Mineração de Dados

Clusterização

Sumário

- Partição:
  - K-Means/K-Medoids
  - Clara/Clarans
- Clusterização Hierárquica
  - Agnes/Diana
- Fuzzy C-Means
- Determinação do Número de Clusters

O que é Clustering?

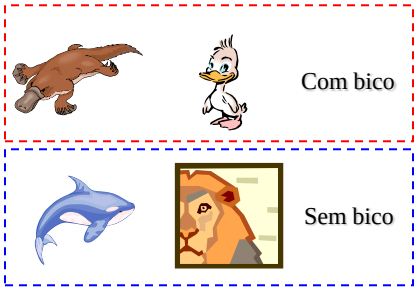
- Dado um conjunto de objetos, colocar os objetos em grupos baseados na similaridade entre eles
- Utilizado para encontrar padrões inesperados nos dados
- Como uma ferramenta autônoma para obter pistas sobre a distribuição de dados
- Como uma etapa de pré-processamento para outros algoritmos

- Como agrupar os seguintes animais?

O que é Clustering?

- Dado um conjunto de objetos, colocar os objetos em grupos baseados na similaridade entre eles
- Utilizado para encontrar padrões inesperados nos dados
- Como uma ferramenta autonoma para obter pistas sobre a distribuição de dados
- Como uma etapa de preprocessamento para outros algoritmos

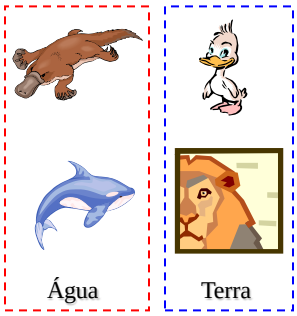
- Como agrupar os seguintes animais?



O que é Clustering?

- Dado um conjunto de objetos, colocar os objetos em grupos baseados na similaridade entre eles
- Utilizado para encontrar padrões inesperados nos dados
- Como uma ferramenta autonoma para obter pistas sobre a distribuição de dados
- Como uma etapa de preprocessamento para outros algoritmos

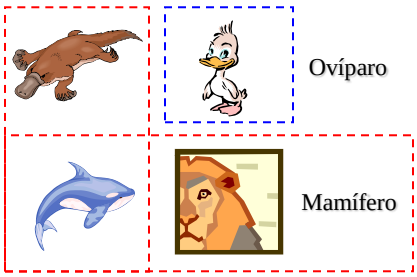
- Como agrupar os seguintes animais?



O que é Clustering?

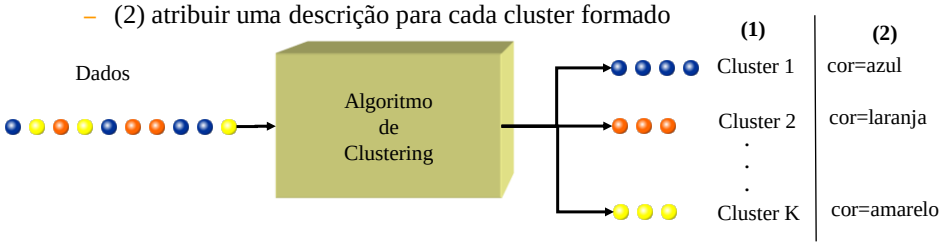
- Dado um conjunto de objetos, colocar os objetos em grupos baseados na similaridade entre eles
- Utilizado para encontrar padrões inesperados nos dados
- Como uma ferramenta autonoma para obter pistas sobre a distribuição de dados
- Como uma etapa de preprocessamento para outros algoritmos

- Como agrupar os seguintes animais?



Descrição do Problema

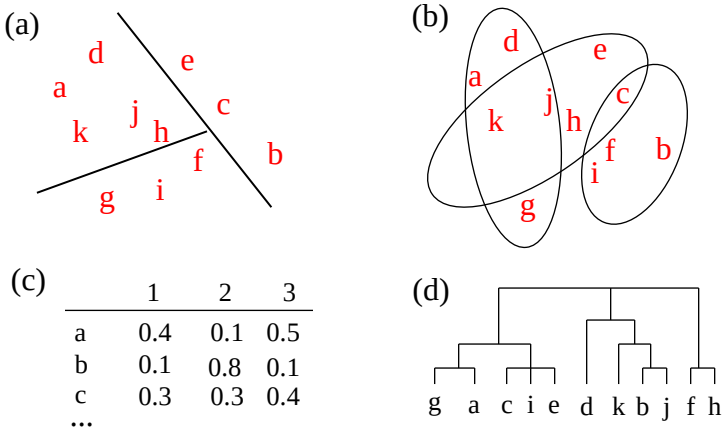
- Clustering (Agrupamento): Aprendizado não Supervisionado
- Dado um conjunto de objetos descritos por múltiplos valores (atributos)
  - (1) atribuir grupos (clusters) aos objetos particionando-os objetivamente em grupos homogêneos de maneira a:
    - Maximizar a similaridade de objetos dentro de um mesmo cluster
    - Minimizar a similaridade de objetos entre clusters distintos
  - (2) atribuir uma descrição para cada cluster formado



