Agrupamento

ACH5504 – Mineração de Dados

Notas de aulas baseadas no livro

"*Introduction to Data Mining*"

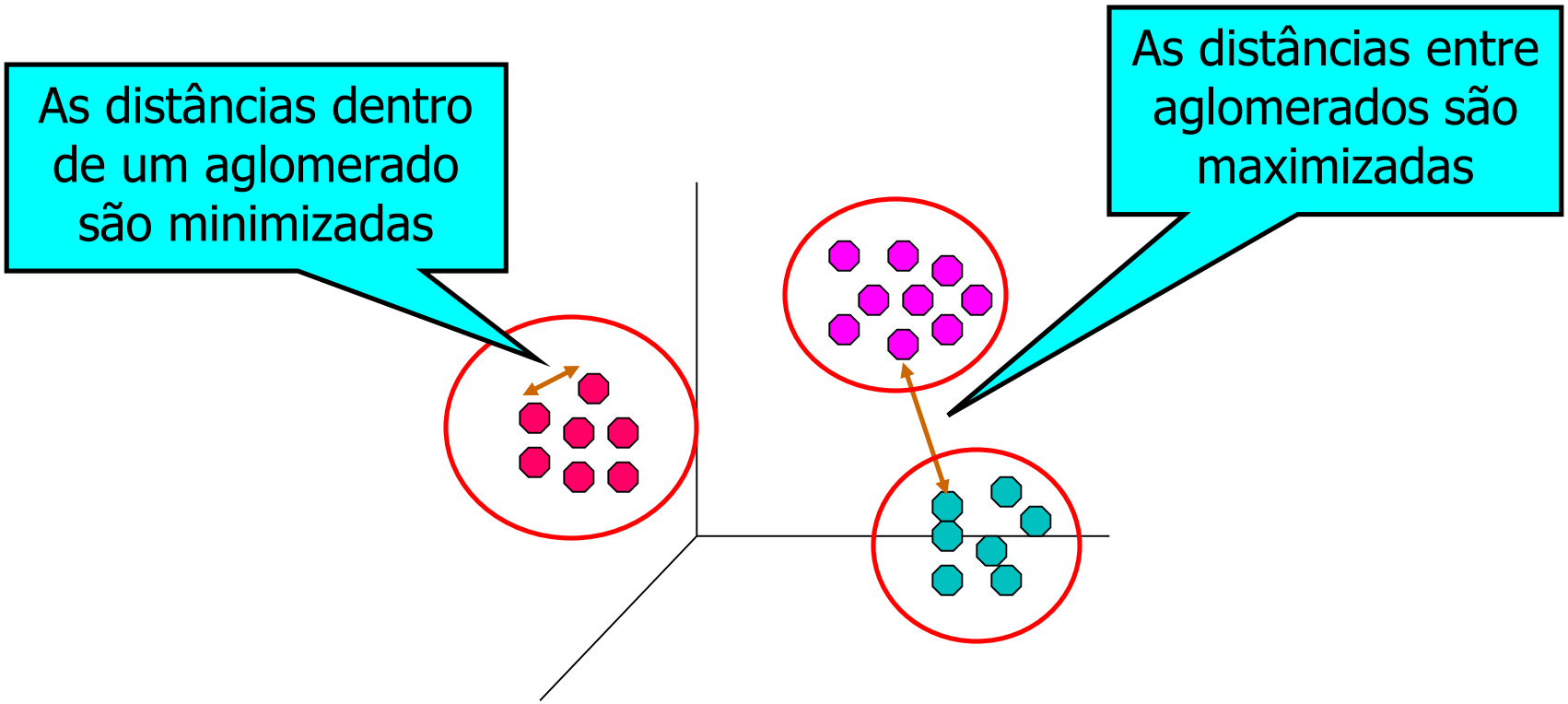
Tan, Steinbach, Karpatne, Kumar

Resumo

- Definição de agrupamento
- Tipos de clusters/aglomerados
- Algoritmos de agrupamento
  - K-means
  - Agrupamento hierárquico
  - Agrupamento por densidade
- Avaliação de agrupamento

O que é análise de agrupamento?

- Localizando grupos de objetos de tal forma que os objetos em um grupo serão semelhantes (ou relacionados) entre si e diferentes de (ou não relacionados a) dos objetos em outros grupos



Agrupamento baseado em densidade

- Clusters são regiões do espaço com alta densidade separados por regiões de baixa densidade.
- Principais características
  - Descoberta de clusters de forma arbitrária
  - Lida bem com ruído
  - Uma passagem pelos dados
  - Precisa de parâmetros de densidade como condição de término

Agrupamento baseado em densidade

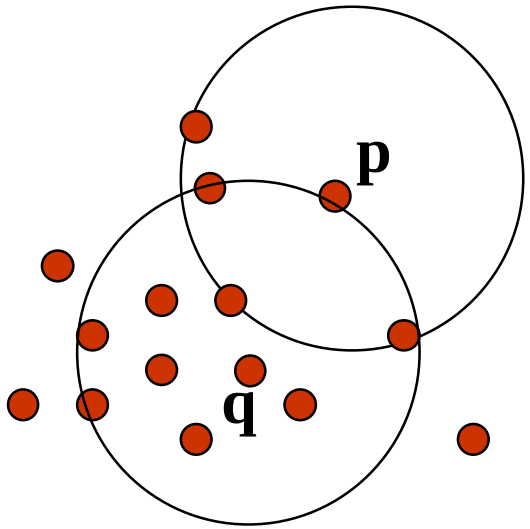
- **DBSCAN**: Density-Based Clustering Based on Connected Regions with High Density
  - Proposto pelo Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu em 1996
  - Para encontrar regiões densas, busca pontos centrais, pontos com densa vizinhança, com muitos pontos próximos.
  - A densidade de um ponto é medido pelo número de pontos próximos, dentro de uma vizinhança.
  - Para determinar a vizinhança, é necessário especificar um parâmetro que indica o raio da vizinhança.

DBSCAN

- **Densidade** – número de pontos dentro de um raio especificado (Eps)
- Um ponto é um ponto central (core point) se ele tem mais que uma quantidade especificada de pontos (MinPts) dentro de Eps
  - Estes são pontos que estão no interior de um cluster
  - Conta o próprio ponto
- Um ponto de borda tem menos do que MinPts dentro de Eps, mas está na vizinhança de um ponto central.
- Um ponto de ruído é qualquer ponto que não é um ponto central nem um ponto de borda.

DBSCAN

- Um ponto *p* é <u>diretamente alcançável por densidade</u> a partir de um ponto central *q* se *p* está dentro do raio Eps a partir de *q*.
  - *q* deve ser ponto central (vizinhos > MinPts)

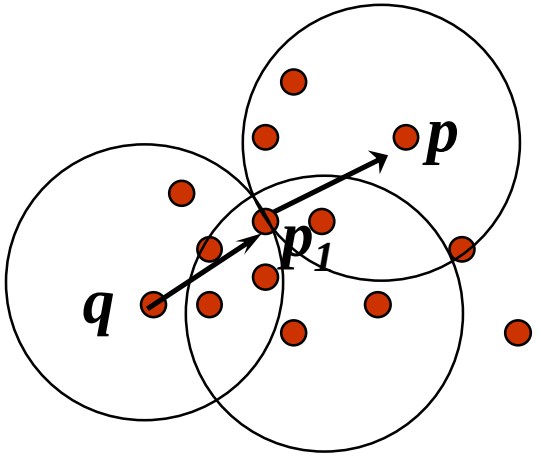


MinPts = 5

Eps = 1 cm

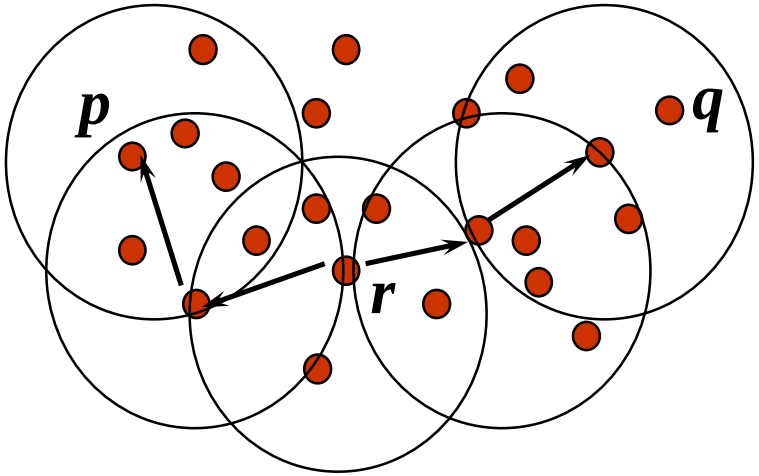
DBSCAN

- Um ponto p é alcançável por densidade a partir de um ponto central q se há uma sequencia de pontos $p_1, \ldots, p_n$, $p_1 = p, p_n = q$ tal que p_{i+1} é diretamente alcançável por densidade de p_i.
  - q deve ser ponto central
  - relação não simétrica

