Mineração de Dados em Biologia Molecular

Agrupamento de dados

Docente: André C. P. L. F. de Carvalho
PAE: Victor Hugo Barella

Tópicos

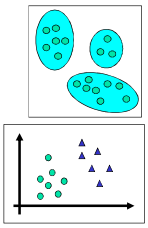
- Aprendizado de máquina descritivo
- Medidas de distância
- Agrupamento de dados
- Algoritmos de agrupamento
- Validação
- Aplicações

Algoritmos de AM descritivos

- Descrevem ou sumarizam um conjunto de dados
- Indução de modelo (treinamento) usa todo o conjunto de dados
  - Geralmente indução ocorre por aprendizado não supervisionado
    - E.X.: Agrupamento de dados

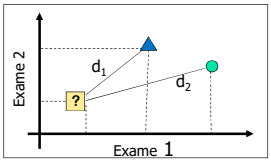
AM e Geometria

- Vários algoritmos de AM são baseados em medidas de distância (MD)
- MD podem ser usadas para:
  - Agrupar dados
    - Ex.: K-médias
  - Predizer a classe de novos dados
    - Ex.: K-NN



Medidas de distância

- Consideram proximidade entre dados
  - Similaridade
  - Dissimilaridade



- Existem várias
  - Euclidiana
  - Norma máxima
  - Bloco-cidade
  - ...

Distância de Minkowski

- Medida de distância generalizada

$$dist(p,q) = (\sum_{k=1}^{m} | p_k - q_k |^r)^{\frac{1}{r}}$$

- Valor de r leva a diferentes distâncias:
  - 1 (L_1): Distância bloco cidade
  - 2 (L_2): Distância Euclidiana

Medidas de distância

- Distância de norma máxima
  - Menor complexidade e exatidão

$$dist(p,q) = MAX(|\, p_k - q_k \,|)$$

- Distância Bloco cidade (Manhattan)
  - Usa r = 1

$$dist(p,q) = \sum_{k=1}^{m} |p_k - q_k|$$

  - Distância de *Hamming,* se valores são binários
    - Ex. *dist (011,101)*

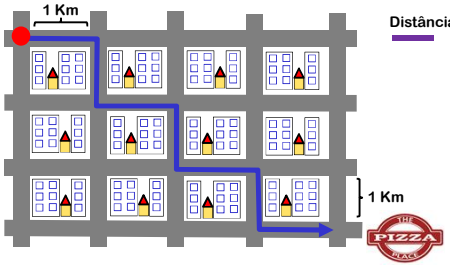
Medidas de distância

- Distância Euclidiana
  - Usa r = 2
  - Sistema de coordenadas cartesianas

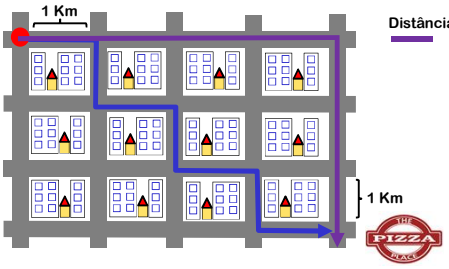
$$dist(p,q) = \sqrt{\sum_{k=1}^{m} (p_k - q_k)^2}$$

  - Medida de distância mais utilizada

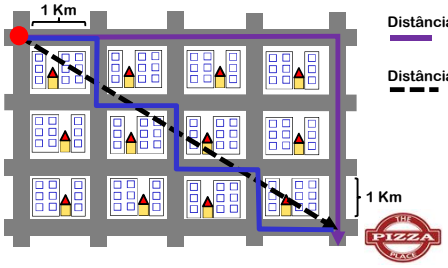
Medidas de distância



Medidas de distância



Medidas de distância



Exercício

- Qual das três medidas resulta na maior e na menor distância entre os exemplos abaixo?
  - Manhattan
  - Euclidiana
  - Norma máxima

Ex1 = (3, 1, 10, 2)

Ex2 = (2, 5, 3, 2)