O que é um bom agrupamento?

- Um bom método de agrupamento fornece grupos de alta qualidade com
  - Alta similaridade intra-grupo
  - baixa similaridade inter-grupo
- A qualidade do resultado de um agrupamento depende tanto da medida de similaridade usada pelo método como da sua implementação.
- A qualidade de um método de agrupamento é também medido pela sua habilidade para descobrir os padrões escondidos.

Representação de Clusters



Principais Etapas da Formação de Agrupamentos

- Construção da Tabela de Dados
- Cálculo da Proximidade
  1) Escolha de um Índice de Proximidade
  2) Construção da Matriz de Proximidades
- Seleção de um Algoritmo de Formação de Grupos em função do tipo de agrupamento desejado
- Análise e Interpretação dos Resultados

A Representação dos Dados

- Matriz de Dados

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Matriz de Dissimilaridade

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Dissimilaridade entre objetos

- Distâncias são normalmente usadas como medida de dissimilaridade entre objetos
- Entre as mais populares: distancia de Minkowski
  - onde i = $(x_{i1}, x_{i2}, \ldots, x_{ip})$ e $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ são dois vetores p-dimensionais, e q é um inteiro positivo

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + \ldots + |x_{i_p} - x_{j_p}|^q)}$$

- Se q = 1, d é a distância de Manhattan

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$$

Medida da Qualidade de um Agrupamento

- Proximidade: é uma função que mede a similaridade ou a dissimilaridade entre um par de observações
- Uma função mede a qualidade de um grupo.
- As funções de proximidade dependem da escala das variáveis: proporcional, intervalar, ordinal, nominal, binária, mista
- Pode-se associar pesos as variáveis como conhecimento do domínio.
- Às vezes, pode ser muito difícil definir o que são dois objetos "bastante similares"
  - a resposta é quase sempre subjetiva

Métodos de Agrupamento

- Distingue-se três grupos de métodos:
  - Técnicas de Otimização
    - Objetivo: obter uma partição.
    - Número de grupos fornecido pelo usuário
  - Técnicas Hierárquicas
    - Objetivo: obter uma hierarquia (ou uma pirâmide)
    - Pode-se obter uma partição "cortando-se" a hierarquia em um determinado nível.
  - Técnicas de Cobertura
    - Objetivo: obter grupos que eventualmente podem partilhar indivíduos.

Principais Métodos de Agrupamento

- Métodos que fornecem uma partição:
  - Construa várias partições que são então avaliadas segundo algum critério
- Métodos Hierárquicos:
  - Fornece uma decomposição hierárquica dos objetos segundo um critério particular
- Métodos de Densidade:
  - Basedos em conectividade e funções de densidade

Métodos que fornecem uma partição

- Produz uma partição de uma base de dados D de n objetos em k grupos
- Dado k, encontre uma partição em k grupos que otimiza um dado critério
- Ótimo global: enumeração exaustiva de todas as partições
- Heuristicas: k-means e k-medoids
- k-means (MacQueen'67): Cada grupo é representado pelo seu centro
- k-medoids ou PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Cada grupo é representado por um objeto no grupo

K-Means

- Associa cada ponto ao centróide mais próximo
- Medidas de distância têm um papel importante
- Algoritmo é simples e rápido
  - Pode ser usado em grandes conjuntos de dados
- Pode não produzir o mesmo resultado em cada execução
- Minimiza a variância intra-cluster mas não a variância global

K-Means

- Objetivo do Algoritmo
  - Minimizar a função

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} \left| x_j - \mu_i \right|^2$$

K-Means

- Objetivo do Algoritmo
  - Minimizar a função

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} \left| x_j - \mu_i \right|^2$$