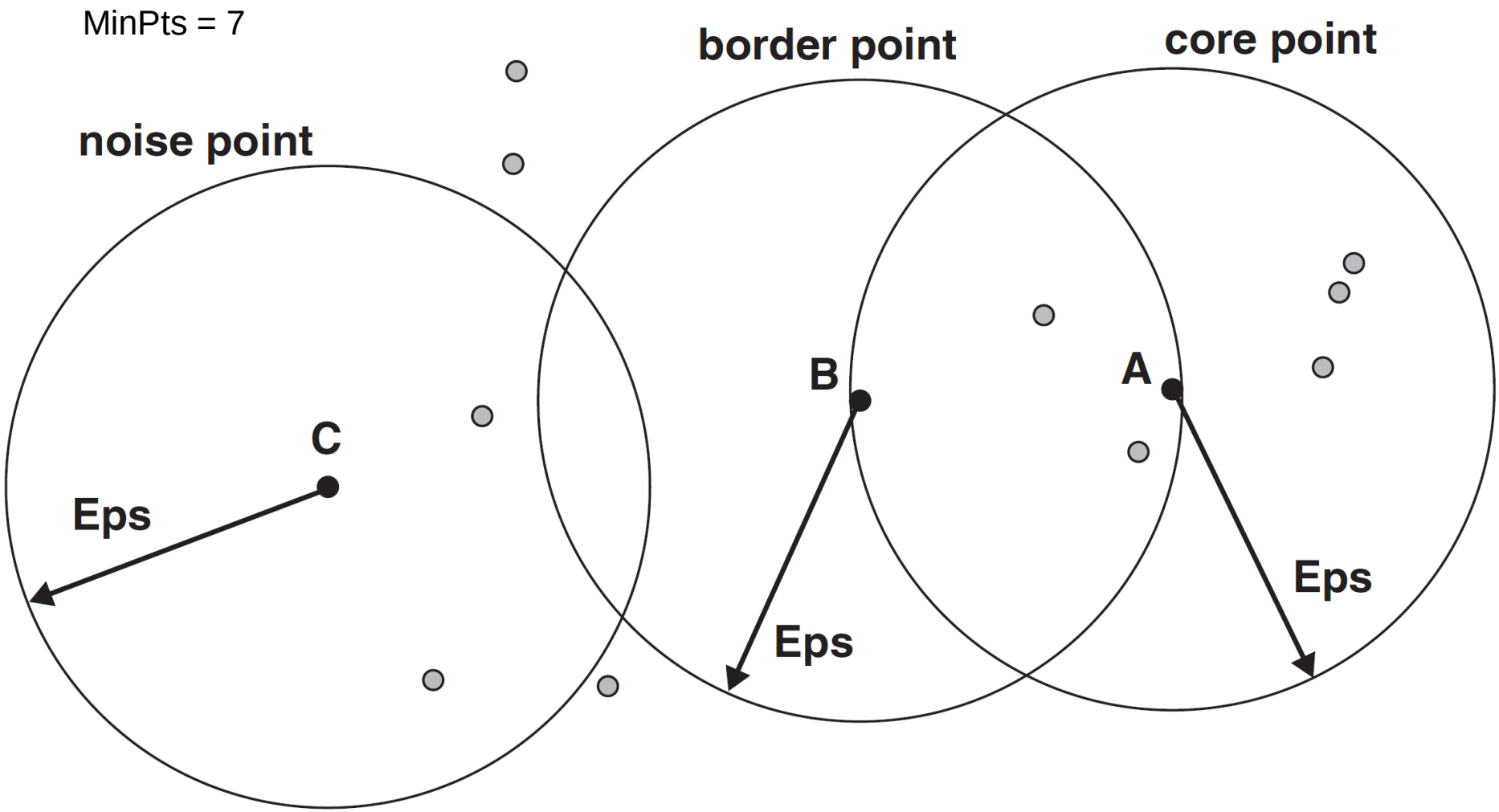


DBSCAN

- Um ponto *q* é conectado por densidade a um ponto partir *p* se há um ponto *r* tal que ambos *p* e *q* são alcançáveis por densidade a partir de *r*



DBSCAN: Ponto central, de borda, e de ruído



DBSCAN: Algoritmo

- Eliminar pontos de ruído
- Realizar agrupamento nos pontos restantes

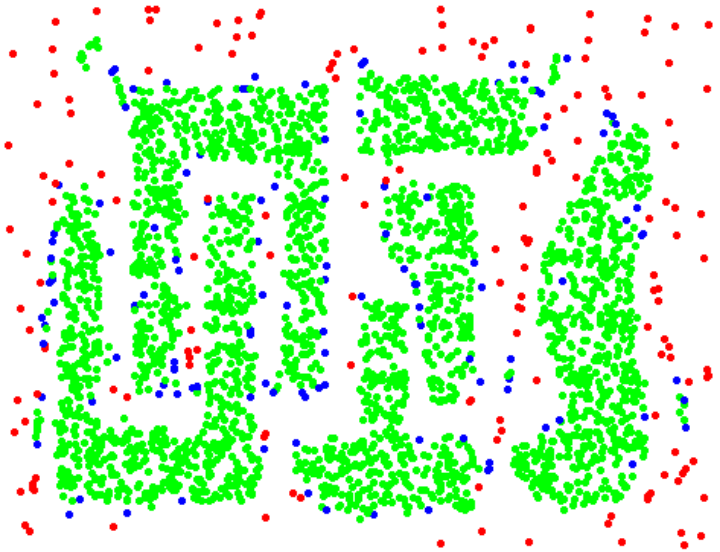
```
current_cluster_label ← 1
for all core points do
  if the core point has no cluster label then
    current_cluster_label ← current_cluster_label + 1
    Label the current core point with cluster label current_cluster_label
  end if
  for all points in the Eps-neighborhood, except i^th the point itself do
    if the point does not have a cluster label then
      Label the point with cluster label current_cluster_label
    end if
  end for
end for
```

DBSCAN: Algoritmo

- Um cluster é um conjunto de pontos que são mutualmente conectados por densidade.
- Se um ponto é conectado por densidade a partir de um ponto em um cluster então ele faz parte do cluster também.
- Um ponto de borda é um ponto que está na vizinhança de um ponto central.
- Um ponto de ruído é um ponto que não é central nem borda.

