

# Mineração de Dados

## Agrupamento



- 1 Análise de Agrupamento
- 2 Métodos de Particionamento
- 3 Métodos Hierárquicos
- 4 Métodos Baseados em Densidade
- 5 Métodos Baseados em Grade
- 6 Avaliando Agrupamentos

# Análise de Agrupamento

# Análise de Agrupamento

- ▶ Grupo ou *Cluster*: uma coleção de objetos
  - ▶ similar (ou relacionado) aos outros objetos do mesmo grupo
  - ▶ dissimilar (ou não relacionados) aos objetos dos outros grupos
- ▶ Agrupamento, *Clustering*, Segmentação dos Dados
  - ▶ Encontrar similaridades entre os dados de acordo com características desses dados e agrupá-los em conjuntos com elementos similares
- ▶ Aprendizado não supervisionado: não há uma indicação de classe ou supervisão
- ▶ Aplicações típicas
  - ▶ Como uma ferramenta *stand-alone* para descobrir relações entre os dados
  - ▶ Como um componente de pré-processamento para outros algoritmos

# Agrupamento para Entendimento dos Dados

- ▶ Biology: taxonomy of living things: kingdom, phylum, class, order, family, genus and species
- ▶ Information retrieval: document clustering
- ▶ Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- ▶ Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults
- ▶ Climate: understanding earth climate, find patterns of atmospheric and ocean
- ▶ Economic Science: market research

# Agrupamento para Pré-processamento

- ▶ Sumarização
  - ▶ Pré-processamento para técnicas de regressão, PCA, classificação e análise de associação
- ▶ Compressão
  - ▶ Processamento de imagem: quantização vetorial
- ▶ Encontrar os k-vizinhos mais próximos
  - ▶ Busca por grupos
- ▶ Detecção de *outlier*
  - ▶ Os *outliers* frequentemente não estão relacionados a nenhum grupo

# Qualidade

- ▶ Uma boa técnica de agrupamento deve produzir uma organização com
  - ▶ Alta similaridade intra-grupo: grupo coesivo
  - ▶ Baixa similaridade inter-grupos: distinção entre grupos
- ▶ A qualidade de uma técnica de agrupamento depende da
  - ▶ medida de similaridade utilizada pelo método
  - ▶ sua implementação
  - ▶ sua habilidade em descobrir alguns ou todos os padrões escondidos nos dados

# Medida de Qualidade do Agrupamento

- ▶ Métrica de similaridade/dissimilaridade
  - ▶ Similaridade é expressa em termos de uma função de distância:  $d(i, j)$
  - ▶ A definição das funções de distância depende do tipo de dado
  - ▶ Pesos podem estar associados a diferentes variáveis com base na aplicação e na semântica dos dados
- ▶ Qualidade de um agrupamento
  - ▶ Normalmente, existe uma função de “qualidade” separada
  - ▶ É difícil definir “suficientemente similar” ou “bom suficiente”
  - ▶ Esta avaliação é tipicamente subjetiva



# Considerações para a Análise de Agrupamento

- ▶ Critério de Particionamento
  - ▶ Único nível
  - ▶ Hierárquico
- ▶ Separação dos grupos
  - ▶ Exclusivo: cada elemento pertence apenas a um grupo
  - ▶ Não exclusivo: um elemento pode pertencer a mais de um grupo
- ▶ Medida de similaridade
  - ▶ Distância: Euclideana, por exemplo
  - ▶ Conectividade: densidade ou contiguidade
- ▶ Espaço de agrupamento
  - ▶ Todo o espaço: comumente adotado quando os dados envolvem baixa dimensionalidade
  - ▶ Subespaços: agrupamento sobre dados de alta dimensionalidade

