Agrupamento

ACH5504 – Mineração de Dados

Notas de aulas baseadas no livro

"*Introduction to Data Mining*"

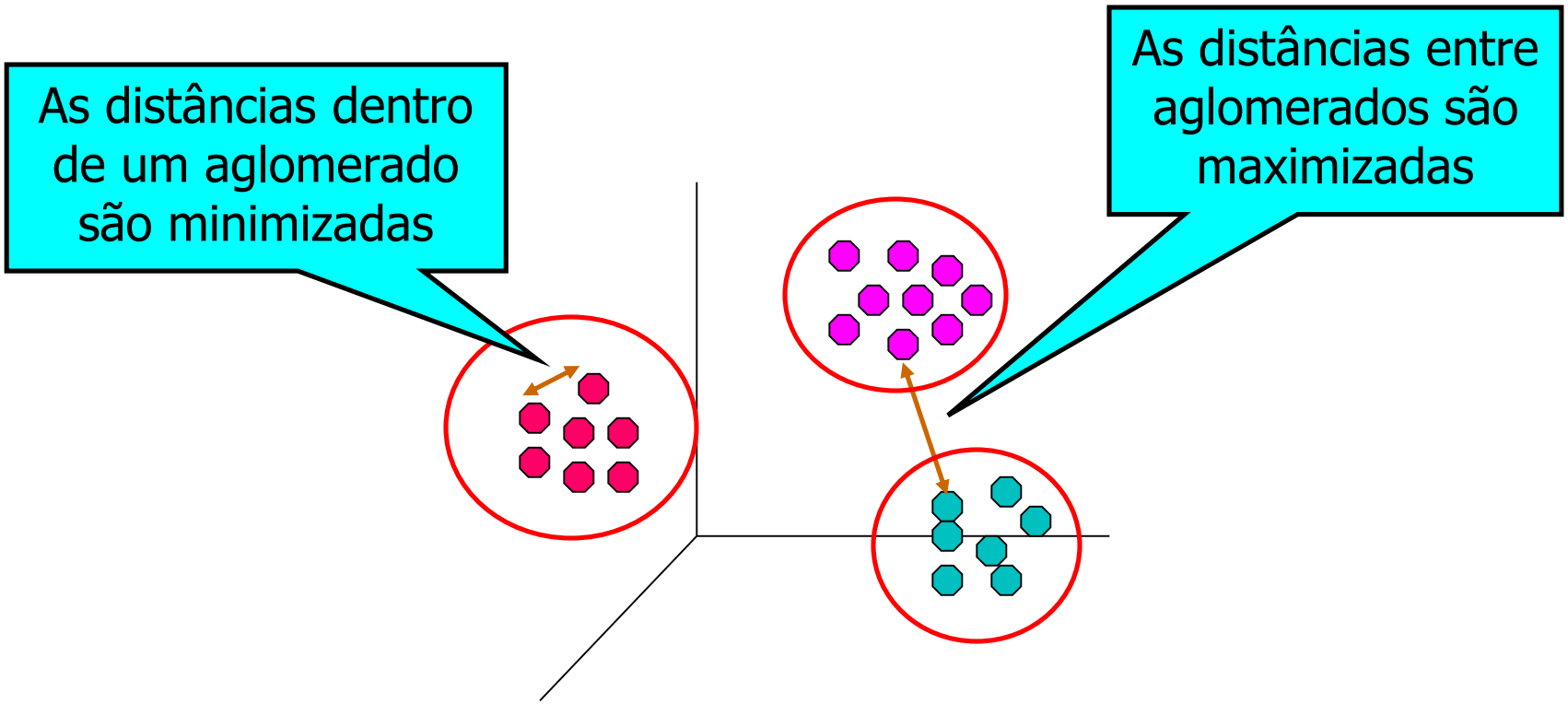
Tan, Steinbach, Karpatne, Kumar

Resumo

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

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



O que não é análise de agrupamento?

- Segmentação simples
  - Dividindo os alunos em grupos de registro diferentes alfabeticamente, por sobrenome
- Resultados de uma consulta ou busca
  - Agrupamento é o resultado de uma especificação externa
  - Clustering é um agrupamento de objetos com base nos dados
- Classificação supervisionada
  - Com informações de rótulo de classe
- Análise de associação
  - Conexões locais versus globais

Noção de um cluster pode ser ambíguo

Quantos clusters?

