Análise de Agrupamentos (Clusters)

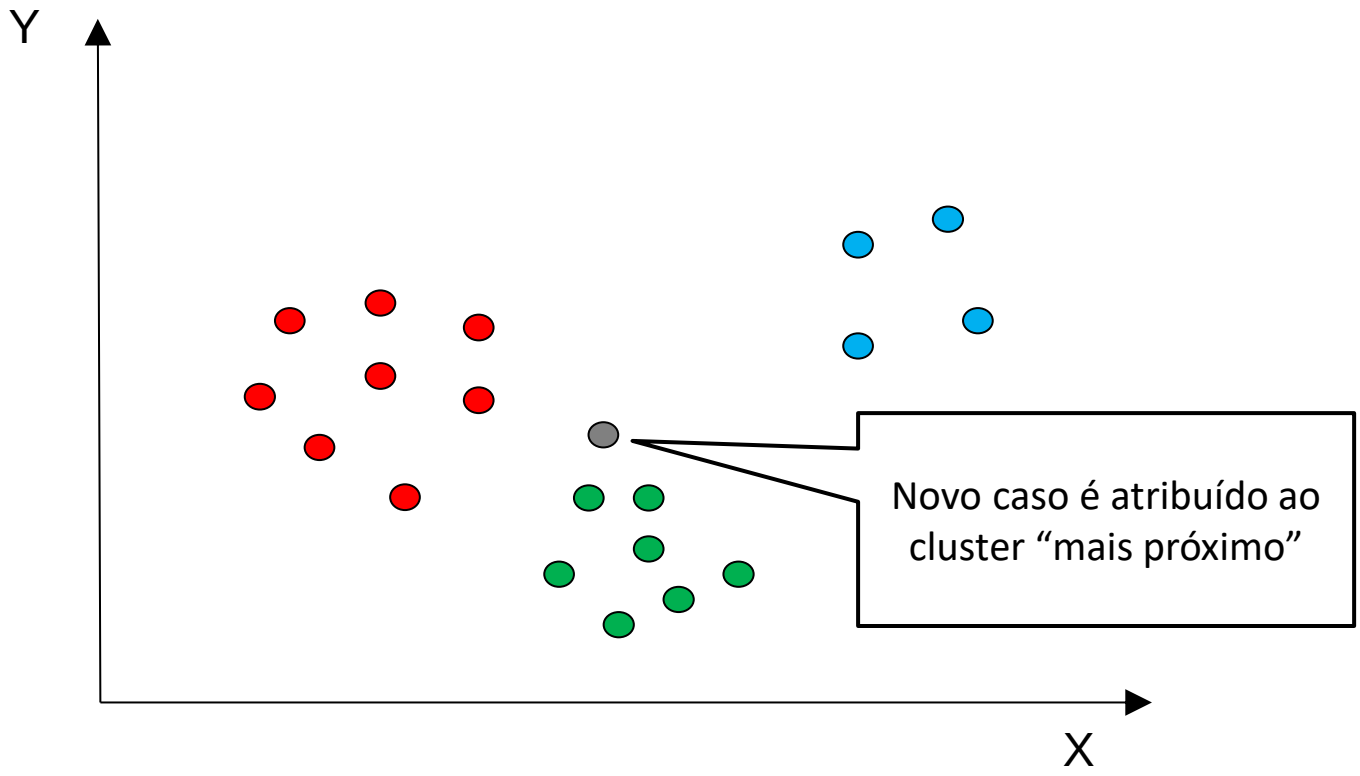
Marcelo Lauretto

Introdução

- Análise de Agrupamentos (Cluster Analysis) é um conjunto de técnicas com o objetivo principal de identificar objetos/entidades com características similares.

- Objetivo: Formação de grupos/classes de objetos com alta homogeneidade interna (intra-cluster) e alta heterogeneidade externa (inter-clusters).