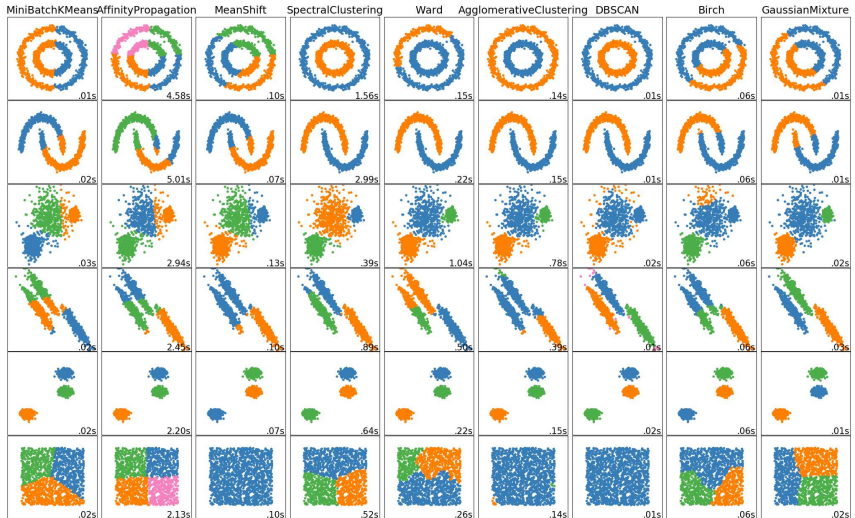
# Requerimentos e Desafios

- ▶ Escalabilidade
  - ▶ Agrupamento de muitos dados
- ▶ Habilidade de lidar com diferentes tipos de atributos
- ▶ Agrupamento com restrições
  - ▶ As restrições são impostas pelo usuário
  - ▶ Utiliza conhecimento do domínio do problema
- ▶ Interpretabilidade e usabilidade
- ▶ Outros
  - ▶ Descoberta de grupos com forma arbitrária
  - ▶ Habilidade de tratar dados com ruído
  - ▶ Agrupamento incremental e insensibilidade à ordem dos dados
  - ▶ Alta dimensionalidade

# Propostas Mais Adotadas

- ▶ **Particionamento**
  - ▶ Várias partições são construídas e depois são avaliadas segundo algum critério; por exemplo, minimizando a soma dos quadrados dos erros
  - ▶ Métodos típicos: k-Médias, k-Medoides, CLARANS
- ▶ **Hierárquica**
  - ▶ Cria uma decomposição hierárquica do conjunto de dados usando algum critério
  - ▶ Métodos típicos: Diana, Agnes, BIRCH, CAMELEON
- ▶ **Baseadas em densidade**
  - ▶ Leva em consideração funções de conectividade de densidade
  - ▶ Métodos típicos: DBSCAN, OPTICS, DenClue
- ▶ **Baseadas em grade**
  - ▶ Organizado numa estrutura de granularidade multinível
  - ▶ Métodos típicos: STING, WaveCluster, CLIQUE

# Agrupamento: Exemplos de Problemas



# Scikit Learn: Agrupamento

Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large <code>n_samples</code> , medium <code>n_clusters</code> with <a href="#">MiniBatch code</a>	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
Affinity propagation	damping, sample preference	Not scalable with <code>n_samples</code>	Many clusters, uneven cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with <code>n_samples</code>	Many clusters, uneven cluster size, non-flat geometry	Distances between points
Spectral clustering	number of clusters	Medium <code>n_samples</code> , small <code>n_clusters</code>	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters	Large <code>n_samples</code> and <code>n_clusters</code>	Many clusters, possibly connectivity constraints	Distances between points
Agglomerative clustering	number of clusters, linkage type, distance	Large <code>n_samples</code> and <code>n_clusters</code>	Many clusters, possibly connectivity constraints, non Euclidean distances	Any pairwise distance
DBSCAN	neighborhood size	Very large <code>n_samples</code> , medium <code>n_clusters</code>	Non-flat geometry, uneven cluster sizes	Distances between nearest points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centers
Birch	branching factor, threshold, optional global clusterer.	Large <code>n_clusters</code> and <code>n_samples</code>	Large dataset, outlier removal, data reduction.	Euclidean distance between points

+ OPTICS

# Scikit Learn: Exemplos Incluídos Aqui

Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large <code>n_samples</code> , medium <code>n_clusters</code> with <a href="#">MiniBatch code</a>	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
Affinity propagation	damping, sample preference	Not scalable with <code>n_samples</code>	Many clusters, uneven cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with <code>n_samples</code>	Many clusters, uneven cluster size, non-flat geometry	Distances between points
Spectral clustering	number of clusters	Medium <code>n_samples</code> , small <code>n_clusters</code>	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters	Large <code>n_samples</code> and <code>n_clusters</code>	Many clusters, possibly connectivity constraints	Distances between points
Agglomerative clustering	number of clusters, linkage type, distance	Large <code>n_samples</code> and <code>n_clusters</code>	Many clusters, possibly connectivity constraints, non Euclidean distances	Any pairwise distance
DBSCAN	neighborhood size	Very large <code>n_samples</code> , medium <code>n_clusters</code>	Non-flat geometry, uneven cluster sizes	Distances between nearest points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centers
Birch	branching factor, threshold, optional global clusterer.	Large <code>n_clusters</code> and <code>n_samples</code>	Large dataset, outlier removal, data reduction.	Euclidean distance between points