Seis clusters

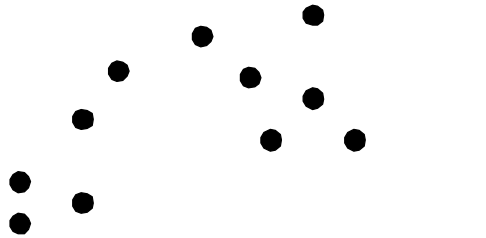
Dois clusters

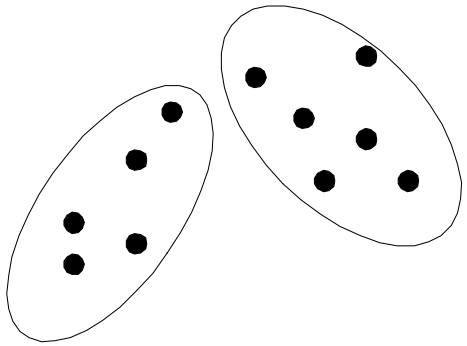
Quatro clusters

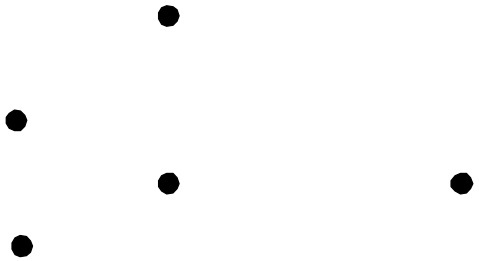
Tipos de agrupamento

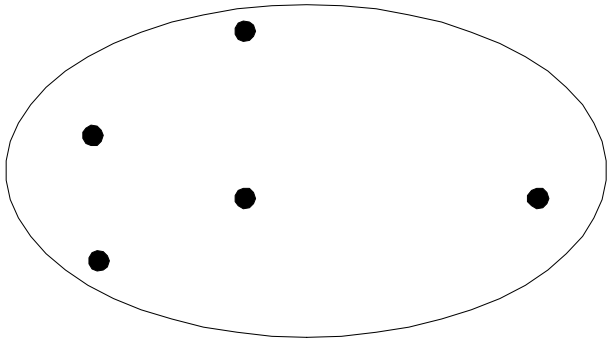
- Um agrupamento é um conjunto de clusters
- Distinção importante entre conjuntos hierárquicos e em partições de clusters
- Agrupamento em partições
  - Uma divisão de objetos de dados em subconjuntos não sobrepostos (clusters), de forma que cada objeto de dados esteja em exatamente um subconjunto
- Agrupamento hierárquico
  - Um conjunto de grupos organizados como uma árvore hierárquica

