Objectivos

- Adquirir a noção de agrupamento hierárquico
- Explicar a estratégia aglomerativa de agrupamento
- Diferenciar os métodos utilizados para o cálculo das distâncias entre grupos no quadro da estratégia aglomerativa

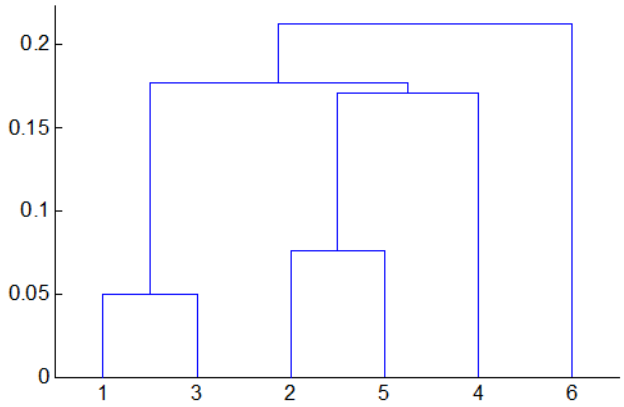
- Sumário:
  - Agrupamento hierárquico.
  - Estratégia aglomerativa

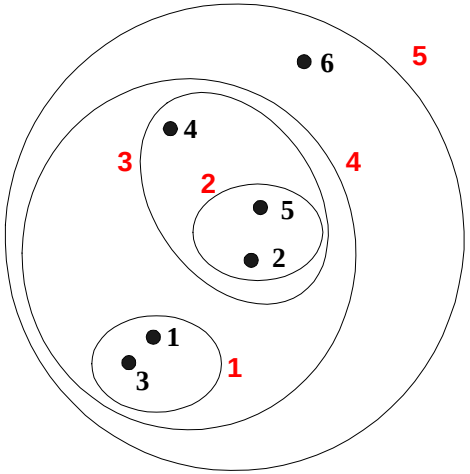
Agrupamento hierárquico (1/2)

- Constitui uma ferramenta popular de análise de dados
- A ideia básica consiste em construir uma árvore binária através da junção (ou separação) sucessiva de conjuntos de dados.

Agrupamento hierárquico (2/2)

- Árvore construída é designada *dendrograma*
- Informação também pode ser apresentada na forma de um *diagrama de Venn*





Agrupamento hierárquico vs. K – médias

K – médias

- Número de clusters a formar, k
- Definição dos centros iniciais
- Medida de semelhança (ou distância) entre os exemplares de dados