# Métodos de Particionamento

# Métodos de Particionamento: Conceito Básico

- ▶ Particionar um banco de dados  $D$  com  $n$  objetos em  $k$  grupos de modo que a soma dos quadrados das distâncias é minimizada como

$$\min E = \sum_{i=1}^k \sum_{p \in C_i} (p - c_i)^2,$$

onde  $c_i$  é o centroide ou medóide do *cluster*  $C_i$

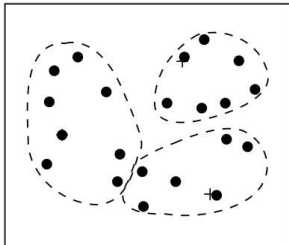
- ▶ Dado  $k$ , encontrar um particionamento de  $k$  grupos que otimizam o critério de particionamento adotado
  - ▶ Solução global: exaustivamente enumera-se todas as partições
  - ▶ Métodos heurísticos: algoritmo k-Médias e k-Medoids
  - ▶ k-Médias: Cada *cluster* é representado pelo seu centro geométrico
  - ▶ k-Medoids: Cada grupos é representado pelo seu objeto central



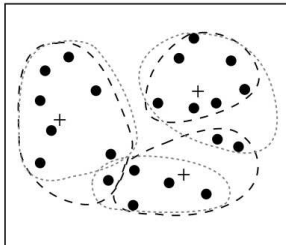
# k-Médias

- ▶ Dado o número de *clusters*  $k$ 
  - ▶ Particiona-se as instâncias em  $k$  conjuntos não vazios
  - ▶ Computa o centro de cada grupo como sendo o ponto médio dos seus elementos
  - ▶ Atribui cada elemento a um *cluster* pela menor distância aos atuais centros
  - ▶ Volta-se ao passo de cômputo dos centros

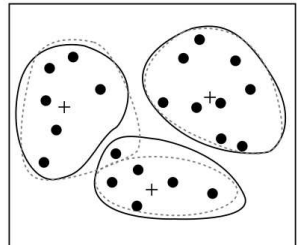
# k-Médias



(a) Initial clustering



(b) Iterate



(c) Final clustering

# k-Médias

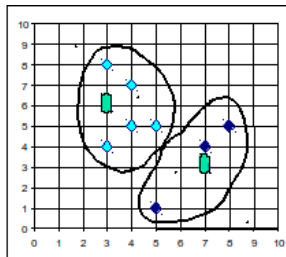
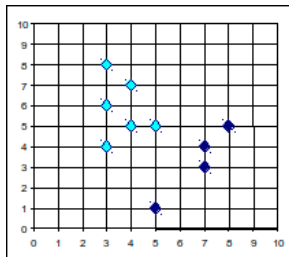
- ▶ Vantagem: eficiência computacional na ordem de  $O(tkn)$ 
  - ▶  $n$  é o número de instâncias,  $k$  é o número de grupos e  $t$  é o número de iterações até a convergência do algoritmo
- ▶ O processo iterativo converge para um mínimo local
- ▶ Fraquezas
  - ▶ Aplicável apenas para o caso contínuo
  - ▶ k-Modas pode ser aplicado ao caso categórico
  - ▶ k-Medoids é mais robusto quando há diferentes tipos de dados
  - ▶ O número de grupos  $k$  deve ser definido *à priori* (apesar de existirem meios de determinar esse parâmetro)
  - ▶ Sensível à ruído e *outliers*
  - ▶ Encontra grupos apenas com organização geométrica convexa

# Variantes do k-Médias

- ▶ A maioria das variantes do k-Médias difere-se em
  - ▶ Seleção dos grupos iniciais
  - ▶ Cálculo da dissimilaridade
  - ▶ Estratégias para cálculo das médias dos grupos
- ▶ Tratamento de dados categóricos: k-Modas
  - ▶ Substitui as médias por modas
  - ▶ Usar novas medidas de dissimilaridades para tratar os dados categóricos
  - ▶ Usar métodos baseados em frequências para atualizar as modas dos grupos
  - ▶ Mistura de dados categóricos e numéricos: k-Protótipos

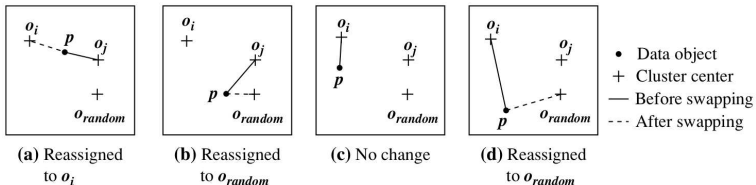
# Variantes do k-Médias

- ▶ O algoritmo k-Médias é sensível à *outliers*
  - ▶ Um objeto com um valor extremamente grande pode distorcer a distribuição dos dados
- ▶ *k-Medoids*
  - ▶ Utiliza o medóide que é o ponto localizado mais ao centro do grupo



# k-Medoids

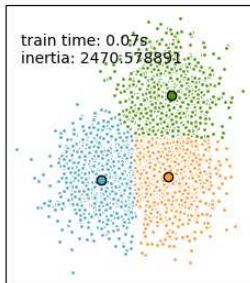
- ▶ Encontra objetos centrais em grupos
  - ▶ PAM (Partition Around Medoids)
  - ▶ Começa com um conjunto inicial de pontos centrais e iterativamente substitui um desses pontos centrais, se houver redução na distância total do grupo resultante
  - ▶ Funciona bem para conjuntos pequenos de dados, mas não escala bem para grandes quantidades de dados



