

Introdução ao Processamento Digital de Imagem

MC920 / MO443

Prof. Hélio Pedrini

Instituto de Computação
UNICAMP

<http://www.ic.unicamp.br/~helio>

1º Semestre de 2019

Roteiro

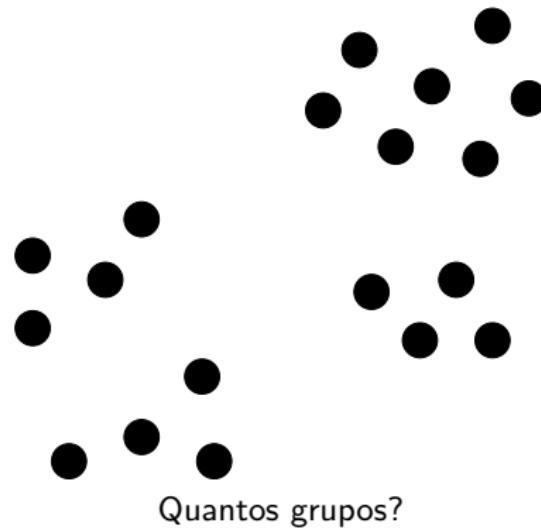
- 1 O que é Agrupamento?
- 2 Tipos de Agrupamento
- 3 Algoritmo K-Means
- 4 Limitações do K-Means
- 5 Validação dos Agrupamentos
- 6 Escolha do Valor k
- 7 Problemas Adicionais do K-Means
- 8 Bisecting K-Means
- 9 Variações do K-Means

O que é Agrupamento?

- **Grupo (*cluster*)**: é um conjunto de dados:
 - ▶ similares aos dados do mesmo grupo.
 - ▶ dissimilares aos dados de outros grupos.
- **Técnicas de Agrupamento (*clustering*)**: conjunto de métodos para partitionar os dados em grupos, segundo um critério de similaridade.
- Técnicas de classificação não supervisionada, ou seja, o partitionamento é realizado com base nos próprios dados, sem os rótulos das classes.

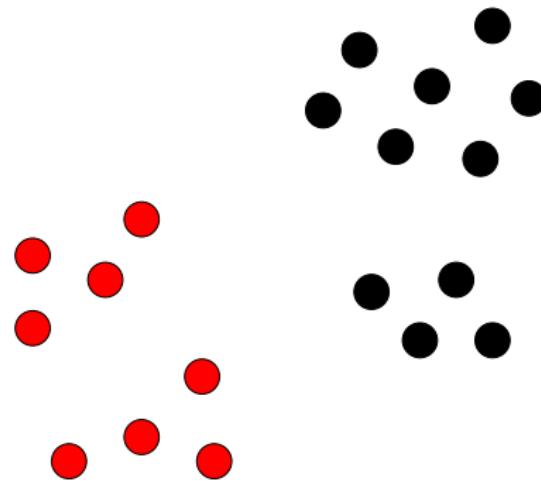
O que é Agrupamento?

A noção de agrupamento pode ser ambígua.



O que é Agrupamento?

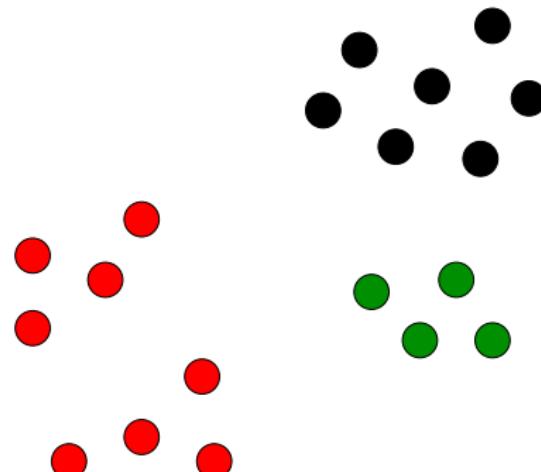
A noção de agrupamento pode ser ambígua.



Quantos grupos? 2

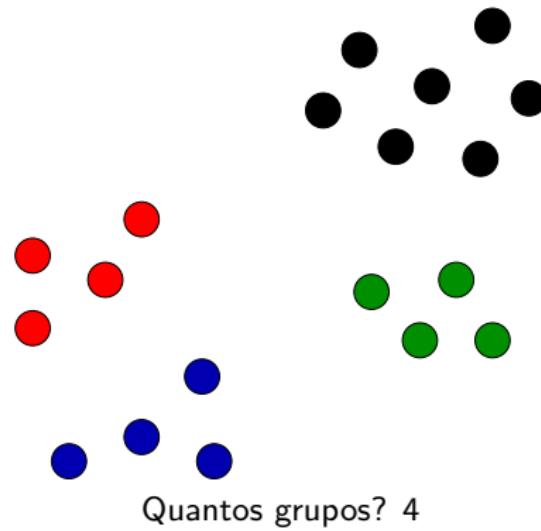
O que é Agrupamento?

A noção de agrupamento pode ser ambígua.



O que é Agrupamento?

A noção de agrupamento pode ser ambígua.



Aplicações

- Ecologia: identifica grupos distintos de espécies.
- Marketing: identifica grupos distintos de clientes.
- Suporte a diagnóstico médico: identifica grupos de pacientes com mesmos sintomas.
- Seguro: identifica grupos de clientes que fazem comunicação de sinistro com alta frequência.
- Planejamento urbano: identifica grupos de casas de acordo com tipo, valor e localização geográfica.

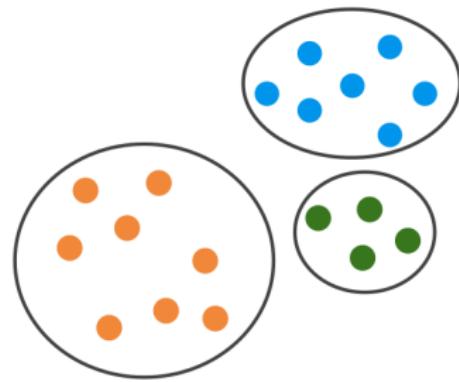
Tipos de Agrupamento

- Agrupamento Particional: divide os dados em subconjuntos sem sobreposição, tal que cada dado está em exatamente um subconjunto.
- Agrupamento Hierárquico: cria uma decomposição hierárquica dos dados, em que cada dado pode pertencer a mais de um subconjunto.

Agrupamento Particional



Dados originais

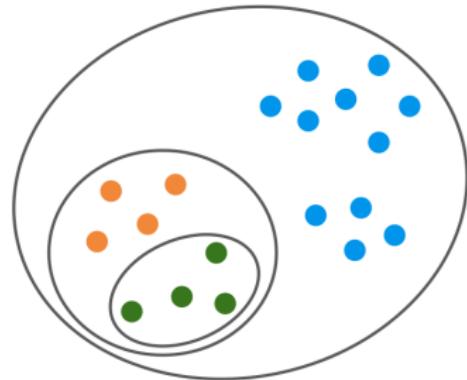


Agrupamento Particional

Agrupamento Hierárquico



Dados originais



Agrupamento Hierárquico

Tipos de Agrupamento

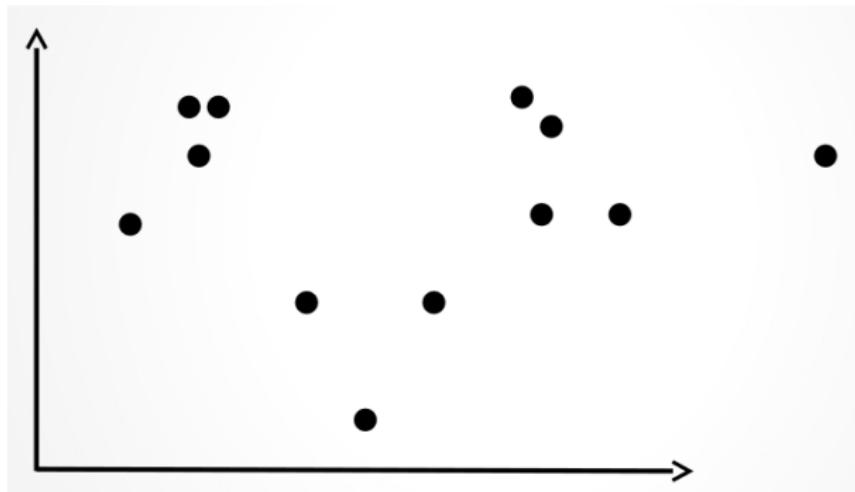
Agrupamento Hierárquico: procura-se construir uma hierarquia de grupos.

- Abordagens normalmente se enquadraram em duas categorias:

- ▶ Aglomerativos (abordagem *bottom-up*):
 - ★ cada ponto inicialmente é um grupo;
 - ★ começa-se com os pontos como grupos individuais e combina-se o par de grupos mais próximos;
 - ★ isto requer a definição de uma métrica de proximidade entre dois grupos.
- ▶ Divisivos (abordagem *top-down*):
 - ★ todos os pontos inicialmente pertencem a um mesmo grupo;
 - ★ começa-se com um único grupo contendo todos os pontos e, em cada etapa, divide-se um grupo até que somente grupos de pontos individuais permaneçam;
 - ★ deve-se decidir qual grupo dividir em cada etapa e como realizar a divisão.