- Em Inteligência Artificial, é comumente considerada como como uma abordagem de Aprendizado Não-supervisionado:
  - Novos casos são atribuídos ao cluster “mais próximo”


Y
X
Novo caso é atribuído ao cluster “mais próximo”

Algumas áreas de aplicação

- Psicologia:
  - Classificação de pessoas de acordo com seus perfis de personalidade
- Biologia:
  - Classificação de espécies
- Medicina:
  - Classificação de sub-tipos de doenças (diabetes, câncer, etc)
- Administração/Marketing
  - Segmentação de clientes de acordo com perfis de consumo

Métodos clássicos para agrupamentos

- Base dos métodos clássicos:
  - Medida de similaridade / dissimilaridade
  - Algoritmos de Agrupamento
  - Definição do número de clusters

Medidas de dissimilaridade

- Distância Euclidiana:

$$x_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,p}]'$$

$$d(x_1, x_2) = \sqrt{\sum_{j=1}^{p} (x_{1,j} - x_{2,j})^2} = \left([x_1 - x_2]'[x_1 - x_2]\right)^{1/2}$$

- Distância *city-block* ou *Manhattan*:

$$d(x_1, x_2) = \sum_{j=1}^{p} | x_{1,j} - x_{2,j} |$$

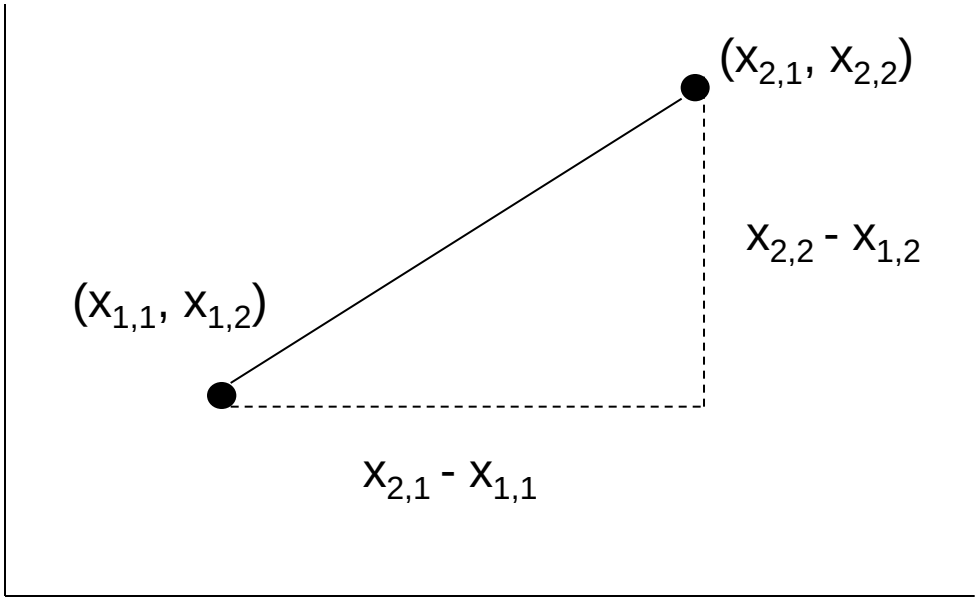
  - Essas medidas são sensíveis à diferença de escalas entre variáveis distintas:
    - Ex: IDH (0 a 100) e PIB (R$ bilhões)

- Distância de Mahalanobis:

$$d(x_1, x_2) = \left([x_1 - x_2]' S^{-1} [x_1 - x_2] \right)^{1/2}$$

onde S é matriz de covariância da amostra

– Distância padronizada: atenua o efeito da diferença de escalas



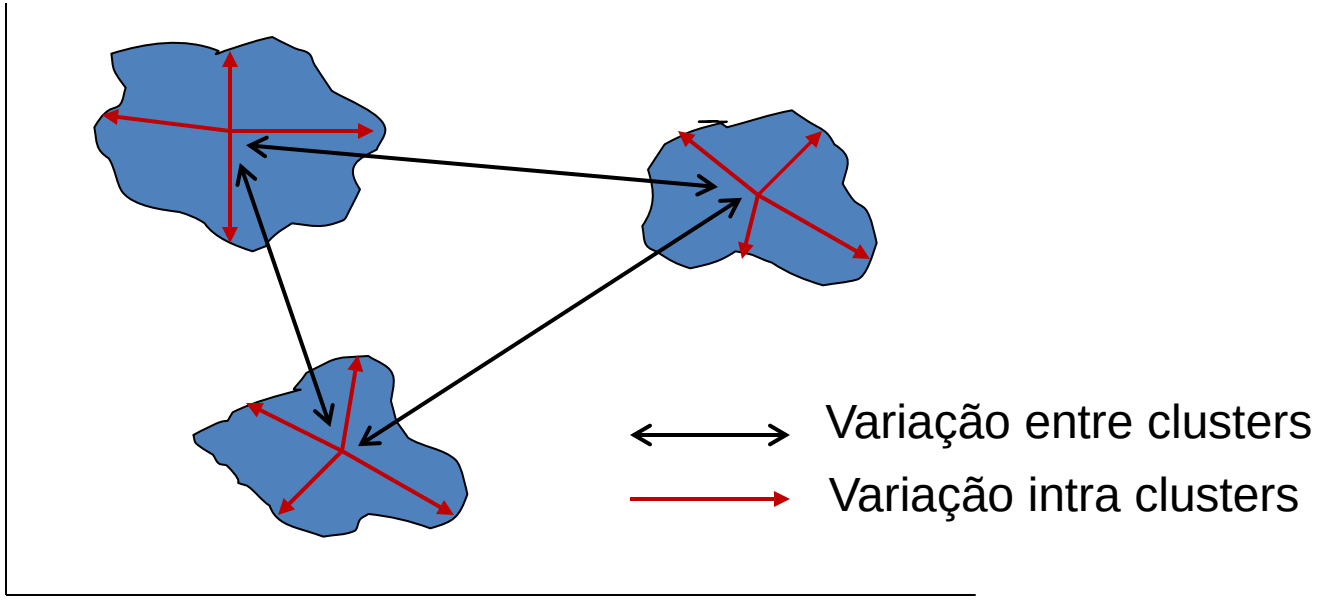
- Coeficiente de correlação linear:

$$\bar{x}_i = \frac{1}{p}\sum_{j=1}^{p} x_{i,j} \quad , \quad s_i = \sqrt{\frac{1}{p}\sum_{j=1}^{p}(x_{i,j} - \bar{x}_i)^2}$$

$$\rho(x_1, x_2) = \frac{1}{n}\frac{\sum_{j=1}^{p}(x_{1,j} - \bar{x}_1)(x_{2,j} - \bar{x}_2)}{s_1 \; s_2}$$

Algoritmos de agrupamento

- Princípio: Formação de clusters buscando-se:
  - maximizar diferenças entre clusters
  - minimizar variações intra-clusters

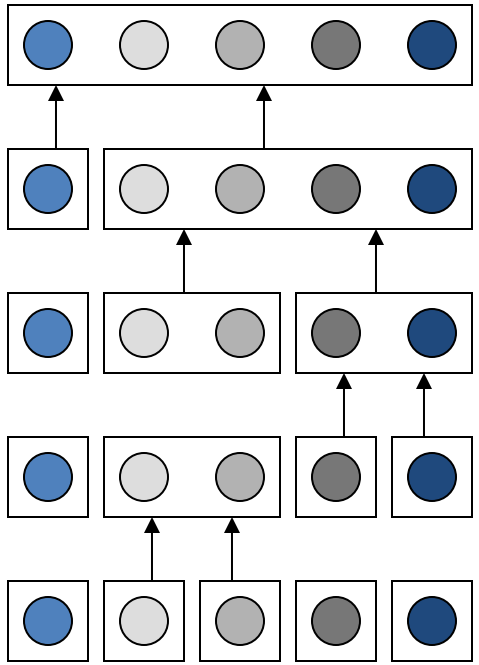


Classes de métodos

- Hierárquicos
  - Aglomerativos
  - Divisivos
- Não-hierárquicos
  - K-medias (K-means)
- Baseados em misturas de distribuições

Métodos aglomerativos

1. O processo começa com *n* clusters, cada um contendo uma observação.
2. A cada iteração, o par de clusters mais próximos entre si são combinados e passam a constituir um novo cluster.
3. O algoritmo pára quando há apenas um cluster contendo todas as observações.