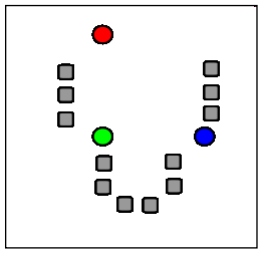
Amostras

Centróides

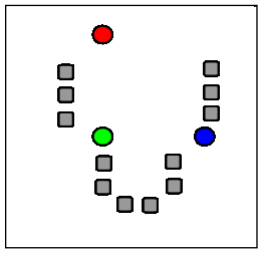
K-Means

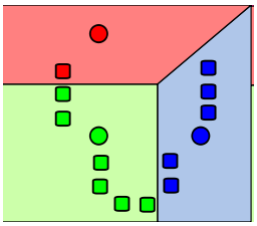
- Algoritmo de Lloyd
  - Escolhe o número de clusters k
  - Gera k clusters aleatoriamente e determina os centróides dos mesmos
    - Opcionalmente pode gerar os k centróides diretamente
  - Associa cada ponto ao centróide mais próximo
  - Recalcula os novos centros dos clusters
  - Repete os dois últimos passos até convergir

K-Means

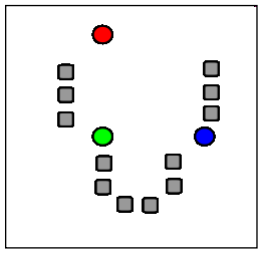


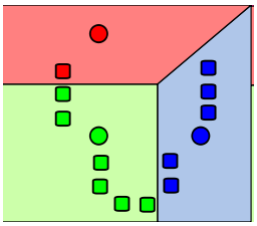
K-Means

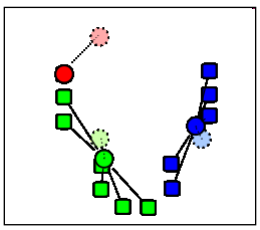




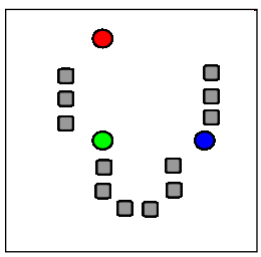
K-Means

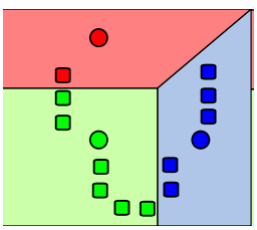


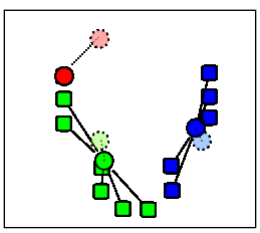


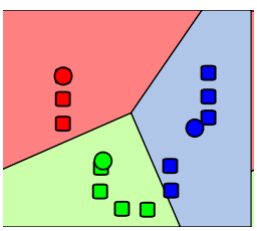


K-Means









k-Means: Exemplo

	X1	X2	X3	X4	X5
Cliente_1	7,000	10,000	9,000	7,000	10,000
Cliente_2	9,000	9,000	8,000	9,000	9,000
Cliente_3	5,000	5,000	6,000	7,000	7,000
Cliente_4	6,000	6,000	3,000	3,000	4,000
Cliente_5	1,000	2,000	2,000	1,000	2,000
Cliente_6	4,000	3,000	2,000	3,000	3,000
Cliente_7	2,000	4,000	5,000	2,000	5,000

Exemplo do uso do *k*-médias, com Correlação de Pearson como medida de proximidade

k-Means: Exemplo

	Cliente_1	Cliente_2	Cliente_3	Cliente_4	Cliente_5	Cliente_6	Cliente_7
Cliente_1	1,000						
Cliente_2	-0,147	1,000	0,000	0,516	-0,408	0,791	-0,516
Cliente_3	0,000	0,000	1,000				
Cliente_4	0,087	0,516	-0,824	1,000			
Cliente_5	0,963	-0,408	0,000	-0,060	1,000	-0,645	0,963
Cliente_6	-0,466	0,791	-0,354	0,699	-0,645	1,000	
Cliente_7	0,891	-0,516	0,165	-0,239	0,963	-0,699	1,000

k-Means: Exemplo

	X1	X2	X3	X4	X5
Cliente_2	9,000	9,000	8,000	9,000	9,000
Cliente_3	5,000	5,000	6,000	7,000	7,000
Cliente_4	6,000	6,000	3,000	3,000	4,000
Cliente_6	4,000	3,000	2,000	3,000	3,000
Centro_1	**6,000**	**5,750**	**4,750**	**5,500**	**5,750**

