Non-supervisée

On veut partitionner n observations en K groupes avec comme objectifs :

1. que les observations dans une même classe soient le plus similaire possible ;
2. que les observations dans des classes différentes soient les moins similaires possibles.

On veut donc définir une fonction, que l'on appelle classifieur, qui prend un numéro d'observation i en entrée et qui donne son numéro de groupe en sortie qui va remplir ces deux objectifs.

$$C : \{1, \dots, n\} \to \{1, \dots, K\} \tag{1}$$
$$i \mapsto C(i) \tag{2}$$

La fonction objectif est

$$W(C) = \sum_{k=1}^{K} \sum_{i:C(i)=k} \sum_{j:C(j)=k} d(x_i, x_j),$$

ou $d(x_i, x_j)$

Problème de grande taille !

Comme on ne peut pas explorer l'espace de toutes les possibilités, nous utiliserons des algorithmes gloutons (greedy algorithm), c'est-à-dire qu'ils vont nous donner une règle C qui minimise $W(C)$ sur un espace restreint et qui ne garantissent pas nous ayons trouvé un minimum global.

Dans notre contexte :

1. p variables sont numériques/ordinales (et habituellement standardisées).
2. La valeur de K est fixé.