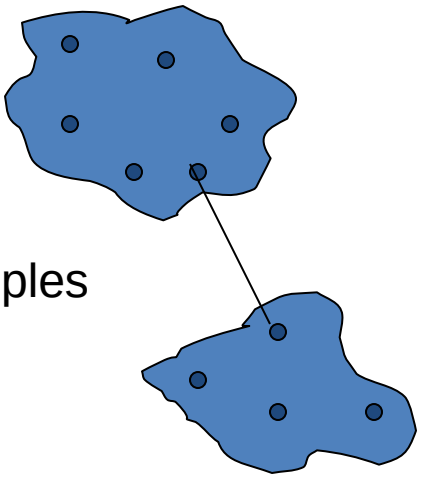


Métodos algomerativos mais comuns:

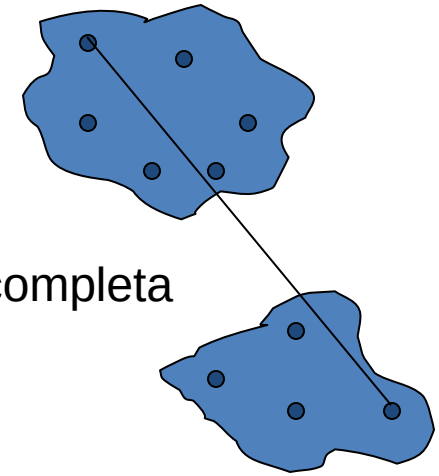
1. Método de ligação simples (Single linkage):
   - Medida de similaridade entre dois clusters é definida pela menor distância de qualquer ponto do 1º cluster para qualquer ponto do 2º cluster.
2. Método de ligação completa (Complete linkage):
   - Medida de similaridade entre dois clusters é definida pela maior distância de qualquer ponto do 1º cluster para qualquer ponto do 2º cluster.

3. Método da média das distâncias (Average linkage):
    - Medida de similaridade entre dois clusters é definida pela média das distâncias de todos os pontos do 1º cluster em relação aos pontos do 2º cluster.
4. Método do centróide (Centroid method):
    - Medida de similaridade entre dois clusters é definida pela distância entre os pontos médios do 1º e 2º clusters.

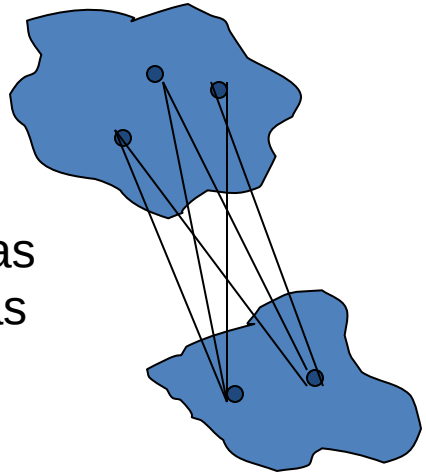
ligação simples



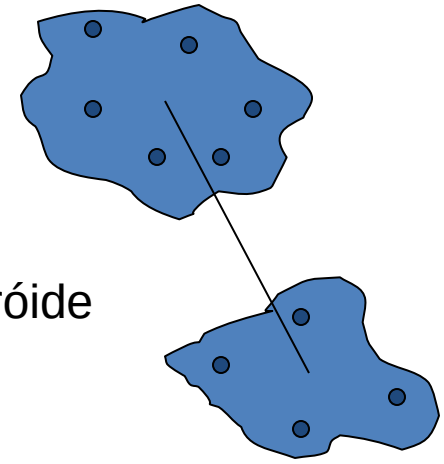
ligação completa



média das
distâncias



centróide



5. Método de Ward (Ward's method):

- Também denominado método da mínima variância.
- Medida de distância entre dois clusters é a soma das distâncias ao quadrado entre os dois clusters:

$x_{l,j,k}$: valor para a variável p na observação j pertencente ao cluster l

SS_l: soma dos erros quadrados dentro do cluster l

$$SS_l = \sum_{k=1}^{n_l}\sum_{j=1}^{p}\left(x_{l,k,j} - \bar{x}_{l,\bullet,j}\right)^2, \quad \bar{x}_{l,\bullet,j} = \frac{1}{n_l}\sum_{k=1}^{n_l} x_{l,k,j}$$

$SST_{l,i}$: soma total dos erros quadrados (agrupando os clusters l e i

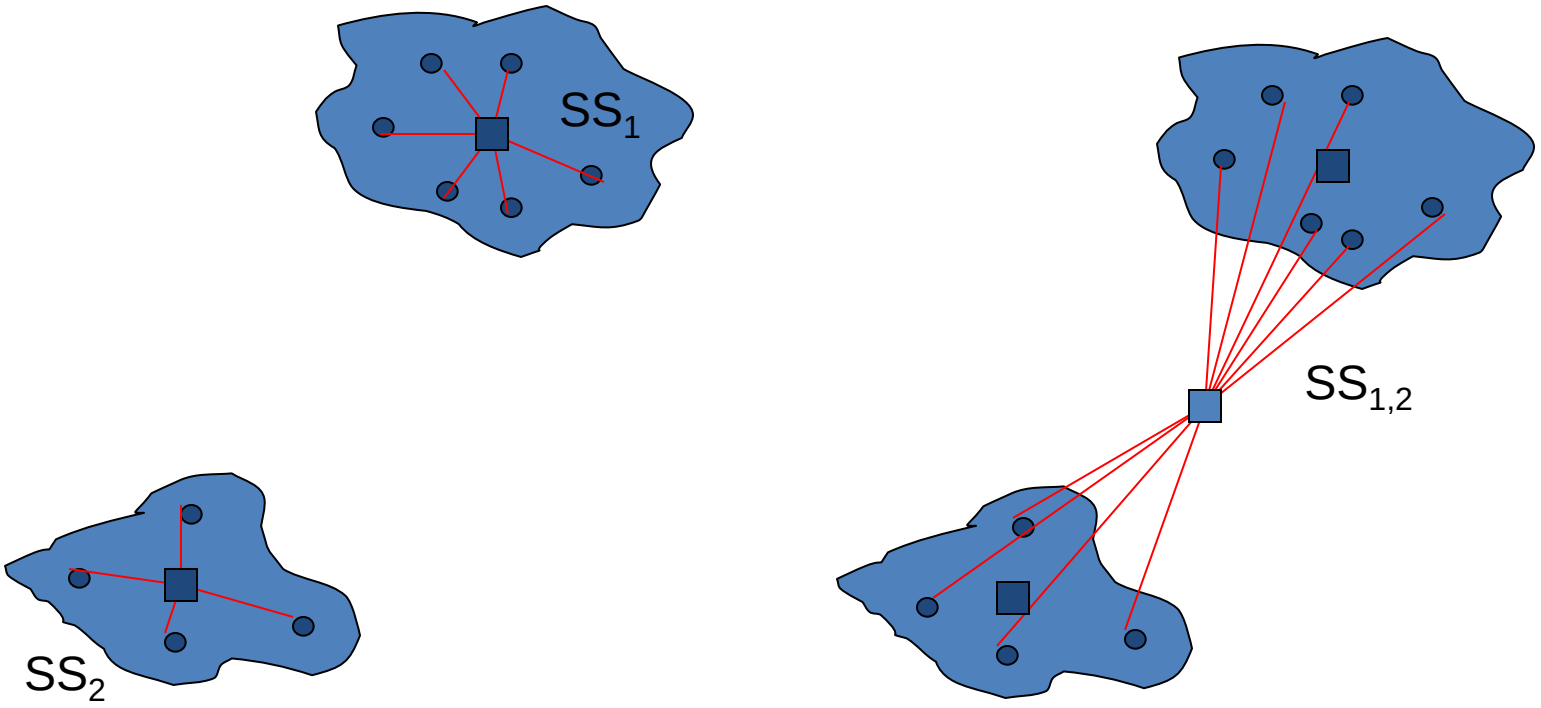
$$SS_{l,i} = \sum_{k=1}^{n_l}\sum_{j=1}^{p}\left(x_{l,k,j} - \bar{x}_j\right)^2 + \sum_{k=1}^{n_i}\sum_{j=1}^{p}\left(x_{i,k,j} - \bar{x}_j\right)^2,$$