	X1	X2	X3	X4	X5
Cliente_1	7,000	10,000	9,000	7,000	10,000
Cliente_5	1,000	2,000	2,000	1,000	2,000
Cliente_7	2,000	4,000	5,000	2,000	5,000
Centro_2	**3,333**	**5,333**	**5,333**	**3,333**	**5,667**

k-Means: Exemplo

	Cliente_1	Cliente_2	Cliente_3	Cliente_4	Cliente_5	Cliente_6	Cliente_7	Centro_1	Centro_2
Cliente_1	1	-0,1474	0	0,087	0,9631	-0,4663	0,8913	-0,1371	0,9723
Cliente_2	-0,1474	1	0	0,516	-0,4082	0,7906	-0,516	0,93	-0,3498
Cliente_3	0	0	1	-0,8242	0	-0,3536	0,1648	-0,2599	0,068
Cliente_4	0,087	0,516	-0,8242	1	-0,0602	0,6994	-0,2391	0,737	-0,0698
Cliente_5	0,9631	-0,4082	0	-0,0602	1	-0,6455	0,9631	-0,3797	0,9926
Cliente_6	-0,4663	0,7906	-0,3536	0,6994	-0,6455	1	-0,6994	0,919	-0,6011
Cliente_7	0,8913	-0,516	0,1648	-0,2391	0,9631	-0,6994	1	-0,4799	0,9723
Centro_1	-0,1371	0,93	-0,2599	0,737	-0,3797	0,919	-0,4799	1	-0,322
Centro_2	0,9723	-0,3498	0,068	-0,0698	0,9926	-0,6011	0,9723	-0,322	1

k-Means: Exemplo

	X1	X2	X3	X4	X5
Cliente_2	9,000	9,000	8,000	9,000	9,000
Cliente_4	6,000	6,000	3,000	3,000	4,000
Cliente_6	4,000	3,000	2,000	3,000	3,000
Centro_1	**6,333**	**6,000**	**4,333**	**5,000**	**5,333**

	X1	X2	X3	X4	X5
Cliente_1	7,000	10,000	9,000	7,000	10,000
Cliente_3	5,000	5,000	6,000	7,000	7,000
Cliente_5	1,000	2,000	2,000	1,000	2,000
Cliente_7	2,000	4,000	5,000	2,000	5,000
Centro_2	**3,750**	**5,250**	**5,500**	**4,250**	**6,000**

k-Means: Exemplo

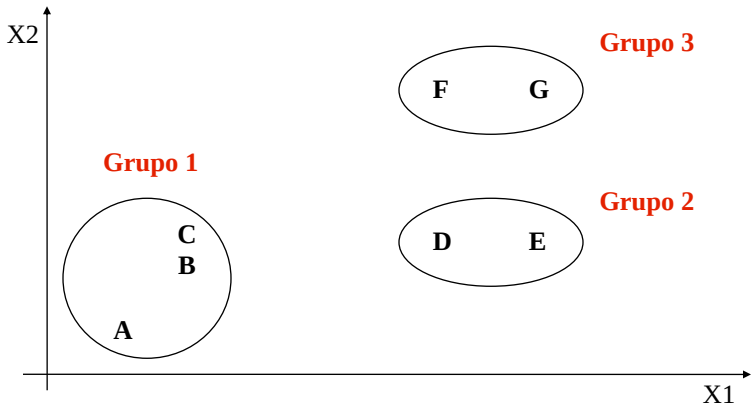
	Cliente_1	Cliente_2	Cliente_3	Cliente_4	Cliente_5	Cliente_6	Cliente_7	Centro_1	Centro_2
Cliente_1	1	-0,1474	0	0,087	0,9631	-0,4663	0,8913	-0,1106	0,9175
Cliente_2	-0,1474	1	0	0,516	-0,4082	0,7906	-0,516	0,75	-0,3323
Cliente_3	0	0	1	-0,8242	0	-0,3536	-0,6281	0,3377	
Cliente_4	0,087	0,516	-0,8242	1	-0,0602	0,6994	-0,2391	0,9389	-0,2939
Cliente_5	0,9631	-0,4082	0	-0,0602	1	-0,6455	0,9631	-0,3062	0,9372
Cliente_6	-0,4663	0,7906	-0,3536	0,6994	-0,6455	1	-0,6994	0,8883	-0,6686
Cliente_7	0,8913	-0,516	0,1648	-0,2391	0,9631	-0,6994	1	-0,4564	0,962
Centro_1	-0,1106	0,75	-0,6281	0,9389	-0,3062	0,8883	-0,4564	1	-0,4473
Centro_2	0,9175	-0,3323	0,3377	-0,2939	0,9372	-0,6686	0,962	-0,4473	1

