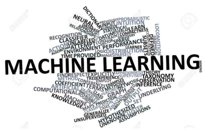
Aprendizado de Máquina

Agrupamento de dados

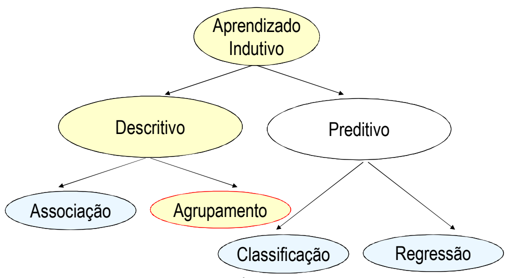
André C. P. L. F. de Carvalho
ICMC-USP



Tópicos

- Agrupamento de dados
- Dificuldades em agrupamento
- Algoritmos de agrupamento
- Validação
- Aplicações

Introdução



Introdução

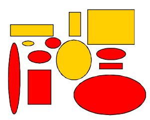
- Nem sempre os dados em um conjunto estão rotulados
  - Custo
  - Impossibilidade
- Conhecimento útil e relevante pode ser extraídos de dados não rotulados
  - Grupos de dados similares

Agrupamento

- Organização de um conjunto de objetos em grupos (clusters)
  - Não existe uma definição precisa
  - Particiona objetos de acordo com alguma relação entre eles
  - Busca partição que maximiza:
    - Similaridade entre objetos de um mesmo grupo e
    - Dissimilaridade entre objetos de grupos diferentes

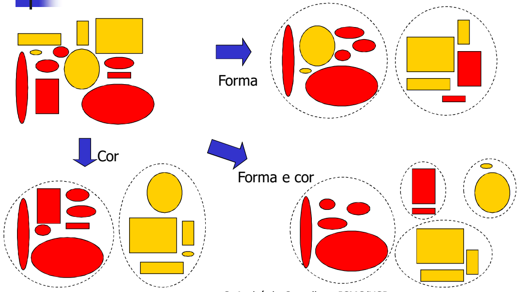
Agrupamento

- Supor os objetos:

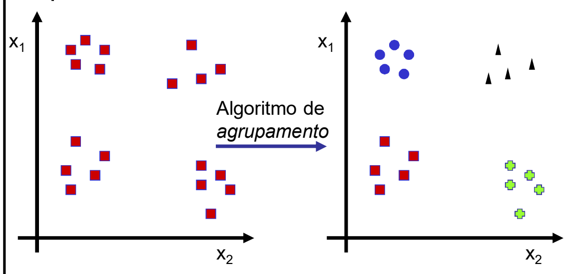


Como particionar?

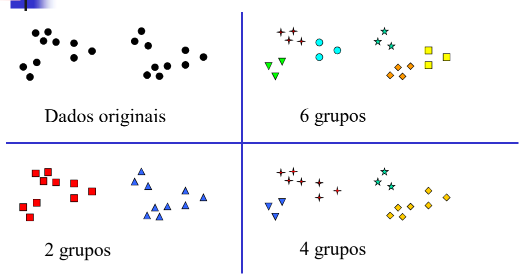
Agrupamento



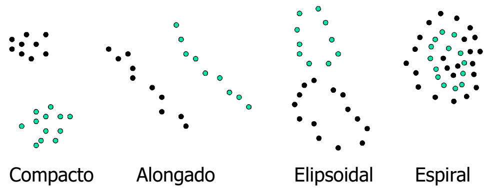
Agrupamento de dados



Quantos grupos?



Possíveis formatos



Tipos de agrupamento

- Seja $X = \{x_1, x_2, ..., x_n\}$ o conjunto de todos os dados
  - Tarefa: colocar cada X_i em um dos k clusters $C_1, C_2, ..., C_k$
- De acordo com a pertinência dos dados, grupamentos podem ser de dois tipos:
  - Tipo 1: duro (crisp)
  - Tipo 2: fuzzy

Tipos de agrupamento

- Agrupamento crisp
  - Cada objeto X_i pertence ou não a cada cluster C_j

$$C_i \neq \emptyset, i = 1, ..., k \qquad \bigcup_{i=1}^{k} C_i = X$$

$$C_i \cap C_j = \emptyset, i \neq j, i, j \in \{1,2, ..., k\}$$

  - Objeto em C_i é mais semelhante a outros objetos em C_i do que àqueles em C_j, $i \neq j$