Agrupamento em partições





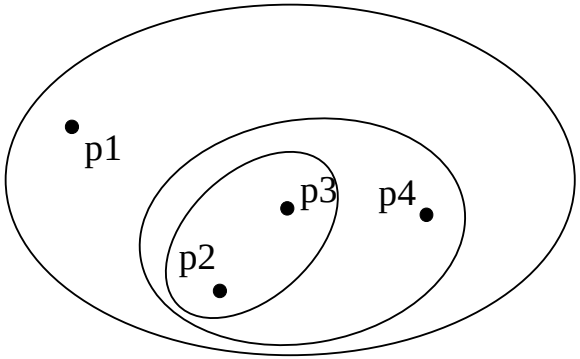




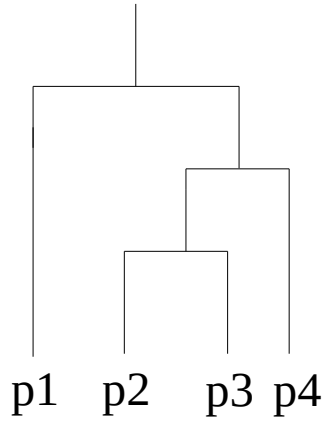
Pontos originais

Um agrupamento em partições

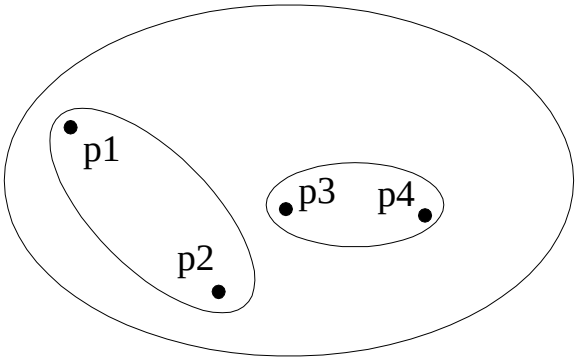
Agrupamento hierárquico



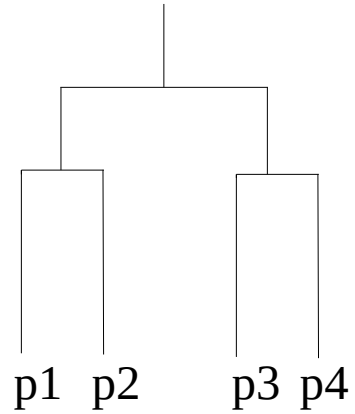
Agrupamento hierárquico tradicional



Dendrograma tradicional



Agrupamento hierárquico não tradicional



Dendrograma não tradicional

Outras distinções entre conjuntos de clusters

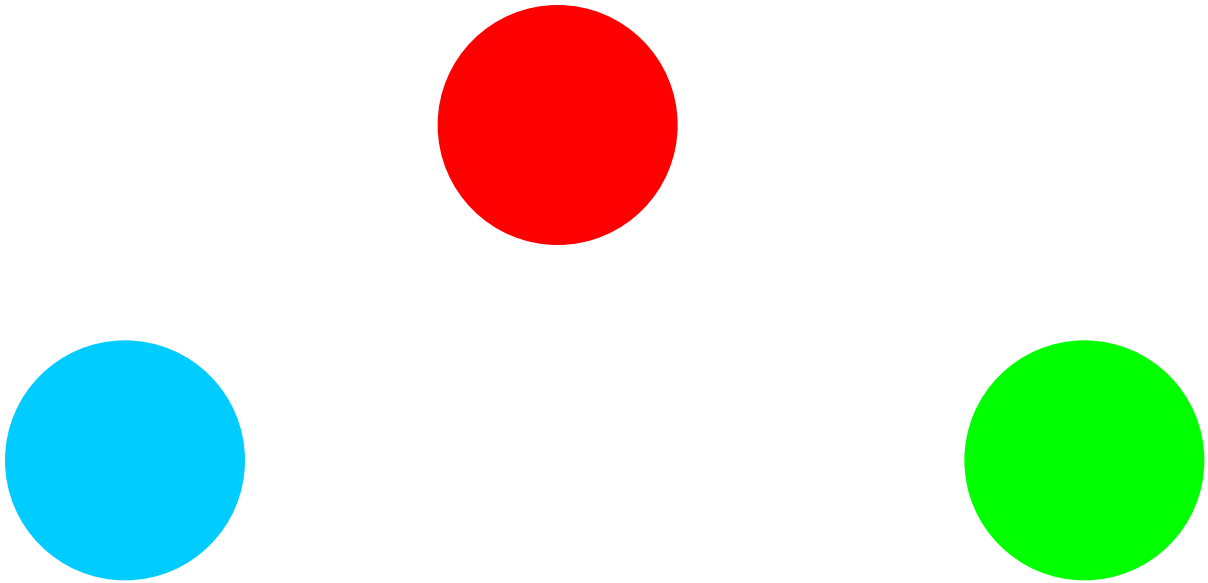
- Exclusivo versus não exclusivo
  - Em agrupamento não-exclusivo, os pontos podem pertencer a vários clusters.
  - Podem representar várias classes ou pontos de “fronteira”
- Fuzzy versus não fuzzy
  - Em agrupamento tipo fuzzy, um ponto pertence a cada cluster com algum peso entre 0 e 1
  - Os pesos devem somar a 1
  - O agrupamento probabilístico tem características semelhantes
- Parcial versus completo
  - Em alguns casos, desejamos agrupar parte dos dados
- Heterogêneo versus homogêneo
  - Grupos de tamanhos, formas e densidades amplamente diferentes

Tipos de clusters

- Clusters bem separados
- Clusters baseados em centros
- Clusters contíguos
- Clusters baseados em densidade
- Propriedade ou Conceitual
- Descrita por uma função objetiva

Tipos de clusters: bem-separados

- Clusters bem separados:
  - Um cluster é um conjunto de pontos de tal forma que qualquer ponto em um cluster é mais próximo (ou mais semelhante) a todos os pontos dentro do cluster do que fora do cluster.



3 clusters bem separados

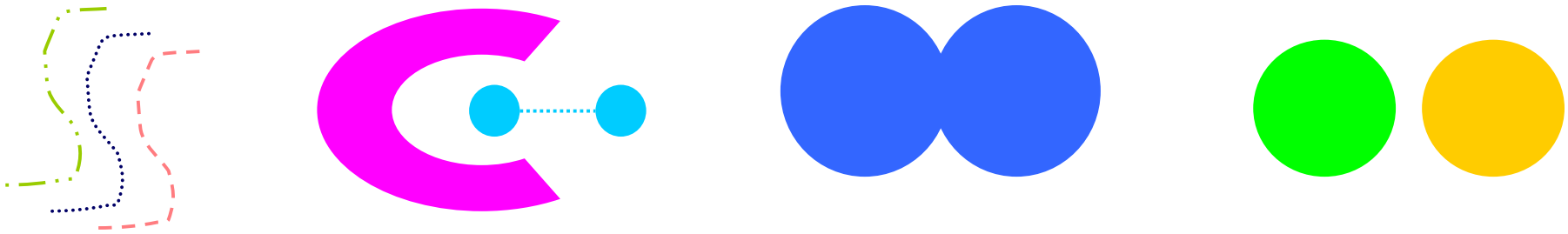
Tipos de clusters: baseado em centros

- Clusters baseados em centros
  - Um cluster é um conjunto de objetos de tal forma que um objeto em um cluster é mais próximo (mais semelhante) para o "centro" do cluster dele, do que para o centro de qualquer outro cluster.
  - O centro do grupo pode ser um centróide, a média de todos pontos do grupo, ou um medóide, o ponto mais “representativo” de um grupo