k-Means: Exemplo

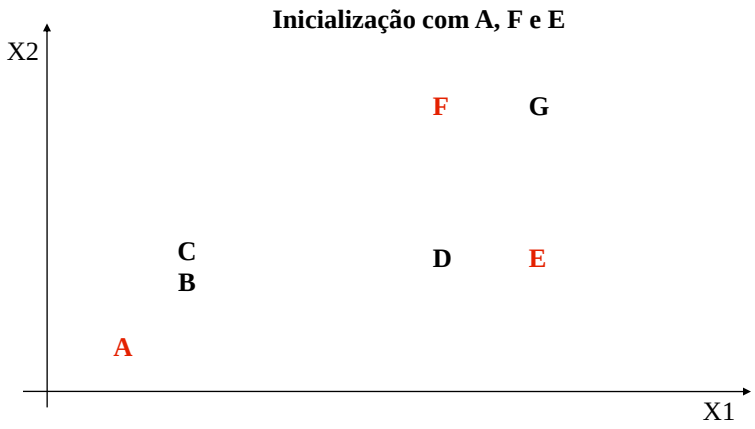
	X1	X2	X3	X4	X5
Cliente_2	9,000	9,000	8,000	9,000	9,000
Cliente_4	6,000	6,000	3,000	3,000	4,000
Cliente_6	4,000	3,000	2,000	3,000	3,000
Centro_1	**6,333**	**6,000**	**4,333**	**5,000**	**5,333**

	X1	X2	X3	X4	X5
Cliente_1	7,000	10,000	9,000	7,000	10,000
Cliente_3	5,000	5,000	6,000	7,000	7,000
Cliente_5	1,000	2,000	2,000	1,000	2,000
Cliente_7	2,000	4,000	5,000	2,000	5,000
Centro_2	**3,750**	**5,250**	**5,500**	**4,250**	**6,000**

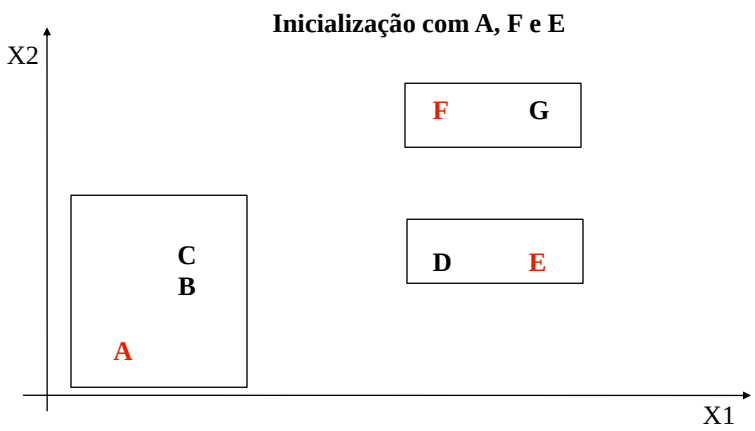
k-means: sensibilidade à condição inicial



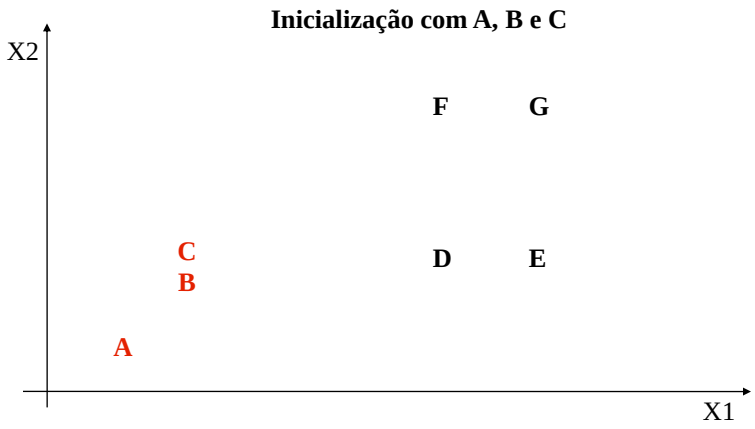
k-means: sensibilidade à condição inicial