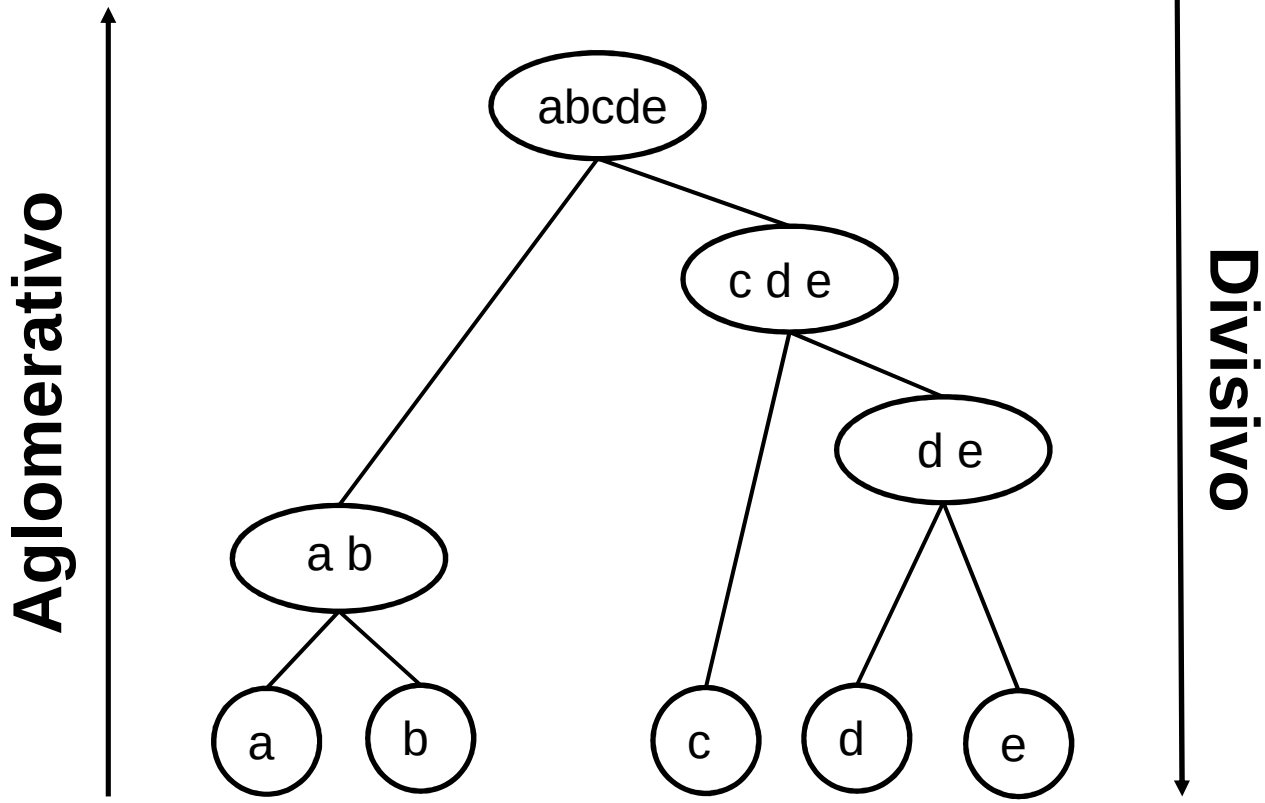
Hierárquico

- Medida de semelhança (ou distância) entre grupos de exemplares de dados

Agrupamento hierárquico: tipos (1/2)

- Existem dois tipos de métodos hierárquicos
  - Aglomerativos (ascendentes, *bottom – up*)
    - Cada exemplar de dados é considerado como um grupo individual
    - Grupos são sucessivamente fundidos até formar um agrupamento final
  - Divisivos (descendentes, *top – down*)
    - Consideram inicialmente todo o conjunto de dados como um único grupo
    - Divide recursivamente o grupo até formar um agrupamento final

Agrupamento hierárquico: tipos (2/2)



Métodos aglomerativos (1/3)

- Algoritmo
  - Formar *n* grupos, contendo cada grupo um objecto de dados individual
  - Repetir
    - Unir os dois grupos mais semelhantes
  - Até que: *todos os objectos de dados sejam incluídos no mesmo grupo*

Métodos aglomerativos (2/3)

- Baseados igualmente no conceito de semelhança (ou distância)
- Necessidade de avaliar a distância entre dois grupos de objectos

Métodos aglomerativos (3/3)

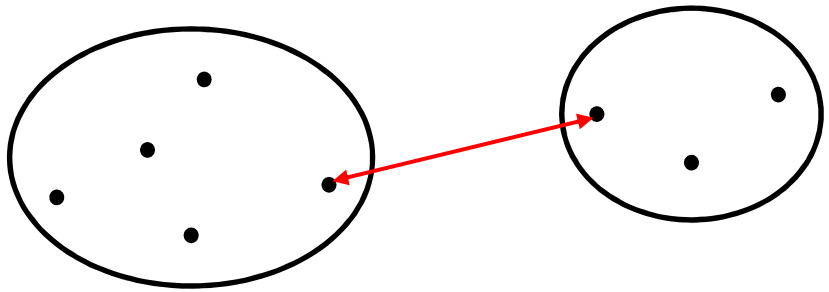
- Existem diferentes variantes do método
- Diferenciam-se quanto à forma de calcular a distância entre grupos

