quantifier la qualité des prédictions d'un modèle.
- Elle permet de quantifier la performance du modèle et de déterminer s'il correspond à nos attentes.

Métriques d'évaluation

Apprentissage supervisé

Apprentissage non-supervisé

Métriques d'évaluation

Le **Clustering** permet de:

Regrouper des points de données comparables en fonction de caractéristiques spécifiques.

Il est essentiel **d'évaluer** la qualité des clusters construits lors de l'utilisation de techniques de clustering.

Métriques d'évaluation

Les **métriques du Clustering** jouent:

Un rôle essentiel dans l'évaluation de l'efficacité des algorithmes conçus pour regrouper des points de données similaires.

Métriques d'évaluation

Il existe **plusieurs** métriques de Clustering, on va présenter quelques unes.

- Le score de silhouette.
- Indice Davies-Bouldin.
- Indice de Dunn.

Métriques d'évaluation

La compréhension et l'application de ces mesures contribuent :

- L'affinement.
- A la sélection des algorithmes de Clustering.

Favorisant ainsi de meilleures connaissances dans les scénarios d'apprentissage non supervisé.

Métriques d'évaluation

Indice de silhouette

- Il est utilisée pour évaluer les clusters bien définis d'un ensemble de données.
- La cohésion et la séparation entre les clusters sont quantifiées.

Métriques d'évaluation

Indice de silhouette

- Les clusters mieux définis sont indiqués par des scores plus élevés, qui vont de -1 à 1.
- Un **objet est bien adapté** à son propre cluster et mal adapté aux clusters voisins si son score est proche de 1.
- Un score d'environ -1 suggère que l'objet pourrait se trouver dans le mauvais cluster.

Métriques d'évaluation

Indice de silhouette

$$S(je) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

$a(i)$: la distance moyenne entre une observation et les autres observation du même cluster.

$b(i)$: la distance moyenne entre l'observation et les observation du plus proche cluster

Métriques d'évaluation

Indice Davies-Bouldin

L'indice Davies-Bouldin (DBI):

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} \left\{ \frac{\Delta(x_i) + \Delta(x_j)}{\delta(x_i, x_j)} \right\}$$

k : le nombre de clusters

$\Delta(x_k)$: la distance intra-classe à l'intérieur du cluster x_k

$\delta(x_i, x_j)$: la distance entre les clusters x_i et x_j

Métriques d'évaluation

Indice Davies-Bouldin

Comment calculer DBI?

1. Calculer la distance moyenne entre les points du cluster et le centre du même cluster.
2. Calculer la distance entre le centroïde d'un cluster avec le centroïde du cluster le plus proche.
3. Calculer (résultat de 1/résultat de 2).
4. Répéter les étapes 1-3.
5. Calculer la moyenne de tous les rapports de tous les clusters.

Métriques d'évaluation

Indice Davies-Bouldin

Interprétation:

Le rapport sera d'autant plus **faible** que les clusters sont compactes et éloignées les unes des autres.

Par conséquent, la partition de meilleure qualité sera celle qui minimisera l'indice de Davies-Bouldin.

Métriques d'évaluation

Indice Davies-Bouldin

Avantage:

L'un des avantages de l'utilisation du DBI comme critère de Clustering est qu'il ne nécessite pas la connaissance du nombre réel de clusters ou des vraies étiquettes de clusters. .

Métriques d'évaluation

Indice Davies-Bouldin

Inconvénient

Il peut être sensible aux valeurs aberrantes et au bruit.

=> Ce qui entraîne une fausse indication d'un mauvais regroupement

Métriques d'évaluation

Indice Davies-Bouldin

Conseils pour l'utilisation de DBI

Il est recommandé de sélectionner une mesure de distance ou de similarité appropriée pour votre type de données et votre domaine, tel que Euclidienne, Manhattan, cosinus ou Jaccard.

Métriques d'évaluation

Indice de Dunn

Il est basé sur l'identification de clusters compacts.

$$D = \frac{d_{min}}{d_{max}}$$

d_{max} la distance maximale entre deux objets de la même classe.

d_{min}
la distance minimale entre deux objets de deux classes différentes.

Métriques d'évaluation

Indice de Dunn

Un bon Clustering est indiqué par des valeurs élevées de l'indice de Dunn.

=> Donc, le but est de maximiser l'indice de Dunn

Métriques d'évaluation

Apprentissage supervisé

Apprentissage non-supervisé

Métriques d'évaluation

Classification

régression

Métriques d'évaluation

Classification

Métriques d'évaluation

Matrice de confusion

Métriques d'évaluation

Pour les problèmes de classification:

- Les métriques consistent globalement à **comparer** les classes réelles aux classes prédites par le modèle.
- L'un des concepts clés de performance pour la classification est la matrice de confusion.