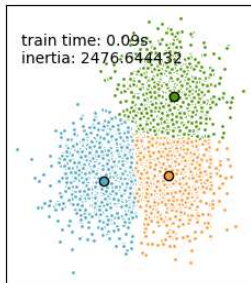
- ▶ Melhora na eficiência do PAM
  - ▶ CLARA: PAM opera sobre amostras
  - ▶ CLARANS: re-amostragem aleatória

## k-Médias / k-Means

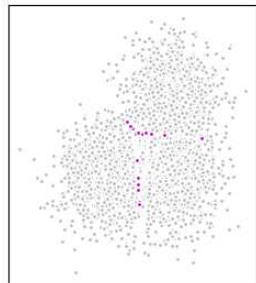
KMeans



MiniBatchKMeans



Difference

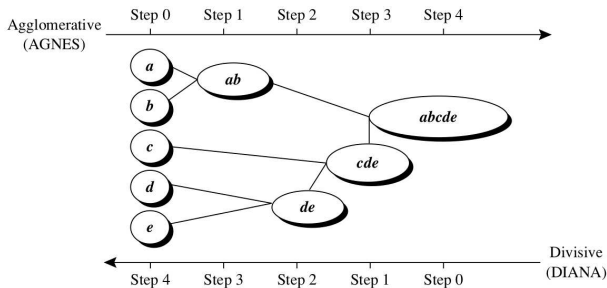


# Métodos Hierárquicos



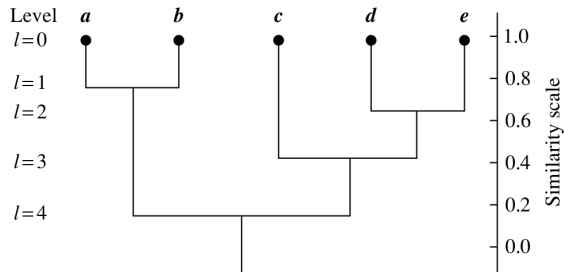
# Agrupamento Hierárquico

- ▶ Usa-se uma matriz de distância como critério de agrupamento
- ▶ Não requer a quantidade  $k$  de grupos, mas requer um critério de parada



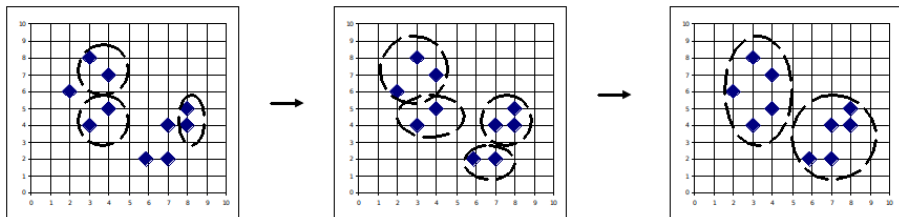
# Agrupamento Hierárquico

Agglomerative and divisive hierarchical clustering on data objects  $\{a, b, c, d, e\}$



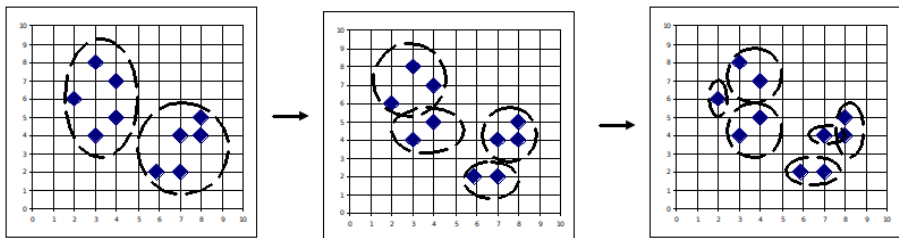
# AGNES (Agglomerative Nesting)

- ▶ Usa o Método de Ligação Única e uma matriz de dissimilaridade
- ▶ Nós que têm uma dissimilaridade mínima são unidos
- ▶ Eventualmente todos os nós pertencem a um único grupo



# DIANA (Divisive Analysis)

- ▶ Opera em ordem inversa do AGNES
- ▶ Eventualmente pode gerar grupos com um único elemento



# Distância entre Grupos

- ▶ Ligação simples
  - ▶ Menor distância entre um elemento de um grupo e um elemento de outro
  - ▶  $dist(C_i, C_j) = \min(t_{ip}, t_{jq})$
- ▶ Ligação completa
  - ▶ Maior distância entre um elemento de um grupo e um elemento de outro
  - ▶  $dist(C_i, C_j) = \max(t_{ip}, t_{jq})$
- ▶ Média
  - ▶ Média das distâncias entre os elementos de um grupo e os elementos de outro
  - ▶  $dist(C_i, C_j) = \text{avg}(t_{ip}, t_{jq})$