DBSCAN: Pontos centrais, de borda e de ruído

Pontos originais

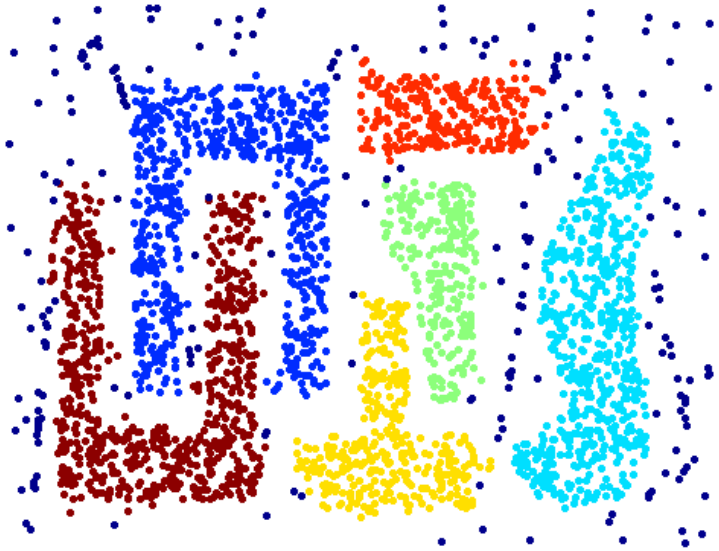


Tipos de ponto: central, de borda e de ruído

Eps = 10, MinPts = 4

Quando DBSCAN funciona bem

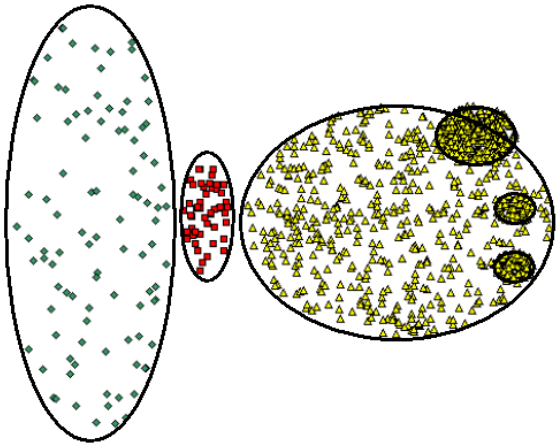
Pontos originais



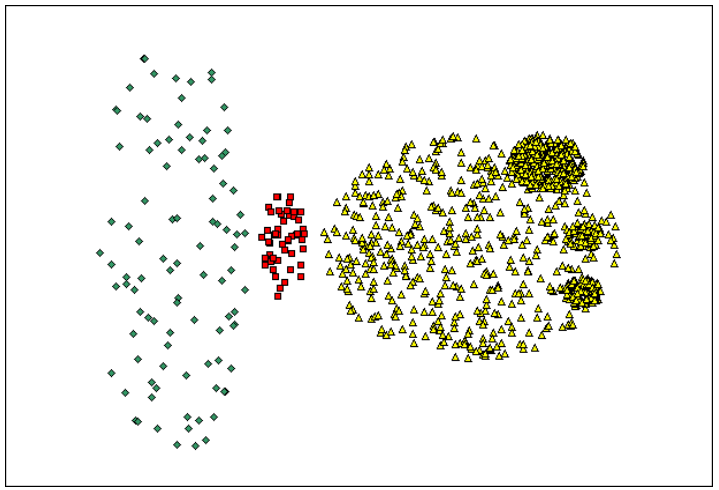
Clusters

- Resistencia a ruído
- Pode lidar com clusters de diferentes formas e tamanhos

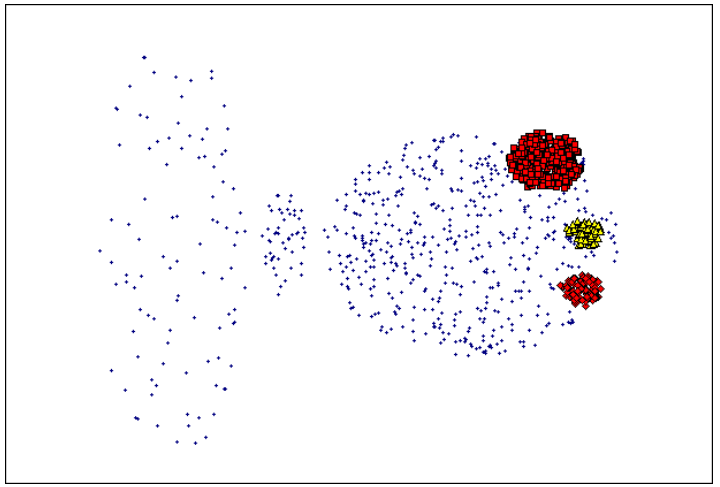
Quando DBSCAN não funciona bem



Pontos originais



(MinPts=4, Eps=9.75).



(MinPts=4, Eps=9.92)

- Densidade variável
- Dados de alta dimensionalidade

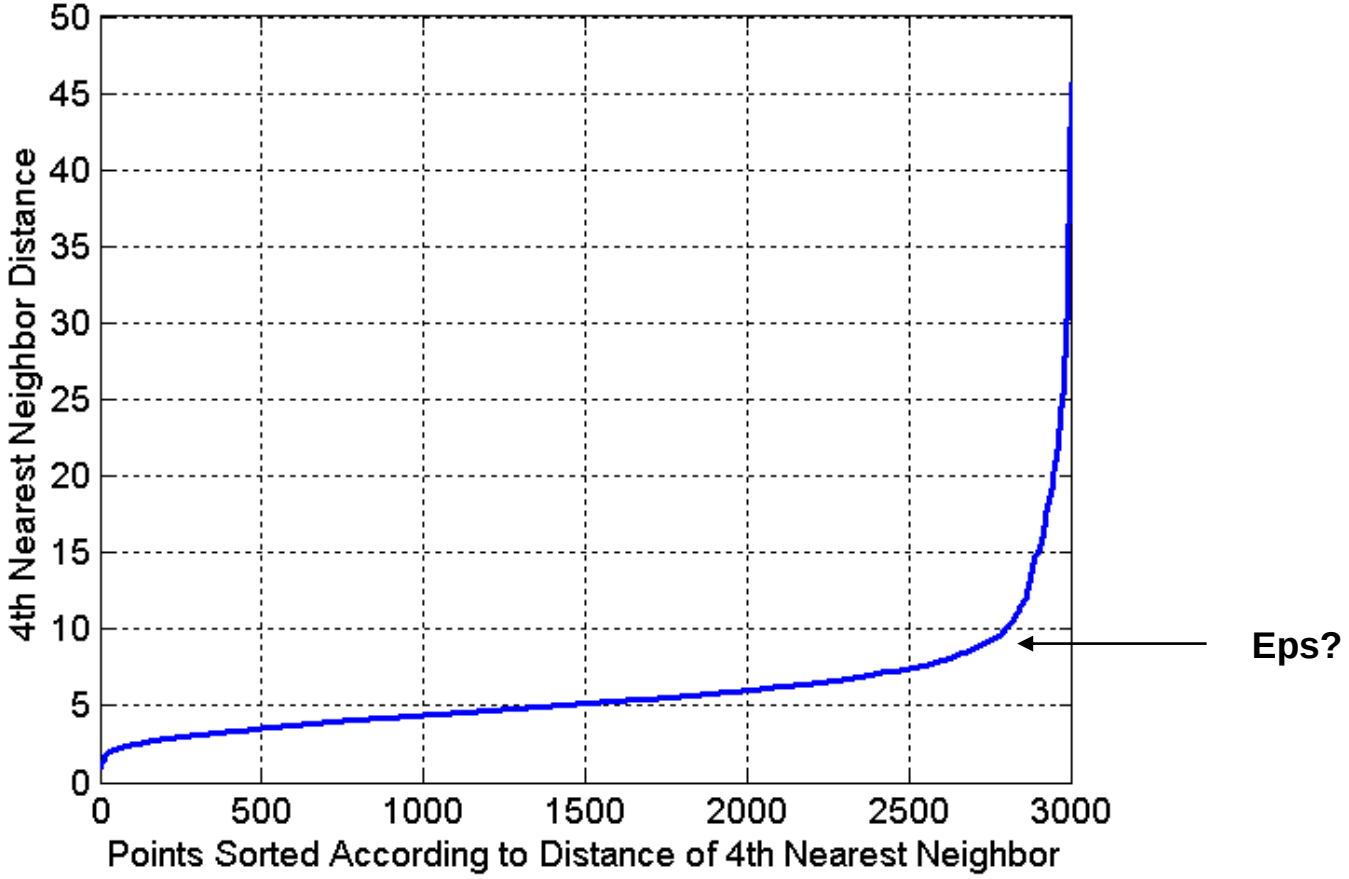
DBSCAN: Escolha de Eps e MinPts

- **MinPts**: MinPts > número de dimensões + 1, e mais dados, maior MinPts

- A ideia é que para pontos dentro de um cluster, o seus k-ésimos vizinhos mais próximos estão aproximadamente na mesma distância
- Pontos de ruído têm o k-ésimo vizinho em uma distância maior