Distância entre grupos

- Dada uma medida de *distância entre dois pontos*, *d(x, y)*
- Existem várias formas de definir a *distância entre dois grupos de pontos*, *D(X, Y)*
- As opções mais populares
  - Distância mínima (*single linkage*)
  - Distância máxima (*complete linkage*)
  - Distância média (*average linkage*)

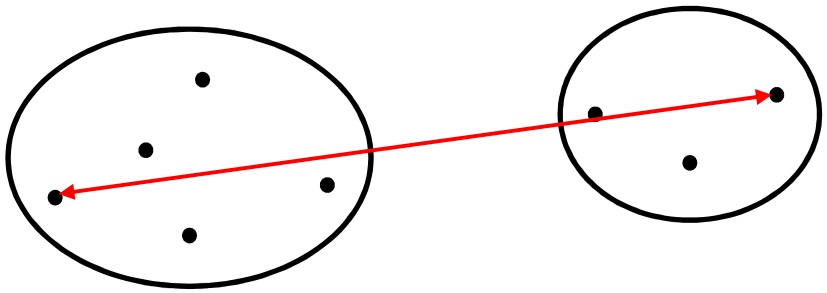
Distância mínima

- A distância entre dois grupos é a distância existente entre os *pontos mais próximos* de um e outro grupo
  - $D_{min}(X, Y) = \min_{x \in X,\, y \in Y} d(x, y)$



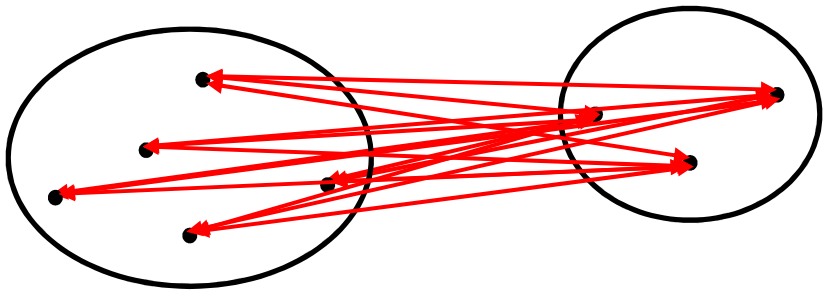
Distância máxima

- A distância entre dois grupos é a distância existente entre os *pontos mais distantes* de um e outro grupo
  - $D_{max}(X, Y) = \max_{x \in X,\, y \in Y} d(x, y)$



Distância média

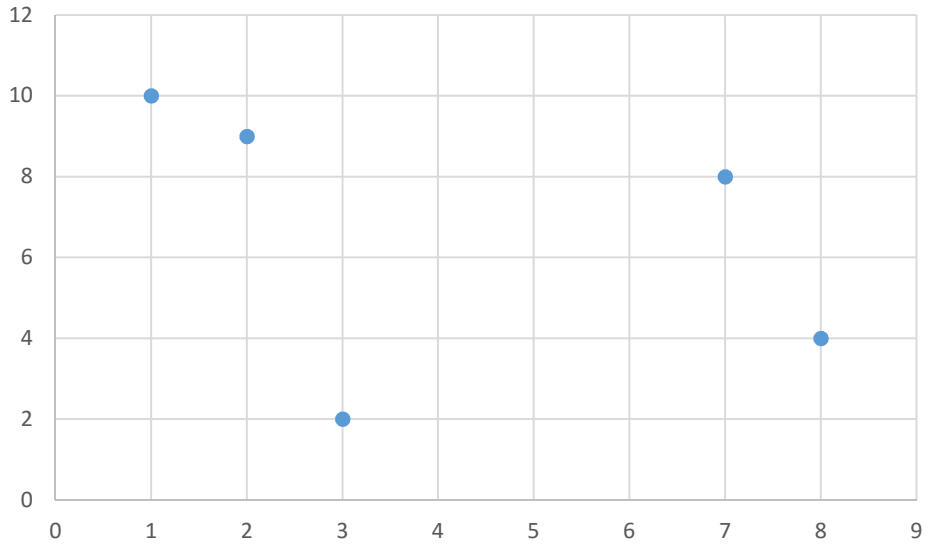
- A distância entre dois grupos é a *distância média entre os pontos* de um e outro grupo
  - $D_{avg}(X,Y) = \frac{1}{|X|.|Y|} \Sigma_{x \in X} \Sigma_{y \in Y} d(x,y)$