# Distância entre Grupos

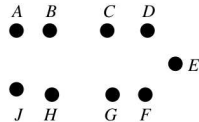
- ▶ Centroide

- ▶ Distância entre os centroides de dois grupos
- ▶  $dist(C_i, C_j) = dist(c_i, c_j)$

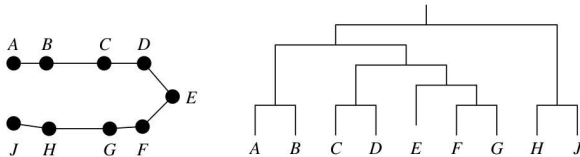
- ▶ Medoid

- ▶ Distância entre os medóides dos dois grupos
- ▶  $dist(C_i, C_j) = dist(m_i, m_j)$ , onde  $m_i$  e  $m_j$  são os elementos mais ao centro dos grupos  $i$  e  $j$ , respectivamente

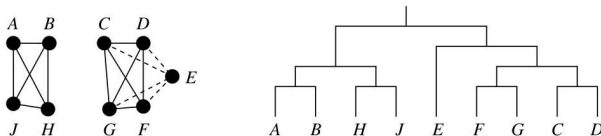
# Agrupamento Hierárquico



(a) Data set



(b) Clustering using single linkage



(c) Clustering using complete linkage

# Extensões para Agrupamentos Hierárquicos

- ▶ Algumas fraquezas dos métodos de agrupamento aglomerativos
  - ▶ Não se pode desfazer o que foi feito antes
  - ▶ Não escala bem: complexidade computacional de ao menos  $O(n^2)$ , onde  $n$  é o número de instâncias



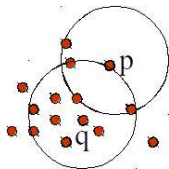
# Métodos Baseados em Densidade

# Métodos Baseados em Densidade

- ▶ Agrupamentos baseados em densidade, tal como pontos conectados por densidade
- ▶ Principais características
  - ▶ Descobrir grupos com forma arbitrária
  - ▶ Trata o ruído
  - ▶ Varredura única
  - ▶ Requer parâmetros de densidade no critério de parada
- ▶ Algumas técnicas
  - ▶ DBSCAN
  - ▶ OPTICS
  - ▶ DENCLUE
  - ▶ CLIQUE

# Conceitos Básicos

- ▶ Parâmetros
  - ▶  $\text{eps}$ : raio máximo de vizinhança
  - ▶  $\text{minPts}$ : quantidade mínima de pontos dentro da vizinhança do ponto
- ▶  $N_{\text{eps}}(p)$ : número de elementos com distância menor ou igual a  $\text{eps}$
- ▶ *Core object*:  $q$  é um *core object* se  $|N_{\text{eps}}(q)| \geq \text{minPts}$
- ▶ Diretamente alcançável pela densidade
  - ▶ Um ponto  $p$  é diretamente alcançável pela densidade de um ponto  $q$  se
    - (i)  $p$  pertence ao  $N_{\text{eps}}(q)$
    - (ii)  $|N_{\text{eps}}(q)| \geq \text{minPts}$

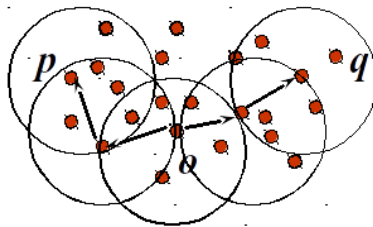
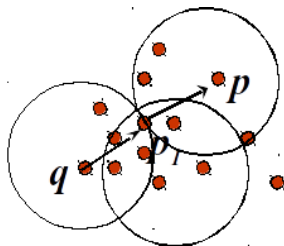


$\text{MinPts} = 5$

$\text{Eps} = 1 \text{ cm}$

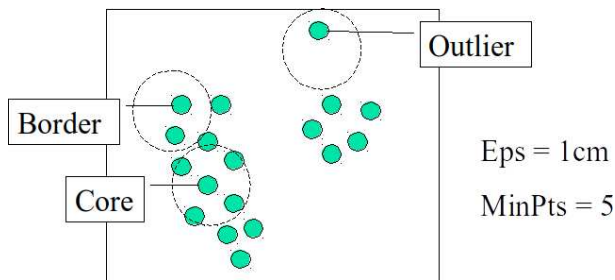
# Alcançável e Conectado pela Densidade

- ▶ Alcançável pela densidade
  - ▶ Um ponto  $p$  é alcançável pela densidade por um ponto  $q$  se há uma sequência de pontos  $q = p_1, \dots, p_n = p$  tal que  $p_{i+1}$  é diretamente alcançável pela densidade por  $p_i$
- ▶ Conectado pela densidade
  - ▶ Dois pontos  $p$  e  $q$  são conectados pela densidade se há um ponto  $o$  tal que  $p$  e  $q$  são alcançáveis pela densidade por ele