DBSCAN: Escolha de Eps e MinPts

- Faça o gráfico da distância de cada ponto para seu k-ésimo vizinho

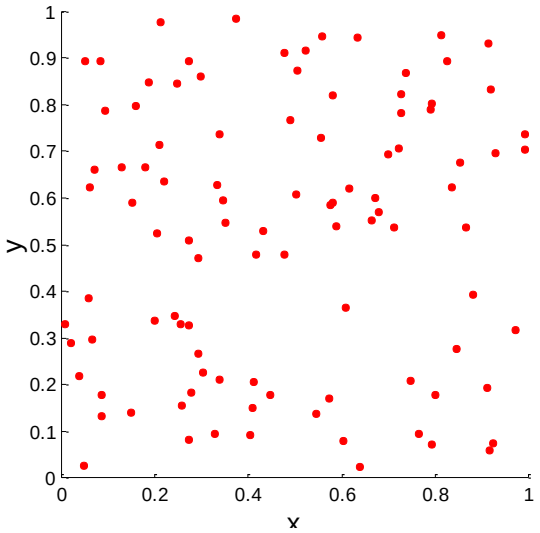


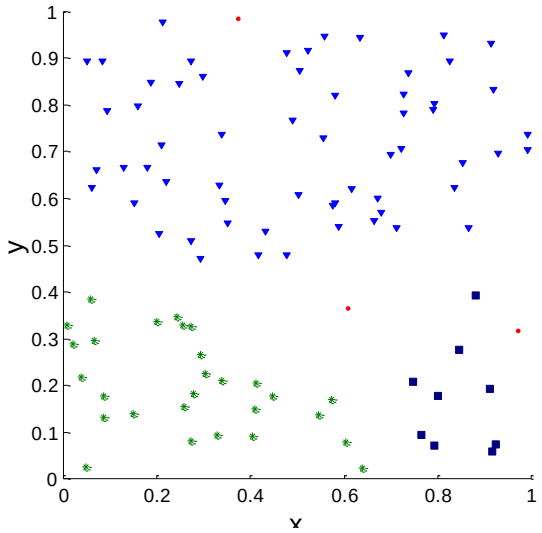
Avaliação de Clusters

- Para classificação, temos várias medidas para avaliar nosso modelo
  - Acurácia, precisão, recall
- Para análise de cluster, como avaliar o quão bom é nosso resultado de agrupamento?
- Geralmente, envolve análise de um especialista
- Porque então queremos avaliar?
  - Para evitar encontrar padrões no ruído.
  - Para comparar algoritmos de agrupamento.
  - Para comparar dois conjuntos de clusters.
  - Para comparar dois clusters.

Clusters encontrados nos dados aleatórios

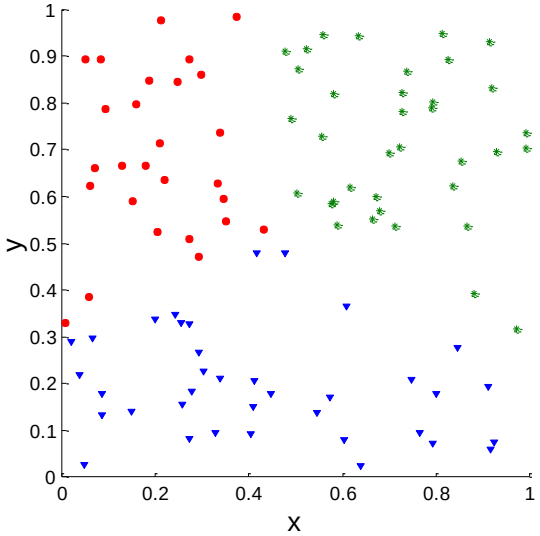
Pontos aleatórios

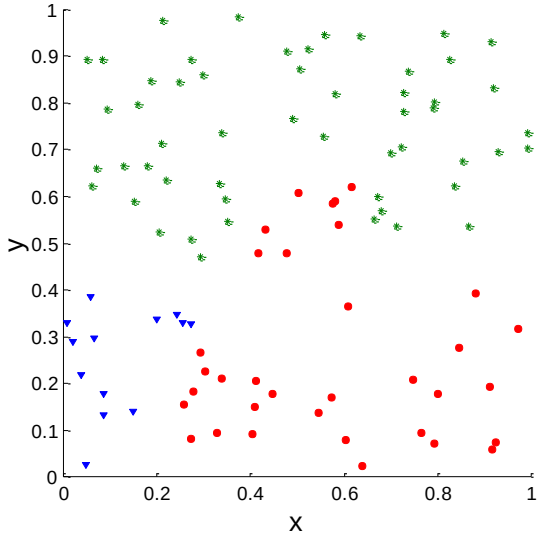




DBSCAN

K-means





Complete Link

Diferentes aspectos da validação de clusters

1. Determinando a tendência de agrupamento de um conjunto de dados, ou seja, distinguir se a estrutura não aleatória realmente existe nos dados.
2. Comparando os resultados de uma análise de clusters com resultados conhecidos externamente, por exemplo, para rótulos de classe fornecidos externamente.
3. Avaliando o quão bem os resultados de uma análise de cluster se encaixam nos dados *sem* referência a informações externas.
   - Utilize apenas os dados
4. Comparando os resultados de dois conjuntos diferentes de análises de cluster para determinar qual é melhor.
5. Determinando o número *'correto'* de clusters.

Para 2, 3 e 4, podemos ainda distinguir se queremos avaliar todo o agrupamento ou apenas clusters individuais.

Como determinar o número correto de clusters?

- Determinação empírica: $k \approx \sqrt{n/2}$, onde *n* é tamanho da base de dados
- Cotovelo (elbow): use a curva de virada da soma da variância intra-cluster
- Validação cruzada: use o conjunto de teste para avaliar a qualidade do agrupamento

Medidas de validade de clusters

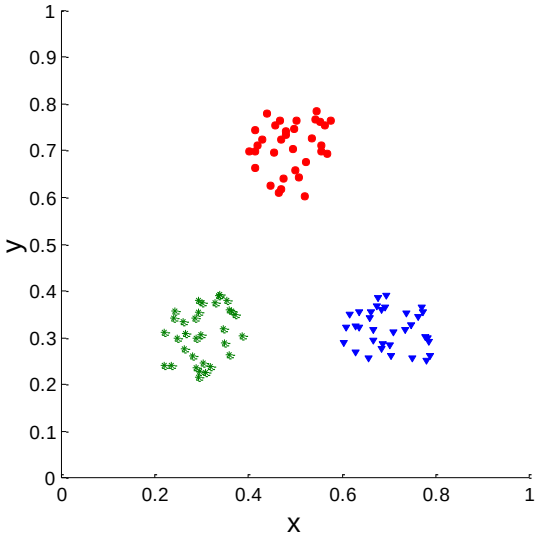
- As medidas numéricas que são aplicadas para julgar vários aspectos da validade de clusters, são classificadas nos três tipos:
  - Índice externo: Usado para medir a medida em que os rótulos de cluster correspondem a rótulos de classe fornecidos externamente.
    - Entropia
  - Índice interno: Usado para medir o quão bom é uma estrutura de agrupamento (o quão separados e compactos estão os clusters) *sem* respeito à informação externa.
    - Soma de erro quadrado (SSE)
  - Índice relativo: Usado para comparar diretamente diferentes agrupamentos, normalmente resultados de diferentes parâmetros do mesmo algoritmo.
    - Um índice externo ou interno é usado para essa função, e.g., SSE ou entropia
- Às vezes, esses são chamados de critérios em vez de índices
  - No entanto, às vezes o critério é a estratégia geral e o índice é a medida numérica que implementa o critério.
  -

Avaliação do agrupamento por correlação

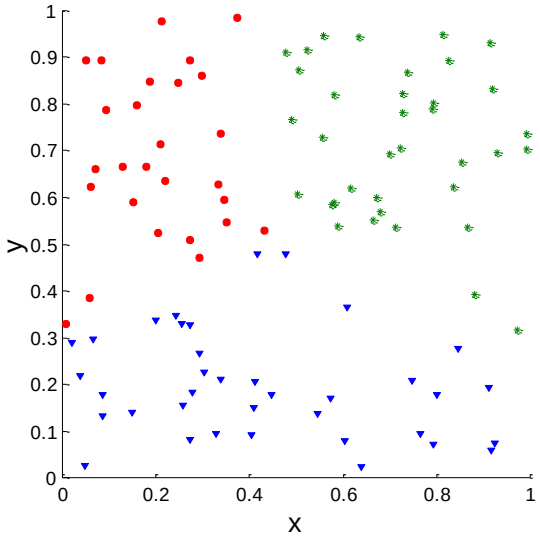
- Duas matrizes
  - Matriz de proximidade
  - Matriz de “incidência”
    - Uma linha e uma coluna para cada instância
    - Se o par de pontos está no mesmo cluster, valor 1
    - Se o par de pontos não está no mesmo cluster, valor 0
- Calcule a correlação entre as duas matrizes
  - São matrizes simétricas então use só parte delas: n(n-1) / 2 elementos.
- Alta correlação indica que pontos do mesmo cluster estão próximos entre si.
- Não é uma boa medida para alguns cluster baseados em densidade ou continuidade.