Tipos de agrupamento

- Agrupamento fuzzy
  - Usa uma função de pertinência para definir o quanto um elemento pertence a um grupo

$Pert_j : x_i \rightarrow [0, 1]$

$Pert_j$ = pertinência ao grupo j
k = número de grupos
n = número de objetos

$$\sum_{j=1}^{k} Pert_j(x_i) = 1, \; i \in \{1, \ldots, n\}$$

$$0 < \sum_{i=1}^{n} Pert_j(x_i) \leq n, \; j \in \{1, \ldots, k\}$$

Objetivo

- Encontrar partição que maximiza similaridade
  - Minimiza dissimilaridade
  - Quanto maior a homogeneidade dentro dos grupos e a diferença entre os grupos, melhor
- Alternativas
  - Busca exaustiva
  - Algoritmos de agrupamento de dados

Busca exaustiva

- Tentar todos os possíveis agrupamentos de *k* grupos (para vários valores de *k*)
- Números de Stirling do segundo tipo
  - Número de formas de particionar *n* dados em *k* subconjuntos não vazios

$$>> \binom{n}{k} \geq \left(\frac{n}{k}\right)^k$$

k = número de grupos
n = número de objetos

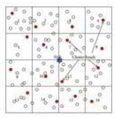
- Impraticável

Agrupamento de dados

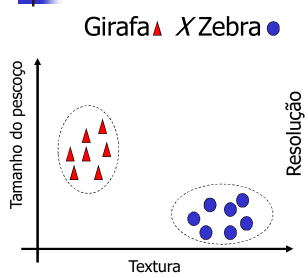
- Diferentes partições podem ser encontradas
  - Por diferentes algoritmos
    - Utilizam critérios diferentes para buscar uma boa partição
  - Pelo mesmo algoritmo
    - Diferentes inicializações
    - Diferentes números de grupos

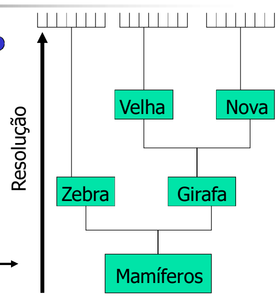
Algoritmos de agrupamento

- Principais abordagens
  - Particionais
    - Protótipos (erro quadrático médio)
    - Densidade
  - Hierárquicos
  - Baseados em grids (grades)
  - Baseados em grafos

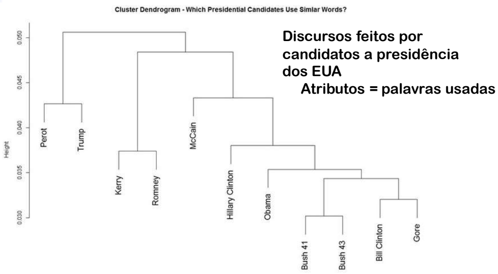


Particional X Hierárquico





Agrupamento hierárquico



Discursos feitos por candidatos a presidência dos EUA
Atributos = palavras usadas

http://www.realclearpolitics.com/articles/2016/10/19/debate_text_mining_trump_has_a_cousin_in_perot__132098.html

Algoritmos particionais

- Principais características
  - Produzem um único agrupamento (partição)
  - A maioria utiliza abordagem "gulosa" (*greedy*)
    - Busca pela melhor alternativa no momento, sem considerar futuras consequências
    - Uma vez tomada uma decisão, não volta atrás
    - Geralmente resultado depende da ordem de apresentação dos exemplos

Algoritmos particionais

- K-médias (K-médias ótimo, K-médias sequencial)
- SOM
- FCM
- DENCLUE
- CLICK
- CAST
- SNN

Algoritmo k-médias

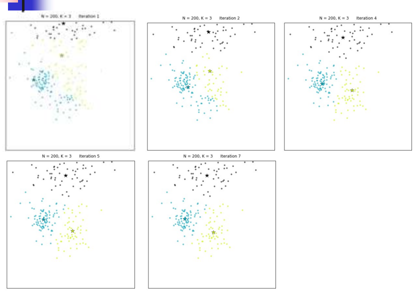
- Supor *n* objetos $x_1, x_2, \ldots, x_n$ a serem agrupados em k clusters, $k < n$
  - Seja μ_j a média dos objetos do cluster C_j
  - Seja *d* uma medida de distância
    - $\boldsymbol{x_p} \in$ cluster C_i se $d(\boldsymbol{x_p}, \mu_i)$ for menor que todas as k–1 distâncias entre $\boldsymbol{x_p}$ e μ_j, j = 1, 2, ..., k e i ≠ j