# DBSCAN

- ▶ DBSCAN: Density-Based Spatial Clustering of Applications with Noise
- ▶ Um *cluster* é definido como o conjunto máximo de pontos conectados pela densidade
- ▶ Descobre grupos de forma arbitrária e em bases com ruído



# DBSCAN

- ▶ Selecciona um ponto  $p$  arbitrário
- ▶ Verifica-se se  $p$  já foi visitado
- ▶ Se  $p$  é um núcleo
  - ▶ Se  $p$  não pertence a um grupo então um novo grupo é formado
  - ▶ Recupera todos os pontos alcançáveis pela densidade de  $p$  (utilizando  $\epsilon$  e minPts)
  - ▶ Esses pontos são também investigados (conectividade)
- ▶ Se  $p$  é um ponto de borda, então nenhum ponto é alcançável pela densidade de  $p$
- ▶ O processo se repete até que todos os pontos sejam visitados

# DBSCAN

- (1) mark all objects as **unvisited**;
- (2) **do**
- (3)     randomly select an unvisited object  $p$ ;
- (4)     mark  $p$  as **visited**;
- (5)     **if** the  $\epsilon$ -neighborhood of  $p$  has at least  $MinPts$  objects
- (6)         create a new cluster  $C$ , and add  $p$  to  $C$ ;
- (7)         let  $N$  be the set of objects in the  $\epsilon$ -neighborhood of  $p$ ;
- (8)         **for** each point  $p'$  in  $N$
- (9)             **if**  $p'$  is **unvisited**
- (10)                 mark  $p'$  as **visited**;
- (11)                 **if** the  $\epsilon$ -neighborhood of  $p'$  has at least  $MinPts$  points,  
                    add those points to  $N$ ;
- (12)                 **if**  $p'$  is not yet a member of any cluster, add  $p'$  to  $C$ ;
- (13)         **end for**
- (14)         output  $C$ ;
- (15)     **else** mark  $p$  as **noise**;
- (16) **until** no object is **unvisited**;

# DBSCAN: Sensibilidade dos Parâmetros

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.

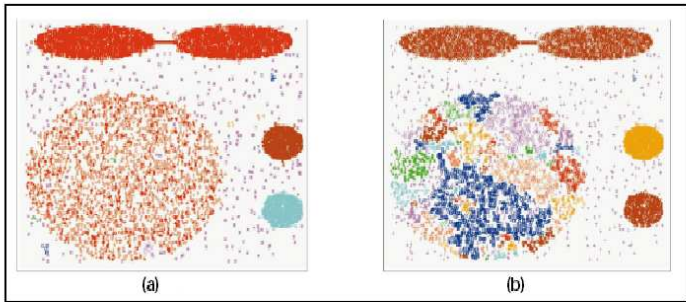
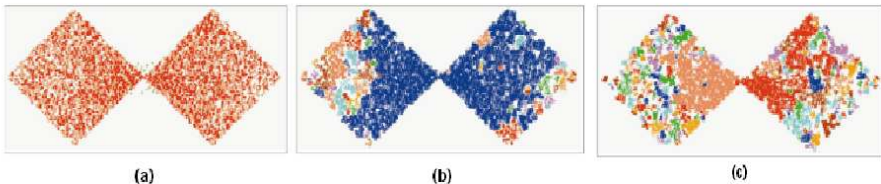


Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.



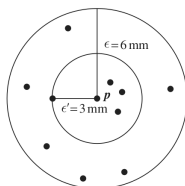
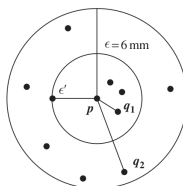


# OPTICS

- ▶ O DBSCAN é muito sensível aos seus parâmetros
- ▶ O Ordering Points To Identify the Clustering Structure (OPTICS) busca diminuir essa dependência
  - ▶ Os parâmetros  $\epsilon$  (aqui maior distância) e minPts ainda são utilizados
- ▶ Nota-se que o agrupamento é monotônico em relação a  $\epsilon$ 
  - ▶ Dado o mesmo MinPts, um grupo gerado com  $\epsilon_2$  é um subgrupo de um gerado com  $\epsilon_1$  quando  $\epsilon_1 < \epsilon_2$
- ▶ Produz uma ordem especial dos dados, que leva em consideração sua estrutura de densidade
- ▶ Os parâmetros da técnica de agrupamento por densidade são determinados com base na estrutura dos dados