Exemplo: distância mínima (1/6)

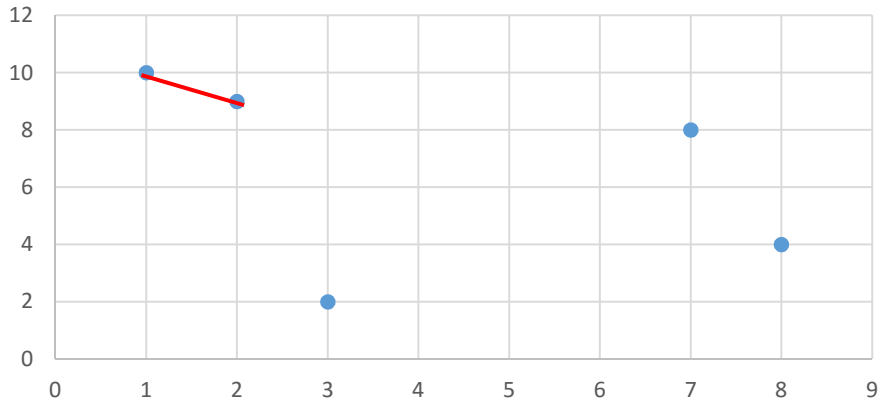
- Suponhamos que temos o seguinte conjunto de dados

	x	y
1	3	2
2	8	4
3	2	9
4	1	10
5	7	8



Exemplo: distância mínima (2/6)

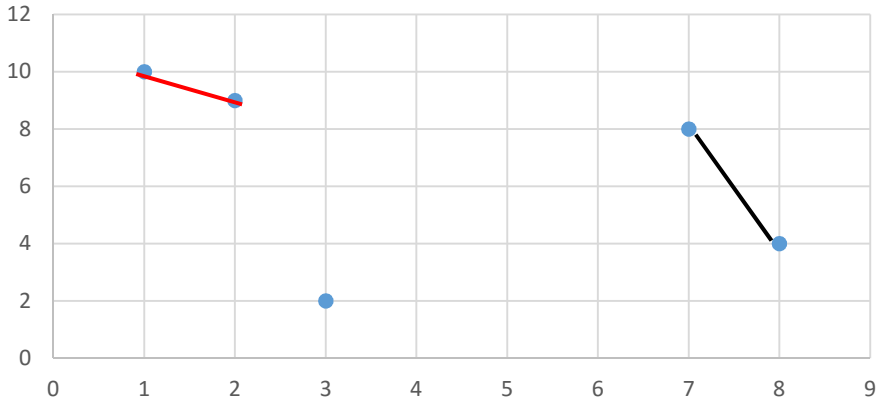
	x	y
1	3	2
2	8	4
3	2	9
4	1	10
5	7	8



	1	**2**	**3**	**4**	**5**
1	0,0				
2	5,4	0,0			
3	7,1	7,8	0,0		
4	8,2	9,2	1,4	0,0	
5	7,2	4,1	5,1	6,3	0,0

Exemplo: distância mínima (3/6)

	1	2	3	4	5
1	0,0				
2	5,4	0,0			
3	7,1	7,8	0,0		
4	8,2	9,2	1,4	0,0	
5	7,2	4,1	5,1	6,3	0,0