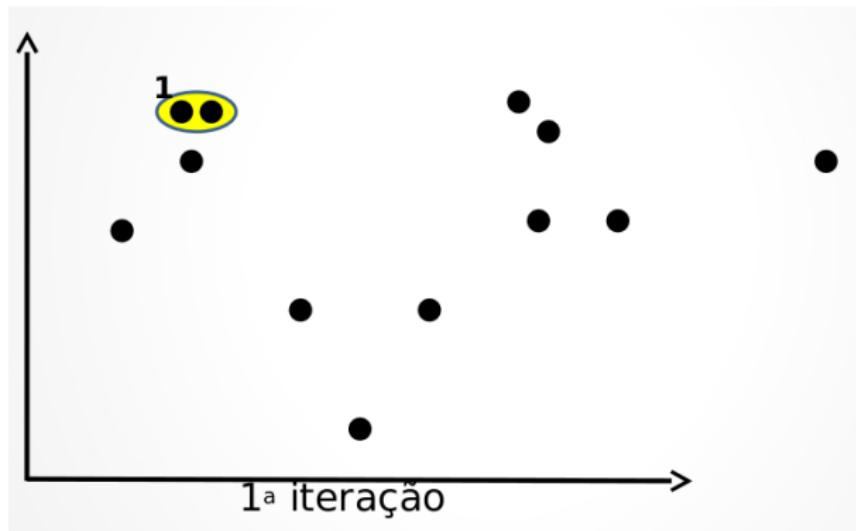
Visualização do Agrupamento Hierárquico

Agrupamento Hierárquico Aglomerativo



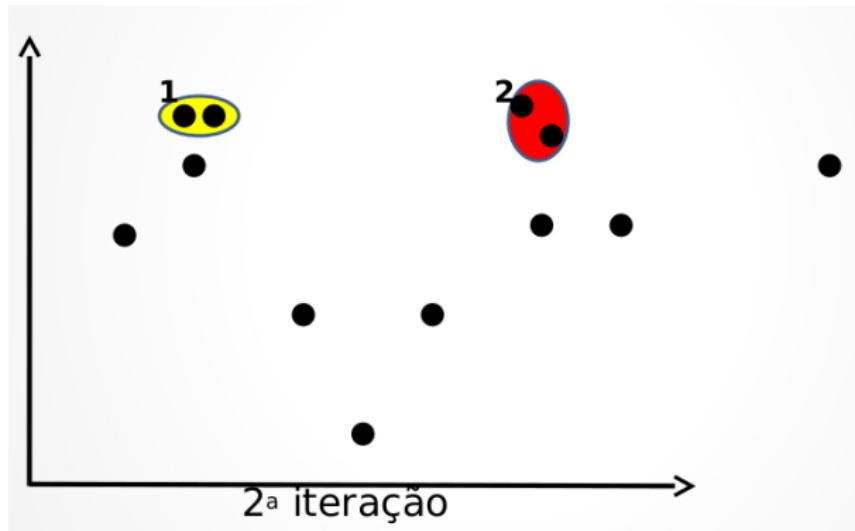
Visualização do Agrupamento Hierárquico

Agrupamento Hierárquico Aglomerativo



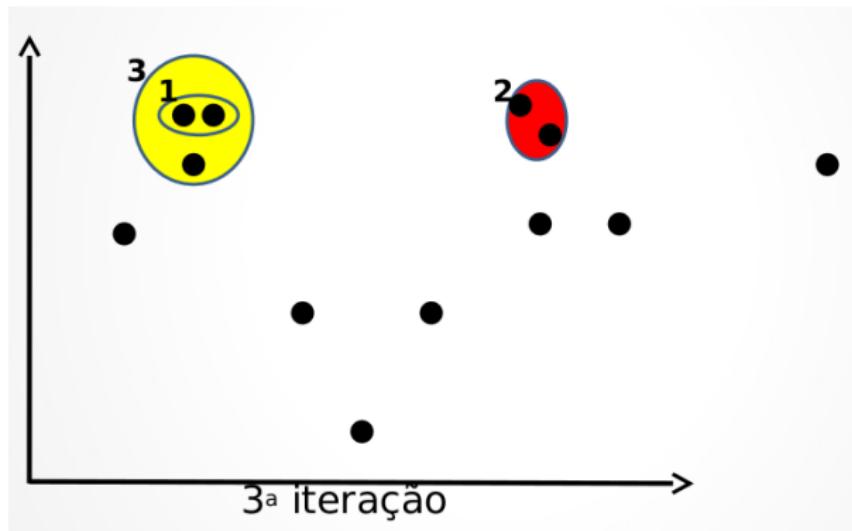
Visualização do Agrupamento Hierárquico

Agrupamento Hierárquico Aglomerativo



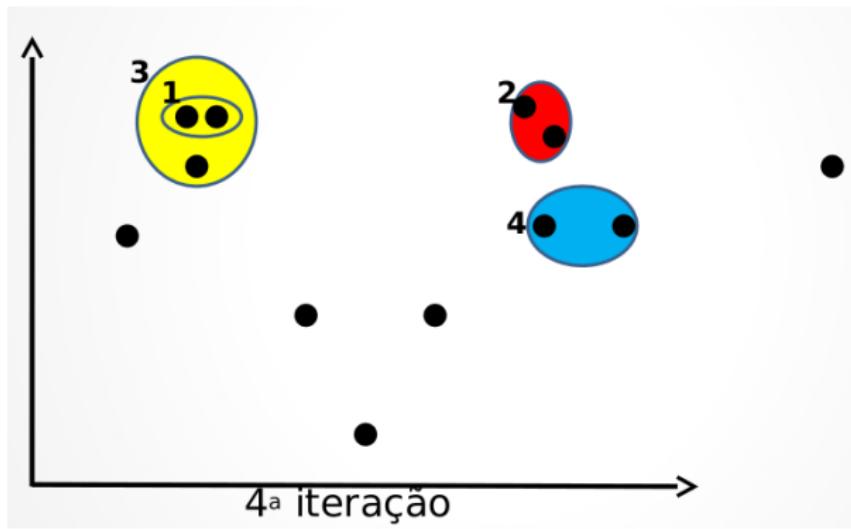
Visualização do Agrupamento Hierárquico

Agrupamento Hierárquico Aglomerativo



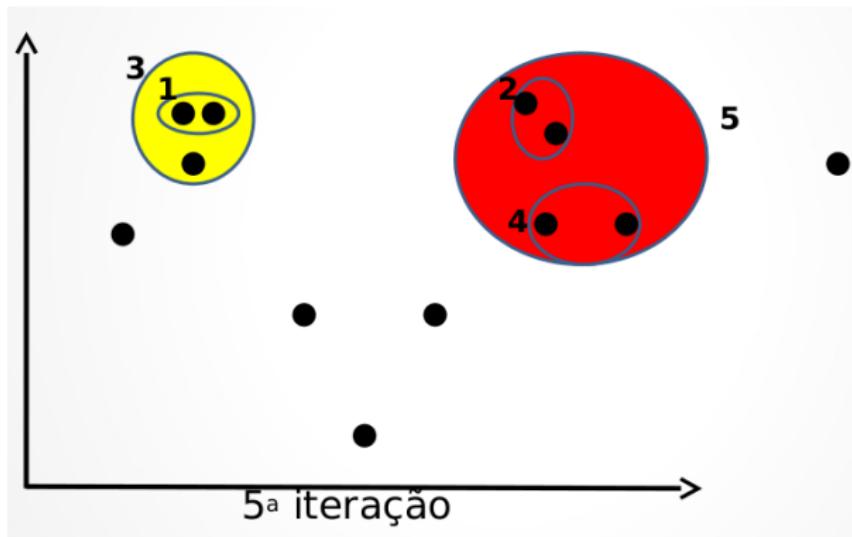
Visualização do Agrupamento Hierárquico

Agrupamento Hierárquico Aglomerativo



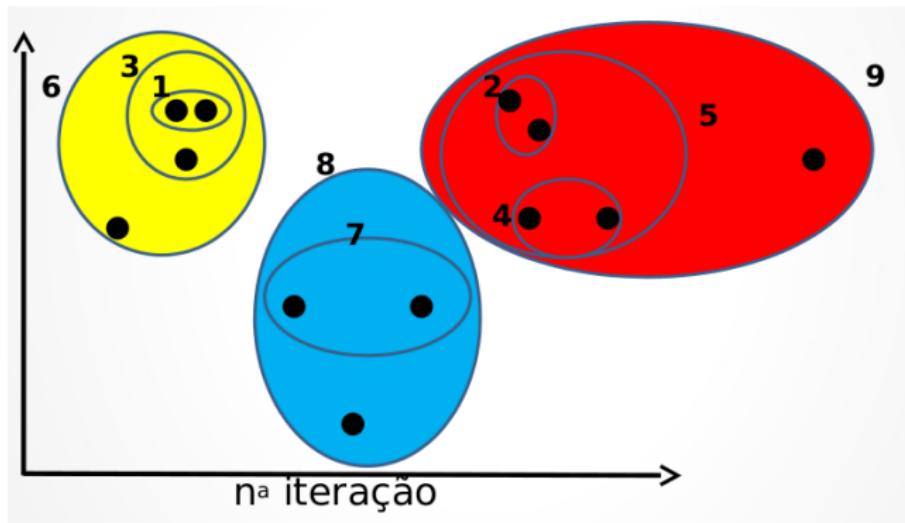
Visualização do Agrupamento Hierárquico

Agrupamento Hierárquico Aglomerativo



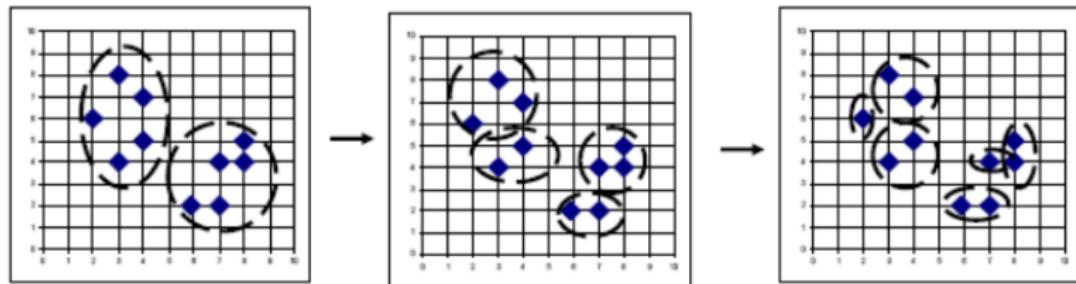
Visualização do Agrupamento Hierárquico

Agrupamento Hierárquico Aglomerativo



Visualização do Agrupamento Hierárquico

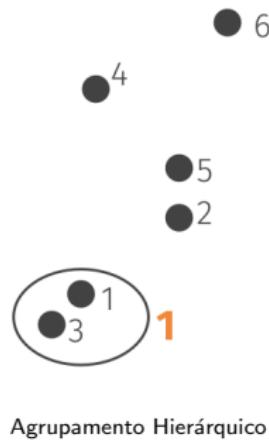
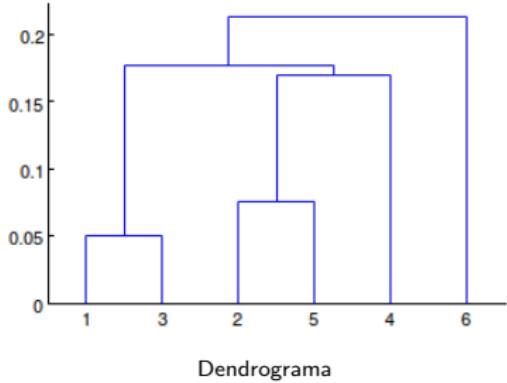
Agrupamento Hierárquico Divisivo



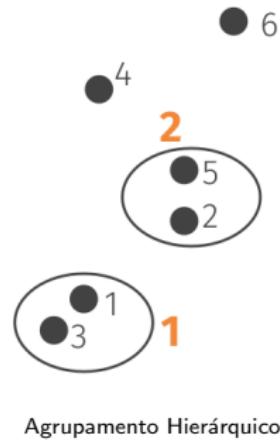
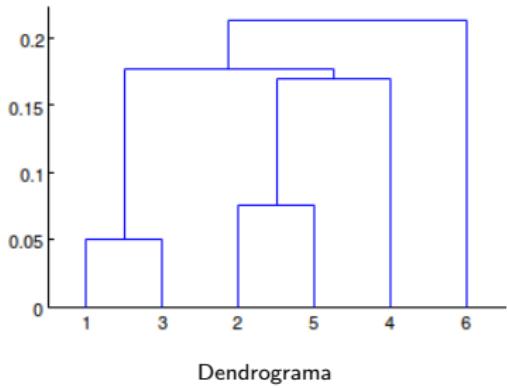
Tipos de Agrupamento

- Em geral, as operações de divisão e fusão são determinadas por uma estratégia *gulosa*.
- Os resultados do agrupamento hierárquico são normalmente apresentados por um *dendrograma*.

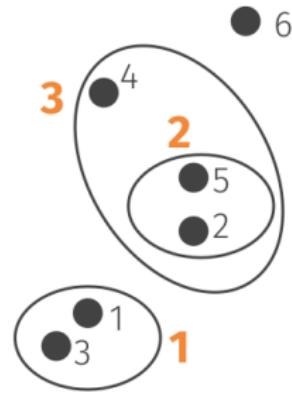
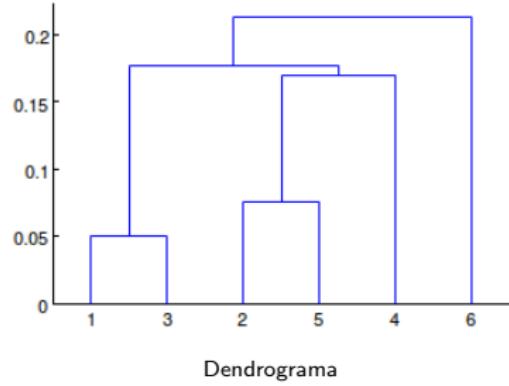
Visualização do Agrupamento Hierárquico



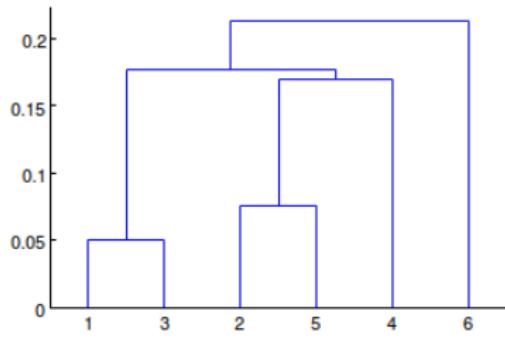
Visualização do Agrupamento Hierárquico



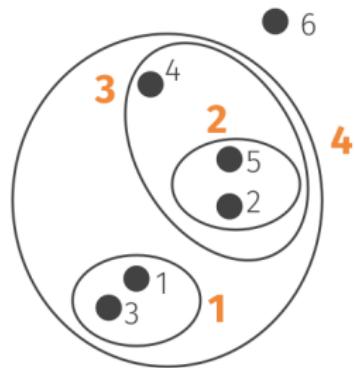
Visualização do Agrupamento Hierárquico



Visualização do Agrupamento Hierárquico

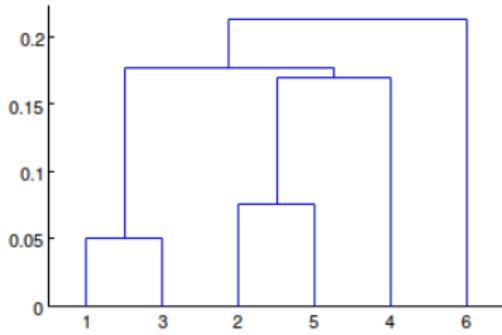


Dendrograma

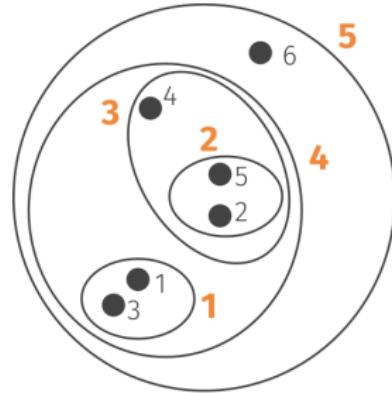


Agrupamento Hierárquico

Visualização do Agrupamento Hierárquico



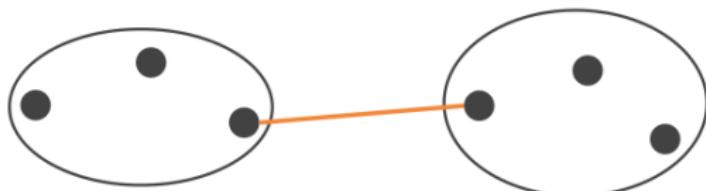
Dendrograma



Agrupamento Hierárquico

Tipos de Agrupamento

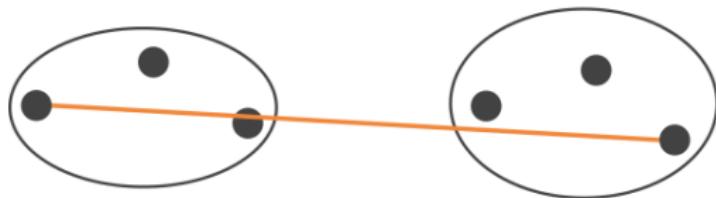
Proximidade entre dois grupos:



- distância mínima (*single-linkage clustering*): a proximidade entre dois grupos é definida como a menor distância entre dois pontos de cada grupo.

Tipos de Agrupamento

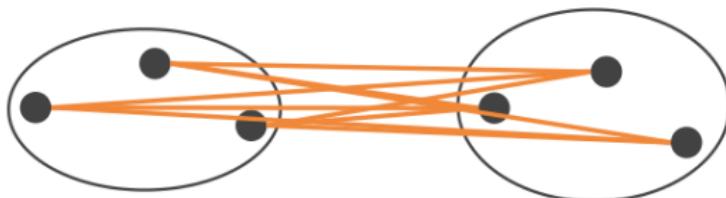
Proximidade entre dois grupos:



- distância máxima (*complete-linkage clustering*): a proximidade entre dois grupos é definida como a maior distância entre dois pontos de cada grupo.

Tipos de Agrupamento

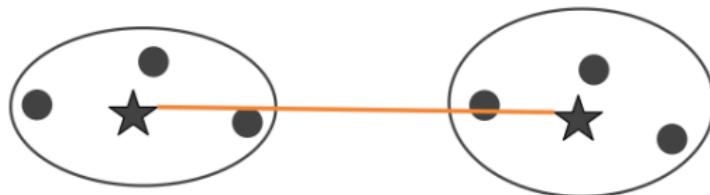
Proximidade entre dois grupos:



- distância média (*average-linkage clustering*): a proximidade entre dois grupos é definida como a distância média de todos os pares de pontos entre cada grupo.

Tipos de Agrupamento

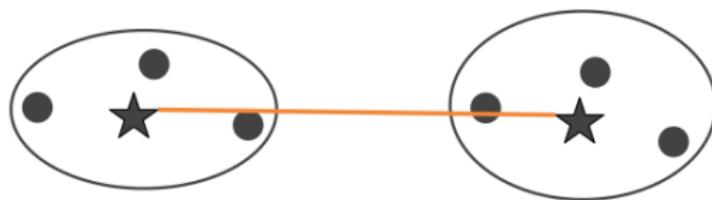
Proximidade entre dois grupos:



- distância entre centroides (*centroid-linkage clustering*): a proximidade entre dois grupos é definida pelos centroides de cada grupo.

Tipos de Agrupamento

Proximidade entre dois grupos:



- distância entre centroides (*Ward-linkage clustering*): a proximidade entre dois grupos também é definida pelos centroides de cada grupo, entretanto, levando-se em conta o aumento da soma dos quadrados das distâncias (SSE) resultante da fusão de dois grupos.

Algoritmo Hierárquico Aglomerativo

Técnica de agrupamento hierárquico mais popular:

1. Calcule a matriz de proximidade.
2. Combine os dois grupos mais próximos.
3. Atualize a matriz de proximidade.
4. Repita os passos 2 e 3 até que apenas um único grupo permaneça.

Uma operação importante é o cálculo da proximidade entre dois grupos.

Algoritmo Hierárquico Aglomerativo

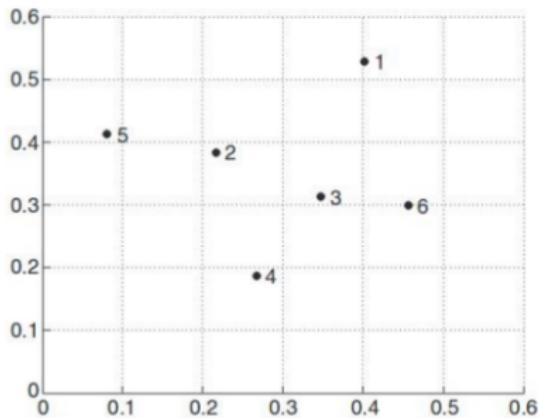
Matriz de dados: linhas (amostras) e colunas (atributos).

$$X_{n \times d} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1d} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2d} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nd} \end{bmatrix}$$

Matriz de dissimilaridade: cada elemento representa a distância entre pares de amostras.

$$D_{n \times n} = \begin{bmatrix} 0 & d(1, 2) & d(1, 3) & \dots & d(1, n) \\ d(2, 1) & 0 & d(2, 3) & \dots & d(2, n) \\ d(3, 1) & d(3, 2) & 0 & \dots & d(3, n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d(n, 1) & d(n, 2) & d(n, 3) & \dots & 0 \end{bmatrix}$$

Algoritmo Hierárquico Aglomerativo



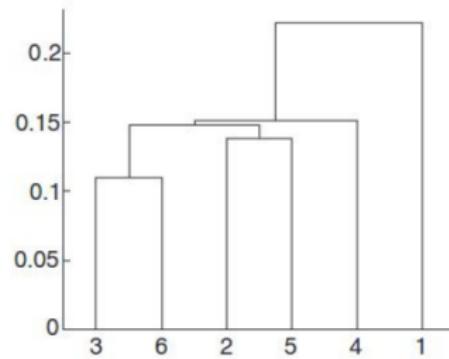
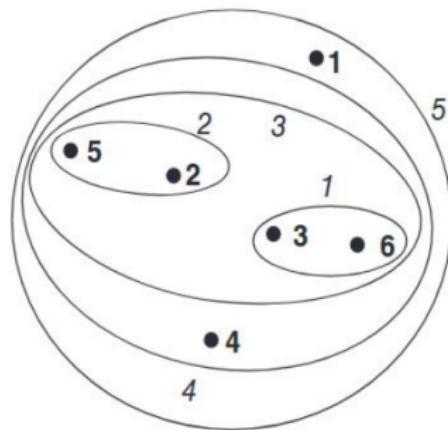
Ponto	Coordenada x	Coordenada y
p_1	0.4005	0.5306
p_2	0.2148	0.3854
p_3	0.3457	0.3156
p_4	0.2652	0.1875
p_5	0.0789	0.4139
p_6	0.4548	0.3022

$$D_{n \times n} = \begin{bmatrix} & p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \\ p_1 & 0.00 & 0.24 & 0.22 & 0.37 & 0.34 & 0.23 \\ p_2 & 0.24 & 0.00 & 0.15 & 0.20 & 0.14 & 0.25 \\ p_3 & 0.22 & 0.15 & 0.00 & 0.15 & 0.28 & 0.11 \\ p_4 & 0.37 & 0.20 & 0.15 & 0.00 & 0.29 & 0.22 \\ p_5 & 0.34 & 0.14 & 0.28 & 0.29 & 0.00 & 0.39 \\ p_6 & 0.23 & 0.25 & 0.11 & 0.22 & 0.39 & 0.00 \end{bmatrix}$$