Exercício

- Utilizando distância de Manhattan, qual dos pares das coordenadas abaixo é mais semelhante ?
  - (2, 5, 1), (3, 0, 6), (4, 1, 3)

Exercício

- Utilizando distância de Manhattan, definir:
  - Qual par dos valores binários abaixo tem a distância mais semelhante à diferença entre seus valores na base decimal
  - 110000, 111001, 000111, 001011, 100111, 101001

Exercício

- Para cada medida de distância
  - Quais são os dois exemplos da tabela abaixo mais próximos e os dois mais distantes?
  - Usar distâncias Euclidiana, bloco cidade e de norma máxima

Estado	Escolaridade	Altura	Salário	Classe
SP	Médio	180	3000	A
RJ	Superior	174	7000	B
RJ	Fundamental	100	2000	A

Similaridade x Dissimilaridade

- Medidas de distância medem dissimilaridade entre objetos
- Alguns algoritmos utilizam similaridade entre objetos
- Quanto mais similares dois objetos, mais distantes e vice-versa
- Ambos calculam a proximidade entre valores

Similaridade x Dissimilaridade

- Similaridade
  - Mede o quão semelhantes são dois objetos
    - Quanto mais parecidos, maior o valor
  - Geralmente valor $\in$ [0.0, 1.0]
- Dissimilaridade
  - Mede o quanto dois objetos são diferentes
    - Quanto mais diferentes, maior o valor
  - Geralmente valor $\in$ [0.0, X]
- Medida de proximidade são usadas nos dois casos

Proximidade entre valores

- Sejam a e b dois valores de um atributo
  - Nominal
    - s = 1 - d

$$d(a,b) = \begin{cases} 1, \text{se a} \neq \text{b} \\ 0, \text{se a} = \text{b} \end{cases}$$

  - Ordinal
    - s = 1 - d

$$d(a,b) = \frac{|pos_a - pos_b|}{n-1}$$ n = #valores

  - Intervalar ou racional
    - s = 1 - d ou s = 1/(1+d)

$$d(a,b) = |a-b|$$

Proximidade entre valores

- Sejam a e b dois valores de um atributo
  - Nominal
    - s = 1 - d

$$d(a,b) = \begin{cases} 1, \text{se a} \neq \text{b} \\ 0, \text{se a} = \text{b} \end{cases}$$

  - Ordinal
    - s = 1 - d

$$d(a,b) = \frac{|pos_a - pos_b|}{n-1}$$ n = #valores

  - Intervalar ou racional
    - s = 1 - d ou s = 1/(1+d)

$$d(a,b) = |a-b|$$

Proximidade entre valores

- Sejam a e b dois valores de um atributo

Tipo	Diferença	Semelhança
Nominal	$diff = \begin{cases} 1, \text{se a} \neq \text{b} \\ 0, \text{se a} = \text{b} \end{cases}$	**p = 1 - d**
Ordinal	$diff = \frac{\lvert pos_a - pos_b \rvert}{n-1}$ *n = #valores*	
Intervalar/ Racional	$diff = \lvert a-b \rvert$	**p = 1 - d ou p =1/(1+d)**

Similaridade entre vetores binários

- Algumas vezes, objetos p e q têm apenas valores binários
  - Ex.: 0110 e 1100
- Similaridades podem ser computadas por:
  - M_{01} = número de atributos em que p = 0 e q = 1
  - M_{10} = número de atributos em que p = 1 e q = 0
  - M_{00} = número de atributos em que p = 0 e q = 0
  - M_{11} = número de atributos em que p = 1 e q = 1

