# Introduction à la classification

Présentée par:
**Zaynab EL KHATTABI**
04/12/2015

# Plan

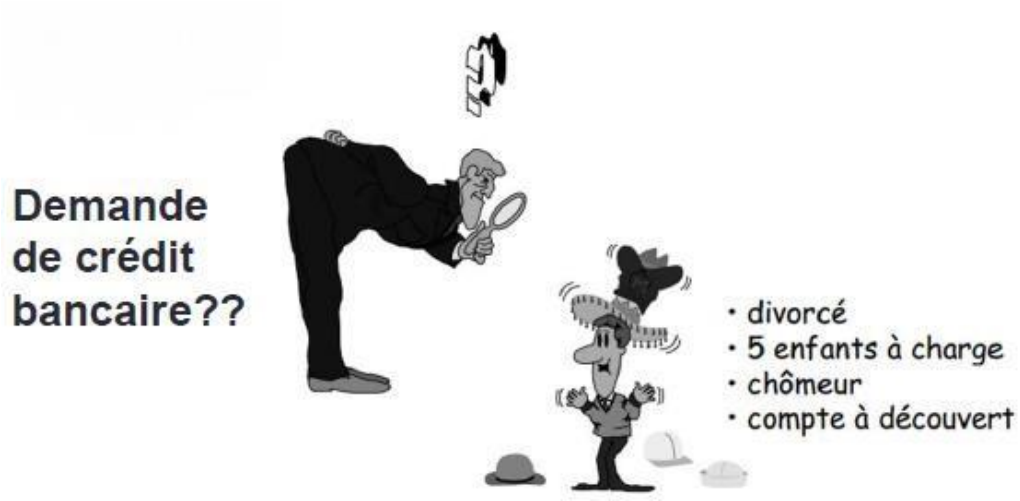
- But de la classification
- Apprentissage supervisé vs non supervisé
- Apprentissage supervisé
  - L'algorithme k-nearest neighbors
- Représentation des données pour la classification
- Apprentissage non supervisé (clustering)
  - L'algorithme k-means

# But de la classification

- Méthodes de l'Analyse des données
- Il ne suffit pas de collecter des montagnes d'informations, de les stocker dans des bases de données, mais il faut les **exploiter** c.-à-d. en **tirer des connaissances**.

- **Domaines d'application:**

  - Toutes les sciences et techniques qui font appel à la statistique multidimensionnelle.
  - Database marketing
  - Speech recognition
  - Detection de spam
  - …



# Apprentissage supervisé Vs.  non supervisé

- **Apprentissage Supervisé**
  - Essayer de prédire une information spécifique
  - Disposer d'un modèle (training set) avec des étiquettes(labels)
  - On peut mesurer la précision directement

- **Apprentissage non supervisé**
  - Essayer de comprendre les données
  - Chercher une structure  ou des modèles communs
  - N'exige pas des données avec des labels.
  - L'évaluation est indirecte

# Apprentissage supervisé Vs. non supervisé

**Apprentissage Supervisé :**

- Decision trees
- Random forest
- K-nearest neighbors
- SVM
- Neural Networks

**Apprentissage non supervisé:**

- K-means
- Expectation-maximization
- Hierarchical clustering
- Self organizing map
- Isodata

# Apprentissage supervisé

- Elle fait partie des techniques **prédictives**
- Les classes sont **prédéterminées**
- La classification consiste à examiner des caractéristiques d'un élément nouvellement présenté afin de l'affecter à une classe d'un ensemble prédéfini.
- Processus en deux étapes:

1. **Construction du modèle** à partir de l'ensemble d'apprentissage (**training data set**).
2. **Utilisation du modèle** pour la classification de nouvelles données

# Apprentissage supervisé

## Exemples de classification supervisée

- **Reconnaissance de caractère manuscrits**

$\in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$?

# Apprentissage supervisé

## Exemples de classification supervisée

- **Reconnaissance d'empreintes digitales:**

∈ {suspect 1, suspect 2, …, suspect  n}?