Avaliação do agrupamento por correlação

- Correlação de matrizes de similaridade ideal e proximidade para agrupamentos K-means de dois conjuntos de dados.



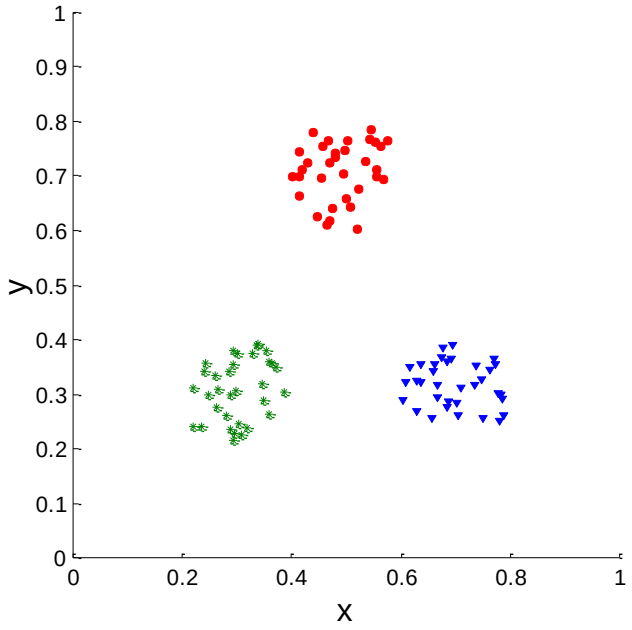
Corr = -0.9235

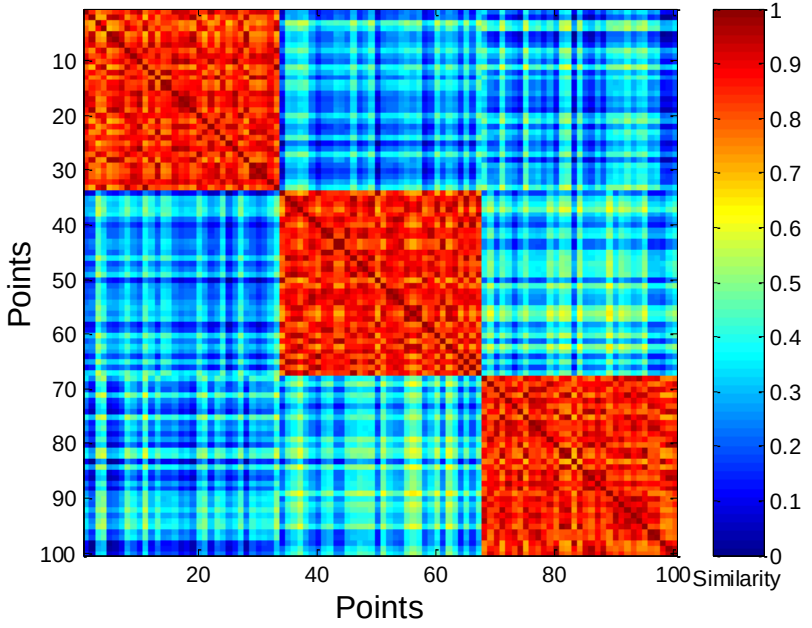


Corr = -0.5810

Usando a matriz de similaridade para avaliação

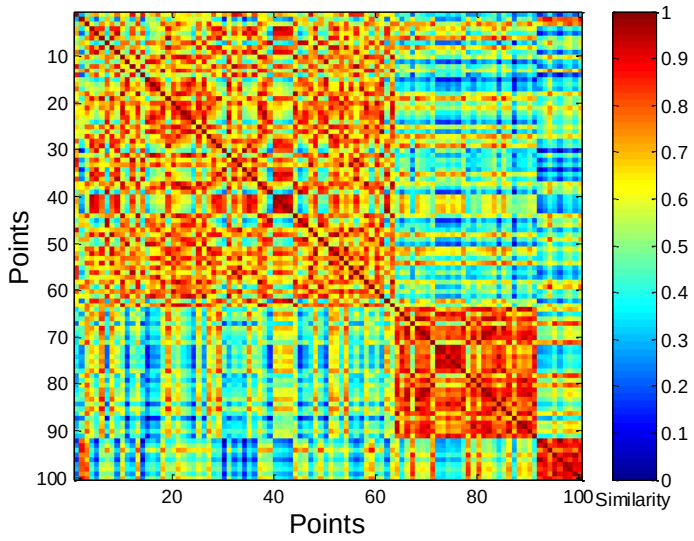
- Ordene a matriz de similaridade com base nos rótulos dos clusters e inspecione visualmente.

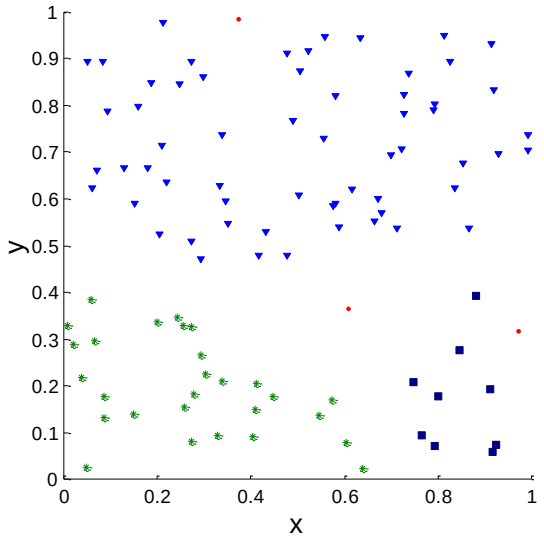




Usando a matriz de similaridade para avaliação

- Clusters em dados aleatórios não são tão nítidos.

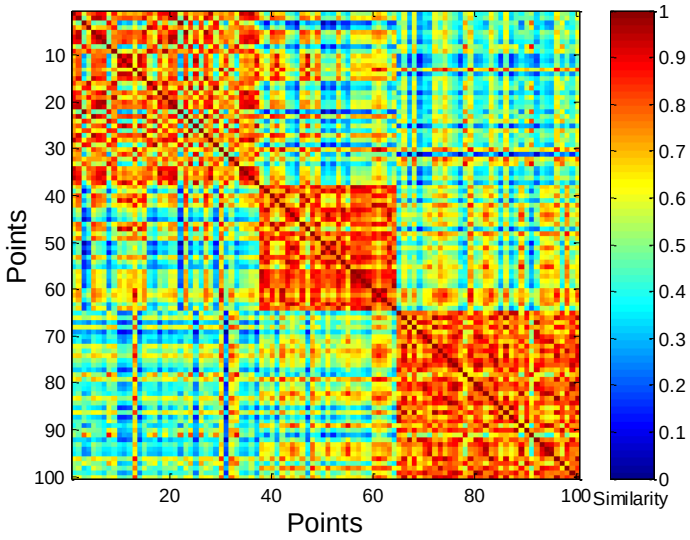


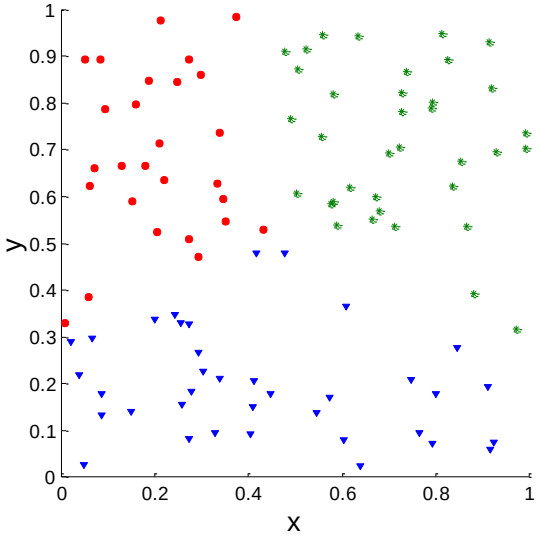


DBSCAN

Usando a matriz de similaridade para avaliação

- Clusters em dados aleatórios não são tão nítidos.