4 grupos baseados em centros

Tipos de clusters: baseado em conexidade

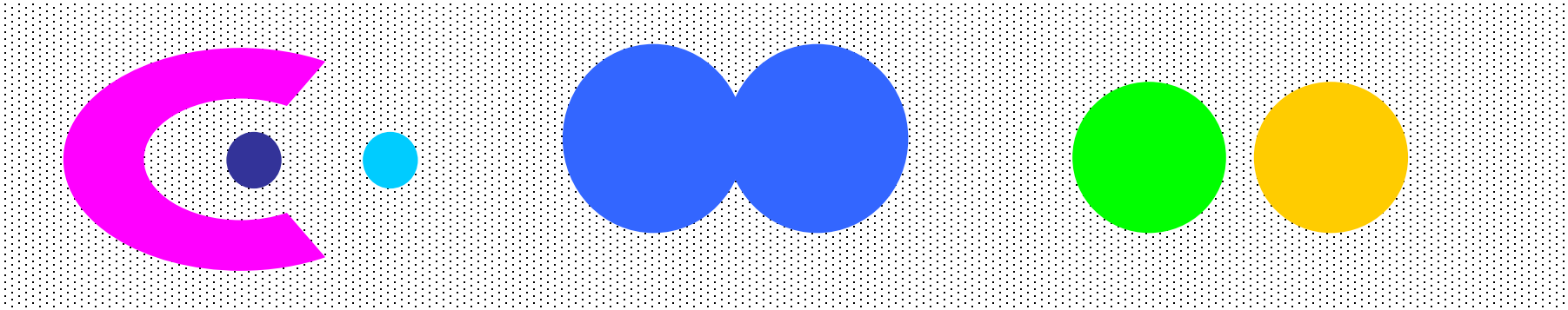
- Clusters contíguos (Nearest neighbor ou Transitivo)
  - Um cluster é um conjunto de pontos de tal forma que um ponto em um cluster é mais próximo (ou mais semelhante) a um ou mais outros pontos no cluster do que a qualquer ponto fora do cluster.



8 clusters contíguos

Tipos de clusters: baseados em densidade

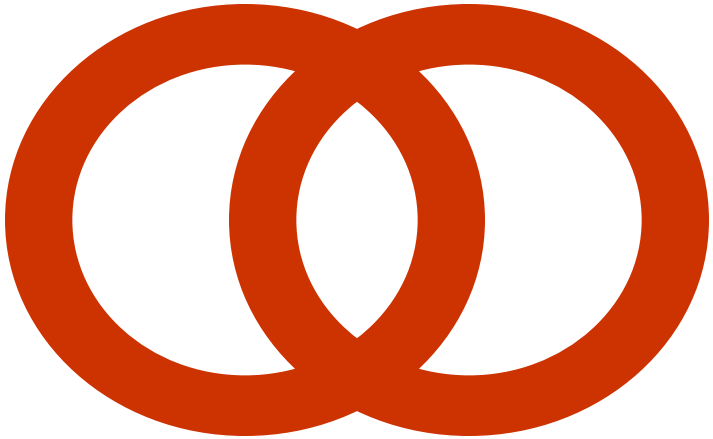
- Clusters baseados em densidade
  - Um cluster é uma região densa de pontos, separada por regiões de baixa densidade, de outras regiões de alta densidade.
  - Usado para clusters irregulares ou entrelaçados, e quando há ruído ou outliers.



6 clusters baseados em densidade

Tipos de clusters: clusters conceituais

- Clusters com propriedade ou conceito comum
  - Localiza clusters que compartilham alguma propriedade comum ou representam um conceito específico.



2 círculos sobrepostos

Tipos de clusters: função objetiva

- Clusters definidos por uma função objetiva
  - Localiza clusters que minimizam ou maximizam uma função objetiva.
  - Enumera todas as formas possíveis de divisão dos pontos em clusters e avalia a "bondade" de cada conjunto potencial de clusters usando a função objetiva dada.
  - Pode ter objetivos globais ou locais.
    - Algoritmos hierárquicos tipicamente tem objetivos locais
    - Algoritmos de partições tipicamente tem objetivo globais
  - A abordagem de variação de função objetivo global é ajustar os cluster a um modelo pré-definido.
    - Parâmetros do modelo são determinados a base dos dados.