# Apprentissage supervisé

## l'algorithme k nearest neighbors KNN

- Un échantillon d'apprentissage dont on connait la classification
- **Objectif** : affecter une classe à une nouvelle instance
- **Données**: un échantillon de **n** enregistrements de dimension **p** :

$$X : \begin{array}{c|cccccc} & X^1 & \dots & X^j & \dots & X^p \\ X_1 & x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \dots & \dots & \dots & \dots & \dots & \dots \\ X_i & x_i^1 & \dots & x_i^j & \dots & x_i^p \\ \dots & \dots & \dots & \dots & \dots & \dots \\ X_n & x_n^1 & \dots & x_n^j & \dots & x_n^p \end{array}$$

- À chaque enregistrement est associée une classe connue à l'avance:
  (Xi, C(Xi))

✓ **Fonction de distance** (pour déterminer les voisins)

✓ **Fonction de choix** de la classe qui dépend de celles des voisins les plus proches.

# Apprentissage supervisé

## l'algorithme k nearest neighbors KNN

- **Input:** **K** le nombre des NN**,** Un enregistrement **Xt**

- **Traitement**:

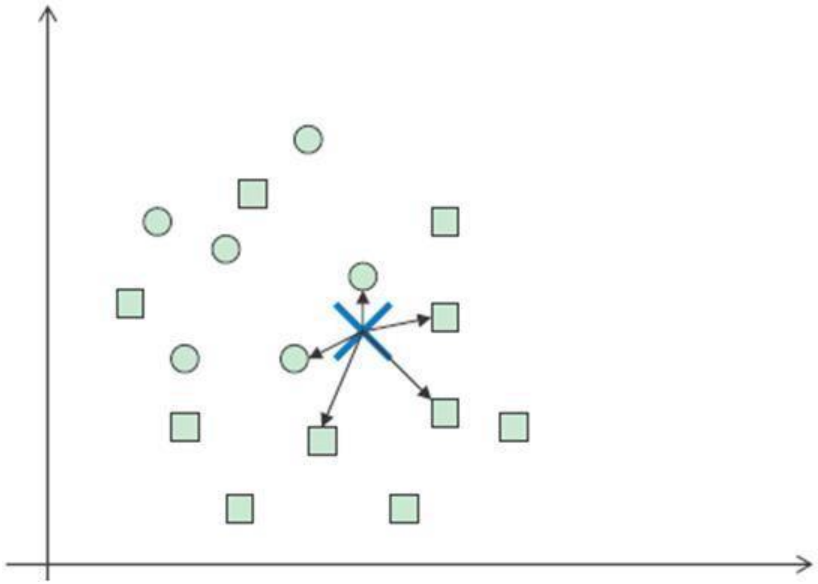
1. Déterminer les **K** plus proches enregistrements (voisins) de Xt
2. On conserve les **K enregistrements les plus proches de Xt**
3. Déterminer la classe **Ct** de l'enregistrement **Xt** en fonction de celles des **K voisins.**

- **sortie**: la classe de Xt est **C(Xt)= Ct**

# Apprentissage supervisé

## l'algorithme k nearest neighbors KNN

- **Exemple:**



- Si k=3, le nouveau point sera associé à la classe des **cercles**
- Si k=5, le nouveau point sera affecter à la classe des **carrés**

