Introduction

Principe :

Le but de la classification est d'obtenir une représentation simplifiée des données.

On cherche à regrouper des individus en classes homogènes

Questions :

- Comment les objets sont définis ?
- Notion de ressemblance entre objets ?
- Qu'est-ce qu'une classe ?
- Comment sont structurées les classes ?
- Juger une classification par rapport à une autre ?

Deux types de démarches :

- on regroupe en classes les objets qui partagent certaines caractéristiques,
- on regroupe en classes les objets qui possèdent des caractéristiques proches.

Deux types de méthodes :

- classifications non-hiérarchiques ou partitions,
- classifications hiérarchiques.

Partition.

Définition:

Soit Ω un ensemble fini de n individus, alors l'ensemble P = {P_1,P_2,...,P_g} de parties non vides de Ω est une partition si :

1. $(\forall k \neq l) P_k \cap P_l = \emptyset$
2. $\cup_{i=1}^{g} P_i = \Omega$

- Chaque élément appartient donc à une classe.
- On peut décrire P à l'aide de la matrice de classification C=(c_{ij}) (1≤i≤n et 1≤j≤g) où c_{ij}=1 si l'individu i appartient à la classe j.
- Si l'on accepte qu'un individu appartienne à plusieurs classes, c_{ij} peut être réel entre 0 et 1.

Objectif de la classification

Objectif :

organisation en classes homogènes de Ω.

Notion de classes homogènes.

On peut utiliser une mesure de similarité (ou dissimilarité). Par exemple imposer qu'un couple d'une classe soit plus proche que deux individus pris dans deux classes différentes.

Recherche de partitions

On utilise un critère numérique qui mesure l'homogénéité.

Problème : Chercher parmi l'ensemble fini des partitions celle qui optimise le mieux ce critère numérique.

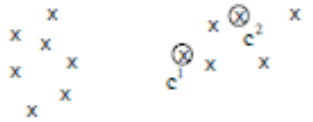
Difficultés : espace de recherche trop important, minimum locaux

Méthode des centres mobiles.

Ω dans $\mathbb{R}^p$ est muni de sa distance euclidienne *d*.

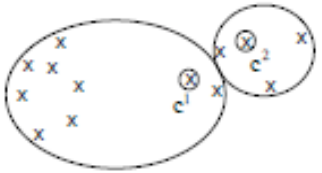
Les points à classer sont des points de Ω, tous d'égale importance (poids affectés aux points = 1)

Initialisation :
tirage au hasard de g points de Ω



while non convergence **do**

| On affecte chaque point à la classe dont il est le plus près du centre.
| Les centres de gravité de chaque classe deviennent les nouveaux centres.

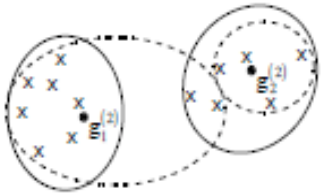


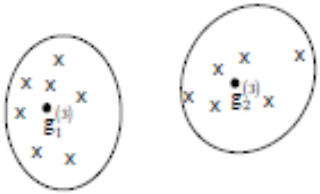
End While

La qualité d'un couple partition-centre est mesurée par la somme des inerties

des classes par rapport à leur centre.

Ce critère est amélioré à chaque étape de l'algorithme.





Généralisation : les nuées dynamiques.

Principe :

Remplacer les centres, élèments de $\mathbb{R}^p$, jouant le rôle de représentants (ou noyaux) de classe, par des éléments de nature diverse adaptés au problème.

Formalisation :

- $L=\{\lambda_i\}$ ensemble des noyaux,
- $D : \Omega \times L \rightarrow \mathbb{R}^+$ mesure de ressemblance entre les élèments de Ω et de L.
- Objectif : trouver la partition qui minimise le critère
- Cette minimisation est réalisée de façon alternée comme pour les centres mobiles.

Généralisation : les nuées dynamiques.

Choix du nombre de classes :

Problème :

Le critère dépend souvent du nombre de classes. Par exemple l'inertie s'annule pour une partition dont chaque point forme une classe, c'est donc la meilleur partition. Il faut donc fixer à priori le nombre de classes.

Solutions utilisées :

- on a une idée du nombre de classes désirées,
- on cherche la meilleur partition pour plusieurs nbr de classes et on étudie la décroissance du critère en fonction du nbr de classes (méthode du coude),
- on définit une fonction f(Ω) qui rend le critère indépendant du nbr de classes,
- on ajoute des contraintes (nbr d'individus par classe, volume de classe...) (méthode Isodata),
- on effectue des tests statistiques sur les classes.