Big Data Analytics

Clustering Hiérarchique

Clustering Hiérarchique

Deux approches :

- Agglomérantes (bottom-up)
- Divisantes (top-down)

Classification Ascendante Hiérarchique (CAH)

Métode Agglomérante

- Chaque élément est dans une classe distincte
- On itère jusqu'à ce qu'on ait le nombre de classes voulues
- On utilise une mesure de dissimilarité inter-classe comme critère d'aggrégation

A chaque itération, on calcule la dissimilarité entre toutes les classes puis on fusionne les plus similaires.

Classification Ascendante Hiérarchique (CAH)

Quelques distances de dissimilarités, après avoir défini une distance D dans l'espace :

- saut minimum : $dissim(C_1, C_2) = \min_{x \in C_1, y \in C_2} D(x, y)$
- saut maximum : $dissim(C_1, C_2) = \max_{x \in C_1, y \in C_2} D(x, y)$
- saut moyen : $dissim(C_1, C_2) = moyenne_{x \in C_1, y \in C_2} D(x, y)$
- distance de Ward qui vise à maximiser l'inertie inter-classe
- ...

$O(n^2) <$ complexité $< O(n^3)$!