Algoritmo Hierárquico Aglomerativo

Single-Linkage Clustering

$$\begin{aligned} \text{dist}(\{3, 6\}, \{2, 5\}) &= \min(\text{dist}(3, 2), \text{dist}(6, 2), \text{dist}(3, 5), \text{dist}(6, 5)) \\ &= \min(0.15, 0.25, 0.28, 0.39) \\ &= 0.15 \end{aligned}$$



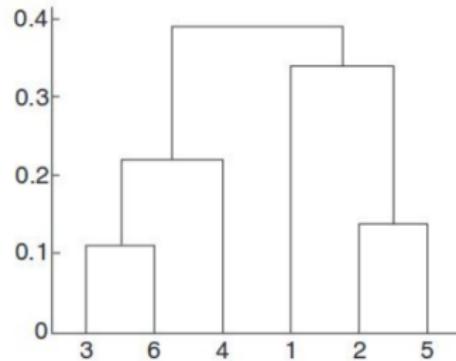
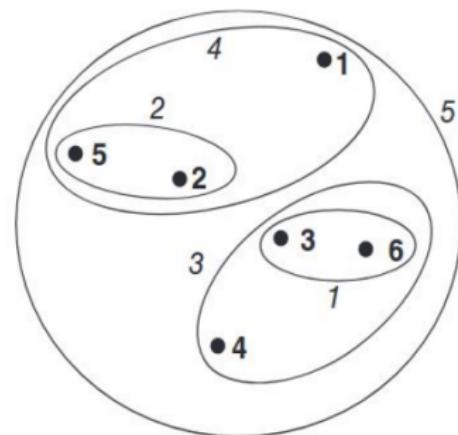
Algoritmo Hierárquico Aglomerativo

Complete-Linkage Clustering

$$dist(\{3, 6\}, \{4\}) = \max(dist(3, 4), dist(6, 4)) = \max(0.15, 0.22) = 0.22$$

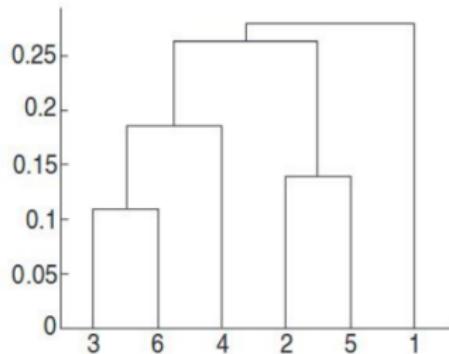
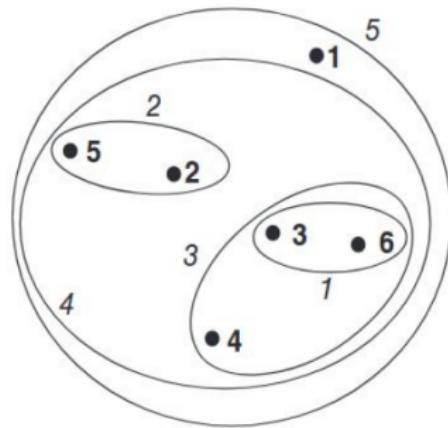
$$\begin{aligned} dist(\{3, 6\}, \{2, 5\}) &= \max(dist(3, 2), dist(6, 2), dist(3, 5), dist(6, 5)) \\ &= \max(0.15, 0.25, 0.28, 0.39) = 0.39 \end{aligned}$$

$$dist(\{3, 6\}, \{1\}) = \max(dist(3, 1), dist(6, 1)) = \max(0.22, 0.23) = 0.23$$



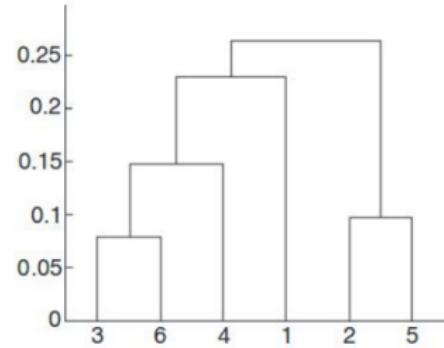
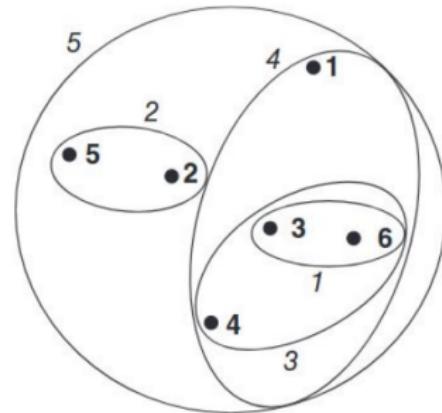
Algoritmo Hierárquico Aglomerativo

Average-Linkage Clustering



Algoritmo Hierárquico Aglomerativo

Ward-Linkage Clustering



Outros Tipos de Agrupamento

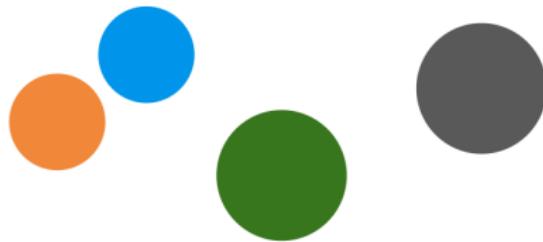
- Exclusivo × Não Exclusivo: em agrupamentos não exclusivos, pontos podem pertencer a vários grupos. Podem representar várias classes ou pontos de fronteira.
- Fuzzy × Não Fuzzy: em agrupamento fuzzy, um ponto pertence a todo grupo com algum peso entre 0 e 1.
- Parcial × Completo: em alguns casos, deseja-se agrupar somente alguns dados em comparação a agrupar todos os dados do conjunto.

Famílias de Algoritmos de Agrupamento

- Baseados em Centroides: descobrir o *centro* de cada grupo. Cada ponto pertence ao(s) grupo(s), cujo(s) centro(s) é (são) mais próximo(s).
- Baseados em Conectividade (ou Agrupamento Hierárquico): pontos e subgrupos próximos uns dos outros devem pertencer ao mesmo super-grupo.
- Baseados em Densidade: grupos são regiões de alta densidade de pontos separados por regiões de baixa densidade.

Famílias de Algoritmos de Agrupamento

- Baseados em Centroides: descobrir o *centro* de cada grupo.
- Cada ponto pertence ao(s) grupo(s), cujo(s) centro(s) é (são) mais próximo(s).

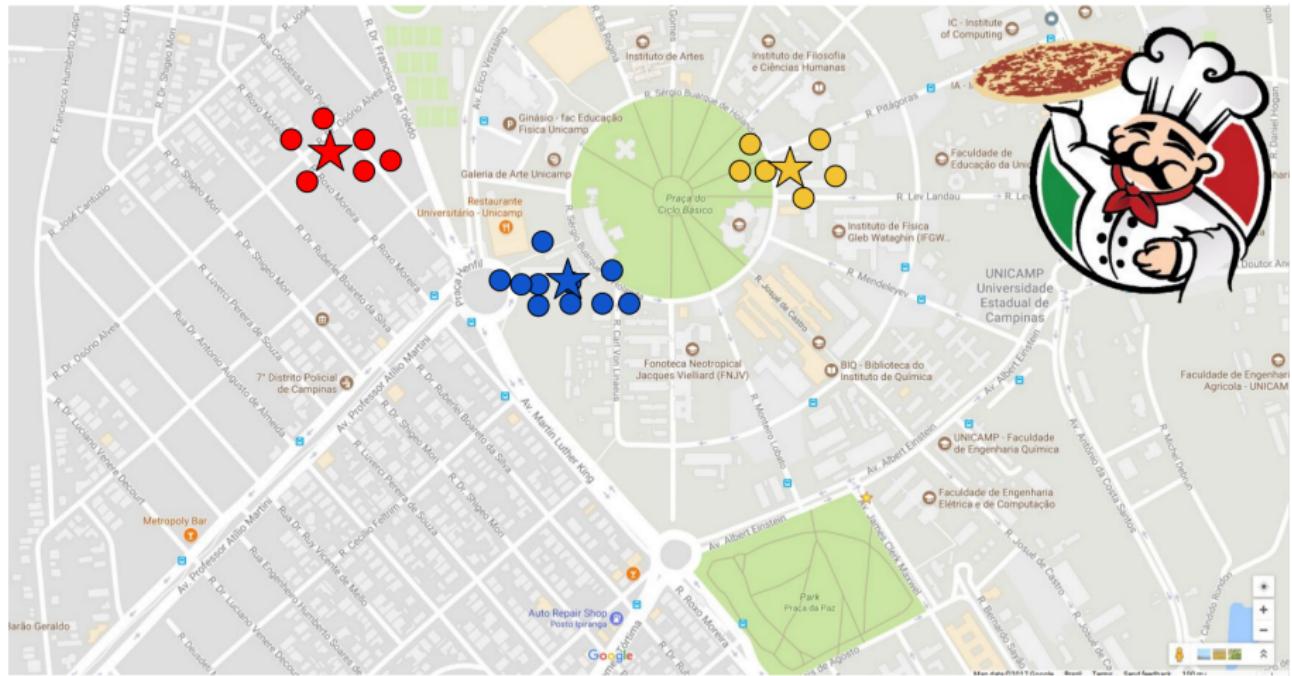


Algoritmos de Agrupamento

Características desejáveis:

- Identificar grupos de tamanhos variados.
- Identificar grupos com forma arbitrária.
- Ser escalável para lidar com qualquer quantidade de amostras.
- Requerer conhecimento mínimo para determinar parâmetros de entrada.
- Encontrar o número adequado de grupos.

K-Means: Particionamento de Dados



K-Means: Segmentação de Imagens

Imagen Original



$K = 10$



$K = 3$



$K = 2$



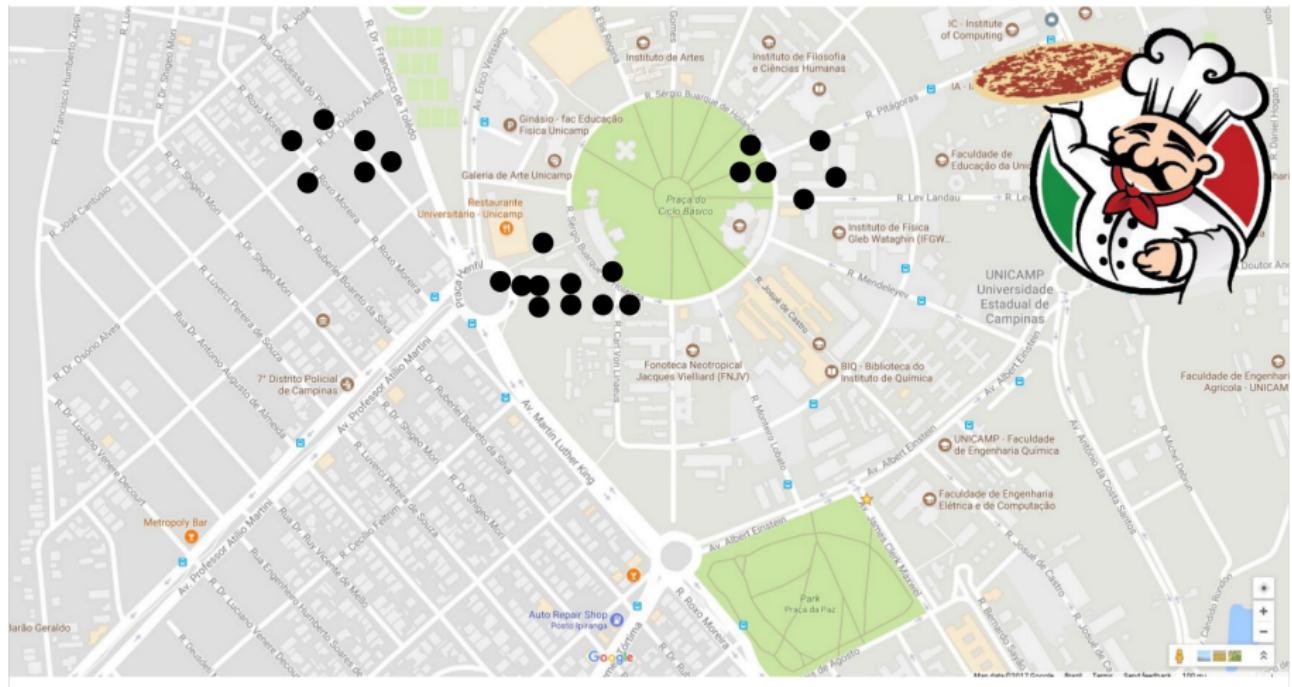
Algoritmo K-Means

1. Defina os k centroides.
 - ▶ Inicialize os centroides aleatoriamente.
2. Encontre o centroide mais próximo e atualize as atribuições do grupo:
 - ▶ Atribua cada dado a um dos k grupos. Cada dado é atribuído ao grupo do centroide mais próximo (distância Euclidiana).
3. Mova os centroides para o centro de seus grupos.
 - ▶ A nova posição de cada centroide é calculada como a média de todos os pontos em seu grupo.
4. Repita os passos 2 e 3 até que o centroide pare de se mover muito em cada iteração.

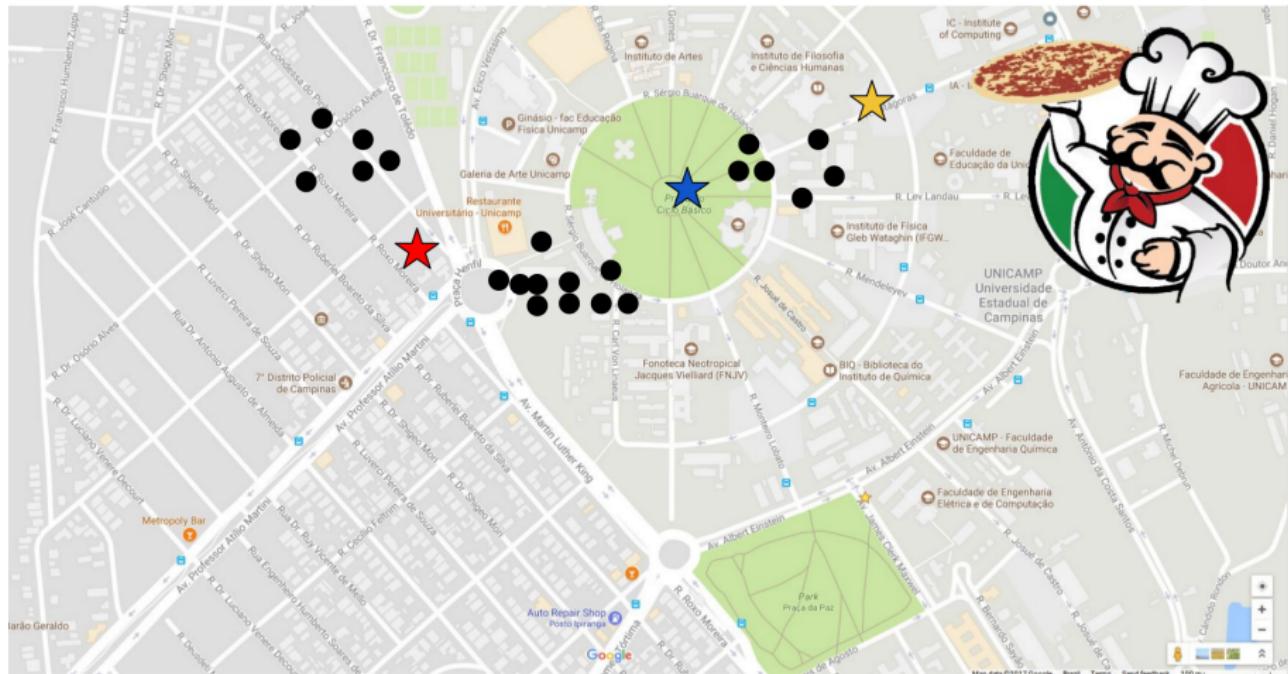
Algoritmo K-Means

1. Defina os k centroides.
 - ▶ Inicialize os centroides aleatoriamente.
2. Encontre o centroide mais próximo e atualize as atribuições do grupo:
 - ▶ Atribua cada dado a um dos k grupos. Cada dado é atribuído ao grupo do centroide mais próximo (distância Euclidiana).
3. Mova os centroides para o centro de seus grupos.
 - ▶ A nova posição de cada centroide é calculada como a média de todos os pontos em seu grupo.
4. Repita os passos 2 e 3 até que o centroide pare de se mover muito em cada iteração.