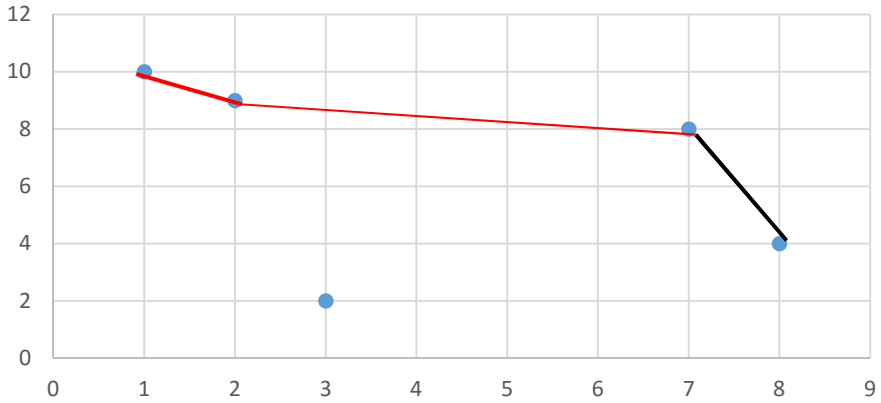


D(1, 34) = min{d(1, 3), d(1, 4)} = 7,1
D(2, 34) = min{d(2, 3), d(2, 4)} = 7,8
D(5, 34) = min{d(5, 3), d(5, 4)} = 5,1

	1	2	(3 4)	5
1	0,0			
2	5,4	0,0		
(3 4)	7,1	7,8	0,0	
5	7,2	4,1	5,1	0,0

Exemplo: distância mínima (4/6)

	1	**2**	**(3 4)**	**5**
1	0,0			
2	5,4	0,0		
(3 4)	7,1	7,8	0,0	
5	7,2	4,1	5,1	0,0

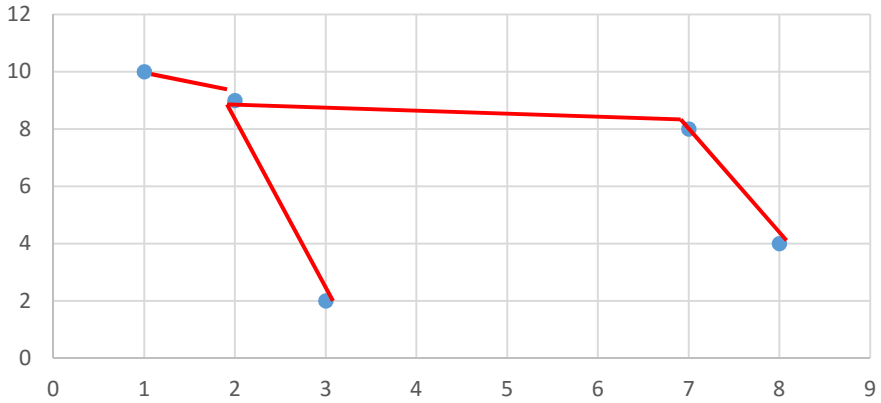


D(1, 25) = min{d(1, 2), d(1, 5)} = 5,4
D(34, 25) = min{d(3, 2), d(3, 5), d(4, 2), d(4, 5)} = 5,1

	1	**(2 5)**	**(3 4)**
1	0,0		
(2 5)	5,4	0,0	
(3 4)	7,1	5,1	0,0

Exemplo: distância mínima (5/6)