## OPTICS

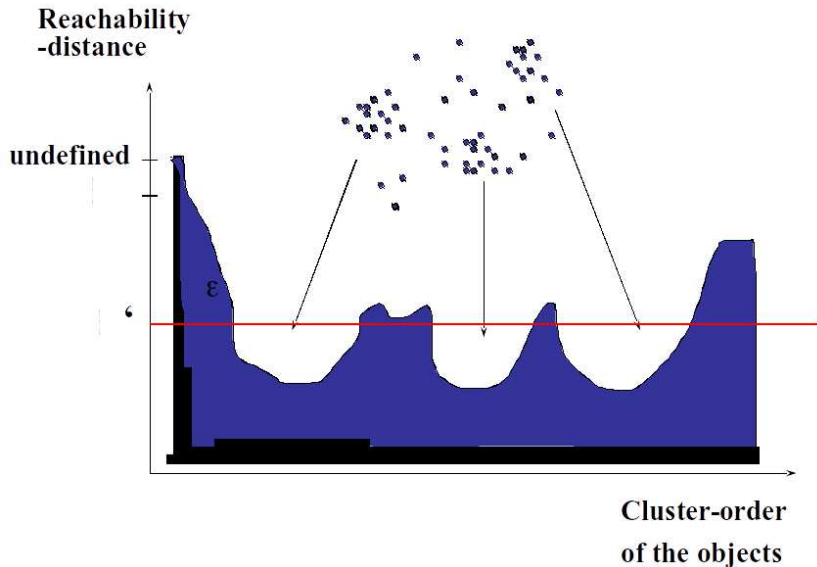
- ▶ Requer as definições de
  - ▶ *core-distance* de  $q$  é o menor valor de  $\epsilon'$  tal que  $N_{eps}(q) \geq \text{MinPts}$
  - ▶ *reachability-distance* de  $q$  para  $p$  é  $\max\{\text{core-distance}(q), \text{dist}(p, q)\}$
  - ▶ Se  $q$  não é um núcleo, então essa distância é indefinida

Core-distance of  $p$ 

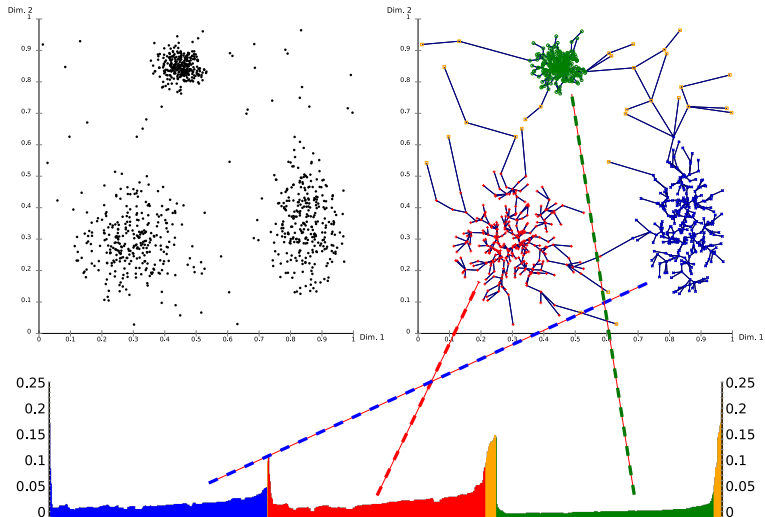
Reachability-distance  $(p, q_1) = \epsilon' = 3 \text{ mm}$   
 Reachability-distance  $(p, q_2) = \text{dist}(p, q_2)$

- ▶ Os objetos são processados numa ordem específica
  - ▶ Objetos com maior  $\epsilon$  aparecem primeiro

## OPTICS



# OPTICS



# Métodos Baseados em Grade


# Métodos Baseados em Grade

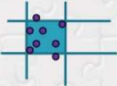
- ▶ Usa uma estrutura de grade multidimensional
- ▶ Alguns métodos
  - ▶ STING: a STatistical INformation Grid approach)
  - ▶ CLIQUE: Clustering In QUEst

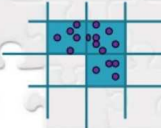
# CLIQUE

- ▶ Pode ser considerado tanto um método de agrupamento baseado em grade quanto em densidade
- ▶ Particiona cada dimensão em intervalos (mesmo número para todas as dimensões) e igualmente espaçados
- ▶ Gera vários retângulos para cada par de dimensões
- ▶ Identifica os subespaços que contém *clusters* usando as densidades de cada dimensão
  - ▶ Identifica unidades densas e as unidades densas conectadas
- ▶ Determina novos grupos pela combinação dos subespaços

# CLIQUE

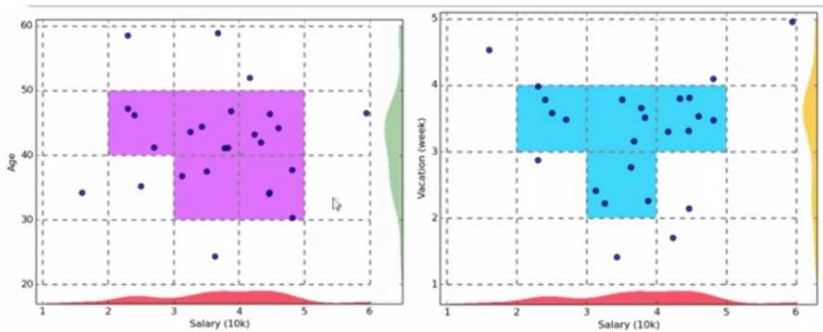
Unit : After forming a grid structure on the space, each rectangular cell is called a Unit. 