Similaridade entre vetores binários

- Coeficiente de Casamento Simples

$$CCS = (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$$

- Coeficiente Jaccard

$$J = (M_{11}) / (M_{01} + M_{10} + M_{11})$$

- Validação de agrupamento de dados

Exercício

- Que medida de similaridade binária gera o maior valor de similaridade entre vetores *p* e *q?*

p = 1 0 0 1 1 0 1 0 1 1 1 0
q = 0 1 0 0 1 1 0 0 1 0 1 1

Similaridade cosseno

- Muito usado quando dados são textos
  - *Bag of words*
    - Grande número de atributos
    - Vetores esparsos
- Sejam *p* e *q* vetores representando documentos
  - $simcos(p, q) = ||p|| ||q|| \cos\theta = (p \bullet q) / (||p|| \, ||q||)$
    - •: vector produto interno entre vetores
    - $|| p ||$: é o tamanho (norma) do vetor *p*

Distância cosseno

- Distância angular entre dois vetores
  - Invariante a escala dos atributos
  - 1 – similaridade cosseno

$$dist_{\cos seno} = 1 - \frac{\sum_{k=1}^{m} p_k \cdot q_k}{\sum_{k=1}^{m} p_k^2 \cdot \sum_{k=1}^{m} q_k^2}$$

Distância de Pearson

- Muito usada em bioinformática e séries temporais
  - 1 – correlação entre dois vetores

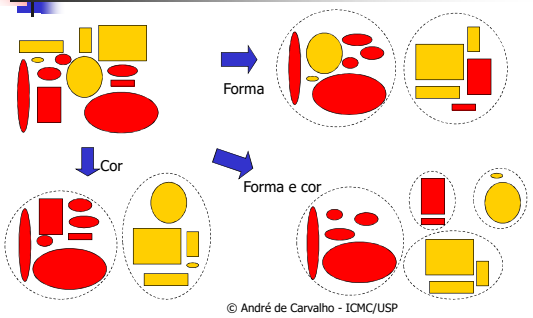
$$dist_{Pearson} = 1 - \frac{\sum_{k=1}^{m} (p_k - \overline{p}).(q_k - \overline{q})}{\sqrt{\sum_{k=1}^{m} (p_k - \overline{p})^2 \cdot \sum_{k=1}^{m} (q_k - \overline{q})^2}}$$

Agrupamento

- Organização de um conjunto de objetos em grupos (clusters)
  - Particionar objetos de acordo com alguma relação entre eles

Como particionar?

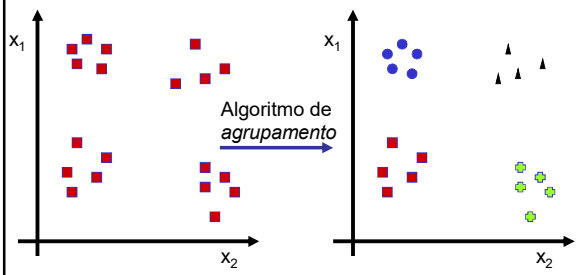
Agrupamento



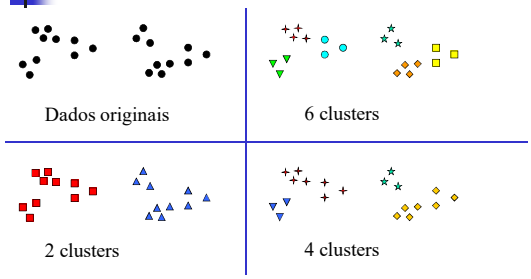
Agrupamento de dados

- Definição do que é um agrupamento
  - Imprecisa
  - Depende de:
    - Natureza dos dados
    - Resultados desejados
  - Existem várias
  - Partições
    - Grupos ou clusters

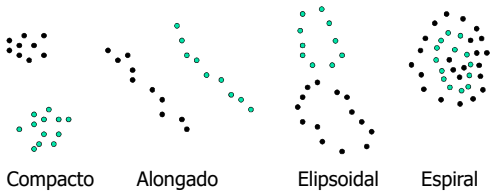
Agrupamento de dados



Possíveis números de clusters



Possíveis formatos de clusters



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

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

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

  - 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

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

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

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

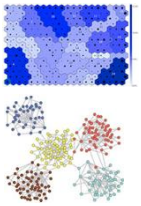
$$>> \binom{n}{k} \geq \left(\frac{n}{k}\right)^k$$