Algoritmo k-médias

Mais formal

1 Sugerir k centros iniciais $\mu_1, \mu_2, \ldots, \mu_k$
2 Repetir
Para i variando de 1 a n # Numero de objetos
Colocar objeto x_i no cluster C_j com média μ_j mais semelhante a ele
Para i variando de 1 a K # Numero de grupos
Substituir μ_i pela média de todos os objetos do cluster C_i
Até nenhuma das médias mudar

Algoritmo k-médias



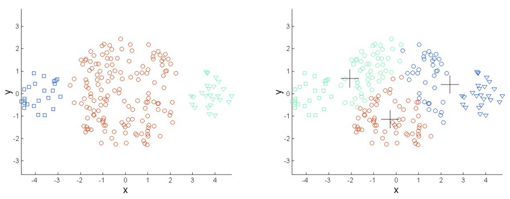
Algoritmo k-médias

- Médias iniciais
  - Objetos (vetores) aleatoriamente gerados
  - Objetos aleatoriamente escolhidos do conjunto de treinamento
  - Objetos bem diferentes entre si

Limitações do k-médias

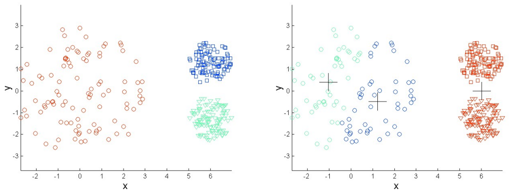
- Escolha do valor de K
  - Tentativa e erro ou automática
- Algoritmos K-médias tem problemas quando:
  - Grupos têm diferentes densidades
  - Grupos têm formatos não hiper-esféricos
  - Atributos estão em diferentes escalas
- Tem problemas também quando os dados contêm *outliers*

Grupos encontrados



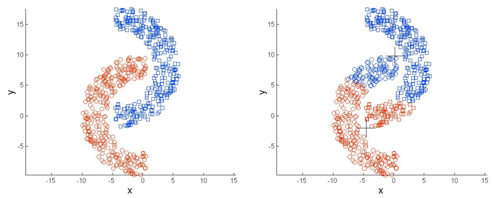
Grupos verdadeiros **K-médias (3 grupos)**

Densidades diferentes



Grupos verdadeiros **K-médias (3 grupos)**

Formatos não hiperesféricos



Grupos verdadeiros **K-médias (2 grupos)**

Exercício

- Agrupar, utilizando k-médias, os dados abaixo em 2 grupos:
  - X_1 = 1, 0, 1, 1
  - X_2 = 0, 1, 0, 0
  - X_3 = 0, 1, 1, 0
  - X_4 = 1, 1, 1, 1
  - X_5 = 0, 1, 0 ,1

Algoritmos hierárquicos

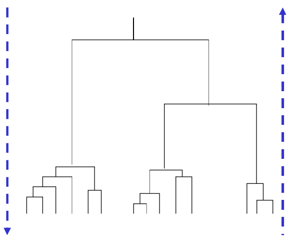
- Utilizam diagrama de árvore (dendograma)
  - Produz uma sequência (hierarquia) de agrupamentos
- Historicamente usados em áreas que trabalham com estruturas hierárquicas
  - Taxonomias
  - Ex.: Biologia e arqueologia

Algoritmos hierárquicos

- Tipos:
  - Aglomerativos: combinam, repetidamente, dois grupos em um
    - A cada passo, combinam os dois grupos atuais mais próximos em um novo grupo
  - Divisivos: Dividem, repetidamente, um grupo em dois
    - A cada passo, dividem o grupo atual menos homogêneo em dois novos grupos

Exemplo

Divisivo



Aglomerativo

Esquema Aglomerativo Generalizado (EAG)

1 Incializar $P_0 = \{\{x_1\}, \ldots, \{x_n\}\}$, $t = 0$
2 Para $t = 1$ até $n - 1$ faça
Encontrar o par de grupos mais próximos (C_i, C_j)
$P_t = (P_{t-1} - \{C_i, C_j\}) \cup \{\{C_i \cup C_j\}\}$
/ atualiza centro*
/ Número de chamadas a $d(C_i, C_j)$ é $O(n^3)$*
/ Esse número pode ser reduzido*