Características dos dados de entrada

- Tipo de medida de proximidade ou densidade
  - Central para clustering
  - Depende dos dados e aplicações

- Características dos dados que afetam a proximidade e/ou densidade
  - Dimensionalidade
    - Esparcidade
  - Tipo de atributo
  - Relações especiais entre os dados
    - Autocorrelação, por exemplo
  - Distribuição de dados

- Ruído e Outliers
  - Frequentemente interfere na operação do algoritmo de clustering

Algoritmos de Clustering

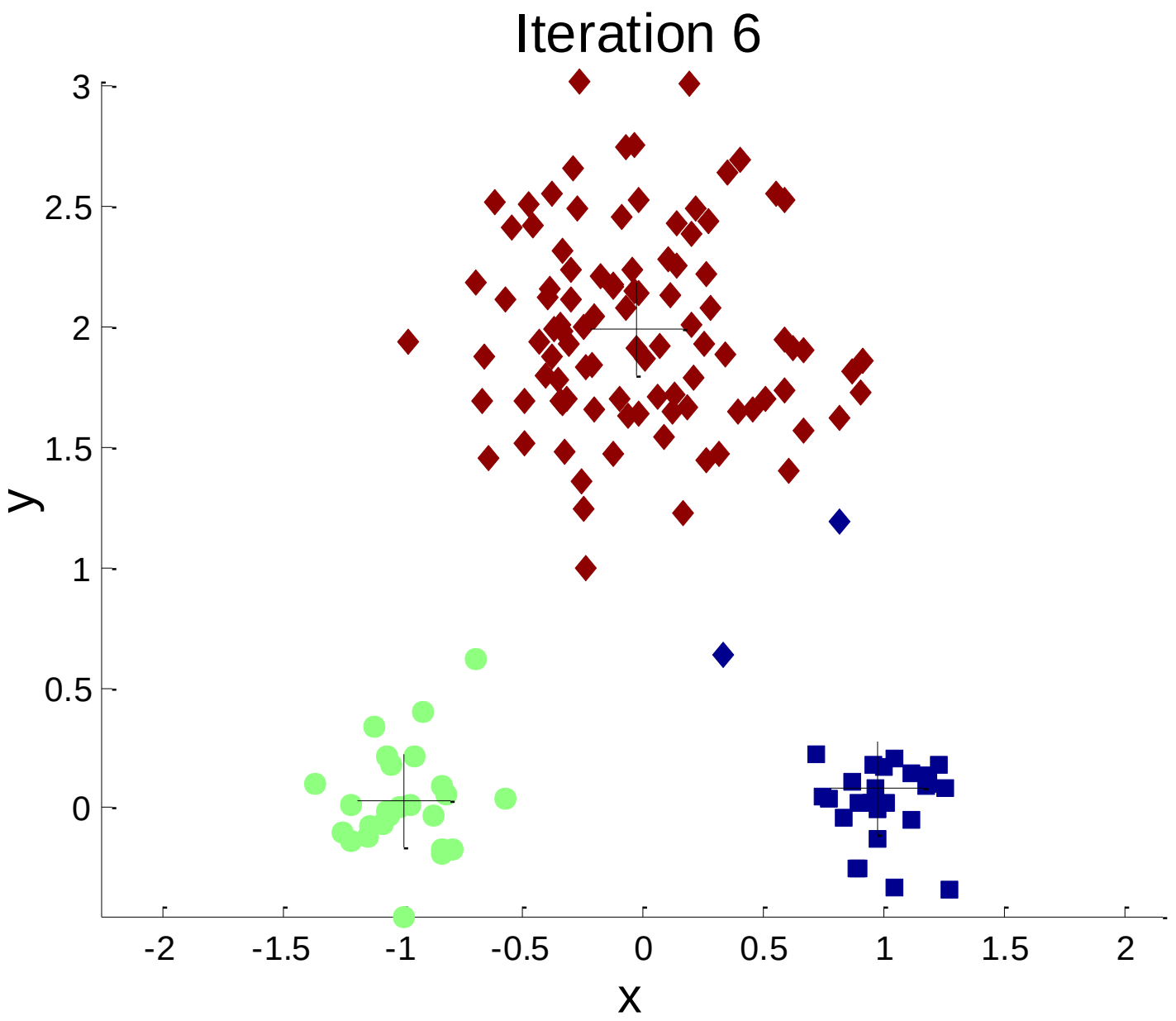
- Baseados em K-means
- Métodos heurísticos
- Baseados em densidade