$$\bar{x}_j = \frac{1}{n_l + n_i}\left(\sum_{k=1}^{n_l} x_{l,k,j} + \sum_{k=1}^{n_i} x_{i,k,j}\right)$$

$$d(C_l, C_i) = SS_{l,i} - (SS_l + SS_i)$$

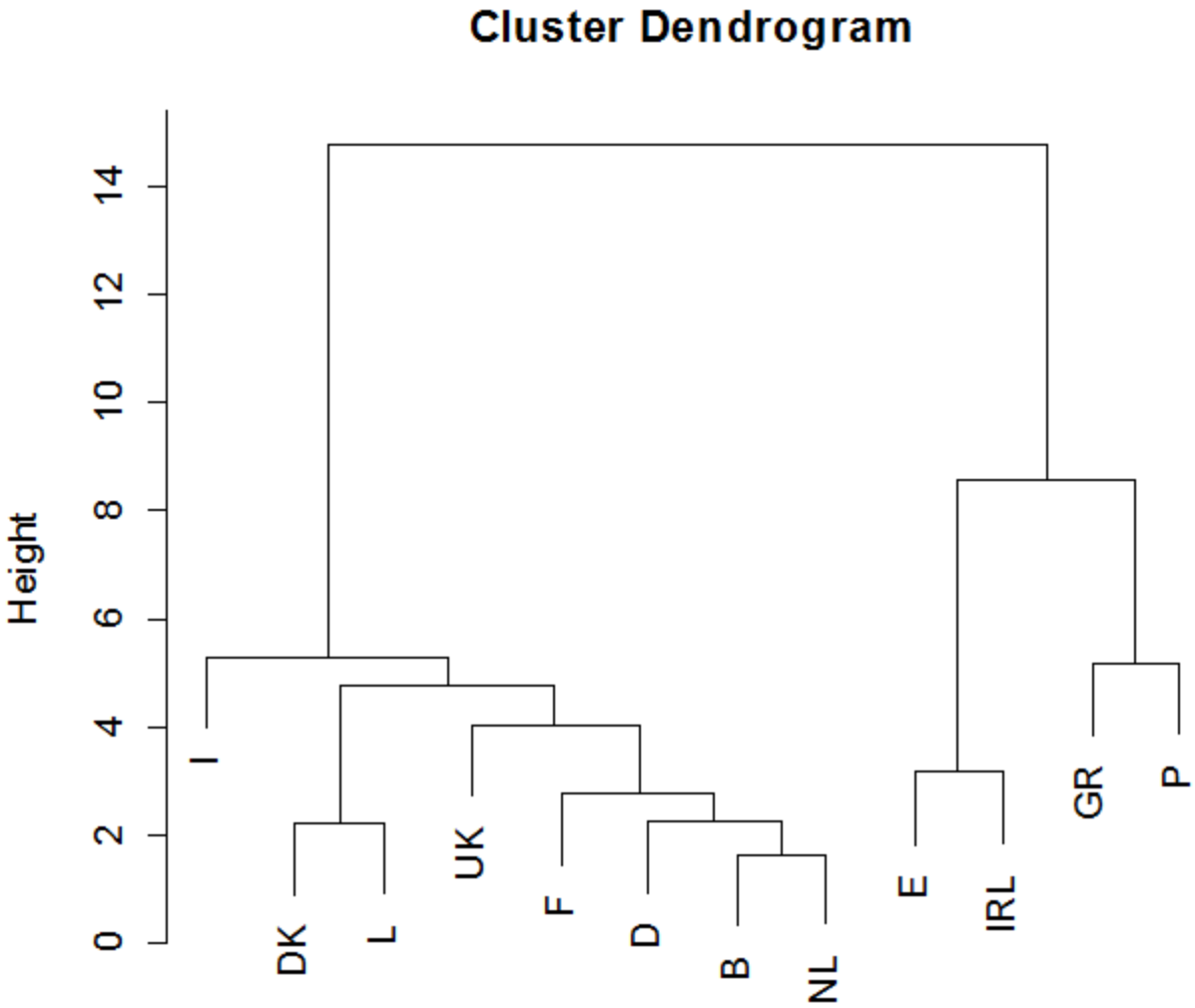
$$= \frac{n_l n_i}{n_l + n_i}\sum_{j=1}^{p}\left(\bar{x}_{l,\bullet,j} - \bar{x}_{i,\bullet,j}\right)^2$$