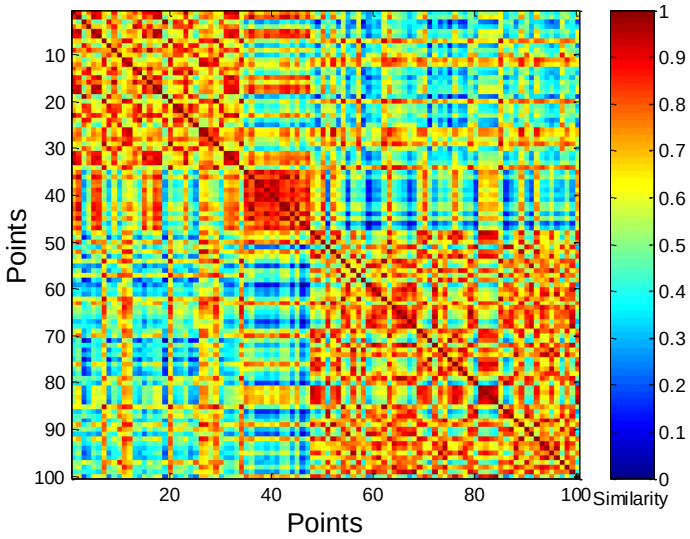


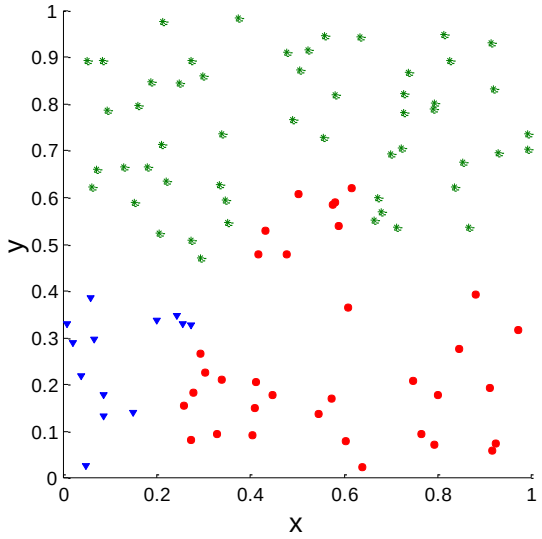


K-means

Usando a matriz de similaridade para avaliação

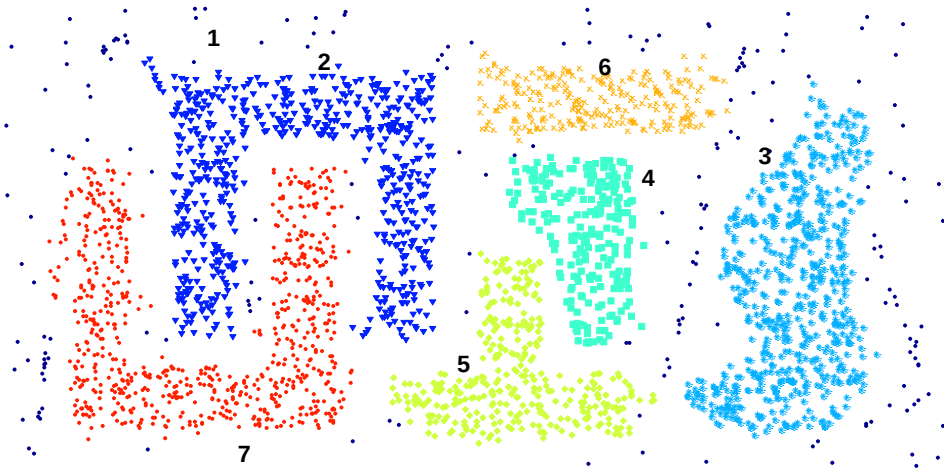
- Clusters em dados aleatórios não são tão nítidos.