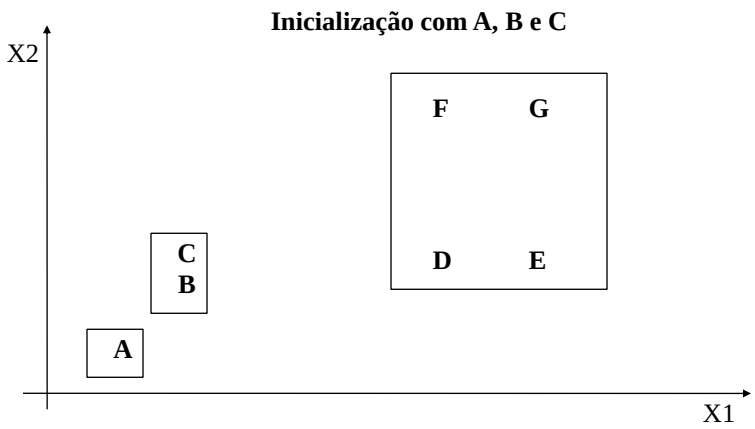
k-means: sensibilidade à condição inicial



k-means: sensibilidade à condição inicial



k-means: sensibilidade à condição inicial



K-Means

- Outros métodos de minimizar a função objetivo ?
  - Algoritmos genéticos
  - Algoritmo de Lloyd costuma ser muito rápido
    - Geralmente muito menos interações do que pontos
  - Alguns conjuntos de pontos podem gerar tempos superpolinomiais

K-Means

- Pontos fortes
  - Relativamente eficiente: O(tkn), onde n é # objetos, k é # grupos, e t é # iterações. Normalmente, k, t << n.
  - Freqüentemente termina em um ótimo local. O ótimo global pode ser encontrado usando técnicas tais como: deterministic annealing e algoritmos genéticos
- Pontos fracos
  - Aplicável apenas quando a média é definida, o que fazer com dados categóricos?
  - É necessário especificar a priori k, o número de grupos
  - É sensível a ruídos e *outliers*
  - Não é apropriado para a descoberta de grupos não esféricos

Variantes do K-Means

- Algumas variantes do k-means diferem em:
  - Seleção das k medias iniciais
  - Cálculo das dissimilaridades
  - Estratégias para calcular as médias dos grupos
- Dados categóricos: k-modas (Huang'98)
  - Troca das médias pelas modas dos grupos
  - Uso de novas medidas de dissimilaridade para tratar dados categóricos
  - Uso de um método baseado na freqüência para atualizar as modas
  - Mistura de dados categóricos e numéricos: método k-prototype

O MétodoK-Medoids

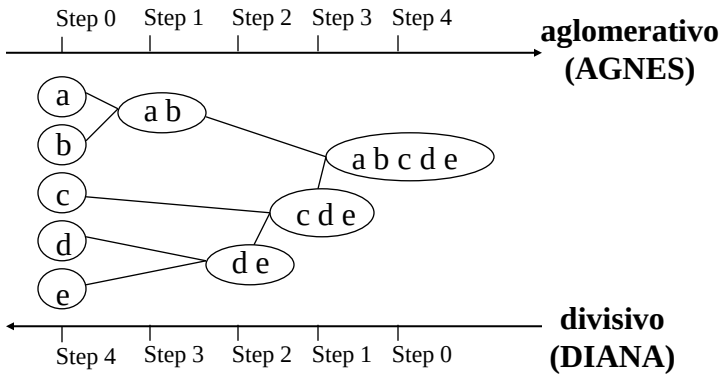
- Encontre objetos representativos, chamados medoids, nos grupos
- PAM (Partitioning Around Medoids, 1987)
  - Inicie com um conjunto inicial de medoids e iterativamente troque um deles por um não medoide verifique se a distancia total do agrupamento melhora
  - PAM funciona para conjuntos de dados pequenos, mas não possui escalabilidade suficiente para os grandes
- CLARA (Kaufmann & Rousseeuw, 1990)
- CLARANS (Ng & Han, 1994): Amostragem aleatória