Método de Ward



$$d(C_l, C_i) = SS_{l,i} - (SS_l + SS_i)$$

- Variáveis:
  - PIB per capita
  - % população trabalhando na agricultura
- Países:
  - B (Belgium)
  - DK (Denmark)
  - D (Germany)
  - GR (Greece)
  - E (Spain)
  - F (France)
  - IRL (Ireland)
  - I (Italy)
  - L (Luxemburg)
  - NL (Netherlands)
  - P (Portugal)
  - UK (U.Kingdom)

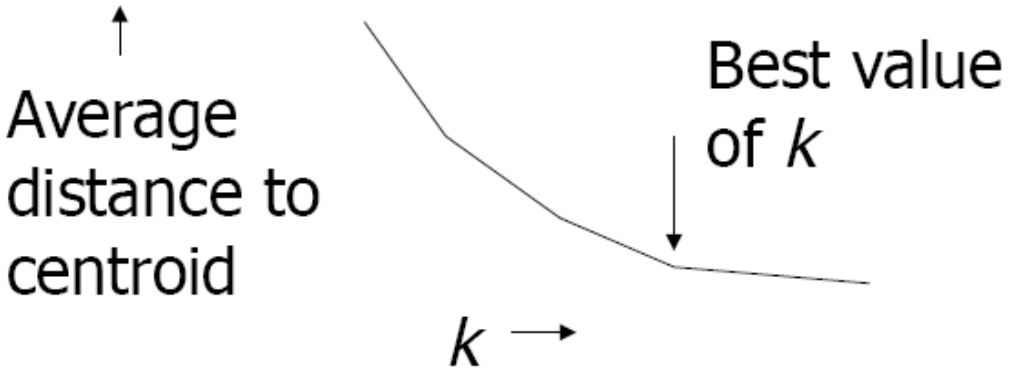

Cluster Dendrogram
Height
0
2
4
6
8
10
12
14
I
DK
L
UK
F
D
B
NL
E
IRL
GR
P

Método não-hierárquico: k-médias

1. Primeiramente escolhem-se *k* centróides, chamados de *sementes* ou *protótipos*, para se inicializar o processo de partição;
2. Cada elemento do conjunto de dados é comparado com cada centróide inicial através da distância desejada (usualmente Euclidiana). O elemento é alocado ao cluster de menor distância
3. Após aplicar o passo 2 para todos os n elementos amostrais, atualiza-se os valores dos centróides de todos os grupos formados, e repete-se o passo 2 considerando os centróides desses novos grupos.
4. Os passos 2 e 3 são repetidos até que nenhum dos elementos amostrais seja realocado.