Algoritmo K-Means



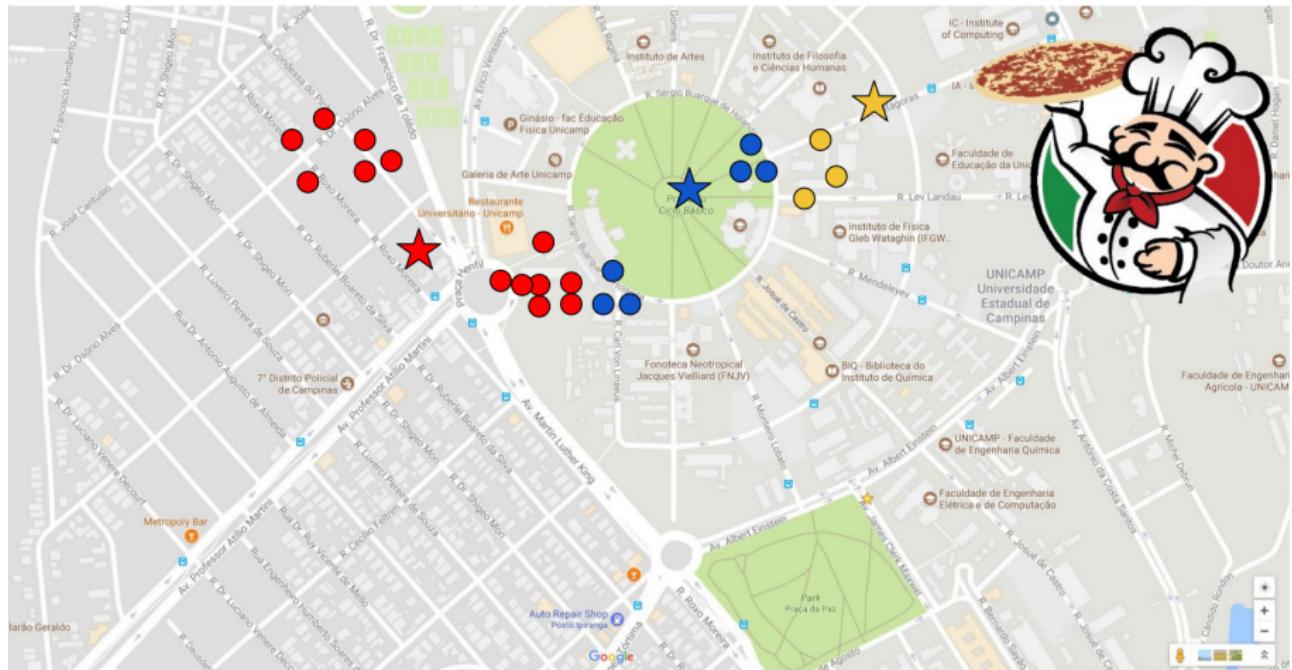
Algoritmo K-Means



Algoritmo K-Means

1. Defina os k centroides.
 - ▶ Inicialize os centroides aleatoriamente.
2. **Encontre o centroide mais próximo e atualize as atribuições do grupo:**
 - ▶ Atribua cada dado a um dos k grupos. Cada dado é atribuído ao grupo do centroide mais próximo (distância Euclidiana).
3. Mova os centroides para o centro de seus grupos.
 - ▶ A nova posição de cada centroide é calculada como a média de todos os pontos em seu grupo.
4. Repita os passos 2 e 3 até que o centroide pare de se mover muito em cada iteração.

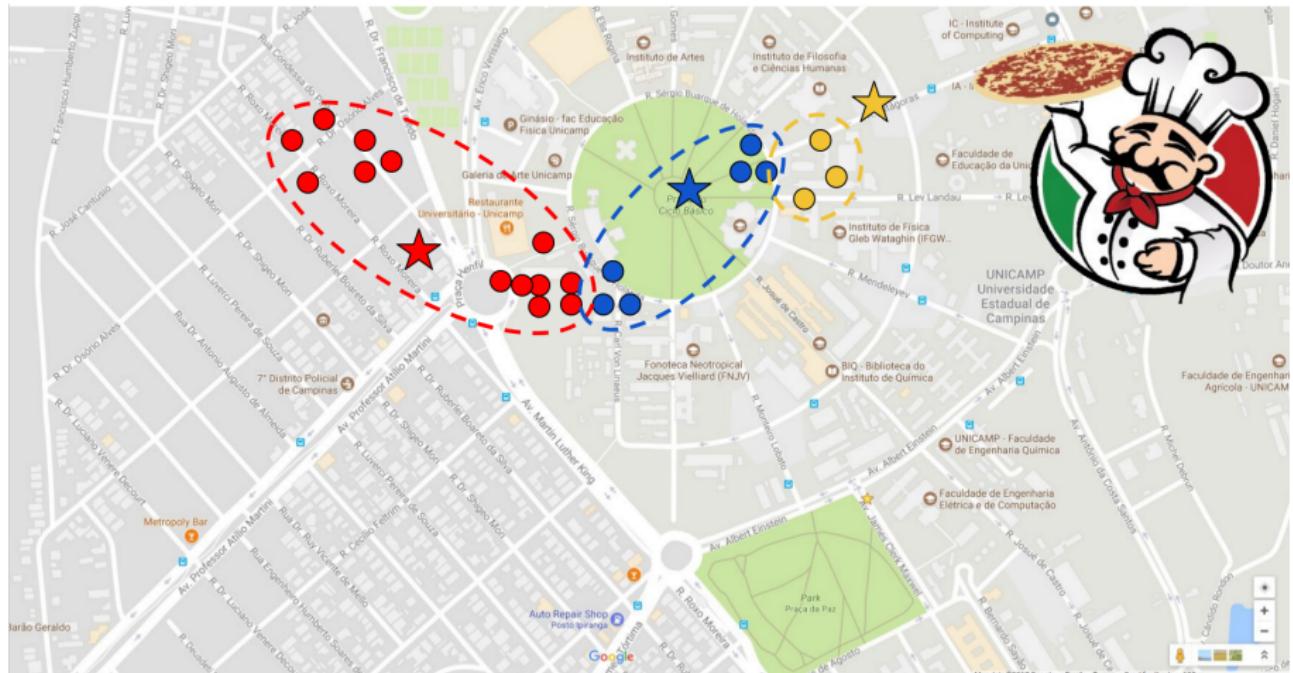
Algoritmo K-Means



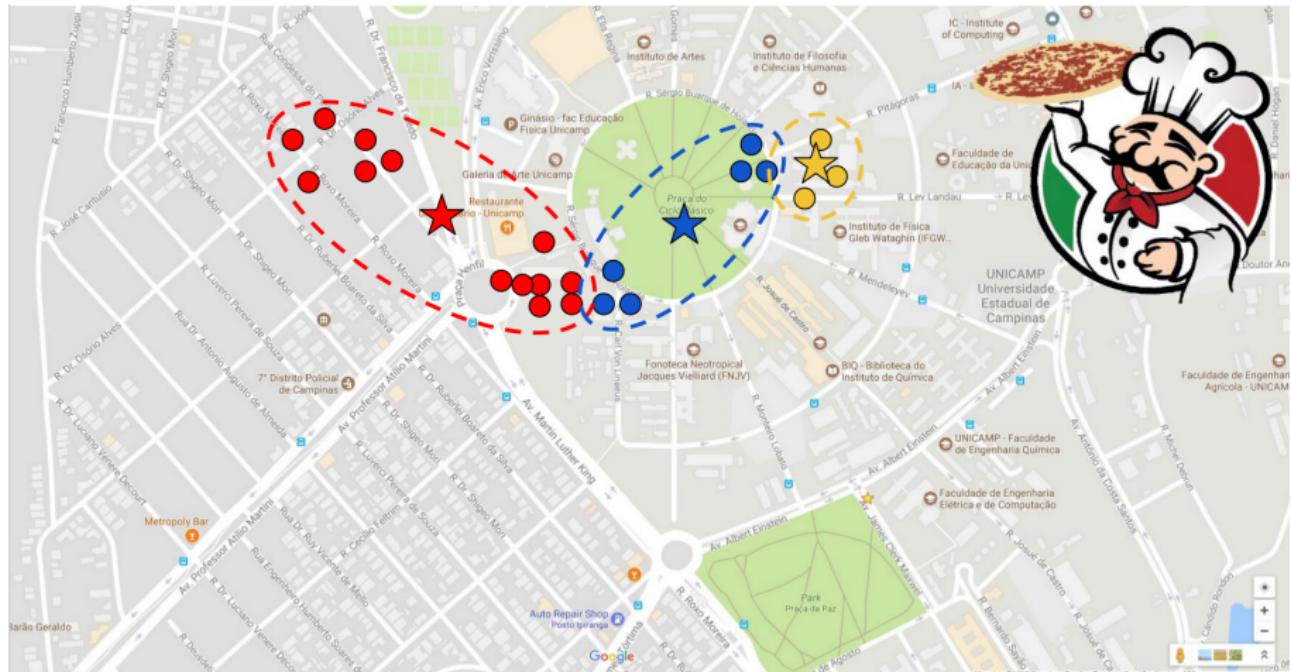
Algoritmo K-Means

1. Defina os k centroides.
 - ▶ Inicialize os centroides aleatoriamente.
2. Encontre o centroide mais próximo e atualize as atribuições do grupo:
 - ▶ Atribua cada dado a um dos k grupos. Cada dado é atribuído ao grupo do centroide mais próximo (distância Euclidiana).
3. Mova os centroides para o centro de seus grupos.
 - ▶ A nova posição de cada centroide é calculada como a média de todos os pontos em seu grupo.
4. Repita os passos 2 e 3 até que o centroide pare de se mover muito em cada iteração.

Algoritmo K-Means



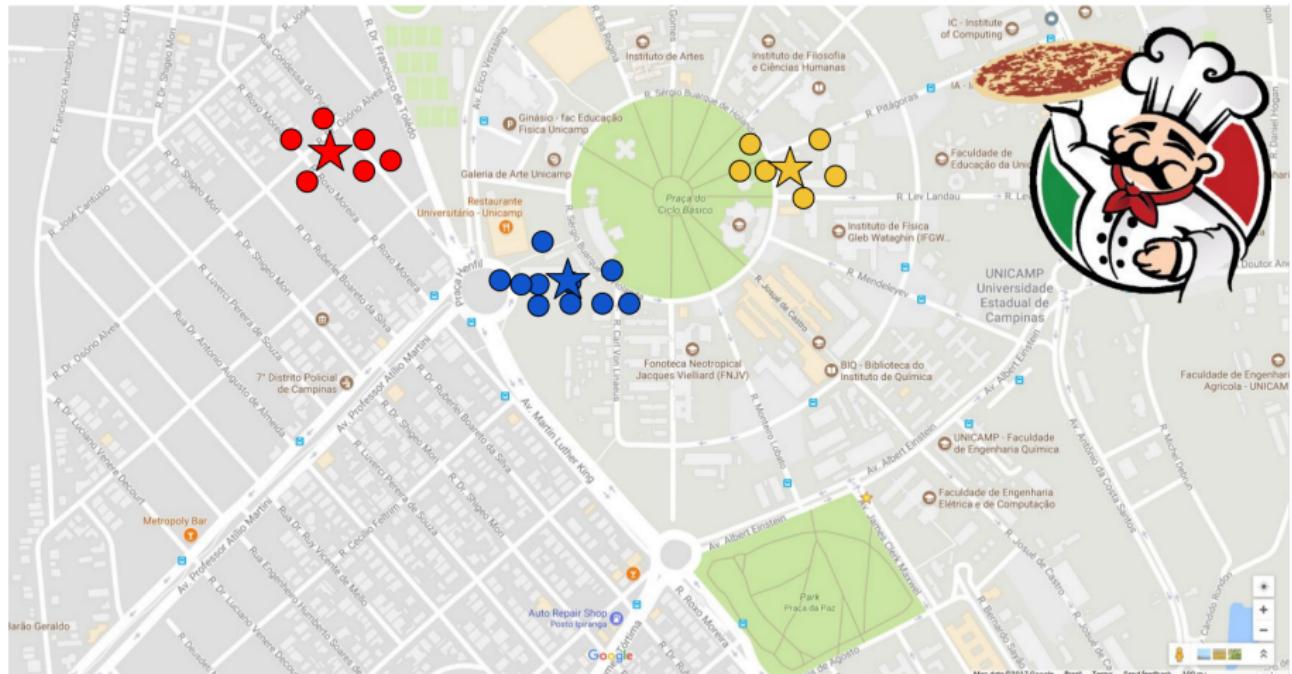
Algoritmo K-Means



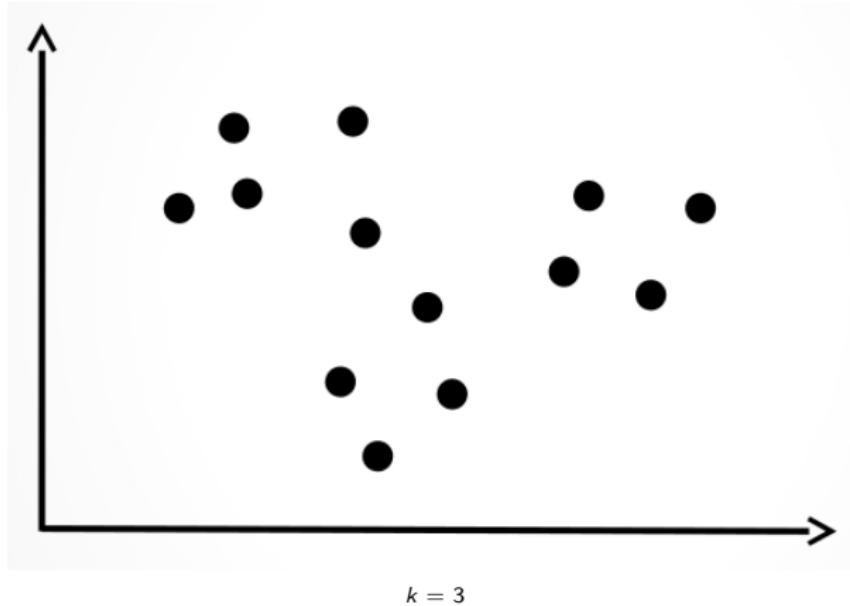
Algoritmo K-Means

1. Defina os k centroides.
 - ▶ Inicialize os centroides aleatoriamente.
2. Encontre o centroide mais próximo e atualize as atribuições do grupo:
 - ▶ Atribua cada dado a um dos k grupos. Cada dado é atribuído ao grupo do centroide mais próximo (distância Euclidiana).
3. Mova os centroides para o centro de seus grupos.
 - ▶ A nova posição de cada centroide é calculada como a média de todos os pontos em seu grupo.
4. Repita os passos 2 e 3 até que o centroide pare de se mover muito em cada iteração.