Clusterização Hierárquica

- Usa uma matriz de distâncias como critério de agrupamento.
- Esse métodos não requerem o número de grupos k como entrada, mas precisa de uma condição de parada
- Cada passo da clusterização é feito usando os clusters do passo anterior
- Processo aglomerativo/divisivo
- Representação baseada em dendrogramas

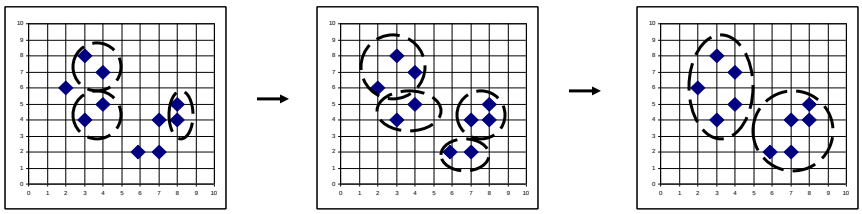
41

Métodos Hierarquicos



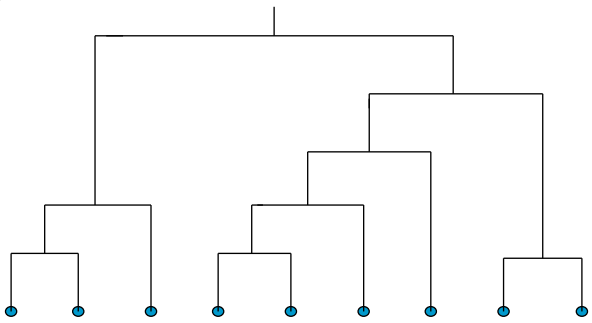
AGNES (Agglomerative Nesting - 1990)

- Fusiona nós que tem as menores dissimilaridades
- Eventualmente todos os nós pertencem ao mesmo grupo



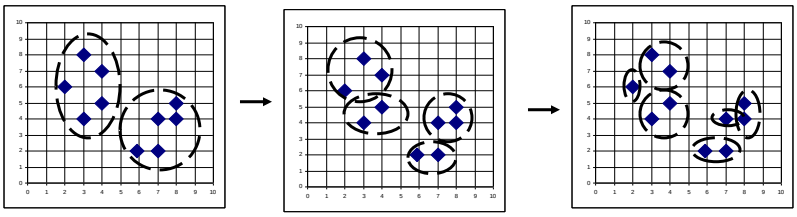
Dendrograma

- Decompõe os objetos em vários níveis de partições embutidas (árvore de grupos, chamado de dendrograma).
- Um agrupamento dos objetos é obtido pelo corte do dendrograma em um nível desejado e então cada componente conectado forma um grupo



DIANA (Divisive Analysis-1990)

- Ordem inversa de AGNES
- Eventualmente cada nó forma um grupo unitário



Procedimentos Hierárquicos de Agrupamento

- Envolvem a construção de uma hierarquia de uma estrutura do tipo árvore
  - Divisivo
  - Aglomerativo
    - Ligação Individual
    - Ligação Completa
    - Ligação Média

Métodos Aglomerativos

- Cada objeto começa como seu próprio grupo (*cluster*)
- Em passos seguintes, os dois grupos (ou objetos) mais próximos (similares) são combinados em um novo agregado
  - O número de grupos é reduzido em uma unidade em cada passo
- Ao final, todos os elementos são reunidos em um grande agregado

Método Hierárquico: Ligação Individual

- Encontra os dois objetos separados pela menor distância (mais similares) e os coloca no primeiro grupo
- Em seguida, a próxima menor distância (ou maior similaridade) é determinada, e um terceiro objeto se junta aos dois primeiros para formar um grupo, ou um novo grupo de dois elementos é formado
  - A distância (similaridade) entre dois grupos quaisquer é **a menor distância** (maior similaridade) de qualquer ponto de um grupo até qualquer ponto do outro
- O procedimento continua até que todos os objetos formem um só agregado

Ligação Individual: Exemplo

	X1	X2	X3	X4	X5
C_1	7,000	10,000	9,000	7,000	10,000
C_2	9,000	9,000	8,000	9,000	9,000
C_3	5,000	5,000	6,000	7,000	7,000
C_4	6,000	6,000	3,000	3,000	4,000
C_5	1,000	2,000	2,000	1,000	2,000
C_6	4,000	3,000	2,000	3,000	3,000
C_7	2,000	4,000	5,000	2,000	5,000