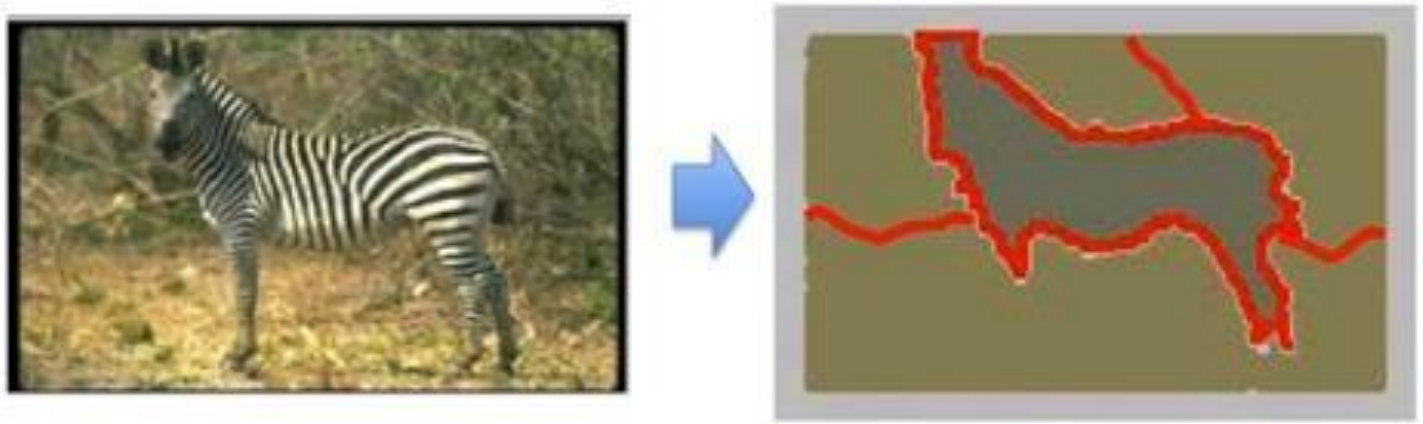
# Représentation des données avant la classification

- **Exemple de reconnaissance de caractères manuscrits:**


**=> Les Pixels peuvent être des attributs**

# Représentation des données avant la classification

- **Exemple de reconnaissance d'objets:**



**=> Les Pixels ne peuvent pas être des attributs!**

# Représentation des données avant la classification

- **Exemple de reconnaissance d'objets:**

- Segmentation de l'image en **regions (Blobs)**
  - Algorithmes : BlobWorld, Normalized cuts…

- Calculer les attributs décrivant la région
  - Convexité , orientation, fréquence de la couleur, texture…

**=> Un vecteur de caractéristiques (Xi) peut être un attribut**

# Apprentissage non supervisé

## Clustering

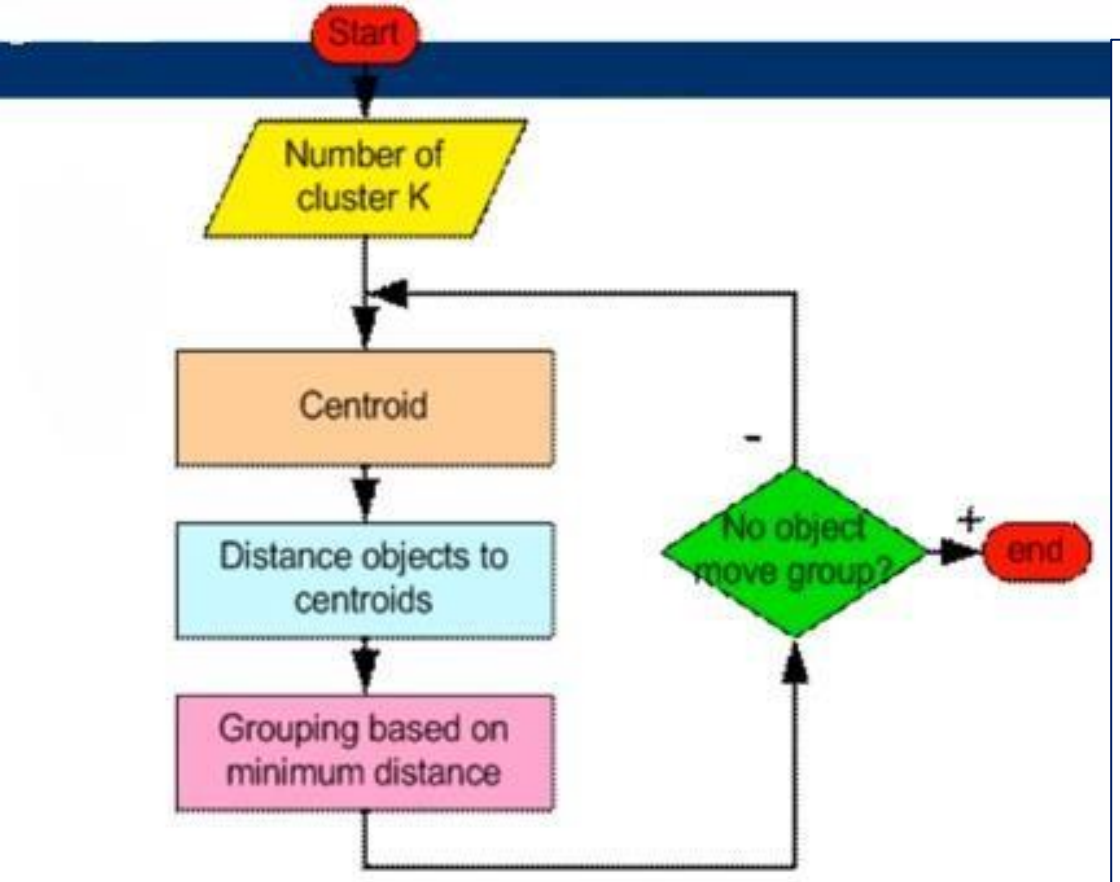
- Signifie **le regroupement des données** ou répartition d'un ensemble de données en **sous ensembles de même similarités** (attributs).