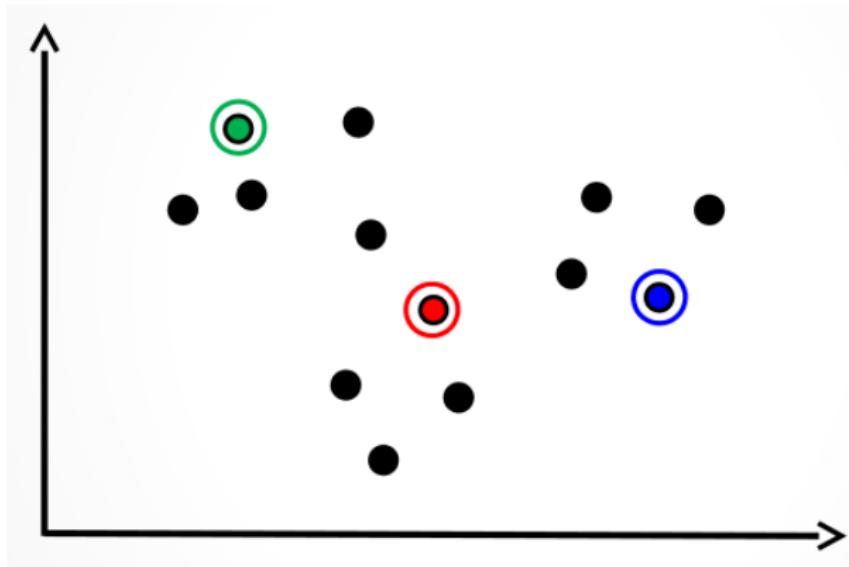
Algoritmo K-Means



Algoritmo K-Means: Exemplo

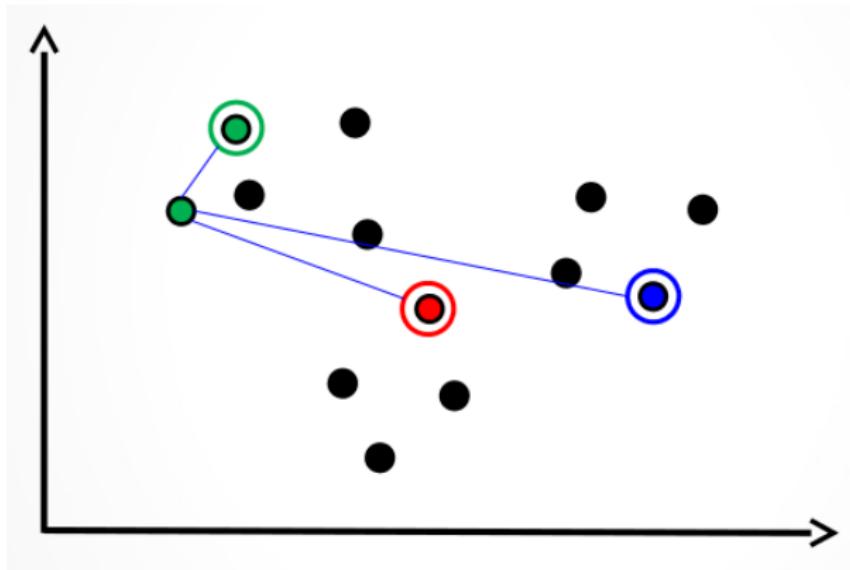


Algoritmo K-Means: Exemplo

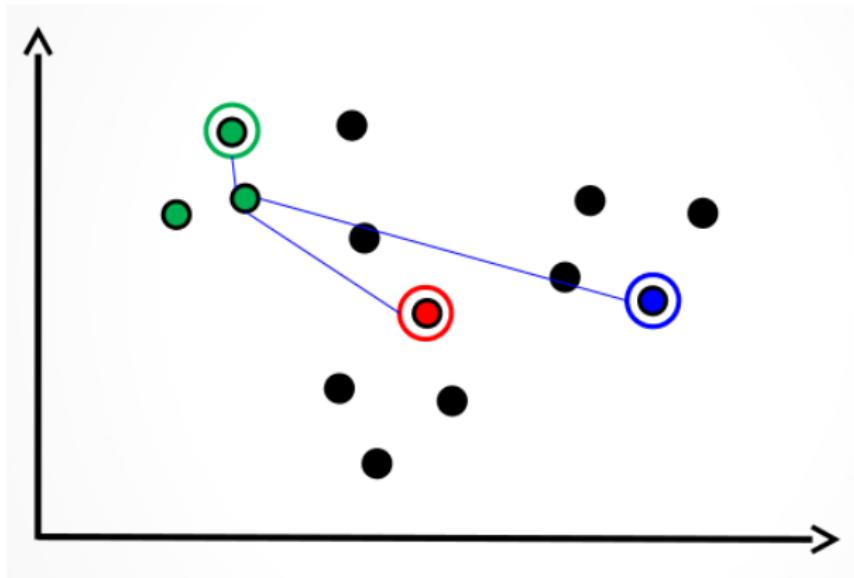


k centroides iniciais são selecionados

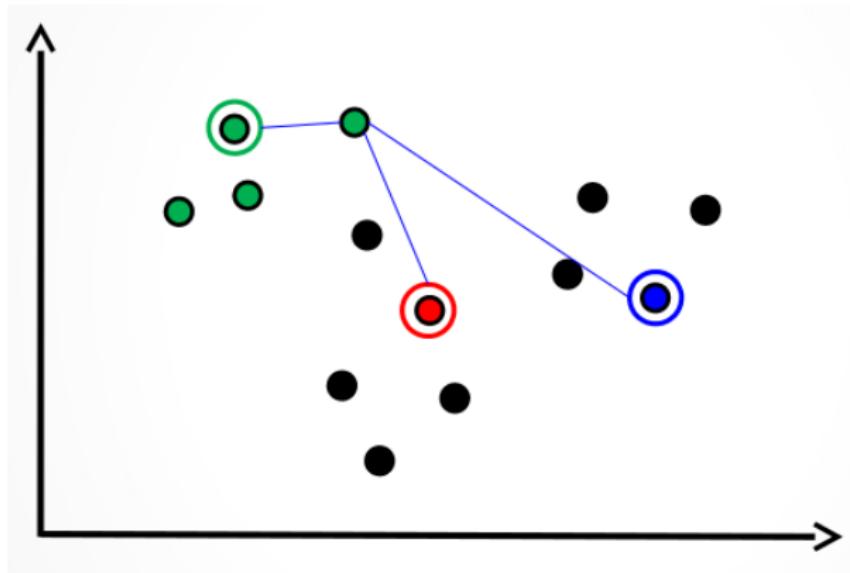
Algoritmo K-Means: Exemplo



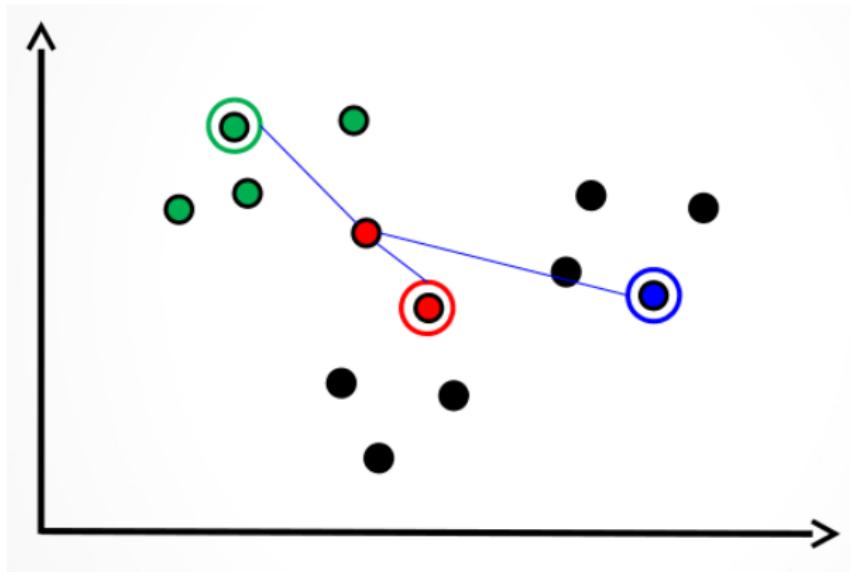
Algoritmo K-Means: Exemplo



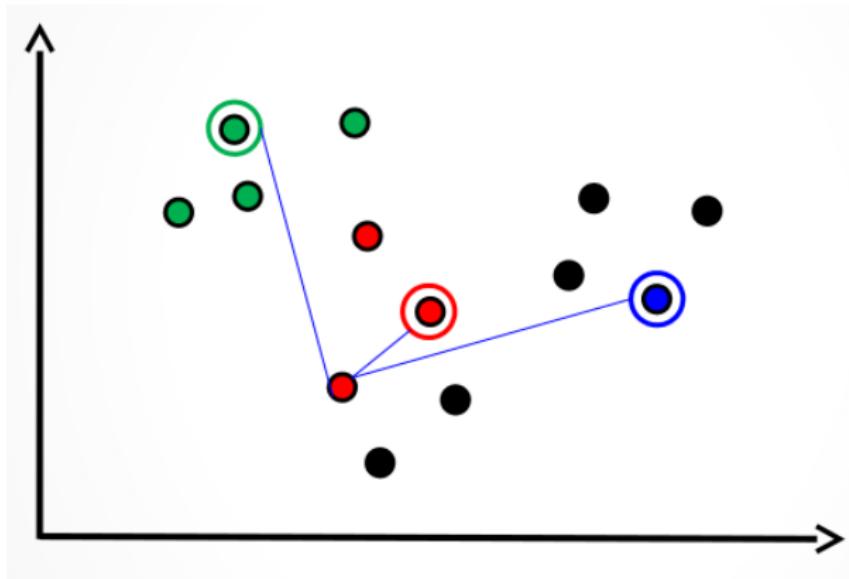
Algoritmo K-Means: Exemplo



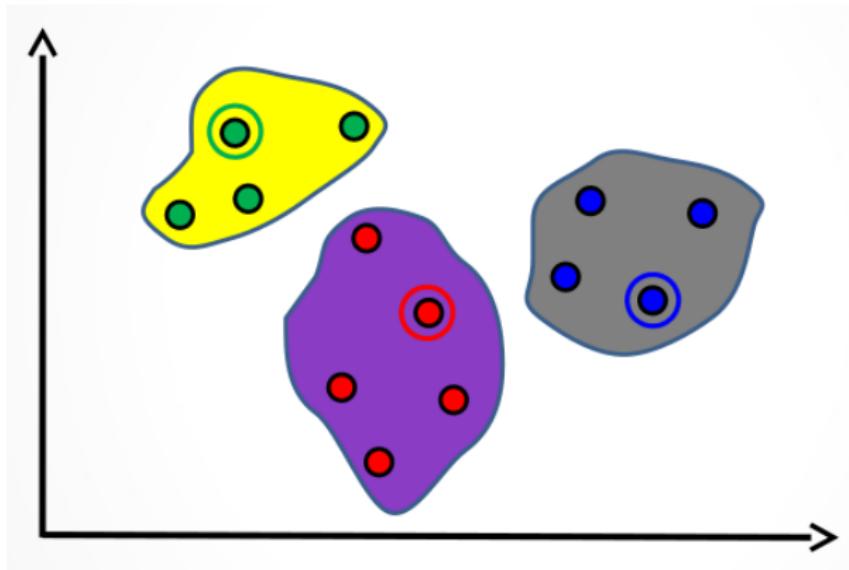
Algoritmo K-Means: Exemplo



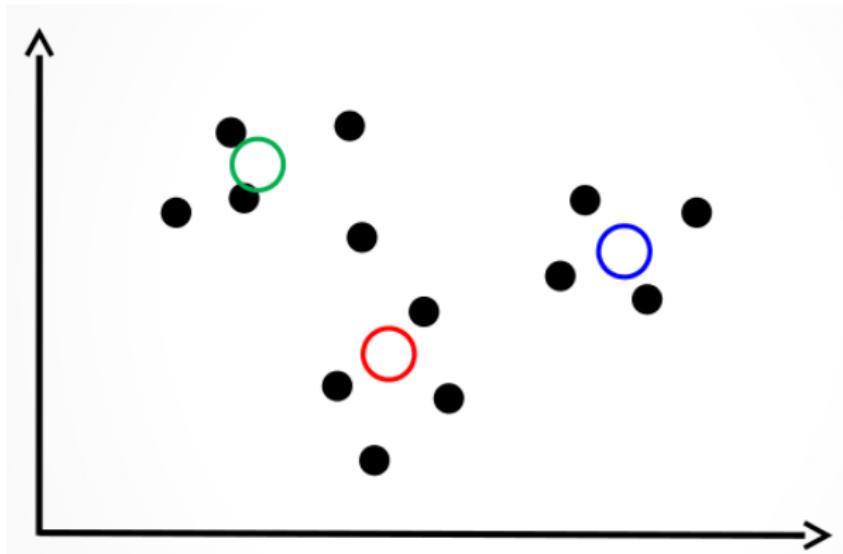
Algoritmo K-Means: Exemplo



Algoritmo K-Means: Exemplo

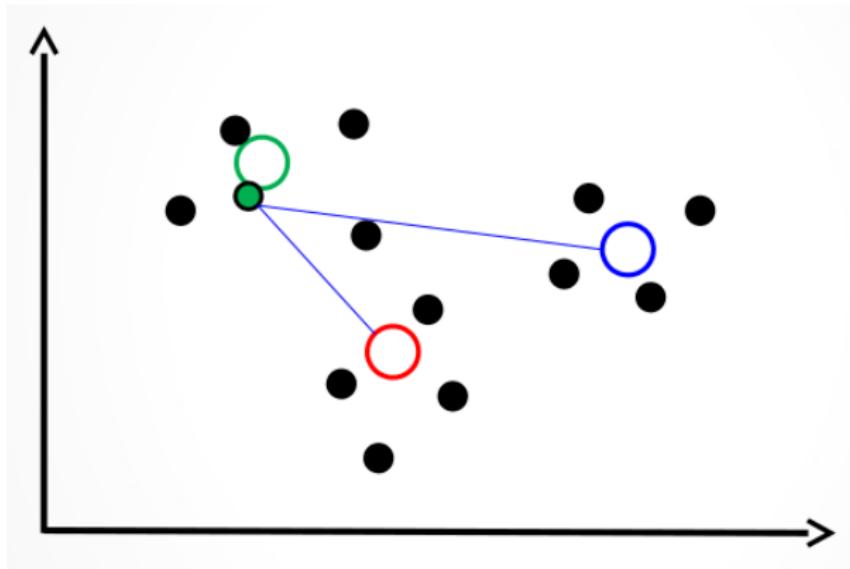


Algoritmo K-Means: Exemplo

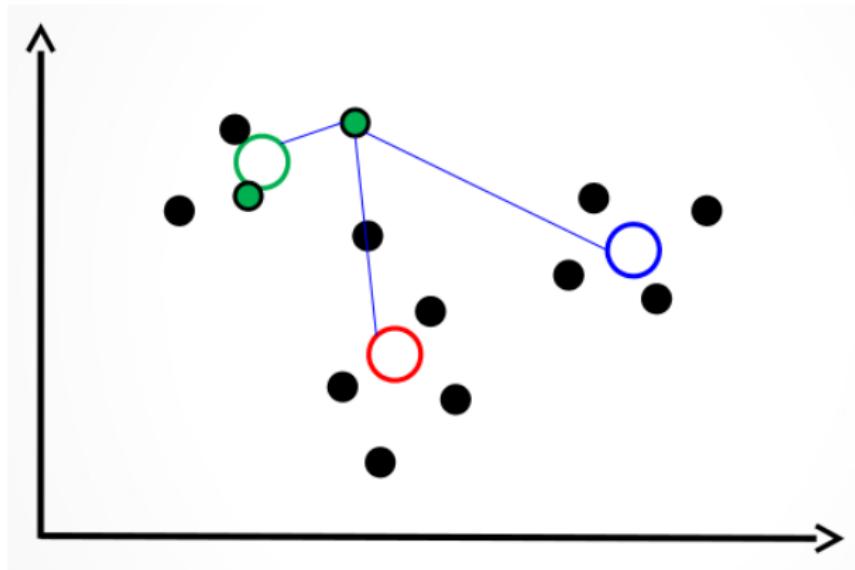


Os passos anteriores são repetidos até que os centroides não se movam mais.