Agrupamento K-means

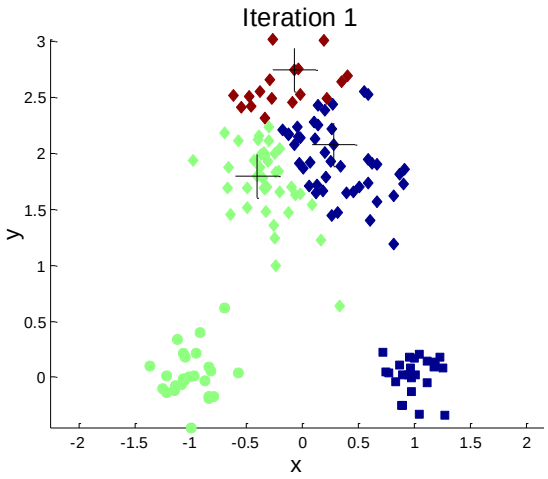
- Abordagem de agrupamento em partições
- Número de clusters, K, precisa ser especificado
- Cada cluster é associado com uma centroide (ponto central)
- Cada ponto é atribuído ao cluster com o centroide mais próximo
- O algoritmo básico é muito simples:

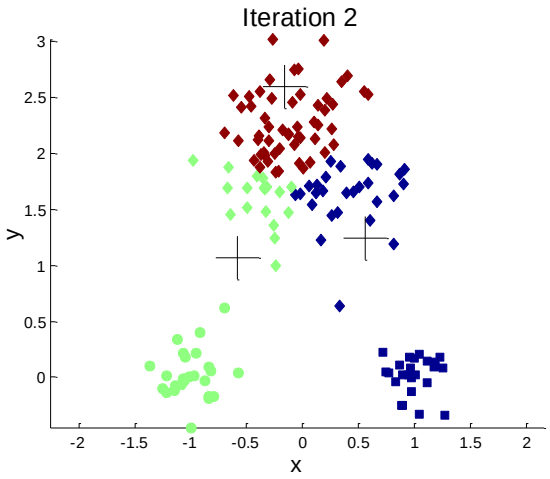
1. Escolhe K pontos como centroides iniciais
2. **Repita**
3. Forma K clusters atribuindo todos os pontos mais próximos a centroide
4. Recalcule o centroide de cada cluster
5. **Até** os centroides não mudam mais.