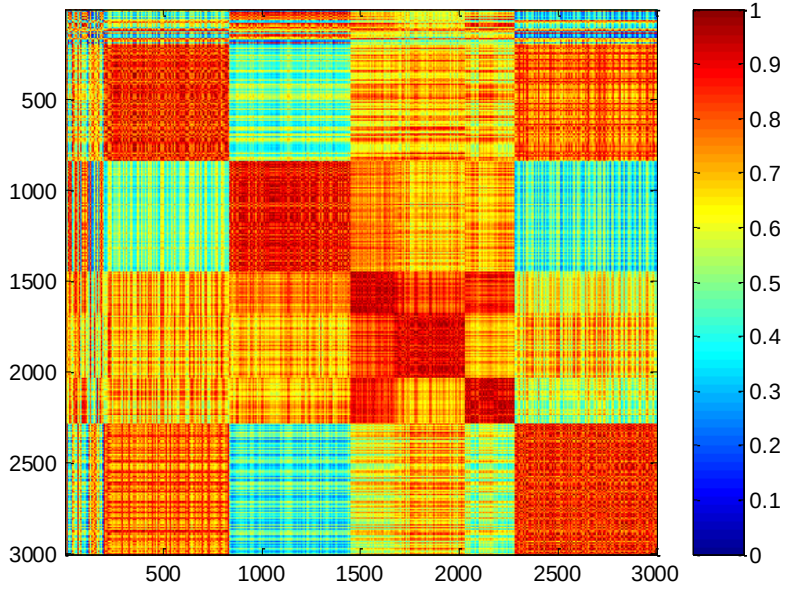




Complete Link

Usando a matriz de similaridade para avaliação

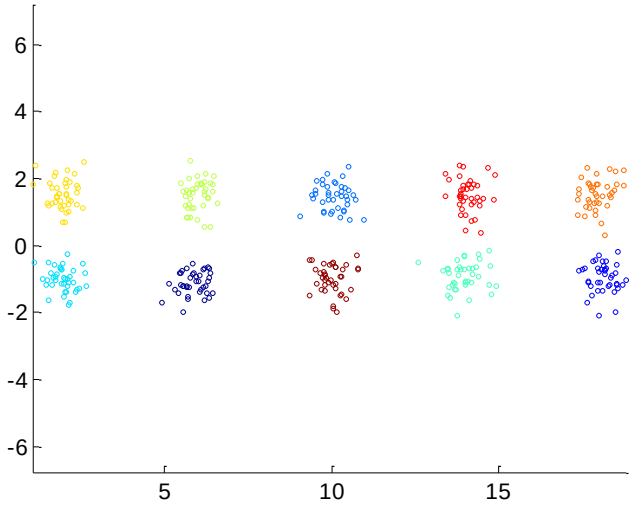


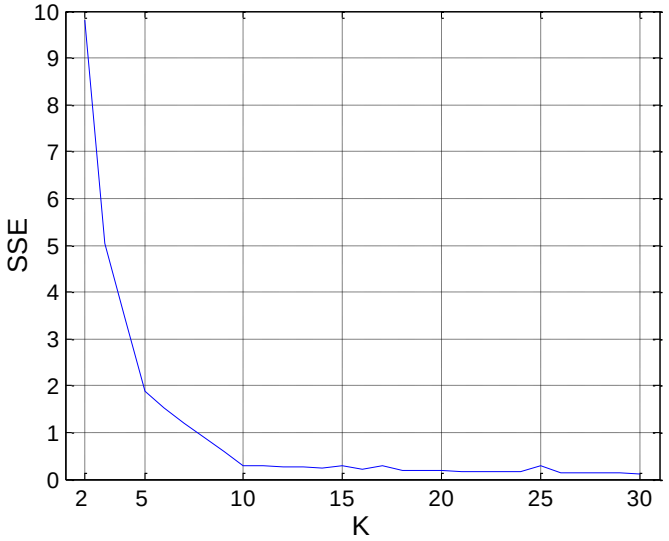


DBSCAN

Avaliação Interna: SSE

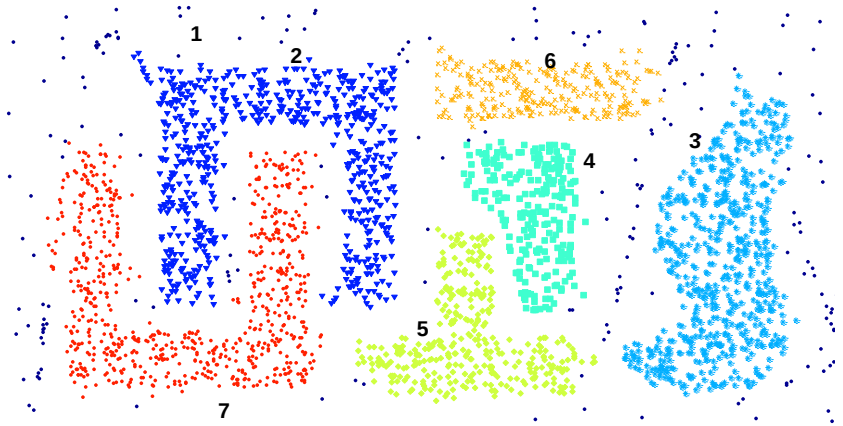
- Clusters em figuras mais complicadas não estão bem separados.
- Índice interno: usado para medir quão bom é uma estrutura de agrupamento sem respeito à informação externa usando SSE.
- SSE é bom para comparar dois agrupamentos ou dois clusters (média de SSE).
- Também pode ser usado para estimar o número de clusters

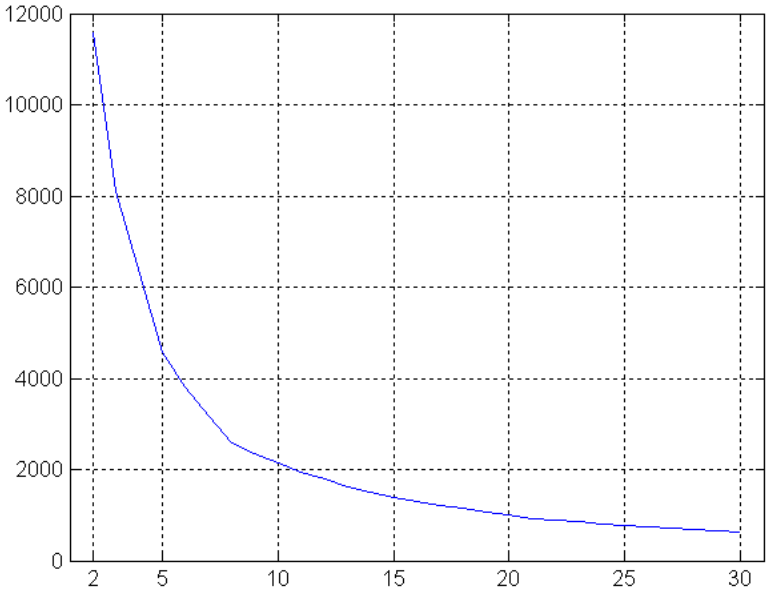




Avaliação Interna: SSE

- Curva SSE para um conjunto de dados mais





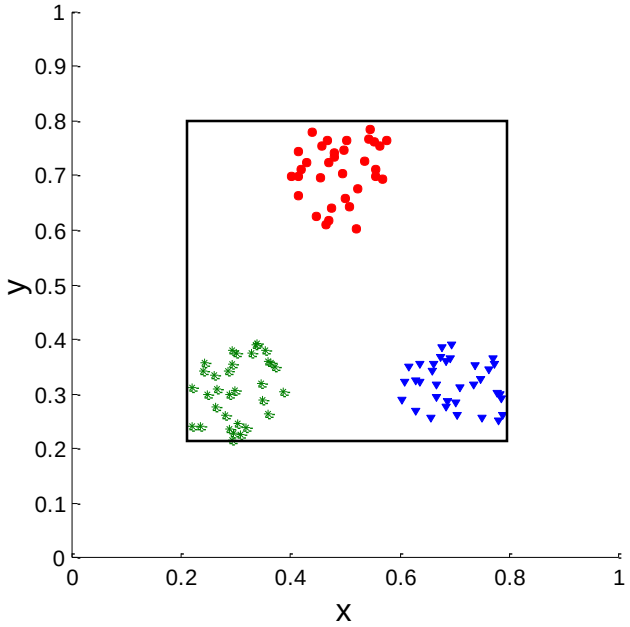
SSE de clusters encontrados usando K-means

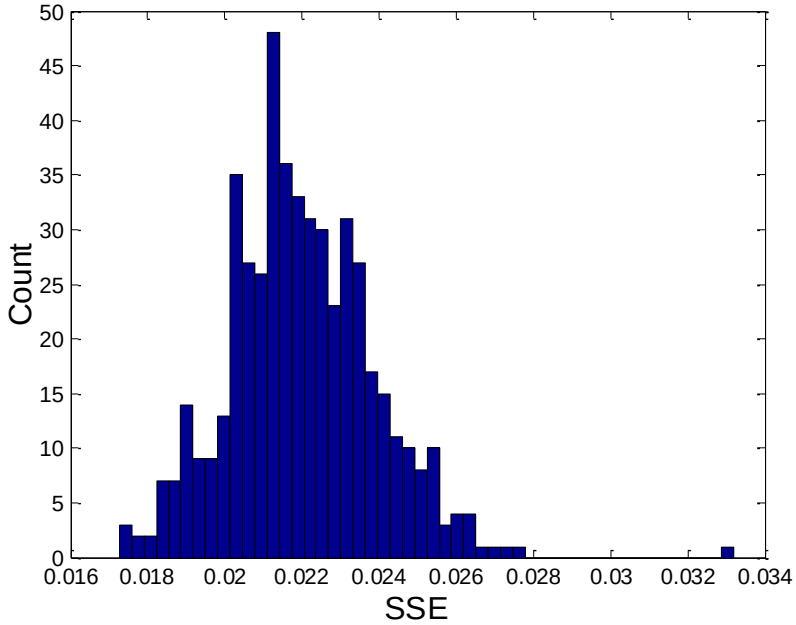
Framework para avaliação de agrupamento

- Precisamos uma ferramenta para interpretar qualquer medida.
  - Por exemplo, se a nossa medida de avaliação tem o valor, 10, isso é bom, justo, ou ruim?
- A estatística fornece uma estrutura para avaliação de clusters
  - Quanto mais "atípico" um resultado de agrupamento é, mais provável que ele representa a estrutura válida nos dados
  - Podemos comparar os valores de um índice para dados aleatórios com os de um resultado de agrupamento.
    - Se o valor do índice for improvável, os resultados do cluster serão válidos
  - Essas abordagens são mais complicadas e mais difíceis de entender.
- Para comparar os resultados de dois agrupamentos diferentes, um framework é menos necessário.
  - No entanto, há a questão de saber se a diferença entre dois valores de índice é significativa

Framework estatístico para SSE

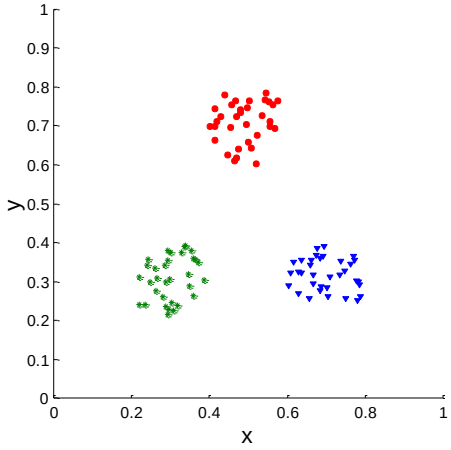
- Exemplo
  - Compare SSE de 0.005 contra três clusters em dados aleatórios
  - Histograma mostra o SSE para três clusters de 500 conjuntos de dados aleatórios distribuídos ao longo do intervalo 0.2 – 0.8 de x e valores y