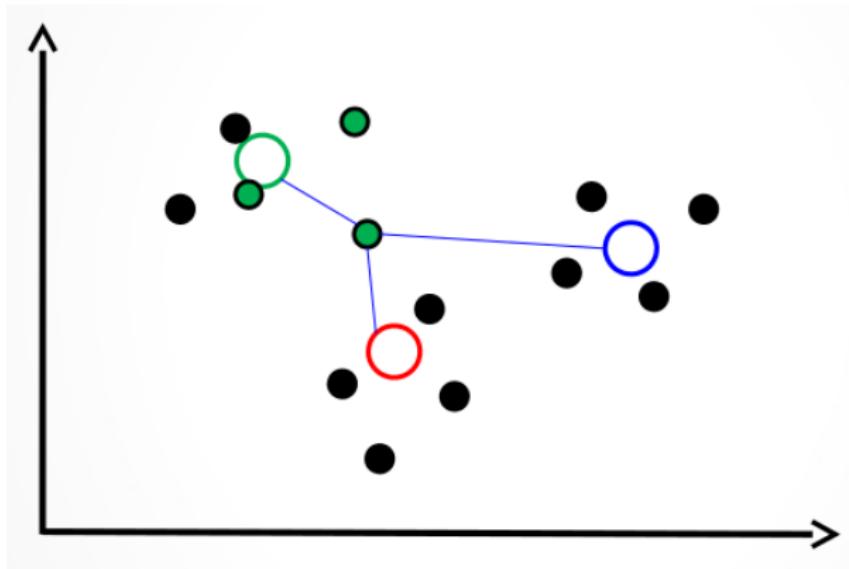
Algoritmo K-Means: Exemplo



Algoritmo K-Means: Exemplo



Algoritmo K-Means: Exemplo



Algoritmo K-Means: Função Objetivo

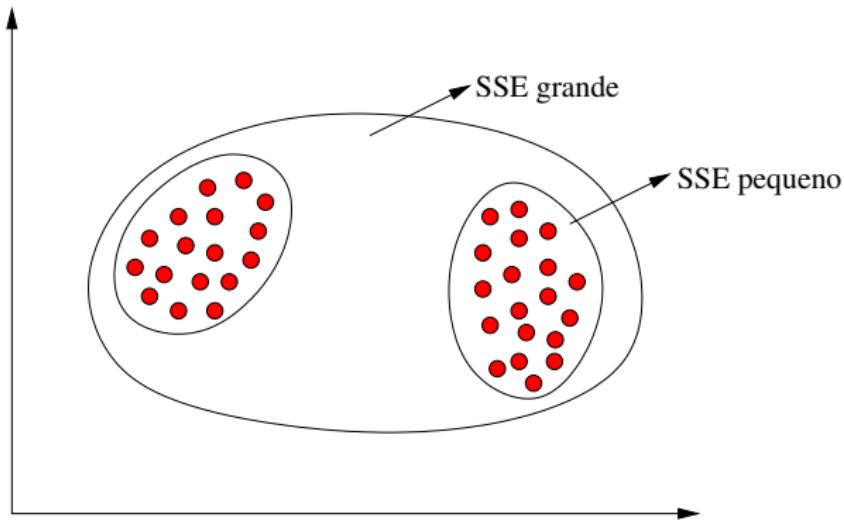
Função objetivo: minimizar a soma dos quadrados das distâncias (SSE), tal que:

$$\text{SSE} = \sum_{i=1}^K \sum_{x \in C_i} \|x - m_i\|^2$$

em que

- K é o número de grupos.
- x é o ponto no espaço representando um dado objeto.
- m_i é o centroide do grupo C_i contendo n_i pontos, dado por: $m_i = \frac{1}{n_i} \sum_{x \in C_i} x$.

Algoritmo K-Means: Função Objetivo



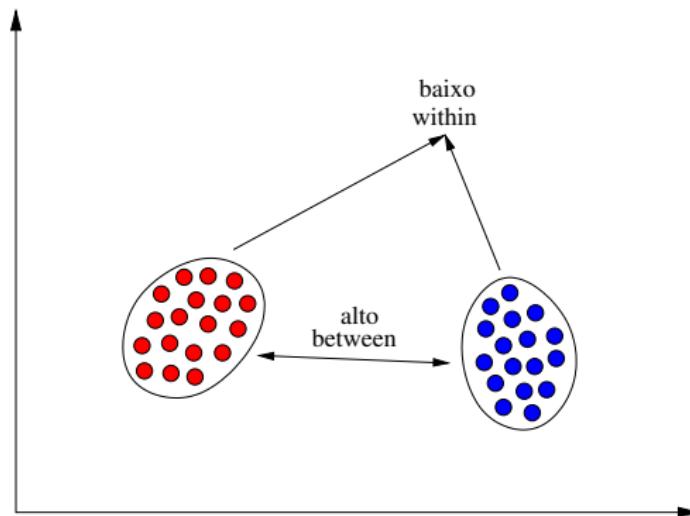
Algoritmo K-Means

Outras medidas: dispersão

- *within-cluster*: dispersão no mesmo grupo.
- *between-cluster*: dispersão entre diferentes grupos.

Algoritmo K-Means

Relação *within-between*

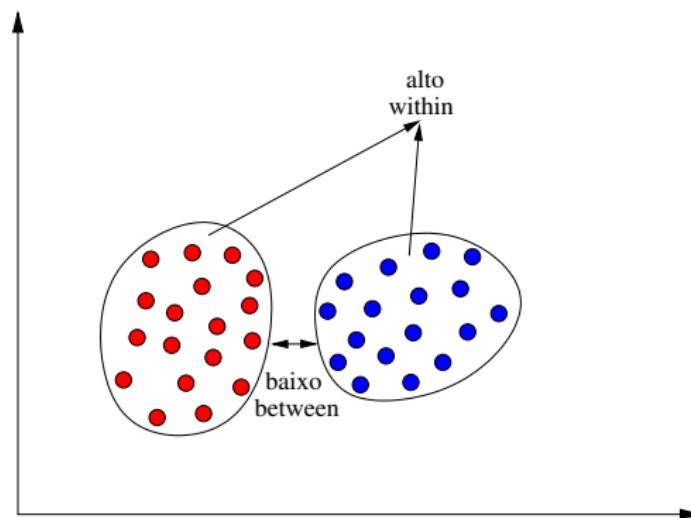


Caso ideal:

- alto *between*: grupos distantes um do outro.
- baixo *within*: boa compactação.

Algoritmo K-Means

Relação *within-between*



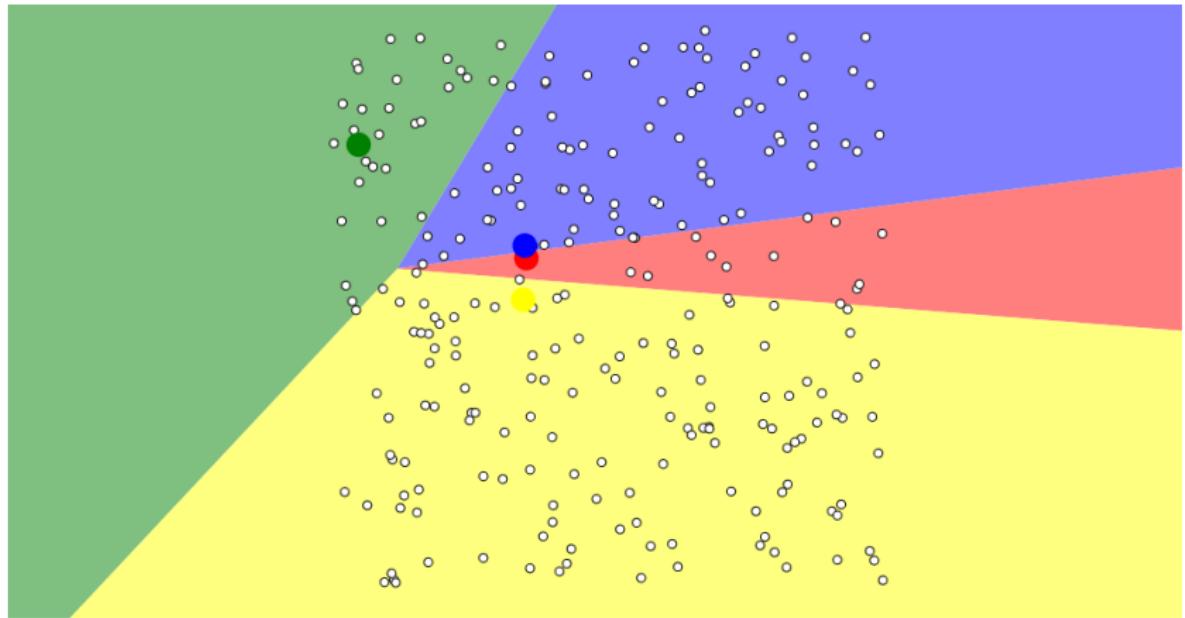
Caso não ideal:

- baixo *between*: grupos próximos um do outro.
- alto *within*: grupos dispersos.

Algoritmo K-Means: Inicialização

1. Defina os k centroides.
 - ▶ Inicialize os centroides aleatoriamente.
2. Encontre o centroide mais próximo e atualize as atribuições do grupo:
 - ▶ Atribua cada dado a um dos k grupos. Cada dado é atribuído ao grupo do centroide mais próximo (distância Euclidiana).
3. Mova os centroides para o centro de seus grupos.
 - ▶ A nova posição de cada centroide é calculada como a média de todos os pontos em seu grupo.
4. Repita os passos 2 e 3 até que o centroide pare de se mover muito em cada iteração.

Algoritmo K-Means: Inicialização

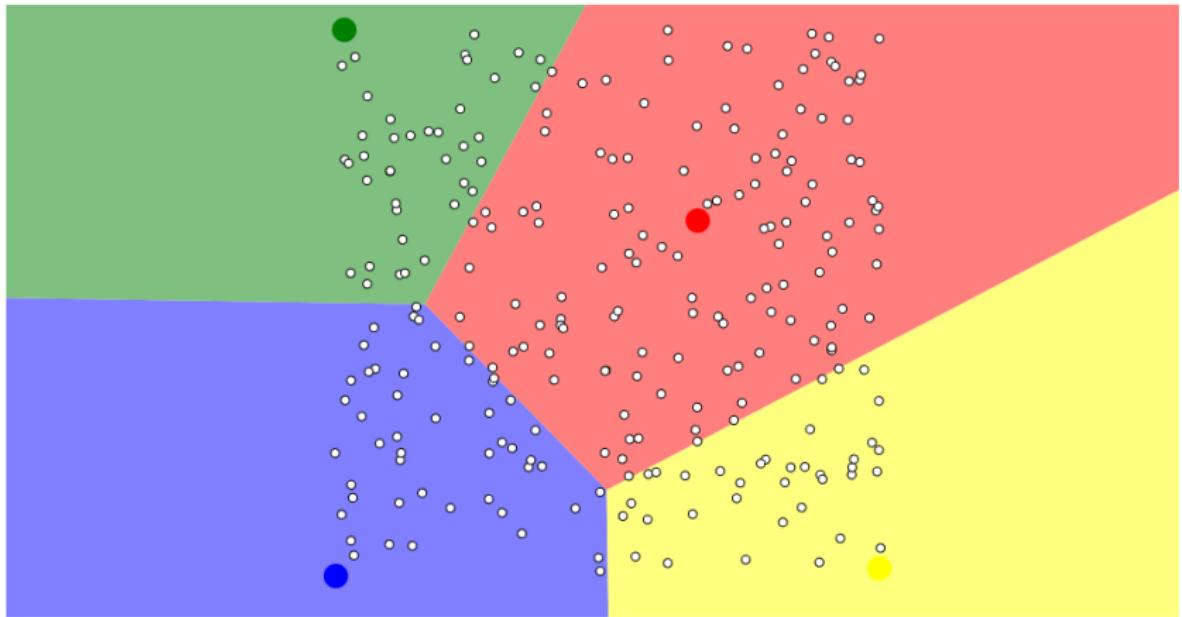


Algoritmo K-Means: Inicialização

Podemos fazer algo melhor?

- Uma ideia para inicializar o algoritmo K-Means é utilizar o ponto mais distante do conjunto de dados, para escolher k pontos que estão longe um do outro.

Algoritmo K-Means: Inicialização



Algoritmo K-Means: Inicialização

Podemos fazer algo melhor?

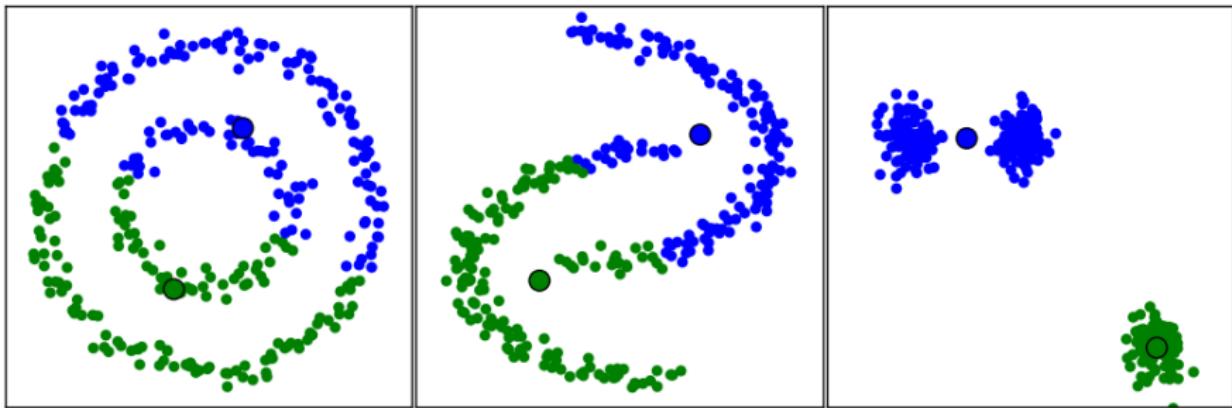
- Uma ideia para inicializar o algoritmo K-Means é utilizar o ponto mais distante do conjunto de dados, para escolher k pontos que estão longe um do outro.
- Problema: Essa estratégia é muito sensível a *outliers*.

Algoritmo K-Means: Inicialização

Podemos fazer algo melhor?

- **K-Means++**
 - ▶ Funciona de forma semelhante à heurística *mais distante*.
 - ▶ Escolhe cada ponto aleatoriamente, com probabilidade proporcional à sua distância ao quadrado dos centros já escolhidos.

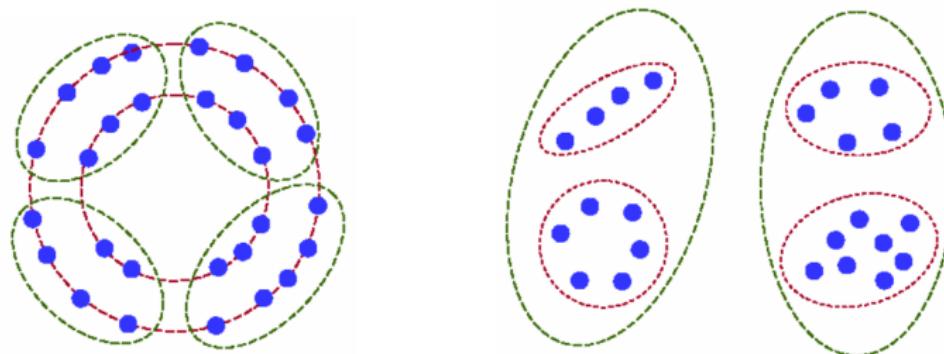
Limitações do K-Means



- K-Means encontra grupos *convexos*.
- K-Means não lida bem com grupos de tamanho e densidade muito diferentes.
- K-Means acaba agrupando dados que não têm estrutura de grupos.

Limitações do K-Means

- Impacto da escolha da métrica de distância:
 - A escolha da métrica de distância afeta o agrupamento final produzido.
 - A validade do agrupamento final é subjetiva.



- Quais os grupos significativos no exemplo?
- Quantos grupos devem ser considerados?

Limitações do K-Means

- A distância é a medida mais natural para dados numéricos.
- Valores baixos indicam maior similaridade.
- Alguns exemplos de medidas de distância:
 - ▶ Minkowski
 - ▶ Manhattan (City-Block)
 - ▶ Euclidiana
 - ▶ Correlação

Validação dos Agrupamentos

- Em classificação supervisionada, há uma grande variedade de medidas para avaliar a eficácia de um modelo: acurácia, precisão, sensibilidade, especificidade, medida F1, curva ROC, entre outras.
- Para análise de agrupamentos, como avaliar a qualidade dos grupos gerados?

Validação dos Agrupamentos

- Avaliar o desempenho de um algoritmo de agrupamento não é tão trivial quanto contar o número de erros ou a precisão de um algoritmo de classificação.