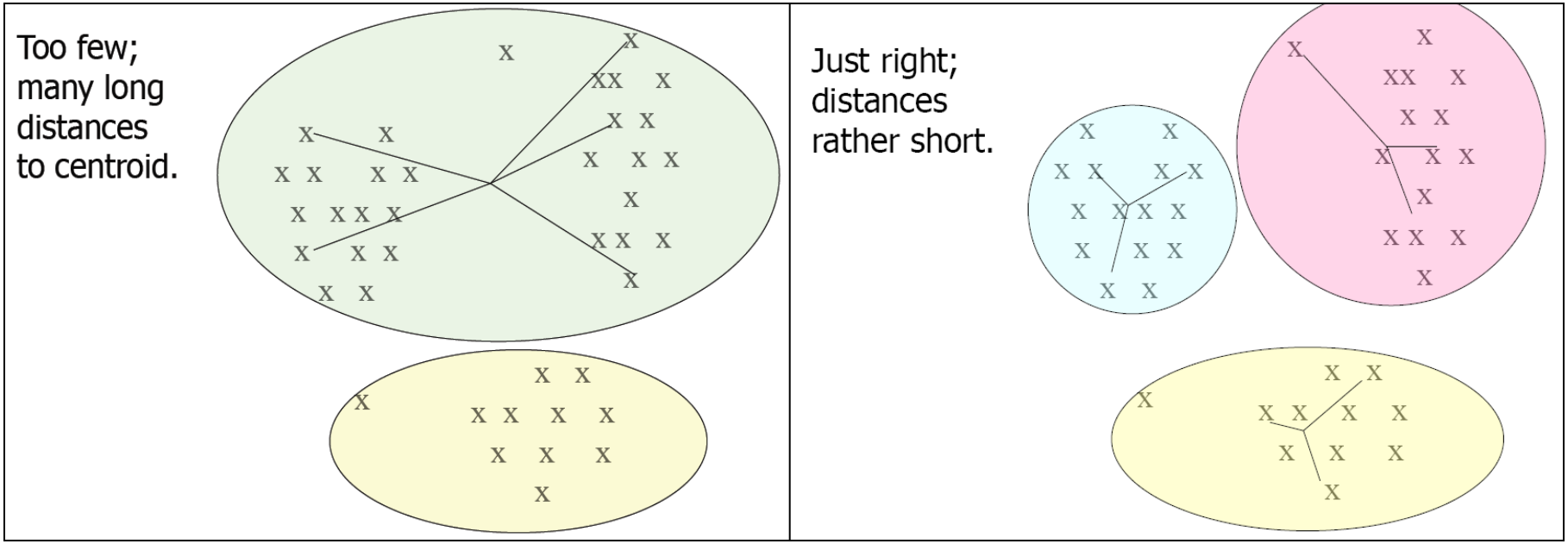
- Critérios para definição das sementes iniciais:
  1. Métodos aglomerativos
     - Utilliza-se um método aglomerativo para obter os k agrupamentos iniciais; em seguida, calculam-se os pontos médios nesses k agrupamentos.
  2. Escolha aleatória
     - *k* elementos amostrais são sorteados para formar as sementes iniciais. P. Ex. função 'kmeans' do R usa essa opção como default.
     - Forma mais robusta:
       - roda-se o algoritmo completo m vezes (cada qual com k sementes iniciais)
       - ao final, escolhe-se o agrupamento com o menor erro quadrado (menor soma dos quadrados das distâncias entre os centróides e os respectivos pontos pertencentes ao cluster correspondente)
  3. Escolha dos k valores mais discrepantes

- Um critério para definição de *k*:
  - Teste diferentes valores de *k*, medindo o decréscimo na distância média dos pontos aos seus respectivos centroides, à medida em que *k* aumenta
  - A distância média cai rapidamente até o valor adequado de *k*; a partir daí se altera pouco.



- Um critério para definição de *k*:

  Exemplo:



Fonte: J. Leskovec, A. Rajaraman. Clustering Algorithms. Stanford University. https://web.stanford.edu/class/cs345a/slides/12-clustering.pdf

Métodos baseados em misturas de distribuições de probabilidade

- Assume-se que os dados provêm de uma ou mais classes
- Assume-se que cada classe possui uma distribuição de probabilidade (p. ex. Normal multivariada) com parâmetros desconhecidos
- Dado um número k de classes, os parâmetros das classes são ajustados através de métodos de máxima verossimilhança ou máxima densidade a posteriori
- Cada ponto (da amostra ou novo) é designado à classe com maior densidade de probabilidade.
- A quantidade de classes é usualmente definida através de medidas de regularidade (AIC, BIC, etc) ou através de testes de hipóteses
- Na linguagem R: pacote **mclust**

- Exemplo: Data set *Iris virginica*
  - Variáveis: comprimento da sépala e comprimento da pétala
  - 49 espécimes observados
  - Problema: uma ou duas subpopulações?