Exemplo do uso do Método Hierárquico Aglomerativo de Ligação Individual, com Correlação de Pearson como medida de proximidade

Ligação Individual: Exemplo

	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_1	1,000				0,963		
C_2	-0,147	1,000			-0,408		
C_3	0,000	0,000	1,000		0,000		
C_4	0,087	0,516	-0,824	1,000	-0,060		
C_5	0,963	-0,408	0,000	-0,060	1,000		
C_6	-0,466	0,791	-0,354	0,699	-0,645	1,000	
C_7	0,891	-0,516	0,165	-0,239	0,963	-0,699	1,000

	(C_1,C_5)	C_2	C_3	C_4	C_6	C_7
(C_1,C_5)	1,000					0,963
C_2	-0,147	1,000				-0,516
C_3	0,000	0,000	1,000			0,165
C_4	0,087	0,516	-0,824	1,000		-0,239
C_6	-0,466	0,791	-0,354	0,699	1,000	-0,699
C_7	0,963	-0,516	0,165	-0,239	-0,699	1,000

Ligação Individual: Exemplo

	(C_1,C_5)	C_2	C_3	C_4	C_6	C_7
(C_1,C_5)	1,000					0,963
C_2	-0,147	1,000				-0,516
C_3	0,000	0,000	1,000			0,165
C_4	0,087	0,516	-0,824	1,000		-0,239
C_6	-0,466	0,791	-0,354	0,699	1,000	-0,699
C_7	0,963	-0,516	0,165	-0,239	-0,699	1,000

	(C_1,C_5,C_7)	C_2	C_3	C_4	C_6
(C_1,C_5,C_7)	1,000	-0,147	0,165	0,087	-0,466
C_2	-0,147	1,000	-0,824	0,516	0,791
C_3	0,165	0,000	1,000	-0,824	-0,354
C_4	0,087	0,516	-0,824	1,000	0,699
C_6	-0,466	0,791	-0,354	0,699	1,000

Ligação Individual: Exemplo

	(C_1,C_5,C_7)	C_2	C_3	C_4	C_6
(C_1,C_5,C_7)	1,000	-0,147	0,165	0,087	-0,466
C_2	-0,147	1,000	-0,824	0,516	0,791
C_3	0,165	0,000	1,000	-0,824	-0,354
C_4	0,087	0,516	-0,824	1,000	0,699
C_6	-0,466	0,791	-0,354	0,699	1,000

	(C_1,C_5,C_7)	(C_2,C_6,C_4)	C_3
(C_1,C_5,C_7)	1,000	0,087	0,165
(C_2,C_6,C_4)	0,087	1,000	0,699
C_3	0,165	0,699	1,000

Ligação Individual: Exemplo

	(C_1,C_5,C_7)	(C_2,C_6,C_4)	C_3
(C_1,C_5,C_7)	1,000	0,087	0,165
(C_2,C_6,C_4)	0,087	1,000	0,699
C_3	0,165	0,699	1,000

	(C_1,C_5,C_7)	(C_2,C_6,C_4,C_3)
(C_1,C_5,C_7)	1,000	0,087
(C_2,C_6,C_4,C_3)	0,087	1,000

Ligação Individual: Exemplo

	(C_1,C_5,C_7)	(C_2,C_6,C_4,C_3)
(C_1,C_5,C_7)	1,000	0,087
(C_2,C_6,C_4,C_3)	0,087	1,000

(C_1,C_5,C_7,C_2,C_6,C_4,C_3)

Ligação Simples: Características

- Gera grupos alongados
- Problemas em lidar com grupos mal delineados
  - Em tais casos, procedimentos de ligação individual geram longas cadeias e eventualmente todos os objetos são colocados em uma sua cadeia
  - Os objetos em extremos opostos de uma cadeia podem ser muito diferentes

Medidas de Distância

- Maior distância entre elementos
  - Ligação Completa
- Menor distância entre elementos
  - Ligação Simples
- Distância média entre elementos
  - Ligação pela Média
- Soma da variância intra-cluster
- Aumento da variância após junção de clusters
  - Critério de Ward
- Matriz de distância auxilia clusterização

Vantagens e Desvantagens

- Vantagens
  - Definição do número de clusters pode ser feita indiretamente
  - Algoritmo é simples de ser implementado
- Desvantagens
  - Escalabilidade baixa – $O(n^2)$