Validação dos Agrupamentos

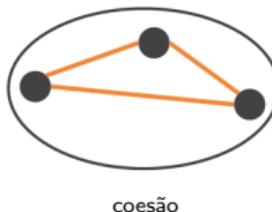
- Em geral, os agrupamentos são avaliados por especialistas de forma subjetiva.
- Então, por que precisamos avaliar agrupamentos?
 - ▶ Para comparar algoritmos de agrupamentos.
 - ▶ Para comparar grupos gerados por mais de um algoritmo.

Validação dos Agrupamentos

- Medidas internas: usam apenas os dados originais.
- Medidas externas: usam informação *extra* sobre os dados, em particular, a classe a que eles pertencem.

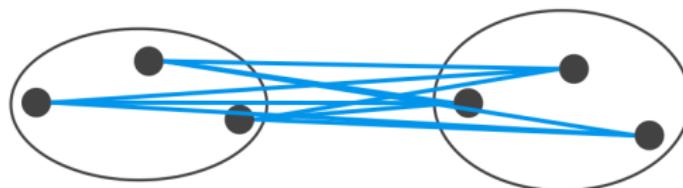
Medidas Internas

- Coesão de um grupo (intra-grupo): os pontos no interior de um mesmo grupo devem estar próximos entre si.



coesão

- Separação dos diferentes grupos (inter-grupo): os grupos devem estar distantes entre si.



separação

Medidas Internas

Em relação ao erro quadrático (SSE):

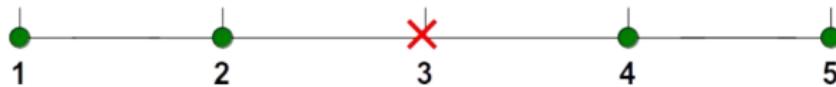
- Coesão é medida pela SSE interna (dentro de um grupo).

$$\text{SSE}_{\text{Grupo}} = \sum_{x \in C_i} \text{dist}(c_i, x)^2 = \frac{1}{2m_i} \sum_{x \in C_i} \sum_{y \in C_i} \text{dist}(x, y)^2$$

- Separação é medida pela soma de quadrados entre grupos.

$$\text{SSB}_{\text{Total}} = \sum_{i=1}^K m_i \text{dist}(c_i, c)^2$$

Medidas Internas



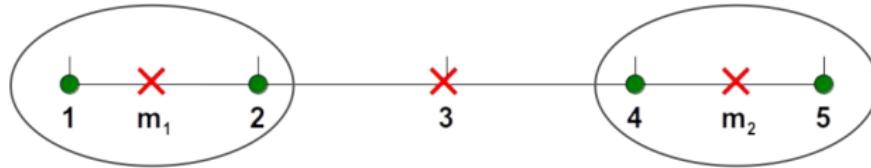
$k = 1$ grupo:

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

$$SSB = 4 \cdot (3 - 3)^2 = 0$$

$$\text{Total} = 10$$

Medidas Internas



$k = 1$ grupo:

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

$$SSB = 4 \cdot (3 - 3)^2 = 0$$

$$\text{Total} = 10$$

$k = 2$ grupos:

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

$$SSB = 2 \cdot (1.5 - 3)^2 + 2 \cdot (4.5 - 3)^2 = 9$$

$$\text{Total} = 10$$

Coeficiente de Silhueta

- Medida que combina coesão (os dados de um mesmo grupo são semelhantes) e separação (os dados de um grupo são distantes em comparação aos outros grupos).
- A silhueta varia de -1 a +1.
 - ▶ Valor alto: a configuração de grupo é apropriada.
 - ▶ Valor baixo: a configuração de grupo pode ter muitos ou poucos grupos.

Coeficiente de Silhueta

- O coeficiente de silhueta é definido **para cada dado** e é composto por duas pontuações:
 - ▶ A distância média a entre um dado e todos os outros dados no mesmo grupo.
 - ▶ A distância média b entre um dado de um grupo e todos os outros dados do grupo mais próximo.

Coeficiente de Silhueta

- O coeficiente de silhueta s para uma única amostra é dado por:

$$s = \frac{b - a}{\max(a, b)}$$

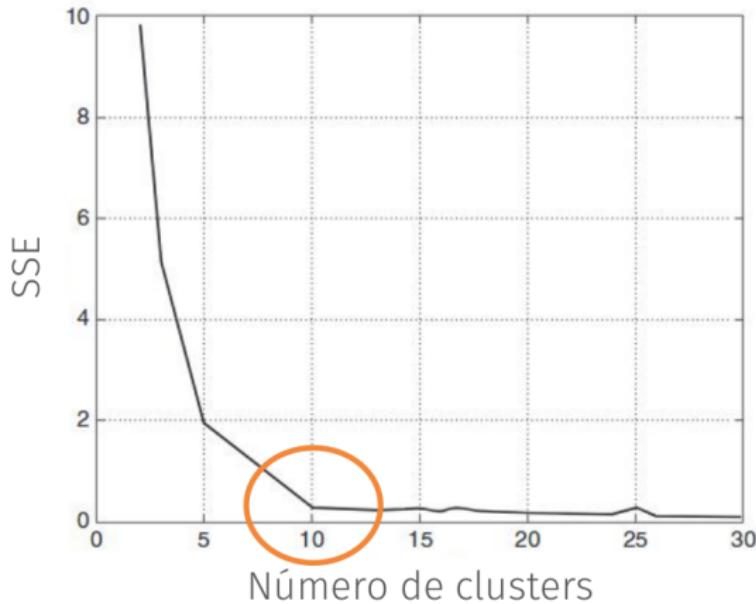
- A pontuação é limitada entre -1 para grupo incorreto e +1 para grupo altamente denso ($a \ll b$).

Algoritmo K-Means

1. Defina os k centroides.
 - ▶ Inicialize os centroides aleatoriamente.
2. Encontre o centroide mais próximo e atualize as atribuições do grupo:
 - ▶ Atribua cada dado a um dos k grupos. Cada dado é atribuído ao grupo do centroide mais próximo (distância Euclidiana).
3. Mova os centroides para o centro de seus grupos.
 - ▶ A nova posição de cada centroide é calculada como a média de todos os pontos em seu grupo.
4. Repita os passos 2 e 3 até que o centroide pare de se mover muito em cada iteração.

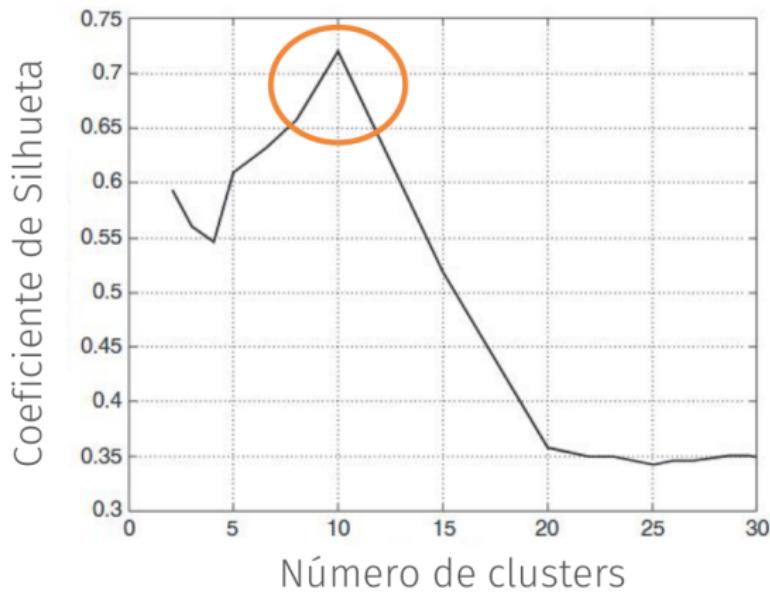
Como Escolher o Valor k ?

Valor de SSE



Como Escolher o Valor k ?

Coeficiente de Silhueta



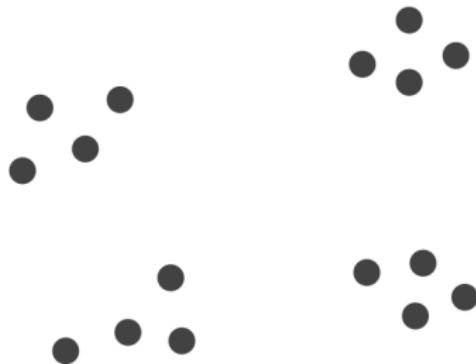
Algoritmo K-Means: Problemas Adicionais

- Reduzir o SSE com pós-processamento.
- Duas estratégias que diminuem o SSE, aumentando o número de grupos, são as seguintes:
 1. Dividir um grupo: o grupo com o maior SSE é geralmente escolhido.
 2. Introduzir um novo centroide: muitas vezes, o ponto que está mais distante de qualquer centro de grupo é escolhido.

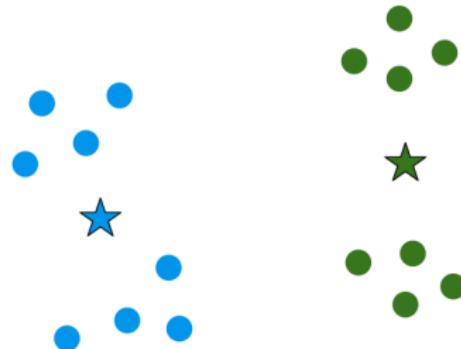
Algoritmo Bisecting K-Means

- Extensão direta do algoritmo K-Means.
- Ideia simples:
 1. Dividir o conjunto de todos os pontos em dois grupos.
 2. Escolher o grupo com o maior SSE para dividir novamente.
 3. Repetir o passo 2 até que os k grupos tenham sido produzidos.
- Menos suscetível a problemas de inicialização do K-Means.

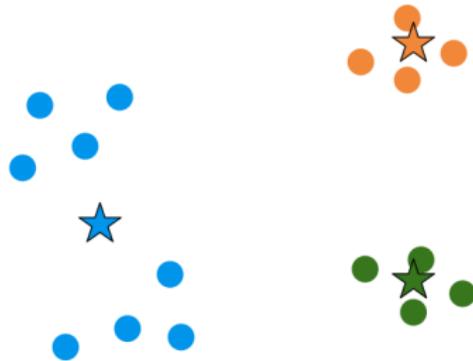
Algoritmo Bisecting K-Means



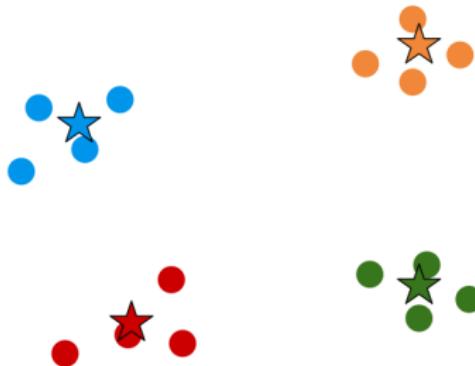
Algoritmo Bisecting K-Means



Algoritmo Bisecting K-Means



Algoritmo Bisecting K-Means



K-Medians

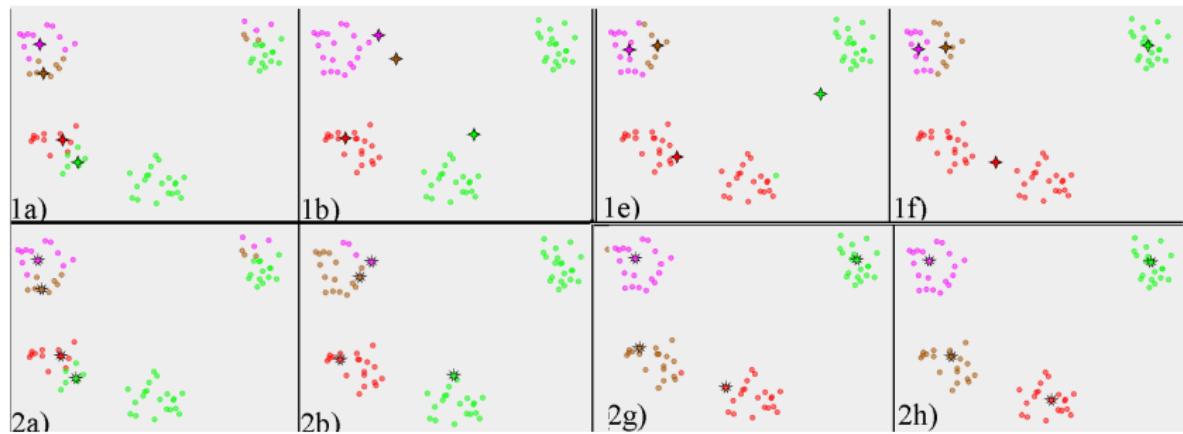
- Em vez de calcular a média de cada grupo para determinar seu centroide, calcula-se a mediana do grupo.
- K-Medians é menos sensível a *outliers*: pontos anômalos não deslocam o centroide para eles.

K-Medoids

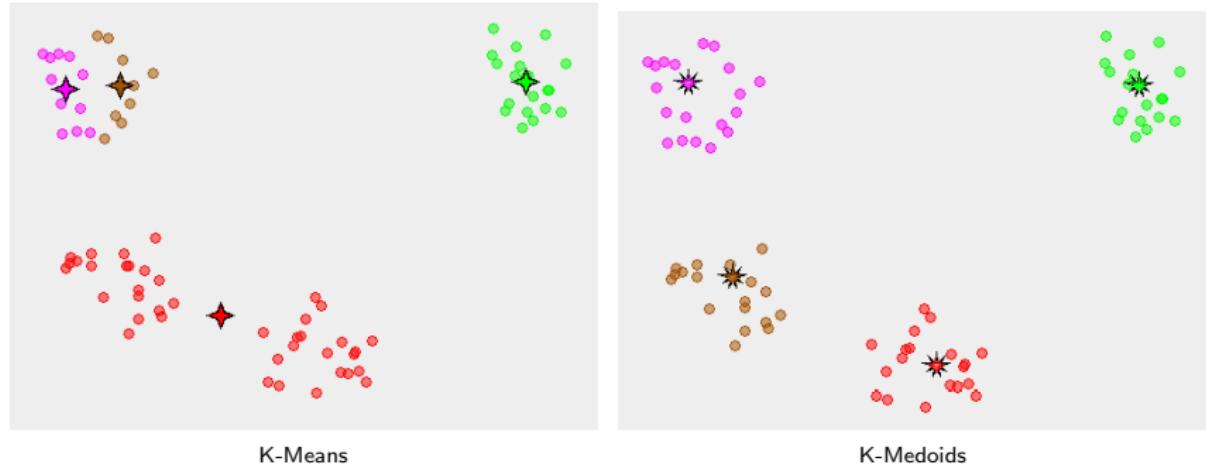
- Em vez de calcular a média de cada grupo para determinar seu centroide, calcula-se o medoide.
- O medoide é o dado do grupo cuja soma das distâncias aos elementos do grupo é a menor.
- Os medoides não são *pontos novos* do espaço, mas sim escolhidos entre os próprios dados.
- Em contraste com o K-Means, K-Medoids escolhe pontos de dados como centroides.

K-Means versus K-Medoids

K-Means (superior) versus K-Medoids (inferior)



K-Means versus K-Medoids



K-Means

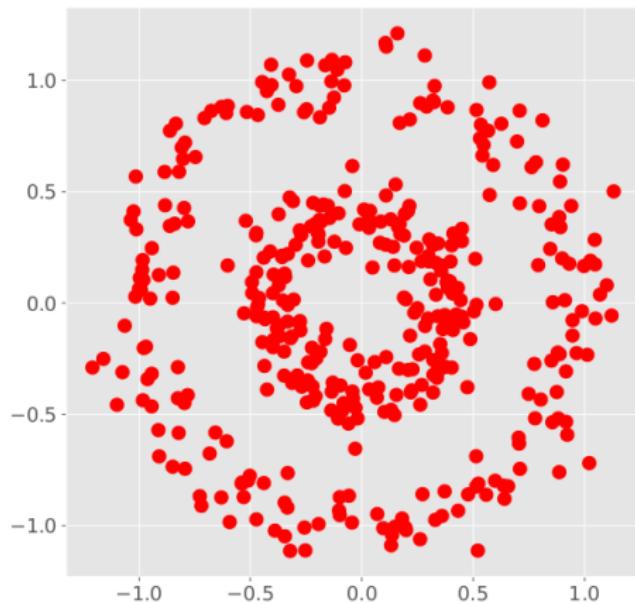
K-Medoids

Fuzzy C-Means

- Fuzzy C-Means: cada ponto pertence a um grupo com *intensidade* proporcional ao inverso de uma potência da distância (normalizada, para que as intensidades somem 1 no final).
- Fuzzy C-Means tem um parâmetro extra (m fuzzyfication) relacionado à potência da distância:
 - ▶ m grande torna as intensidades dos pontos aproximadamente similares e, portanto, os grupos têm grande intersecção.
 - ▶ $m = 1$ faz a intensidade ser 1 para o grupo mais próximo e 0 para os outros, portanto, o algoritmo K-Means tradicional.
 - ▶ $m = 2$ é normalmente usado.
- O centroide é a média ponderada (pela intensidade) dos pontos.