Dense: A unit is dense, if the fraction of total data points contained in the unit exceeds the input model parameter. 

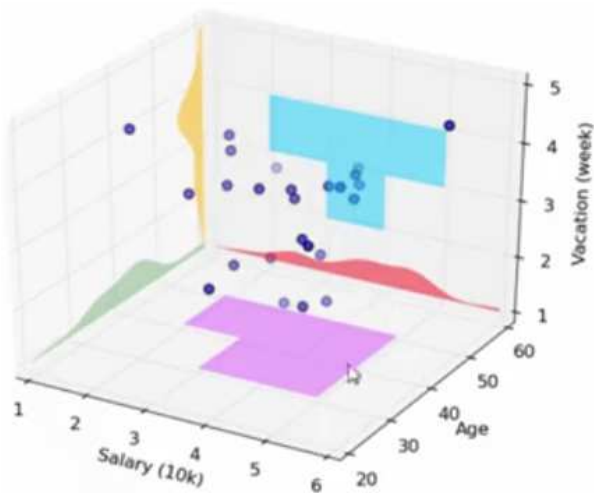
Cluster: A cluster is defined as a maximal set of connected dense units. 



## CLIQUE



## CLIQUE



# Avaliando Agrupamentos

# Tendência do Agrupamento

- ▶ Identificar se uma estrutura não aleatória existe nos dados medindo a probabilidade dos dados serem geradas por uma distribuição uniforme
- ▶ Teste de aleatoriedade espacial via teste estatístico: Estatística de Hopkins
  - ▶ Amostre  $n$  pontos  $p_1, \dots, p_n$  uniformemente de  $D$
  - ▶ Para cada  $p_i$ , encontre o vizinho mais próximo em  $D$ :  
 $x_i = \min dist(p_i, v)$ , onde  $v \in D$
  - ▶ Amostre  $n$  pontos  $q_1, \dots, q_n$  uniformemente no mesmo intervalo de  $D$
  - ▶ Para cada  $q_i$ , encontre o vizinho mais próximo em  $D - \{q_i\}$ :  
 $y_i = \min dist(q_i, v)$ , onde  $v \in D$  e  $v \neq q_i$
  - ▶ Calcule  $H = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i + \sum_{i=1}^n y_i}$
  - ▶ Quando  $H \approx 1$  então o conjunto de dados é considerado agrupável

# Determinar o Número de Grupos

- ▶ Método empírico
  - ▶ Número de grupos  $\approx \sqrt{n}/2$  para um banco de dados de  $n$  pontos
- ▶ Método de Elbow
  - ▶ Use o ponto em que há mudança de tendência da variância da soma intragrupo
- ▶ Validação cruzada
  - ▶ Divide-se um conjunto de dados em  $m$  partes
  - ▶  $m - 1$  partes são usadas para obter um modelo de agrupamento
  - ▶ Os demais dados são usados para testar a qualidade do agrupamento
  - ▶ A qualidade do modelo é avaliada pela soma dos quadrados das distâncias entre os pontos de teste e o centroide mais próximo
  - ▶ Para todo  $k > 0$ , o processo é repetido  $m$  vezes e o melhor  $k$  é assumido como aquele que apresenta os melhores resultados

# Qualidade do Agrupamento

- ▶ Extrínsecos
  - ▶ Avalia o modelo com base em alguma supervisão
- ▶ Intrínsecos
  - ▶ Avalia o modelo considerando: separação entre os grupos e compactação dos grupos

## Método Extrínseco

- ▶ Índice aleatório ajustado (ARI): considera todos os pares de amostras entre dois agrupamentos e avalia quais estão no mesmo grupo
  - ▶ `adjusted_rand_score`
  - ▶ Vale 1 quando os agrupamentos são idênticos
- ▶ Homogeneidade dos grupos: avalia se os dados de um grupo pertencem a uma única classe
- ▶ Completude dos grupos: avalia se os dados de uma classe são elementos do mesmo grupo
- ▶ Medida-V: média harmônica entre homogeneidade e completude

## Coeficiente da Silhueta

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ se } |C_i| > 1, \text{ e } 0, \text{ caso contrário}$$

- ▶ Calcula-se a média dos  $s(i)$  para todos os dados
- ▶ Comum na escolha do número de grupos (maior média)



# Avaliação de Modelos de Agrupamento