	1	**(2 5)**	**(3 4)**
1	0,0		
(2 5)	5,4	0,0	
(3 4)	7,1	5,1	0,0

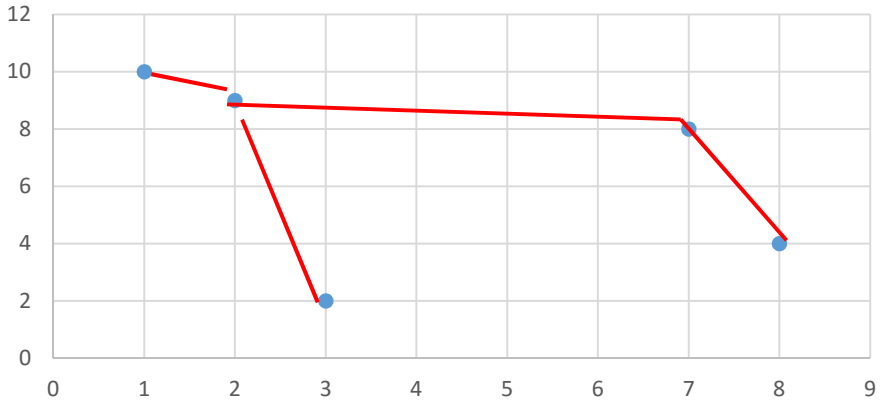


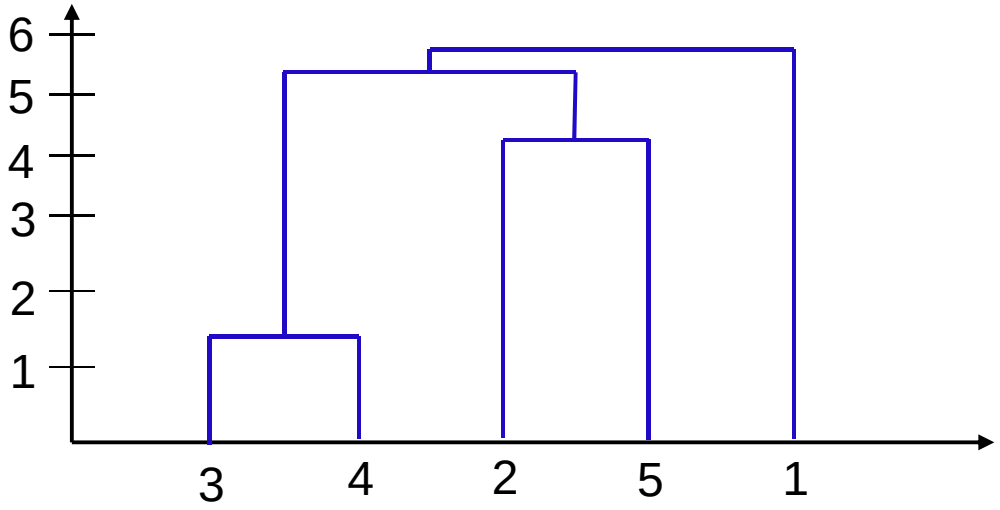
D(1, 2534) = min{d(1, 2), d(1, 5), d(1, 3), d(1, 4)} = 5,4

	1	**(2 5 3 4)**
1	0,0	
(2 5 3 4)	5,4	0,0

Exemplo: distância mínima (6/6)

	1	**(2 5 3 4)**
1	0,0	
(2 5 3 4)	5,4	0,0





Distância entre grupos: propriedades (1/2)

- Distância mínima
  - Pode produzir cadeias de alinhamento entre instâncias distantes (*chaining*)
  - Tende a formar clusters alongados
  - Os clusters formados algumas vezes viola a propriedade de *compacidade*
- Distância máxima
  - Favorece a compacidade
  - Tende a formar clusters com forma esférica
  - Os clusters formados algumas vezes violam a propriedade de *proximidade*

Distância entre grupos: propriedades (2/2)

- Distância média
  - Constitui um compromisso natural entre os dois métodos anteriores
  - Tende a produzir grupos relativamente compactos que se encontram relativamente separados uns dos outros

Agrupamento hierárquico: precauções

- Diferentes decisões sobre a distância entre grupos podem conduzir a agrupamentos muito diferentes
- O algoritmo impõe uma estrutura hierárquica aos dados até nos casos em que esta estrutura não é adequada
- O agrupamento é feito com base em decisões locais; uma vez tomadas não podem ser reavaliadas
- Não é robusto ao ruído
- Complexidade de tempo e espaço

Bibliografia

- Witten, pg. 274 – 279
- Hastie, T., Tibshirani, R. e Friedman, J., The Elements of Statistical Learniing, pg. 520 – 528