k = número de grupos
n = número de objetos

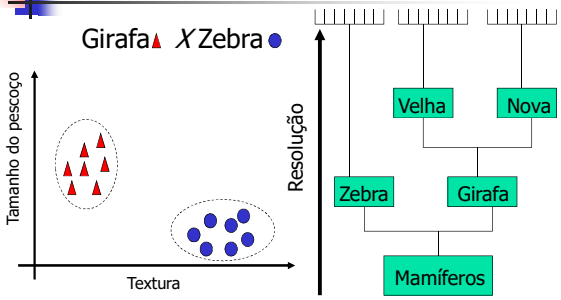
- Impraticável

Algoritmos de agrupamento

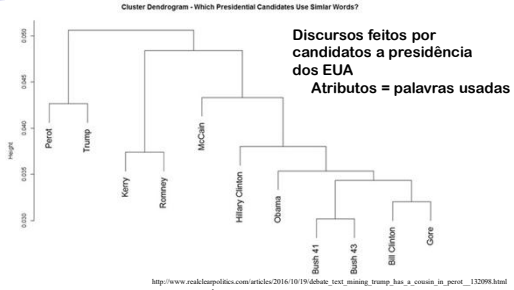
- Principais abordagens
  - Particionais
    - Protótipos (erro quadrático médio)
    - Densidade
  - Hierárquicos
  - Baseados em grids (grades)
  - Baseados em grafos



Particional X Hierárquico



Agrupamento hierárquico



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 $x_1, x_2, ..., x_n$ a serem agrupados em k clusters, $k < n$
  - Seja μ_i a média dos objetos do cluster C_i
  - Seja d uma medida de distância
    - $\boldsymbol{x_p} \in$ cluster C_i se $d(\boldsymbol{x_p}, \boldsymbol{\mu_i})$ for menor que todas as k–1 distâncias entre $\boldsymbol{x_p}$ e $\boldsymbol{\mu_j}$, j = 1, 2, ..., k e i ≠ j
      - $\boldsymbol{x_p}$ é colocado no cluster mais próximo

Algoritmo k-médias

Mais formal

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

Algoritmo k-médias

Mais informal

1 Sugerir centros iniciais $\mu_1, \mu_2, ..., \mu_k$
2 Repetir
Usar as médias sugeridas para colocar cada objeto no cluster C_i mais próximo
Substituir μ_i de cada cluster C_i pela média de todos os objetos de C_i
Até nenhuma das médias mudar

Algoritmo k-médias

- Médias iniciais
  - Objetos (vetores) aleatórios
  - Objetos aleatoriamente escolhidos do conjunto de treinamento
  - Objetos bem diferentes

Limitações do k-médias

- Escolha do valor de K
- K-médias tem problemas quando:
  - Grupos têm diferentes densidades
  - Grupos possuem formatos não hiper-esféricos
  - Conjunto de dados contém *outliers*

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

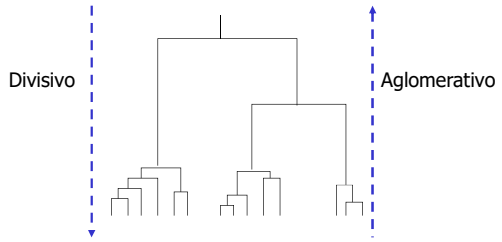
Algoritmos hierárquicos

- Utilizam diagrama de árvore (dendograma)
  - Produz uma sequência (hierarquia) de agrupamentos
- Historicamente usados em áreas que empregam estrutura hierárquica
  - Ex.: Biologia e arqueologia

Algoritmos hierárquicos

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

Exemplo



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 centros*
/ 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

- 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 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 dois grupos usando o algoritmo K-médias e medida de silhueta
  - Usar k = 2 e k = 3
  - Informação sobre a classe não deve ser usada
- Em que grupos seriam colocados os novos casos?
  - (Luis, não, não, pequenas, sim)
  - (Laura, sim, sim, grandes, sim)

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?