Métriques d'évaluation

Pour les problèmes de classification:

- Les métriques consistent globalement à **comparer** les classes réelles aux classes prédites par le modèle.
- L'un des concepts clés de performance pour la classification est la matrice de confusion.

Matrice de confusion

- Un outil qui permet de savoir à quel point le modèle de Machine Learning se trompe.

Il s'agit d'un tableau avec en colonne les différents cas réels et en ligne les différents cas d'usage prédits.

Métriques d'évaluation

Matrice de confusion

La matrice de confusion est:

- Une visualisation, sous forme de tableau, des prédictions du modèle par rapport aux vrais labels.

		Classes prédites			
		1	2	3	4
Classes réelles	1	52	3	7	2
	2	2	28	2	0
	3	5	2	25	12
	4	1	1	9	40

Métriques d'évaluation

Matrice de confusion

- Chaque ligne de la matrice de confusion représente les instances d'une classe réelles et chaque colonne représente les instances d'une classe prédite.

Classes réelles

Classes prédites

	1	2	3	4
1	52	3	7	2
2	2	28	2	0
3	5	2	25	12
4	1	1	9	40

Métriques d'évaluation

Matrice de confusion

Exemple:

- Une classification binaire, où l'on dispose de 100 instances positives et 70 instances négatives.

Classes prédites			
Classes réelles		Positif	Négatif
	Positif	90	10
	Négatif	13	57

La matrice de confusion

Métriques d'évaluation

Matrice de confusion

Elle permet d'obtenir une vue d'ensemble des prédictions justes et des prédictions fausses.

Le taux de bonnes prédictions ou *Accuracy*.

$$(90+57)/170 = 0.86$$

Classes prédites			
Classes réelles		Positif	Négatif
	Positif	90	10
	Négatif	13	57

Métriques d'évaluation

Matrice de confusion

Exemple:

Prenons l'exemple d'un test médical.

		REEL	
		<i>Si le patient est atteint ou non</i>	
		Est atteint	N'est pas atteint
PREDICTION <i>Ce que notre modèle prédisait</i>	Est atteint	Nombre de Vrai positif	Nombre de Faux positif
	N'est pas atteint	Nombre de Faux négatif	Nombre de Vrai négatif

Métriques d'évaluation

Matrice de confusion

On obtient donc les quatre valeurs suivantes :

- **Vrai positif (VP)**, les valeurs réelles et prédites sont identiques et positives. Le patient est malade et le modèle le prédit.

		REEL	
		<i>Si le patient est atteint ou non</i>	
		Est atteint	N'est pas atteint
PREDICTION <i>Ce que notre modèle prédisait</i>	Est atteint	Nombre de Vrai positif	Nombre de Faux positif
	N'est pas atteint	Nombre de Faux négatif	Nombre de Vrai négatif

Métriques d'évaluation

Matrice de confusion

On obtient donc les quatre valeurs suivantes :

- **Vrai négatif (VN)**, les valeurs **réelles** et **prédites** sont identiques et négatives. Le patient n'est pas malade et le modèle prédit qu'il ne l'est pas.

		REEL	
		<i>Si le patient est atteint ou non</i>	
		Est atteint	N'est pas atteint
PREDICTION <i>Ce que notre modèle prédisait</i>	Est atteint	Nombre de Vrai positif	Nombre de Faux positif
	N'est pas atteint	Nombre de Faux négatif	Nombre de Vrai négatif

Métriques d'évaluation

Matrice de confusion

On obtient donc les quatre valeurs suivantes :

- **Faux positif (FP)**, les valeurs **réelles et prédites** sont **différentes**. Le patient n'est pas malade, mais le modèle prédit qu'il l'est.

		REEL	
		<i>Si le patient est atteint ou non</i>	
		Est atteint	N'est pas atteint
PREDICTION <i>Ce que notre modèle prédisait</i>	Est atteint	Nombre de Vrai positif	Nombre de Faux positif
	N'est pas atteint	Nombre de Faux négatif	Nombre de Vrai négatif

Métriques d'évaluation

Matrice de confusion

On obtient donc les quatre valeurs suivantes :

- **Faux négatif (FN)**, les valeurs **réelles et prédites** sont **différentes**. Le patient est malade, mais le modèle prédit qu'il ne l'est pas.