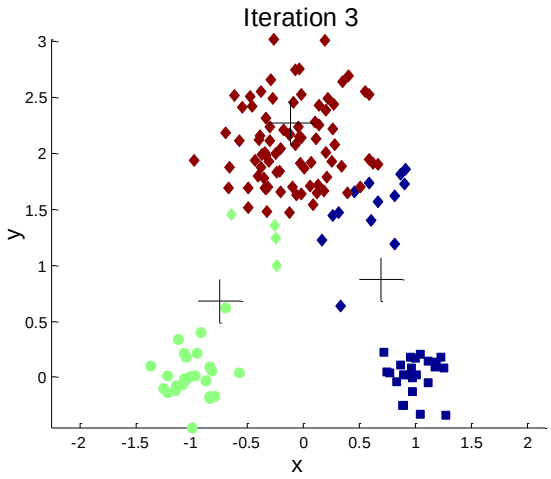
Exemplo de K-means Clustering

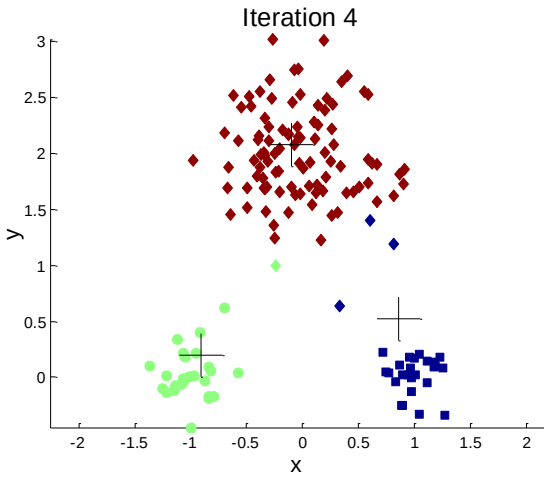


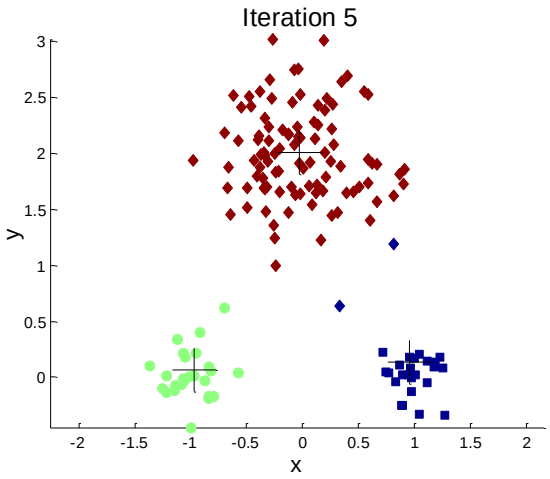
Exemplo de K-means Clustering

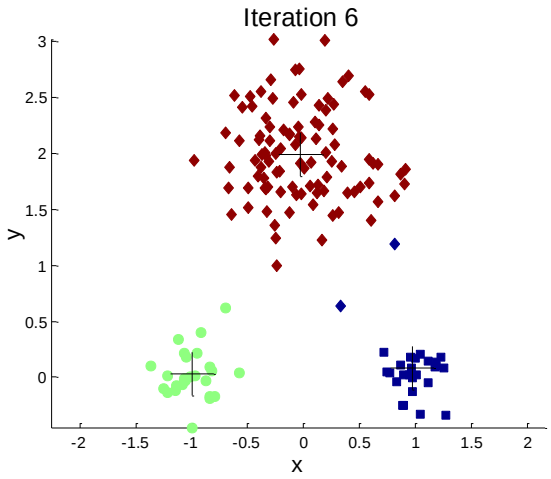












K-means Clustering – Detalhes

- Centroides iniciais são escolhidos aleatoriamente.
  - Clusters produzidos variam entre iterações.
- O centroide é (normalmente) a média dos pontos no cluster.
- Proximidade é medida usando a distância euclidiana, similaridade de cosseno, correlação, etc.
- K-means convergirá para medidas de similaridade acima.
- A maior parte da convergência ocorre nas primeiras iterações.
  - Frequentemente, a condição de parada é alterada para "Até que relativamente poucos pontos alterem clusters"
- Complexidade é **O(n × K × I × d)**
  - n = número de pontos, K = número de clusters, I = número de iterações, d = número de atributos

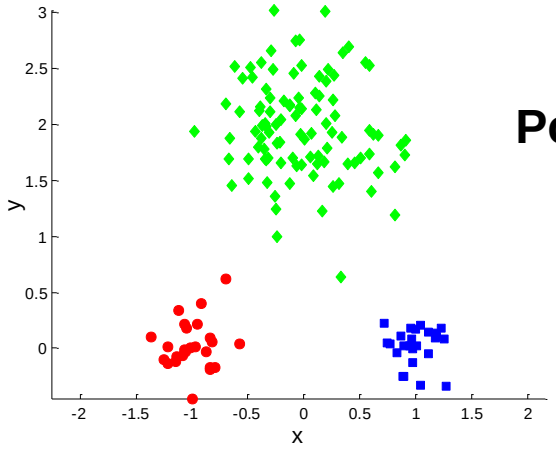
Avaliando Clusters produzidos por K-means

- A medida mais comum é soma do erro quadrático (Sum of Squared Error – SSE)
  - Para cada ponto, o erro é a distância do cluster mais próximo
  - Para obter o SSE, agrupamos esses erros e os somamos.

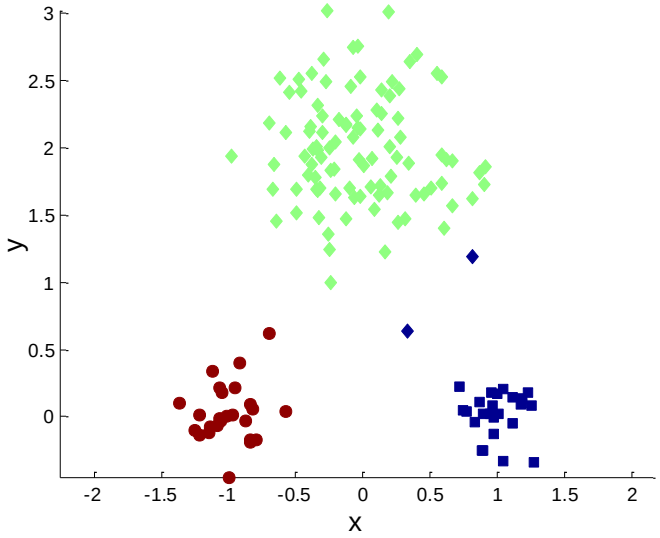
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - *x* é um ponto em cluster C_i e m_i é o ponto representativo de cluster C_i
    - Podemos mostrar que m_i corresponde o centro (média) do cluster
  - Dados dois conjuntos de clusters, a preferência é para o com menor erro
  - Um jeito simples de reduzir SSE é aumentar K, o número de clusters
    - Um bom agrupamento com K pequeno pode ter SSE menor do que um agrupamento ruim com K alto