- Tâche non supervisée, **aucune prédiction spécifique**.

- Quels sont les sous-populations qui existent dans les données?
  - Leurs tailles?
  - Sont ils cohésives?
  - Les éléments d'une sous population partagent des propriétés communes?
  - Y a-t-il des aberrantes?
  - …

# Apprentissage non supervisé

## L'algorithme k-means




**Input:** **k, ensemble de points x1, x2,…,xi**

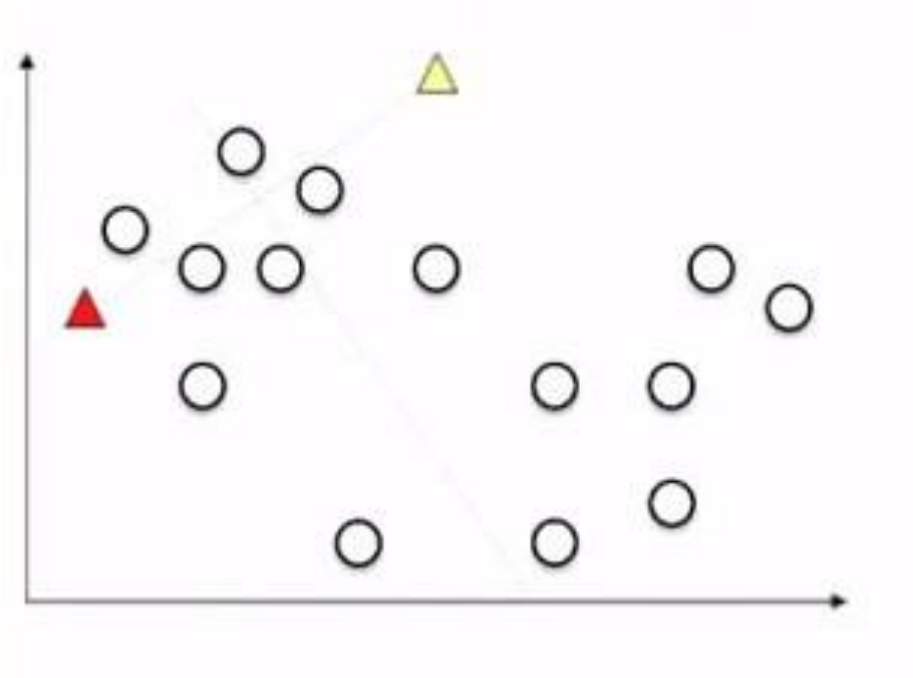
1. Placer les centres des clusters (centroids) à des locations aléatoires.
2. Pour chaque **Point (xi)**:
   - Trouver **le plus proche centroid cj**. (D(xi,cj) distance euclidien)
   - **Assigner le point xi au cluster j.**
3. Pour chaque **Cluster (j)**:
   - **Nouveau centroid cj=moyenne de tous les pts xi assignés au cluster j dans l'étape précédente.**
3. Répéter (2) jusqu'à aucun changement de centroid n'est possible.