		REEL	
		<i>Si le patient est atteint ou non</i>	
PREDICTION	<i>Ce que notre modèle prédisait</i>	Est atteint	N'est pas atteint
		Est atteint Nombre de Vrai positif	N'est pas atteint Nombre de Faux positif
	N'est pas atteint	Nombre de Faux négatif	Nombre de Vrai négatif

Processus du ML





Le vrai positif (VP)	Une sortie prédite qui appartienne à une classe et qui appartienne réellement à cette classe.
Le vrai négatif (VN)	Une sortie prédite qui n'appartienne pas à une classe et qui n'appartienne pas réellement à cette classe.
Le faux positif (FP)	Une sortie prédite qui appartienne à une classe et qui n'appartienne pas réellement à cette classe.
Le faux négatif (FN)	Une sortie prédite qui n'appartienne pas à une classe et qui appartienne réellement à cette classe.

Métriques d'évaluation

		<i>La classe prédite</i>		
		Positive	Négative	
<i>La classe réelle</i>	<i>Positive</i>	TP True Positive	FN False negative(Erreur)	<i>Sensitivity</i> TP/(TP+FN)
	<i>Négative</i>	FP False positive (Erreur)	TN True negative	<i>Specifity</i> TP/(TN+FP)
		<i>Précision</i> TP/(TP+FP)	<i>Negative Predictive Value</i> TN/(TN+FN)	<i>Accuracy</i> (TP+TN)/ (TP+FN+FP+TN)

Métriques d'évaluation

	Actually an Orange 106	Actually Not an Orange 60
Predicted Orange 115	 <p>True Positive 105</p>	 <p>False Positive 10</p>
Predicted Not Orange 51	 <p>False Negative 1</p>	 <p>True Negative 50</p>

		PREDICTIVE VALUES	
		POSITIVE (CAT)	NEGATIVE (DOG)
ACTUAL VALUES	POSITIVE (CAT)	<p>TRUE POSITIVE</p>  <p>3</p> <p>Actual: 3, Predicted: 3</p>	<p>FALSE NEGATIVE</p>  <p>1</p> <p>Actual: 4, Predicted: 3</p> <p>TYPE II ERROR</p>
	NEGATIVE (DOG)	<p>FALSE POSITIVE</p>  <p>2</p> <p>Actual: 1, Predicted: 3</p> <p>TYPE I ERROR</p>	<p>TRUE NEGATIVE</p>  <p>4</p> <p>Actual: 5, Predicted: 5</p>

Métriques d'évaluation

Accuracy

Accuracy:

- Elle mesure **l'exactitude globale** des prédictions d'un modèle.
- Elle calcule le **rapport** entre les échantillons **correctement classés** et le **nombre total** d'échantillons.

Métriques d'évaluation

Accuracy

- Elle calcule le **rapport** entre les échantillons **correctement classés** et le **nombre total** d'échantillons.

$$(TP+TN) / (TP+FN+FP+TN)$$

		La classe prédite		
		Positive	Négative	
La classe réelle	Positive	TP True Positive	FN False negative (Erreur)	Sensitivity TP/(TP+FN)
	Négative	FP False positive (Erreur)	TN True negative	Specifity TP/(TN+FP)
		Précision TP/(TP+FP)	Negative Predictive Value TN/(TN+FN)	Accuracy (TP+TN)/ (TP+FN+FP+TN)

Métriques d'évaluation

Accuracy

- Cela ne fonctionne bien que s'il y a un nombre égal d'échantillons appartenant à chaque classe
- Elle peut ne pas convenir aux ensembles de données présentant des distributions de classes déséquilibrées.

Métriques d'évaluation

Accuracy

- Elle peut ne pas convenir aux ensembles de données présentant des distributions de classes déséquilibrées.
- Dans de tels cas, où une classe l'emporte largement sur l'autre, la précision peut être trompeuse.

Métriques d'évaluation

Accuracy

- Elle peut ne pas convenir aux ensembles de données présentant des distributions de classes déséquilibrées.
- Dans de tels cas, où une classe l'emporte largement sur l'autre, la précision peut être trompeuse.

Métriques d'évaluation

Précision

Elle ne prend pas en considération les vrais négatifs

- Elle quantifie le rapport entre les **vrais positifs** et le nombre total de **prédictions positives**.

$$\text{TP}/(\text{TP}+\text{FP})$$

		La classe prédite		
		Positive	Négative	
La classe réelle	Positive	TP True Positive	FN False negative(Erreur)	<i>Sensitivity</i> TP/(TP+FN)
	Négative	FP False positive (Erreur)	TN True negative	<i>Specificity</i> TN/(TN+FP)
		<i>Précision</i> TP/(TP+FP)	<i>Negative Predictive Value</i> TN/(TN+FN)	<i>Accuracy</i> (TP+TN)/ (TP+FN+FP+TN)

Métriques d'évaluation

Précision

- Elle calcul le **taux d'erreurs** du modèle quand il fait une **prédiction positive**.
- Plus ce taux est **élevé**, plus le modèle est **précis**.