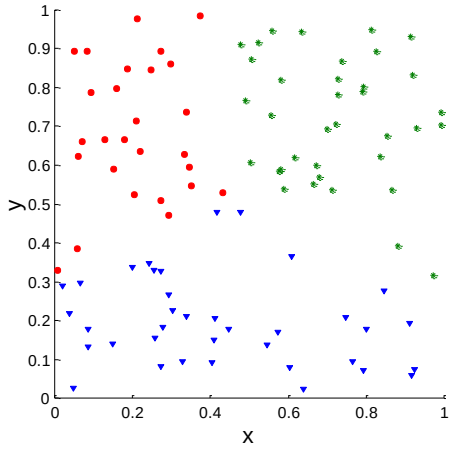


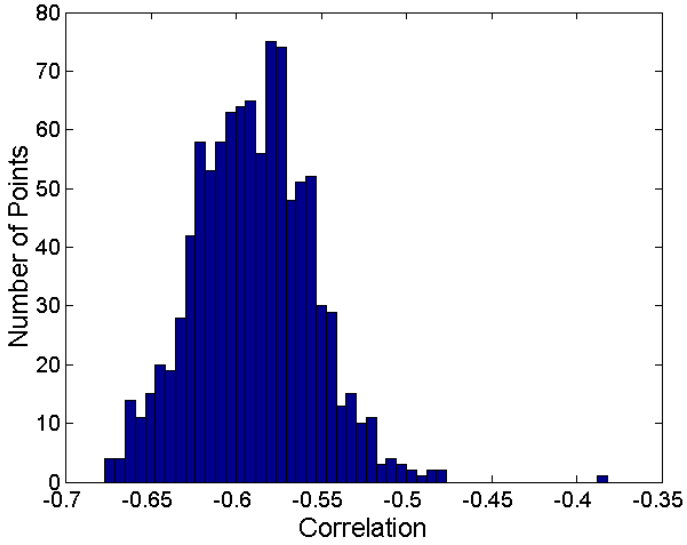


Framework estatístico para SSE

- Correlação de matrizes de similaridade ideal e de proximidade para agrupamento K-means de dois conjuntos de dados.







Corr = -0.9235

Corr = -0.5810

Avaliação Interna: Coesão e Separação

- Coesão do cluster: quão próximos estão os objetos dentro de um cluster
  - Exemplo: SSE
- Separação dos clusters: quão separados são os clusters
- Exemple: Erro quadrático
  - Coesão é medida pelo soma dos quadrados dentro do cluster (SSE)

$$SSE = WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)^2$$

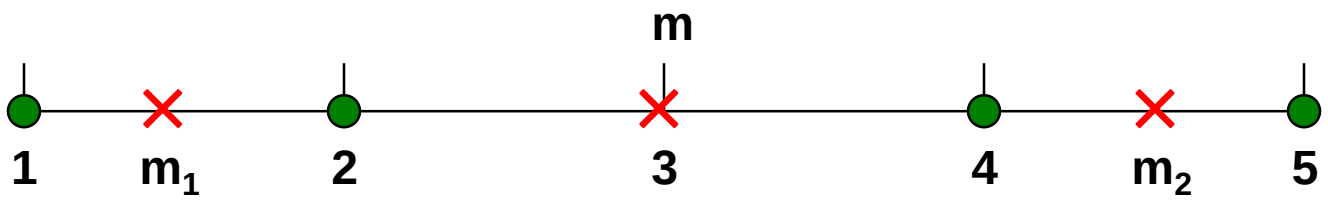
  - Separação é medida pela soma dos quadrados entre clusters

$$BSS = \sum_{i} |C_i| (m - m_i)^2$$

    - $|C_i|$ é o tamanho do cluster i

Avaliação Interna: Coesão e Separação

- Exemplo: SSE
  - BSS + WSS = constante



K=1 cluster:

$$SSE = WSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$BSS = 4 \times (3-3)^2 = 0$$

$$Total = 10 + 0 = 10$$

K=2 clusters:

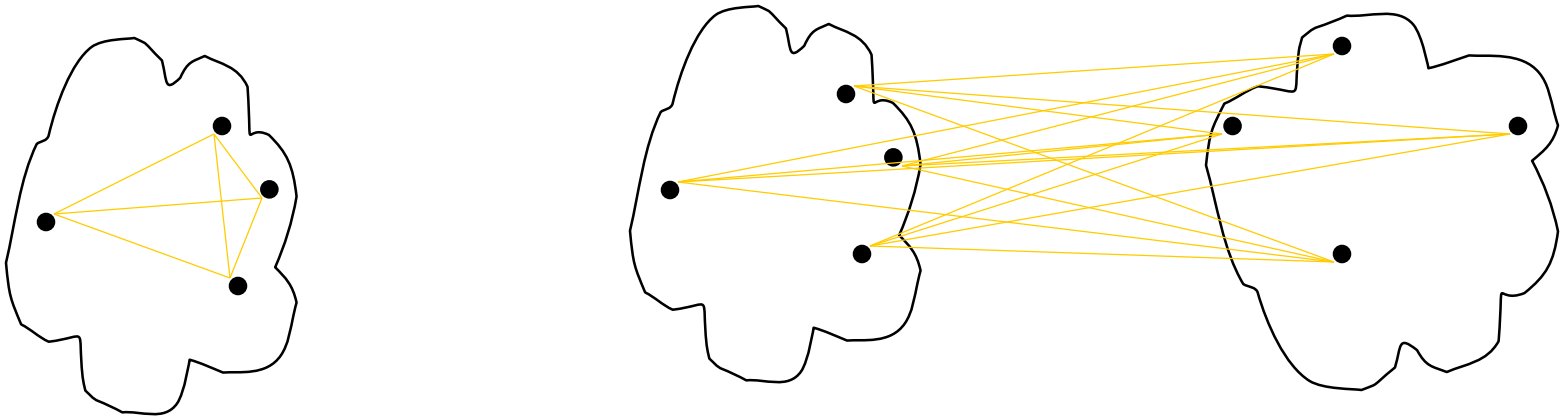
$$SSE = WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$

$$Total = 1 + 9 = 10$$

Avaliação Interna: Coesão e Separação

- Uma abordagem baseada no gráfico de proximidade também pode ser utilizada para coesão e separação.
  - A coesão do cluster é a soma de todos os links dentro de um cluster.
  - A separação dos clusters é a soma dos pesos entre elementos de clusters separados.



coesão

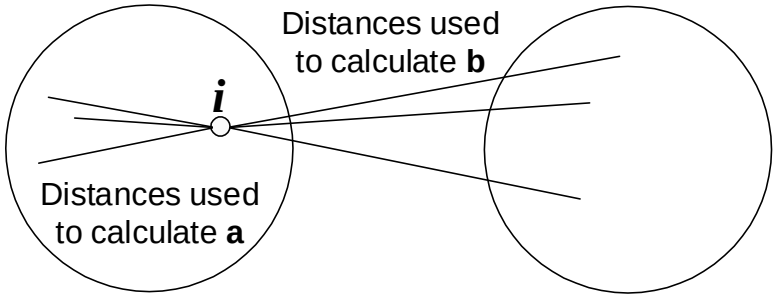
separação

Avaliação Interna: Coeficiente de Silhouette

- Coeficiente de Silhouette combina coesão e separação, mas para pontos individuais, assim como clusters e conjunto de clusters
- Para um ponto individual, i
  - Calcule ***a*** = distância média de *i* para os pontos no seu cluster
  - Calcule ***b*** = min (distância média de *i* para pontos em outro cluster)
  - O coeficiente de Silhouette para um ponto é dado por

    s = (b – a) / max(a,b)

  - Tipicamente entre 0 e 1.
  - Mais próximo de 1 é melhor.



- Podemos calcular o coeficiente médio para um cluster ou conjunto de clusters

Comentário final sobre a avaliação de clusters

"A avaliação de estruturas de agrupamento é a parte mais difícil e frustrante da análise de clusters.

Sem um forte esforço nesta direção, a análise de clusters continuará a ser uma arte negra acessível apenas àqueles verdadeiros crentes que têm experiência e grande coragem."

Algorithms for Clustering Data, Jain e Dubes