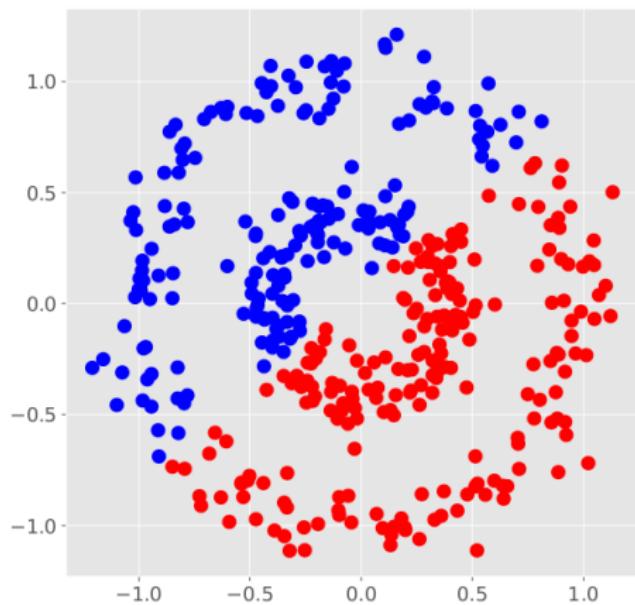
Agrupamento Espectral

- Certos conjuntos de dados apresentam agrupamentos não óbvios quando mensurados por uma medida de similaridade.



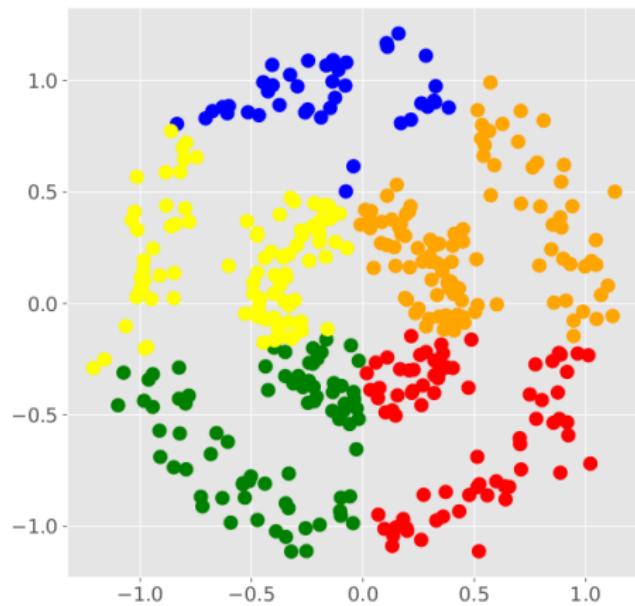
Agrupamento Espectral

- A técnica de agrupamentos K-Means é insuficiente para encontrar dois grupos existentes:



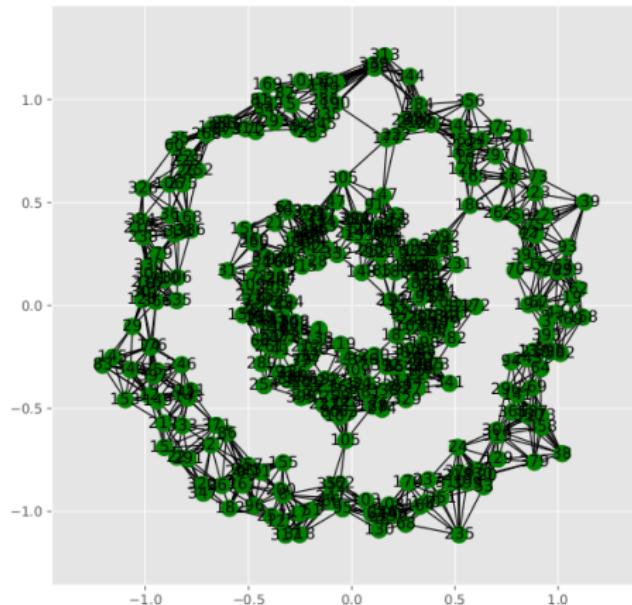
Agrupamento Espectral

- Mesmo com um número maior de grupos, a técnica K-Means apresenta dificuldades para representar a estrutura do conjunto de dados.



Agrupamento Espectral

- Uma forma de representar os pontos de dados é por meio de um grafo, em que os k pontos mais próximos de um determinado ponto formam uma aresta com ele.



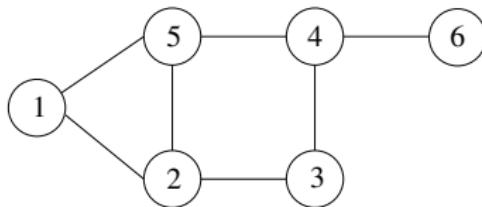
Agrupamento Espectral

- Os dados podem ser representados pela matriz Laplaciana, dado por:

$$L = D - A$$

em que A é a matriz de adjacência e D uma matriz diagonal com os elementos da diagonal igual ao grau do nó correspondente.

- Exemplo:



$$L = \begin{bmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$

Agrupamento Espectral

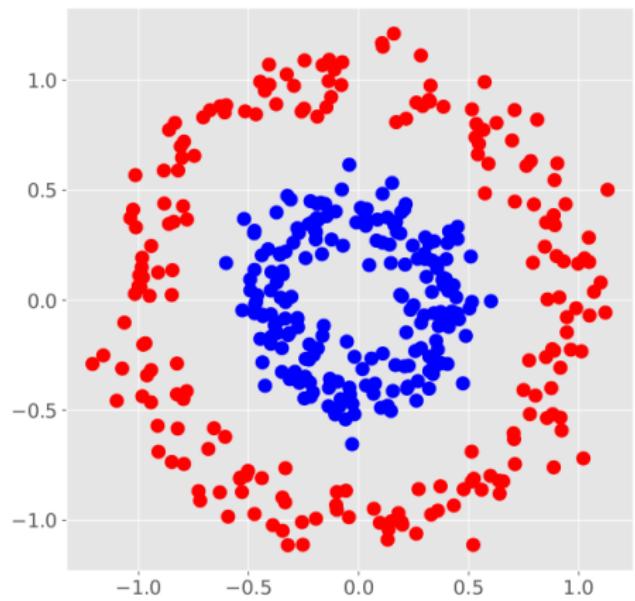
- A matriz Laplaciana tem algumas propriedades interessantes:
 - ▶ O número de autovalores iguais a 0 é igual ao número de componentes conexos.
 - ▶ Os autovetores correspondentes aos autovalores iguais a 0 representam um grupo, sendo os nós pertencentes a esse grupo com valores positivos e todo o restante igual a 0.
 - ▶ Os autovetores restantes representam diversas formações de possíveis agrupamentos.

Agrupamento Espectral

- Pode-se calcular a matriz Laplaciana de uma base de dados e utilizar os k primeiros autovetores dessa matriz (com autovalores diferentes de 0) e gerar (de forma similar ao PCA) uma matriz $n \times k$ contendo a informação dos grupos.
- Nesse ponto, tem-se a projeção dos dados em um espaço de dimensão reduzida.
- Pode-se agora aplicar a técnica K-Means utilizando essa representação para se obter os grupos.

Agrupamento Espectral

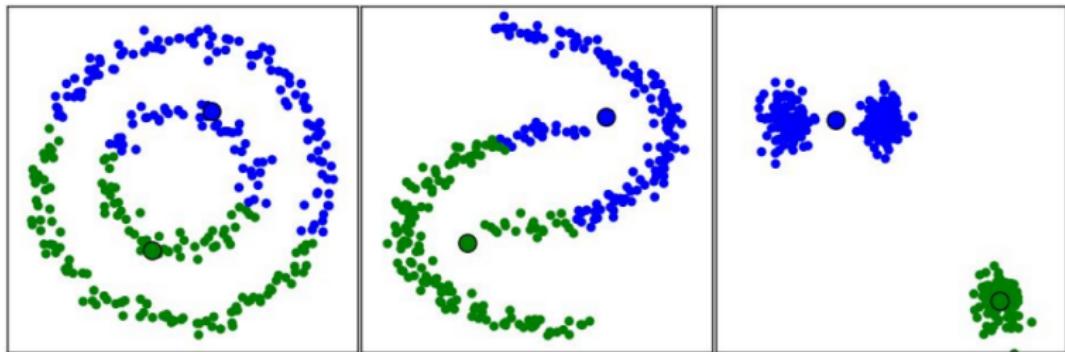
- A partir dessas ideias, os grupos são corretamente identificados no exemplo.



DBSCAN

Density Based Spatial Clustering of Application with Noise (DBSCAN)

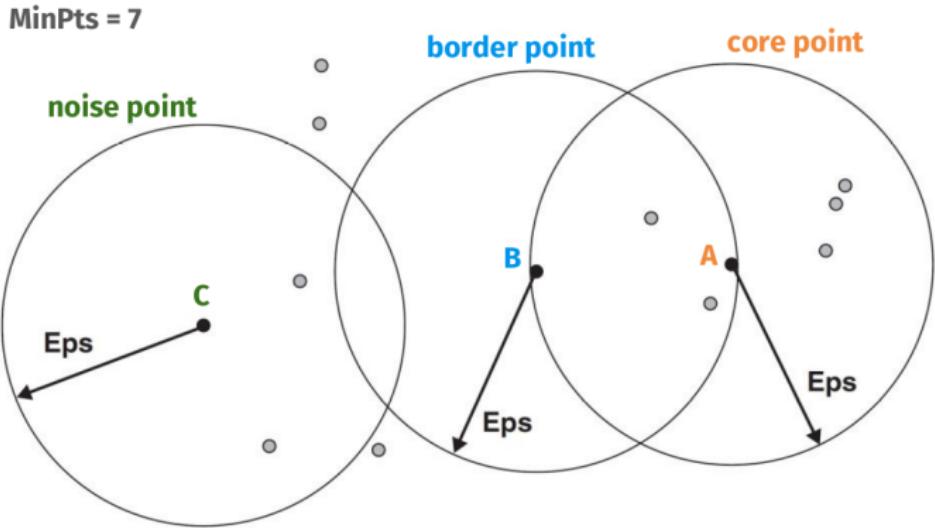
- Algoritmo de agrupamento baseado em densidade: grupos são regiões de alta densidade de pontos separados por regiões de baixa densidade.



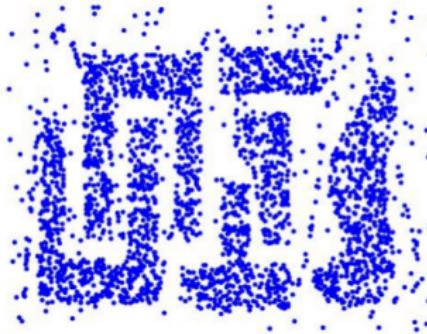
limitações do K-Means

- Densidade é o número de pontos no interior de um raio específico (Eps).
- Se o raio for grande o suficiente, então todos os pontos terão uma densidade igual a m (o número de pontos no conjunto de dados).
- Da mesma forma, se o raio for muito pequeno, todos os pontos terão uma densidade igual a 1.

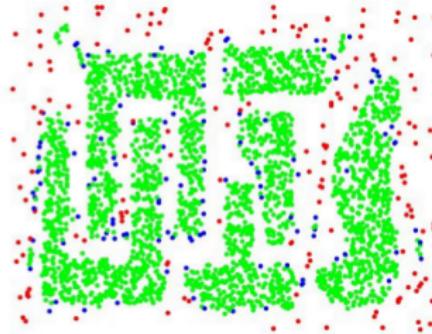
- Um ponto de centro (*core point*) tem um número mínimo de pontos especificado pelo usuário (MinPts) no interior do raio (Eps).
- Um ponto de fronteira (*border point*) fica localizado na vizinhança de um ponto de centro.
- Um ponto de ruído (*noise point*) é qualquer ponto que não se classifica como ponto de centro nem como ponto de fronteira.



1. Arbitrariamente, selecione um ponto p .
2. Identifique todos os pontos densamente conectados a p com relação aos parâmetros Eps e MinPts.
 - ▶ Se p é um ponto de centro, um grupo é formado.
 - ▶ Se p é um ponto de fronteira e não há pontos densamente conectados a p , DBSCAN visita o próximo ponto do conjunto de dados.
 - ▶ Continua o processo até que todos os pontos do conjunto de dados tenham sido analisados.



Pontos originais

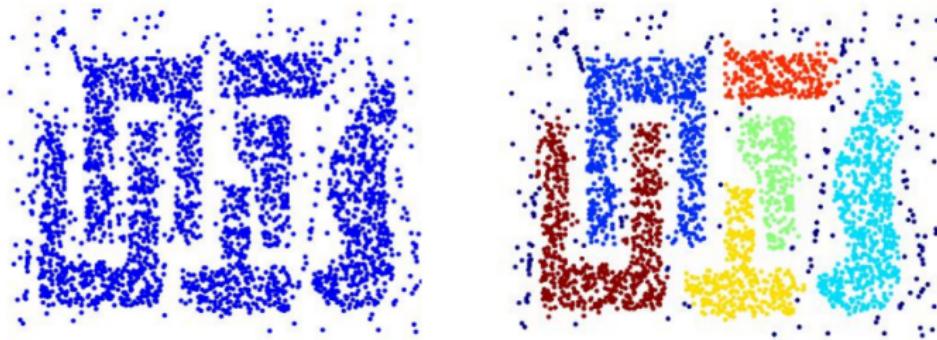


Core, border e noise

Eps = 10, MinPts = 4

DBSCAN

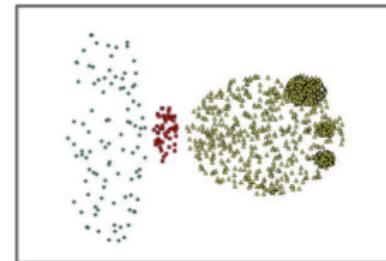
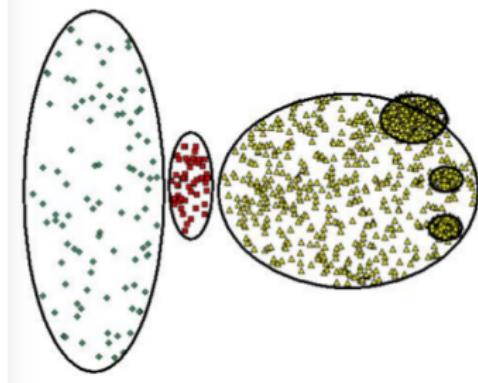
Quando DBSCAN funciona bem?



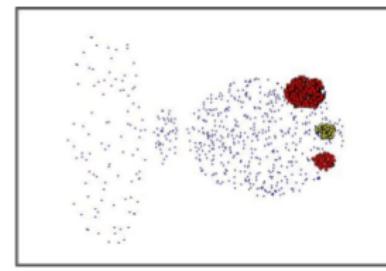
- Na presença de ruído.
- Na geração de grupos com diferentes formatos e tamanhos.

DBSCAN

Quando DBSCAN não funciona bem?



(MinPts=4, Eps=9.75)



(MinPts=4, Eps=9.92)

- Variação na densidade dos pontos.
- Dados com alta dimensionalidade.

Como escolher Eps e MinPts?

- Determinar, para cada ponto, a distância do k -ésimo vizinho mais próximo.
- Recomendação: $k = 4$ ou $k = 5$ (esse será o MinPts).
- Mostre em um gráfico a distância ordenada do k -ésimo vizinho mais próximo de cada dado.
- Deve haver um ponto de inflexão (*cotovelo* a partir do qual a distância do k -ésimo vizinho aumenta (essa distância será o Eps)).