Métriques d'évaluation

Précision

- Cependant, il ne prend pas en compte les faux négatifs, ce qui pourrait conduire à des résultats trompeurs.

$$TP/(TP+FP)$$

Métriques d'évaluation

Spécificité

- Elle mesure la proportion de **négatifs** correctement prédits par rapport au nombre total de **négatifs** réels. $TP/(TN+FP)$

		La classe prédite		
		Positive	Négative	
La classe réelle	Positive	TP True Positive	FN False negative (Erreur)	Sensitivity TP/(TP+FN)
	Négative	FP False positive (Erreur)	TN True negative	Specifity TP/(TN+FP)
		Précision TP/(TP+FP)	Negative Predictive Value TN/(TN+FN)	Accuracy (TP+TN)/ (TP+FN+FP+TN)

Métriques d'évaluation

Sensibilité (Recall)

- Appelée aussi: Recall (Rappel). Elle quantifie la proportion de **positifs correctement prédits** par rapport au nombre total de positifs réels. $TP/(TP+FN)$

		La classe prédite		
		Positive	Négative	
La classe réelle	Positive	TP True Positive	FN False Negative (Erreur)	Sensitivity $TP/(TP+FN)$
	Négative	FP False positive (Erreur)	TN True Negative	Specificity $TP/(TN+FP)$
		Précision $TP/(TP+FP)$	Negative Predictive Value $TN/(TN+FN)$	Accuracy $(TP+TN)/(TP+FN+FP+TN)$

Métriques d'évaluation

Spécificité et sensibilité

- Des mesures pour les ensembles de données déséquilibrés.
- Elles offrent des informations supplémentaires sur les performances d'un modèle.

Métriques d'évaluation

Spécificité et sensibilité

- Les ensembles de données déséquilibrés, dans lesquels une classe est nettement plus nombreuse que l'autre, posent des défis pour les mesures d'évaluation telles que l'exactitude, la précision et le rappel.

Métriques d'évaluation

Spécificité et sensibilité

- Utiliser les métriques est essentiel pour **évaluer les performances d'un modèle de Machine Learning.**
- Choisir la métrique correcte selon le modèle permet de prendre les bonnes décisions quant à la manière de l'améliorer.

Métriques d'évaluation

Spécificité et sensibilité

- Utiliser les métriques est essentiel pour évaluer les performances d'un modèle de Machine Learning.
- Choisir la métrique correcte selon le modèle permet de prendre les bonnes décisions quant à la manière de l'améliorer.

Métriques d'évaluation

Spécificité et sensibilité

- En fonction du type de modèle (**modèle de classification ou de régression**), du contexte et du type des données, certaines métriques seront préférables à d'autres, et il est important de comprendre les avantages et les inconvénients de chaque métrique pour utiliser celle qui correspondra le mieux à votre problématique.

Métriques d'évaluation

F1-Score

- Le F1-Score est la moyenne des deux taux: Précision et Recall (sensibilité).

$$F1 - Score = 2x((Précision \times Recall) / (Précision + Recall))$$

- Cette métrique varie entre 0 et 1. Plus elle est proche de 1, meilleur est le modèle.

Métriques d'évaluation

Classification

régression

Métriques d'évaluation

Régression

Deux des principales métriques de régression :

1. L'erreur quadratique moyenne MSE.
2. L'erreur absolue moyenne **MAE** (Mean Absolute Error).
3. Erreur quadratique moyenne racine **RMSE** (Root Mean Squared Error)

Métriques d'évaluation

Régression

- L'erreur quadratique moyenne (**MSE**) est définie comme suit :

$$\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

où **N** est le nombre d'observations.

y_i est la valeur réelle.

\hat{y}_i est la prédiction réalisée.

Métriques d'évaluation

Régression

- L'erreur absolue moyenne (**MAE**) est définie comme suit :

$$\frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

où **N** est le nombre d'observations.

y_i est la valeur réelle.

\hat{y}_i est la prédiction réalisée.

Métriques d'évaluation

Régression

- L'erreur absolue moyenne est moins sensible aux grandes différences que l'erreur quadratique moyenne.
- Elle nous donne la mesure de la distance entre les prévisions et la sortie réelle.

Métriques d'évaluation

Régression

- Cependant, elle ne nous donne aucune idée de la direction de l'erreur, c'est-à-dire si nous sommes sous-prédits ou sur-prédits.

Métriques d'évaluation

Régression

L'erreur quadratique moyenne racine RMSE:

C'est une bonne mesure de la précision pour la comparaison entre les erreurs de prédiction de différents modèles.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^n (S_i - O_i)^2}$$

Merci pour votre attention