Algoritmos hierárquicos

- Existe uma grande variedade de algoritmos hierárquicos
  - Geralmente diferem na forma de calcular distância entre grupos

$$d_{AB} = \min_{\substack{i \in A \\ j \in B}}(d_{ij})$$ Por ligação simples (single-link)

$$d_{AB} = \max_{\substack{i \in A \\ j \in B}}(d_{ij})$$ Por ligação completa (complete-link)

$$d_{AB} = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{j \in B} d_{ij}$$ Pela média do grupo (average-link)

Validação de agrupamentos

- Como avaliar os clusters gerados por um algoritmo de agrupamento?
  - Especialista no domínio dos dados
    - Demorado para grandes conjuntos de dados
    - Subjetivo
  - Existem várias medidas de validação para agrupamento de dados
    - Julgam aspectos diferentes

Medidas de validação

- Podem ser divididas em três grupos
  - Índices ou critérios internos
    - Medem a qualidade da partição obtida sem considerar informações externas
  - Índices ou critérios relativos
    - Usados para comparar duas partições ou grupos
  - Índices ou critérios externos
    - Medem o quanto os rótulos dos grupos coincidem com a classe verdadeira

Medidas internas

- Muitas são baseadas em:
  - Coesão de clusters
    - Mede o quão próximos estão os objetos dentro de um cluster
  - Separação de clusters
    - Mede o quão separado cada cluster está dos demais clusters

Silhueta

- Combina coesão com separação
- Calculada para cada objeto que faz parte de um agrupamento
  - Baseada em:
    - Distância entre os objetos de um mesmo cluster e
    - Distância dos objetos de um cluster ao cluster mais próximo

Silhueta

- Para cada objeto x_i de um conjunto de dados
  - $a(x_i)$: distância média de x_i aos outros objetos de seu cluster
  - $b(x_i)$: min (distância média de x_i a todos os objetos nos outros clusters)

$$s(x_i) = \begin{cases} 1 - a(x_i)/b(x_i), & \text{se } a(x_i) < b(x_i) \\ 0, & \text{se } a(x_i) = b(x_i) \\ b(x_i)/a(x_i) - 1, & \text{se } a(x_i) > b(x_i) \end{cases}$$

- Largura média da silhueta
  - Média de *s(x)* de todos os objetos do conjunto de dados
  - Valor entre -1 e 1 (quanto mais próximo de 1, melhor)

Exemplo

- Agrupar, utilizando k-médias, os dados abaixo em 2 grupos e em 3 grupos:
  - X_1 = 1, 0, 1, 1
  - X_2 = 0, 1, 0, 0
  - X_3 = 0, 1, 1, 0
  - X_4 = 1, 1, 1, 1
  - X_5 = 0, 1, 0 ,1
- Calcular valor da silhueta para as duas partições

Exercício

- Seja o seguinte cadastro de pacientes:

Nome	Febre	Enjôo	Manchas	Dores	Diagnóstico
João	sim	sim	pequenas	sim	doente
Pedro	não	não	grandes	não	saudável
Maria	sim	sim	pequenas	não	saudável
José	sim	não	grandes	sim	doente
Ana	sim	não	pequenas	sim	saudável
Leila	não	não	grandes	sim	doente

Exercício

- Agrupar os dados em grupos usando o algoritmo K-médias e medida de silhueta
  - Usar k = 2 e k = 3
  - Informação sobre a classe não deve ser usada
  - Usar distância bloco-cidade
- Em que grupos seriam colocados os exemplos abaixo?
  - (Luis, não, não, pequenas, sim)
  - (Laura, sim, sim, grandes, sim)

Exercício

- Agrupar os dados usando o algoritmo hierárquico aglomerativo visto em aula
  - Usar distância bloco-cidade
  - Informação sobre a classe não deve ser usada
- Em que grupos das partições com 2 e 3 grupos ficariam os exemplos abaixo?
  - (Luis, não, não, pequenas, sim)
  - (Laura, sim, sim, grandes, sim)

Tarefa

- Usando o algoritmo de agrupamento hierárquico aglomerativo, construir um equivalente, só que divisivo
- Aplicar k-médias e um algoritmo de agrupamento hierárquico ao conjunto de dados do kaggle usado nas tarefas anteriores

Considerações finais

- Agrupamento de dados é umas das principais tarefas de AM
  - Várias definições de agrupamento
  - Diversos algoritmos
- Validação das partições encontradas
- Custo de rotular exemplos
  - Aprendizado semi-supervisionado
  - Aprendizado ativo

Perguntas