**Sortie:** **Ensemble de clusters**

# Apprentissage non supervisé

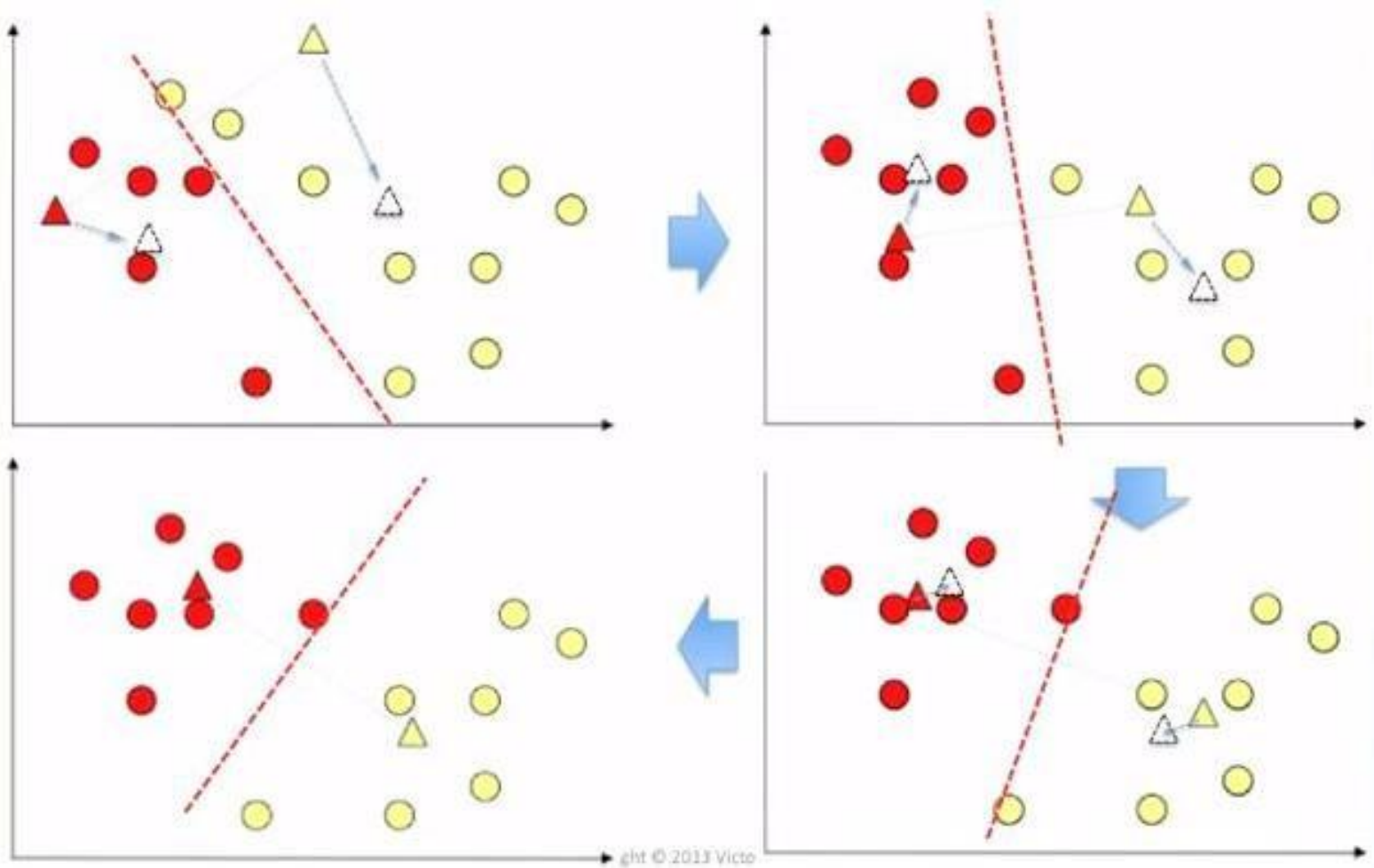
## L'algorithme k-means

- **Exemple:**



# Apprentissage non supervisé
## L'algorithme k-means

- **Exemple:**



# Discussion