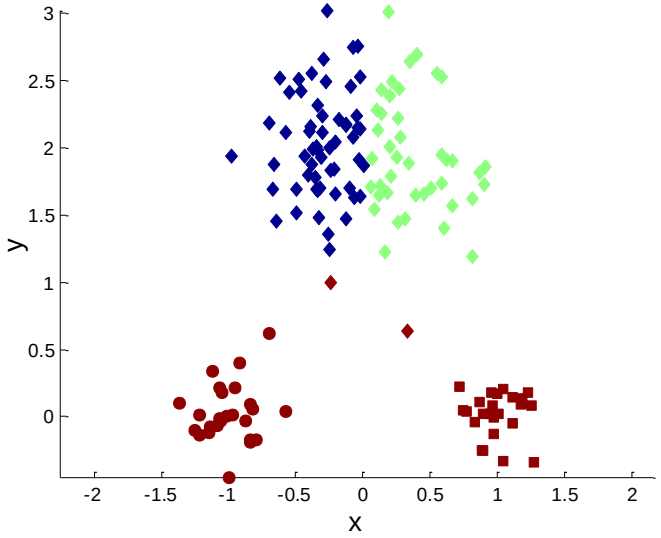
Dois agrupamentos K-means diferentes



Pontos originais



Agrupamento ótimo



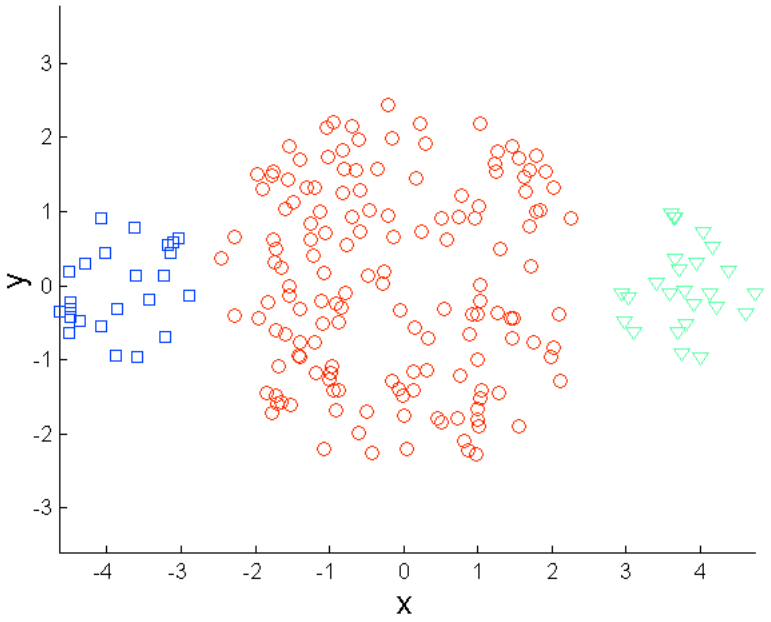
Agrupamento sub-ótimo

Limitações de K-means

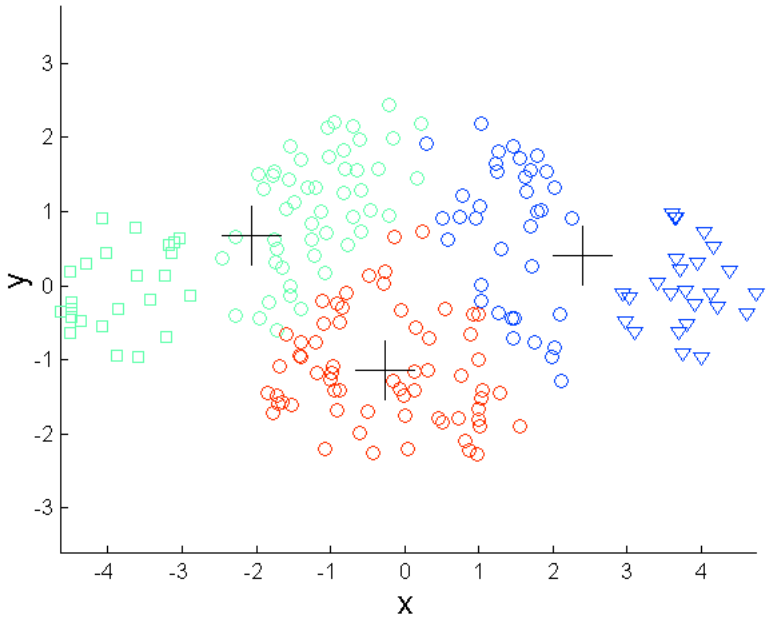
- K-means possui problemas quando os clusters tem
  - Tamanhos diferentes
  - Densidades diferentes
  - Formas não globulares

- K-means tem problemas quando os dados contêm valores discrepantes ou outliers.

Limitações de K-means: Tamanhos diferentes

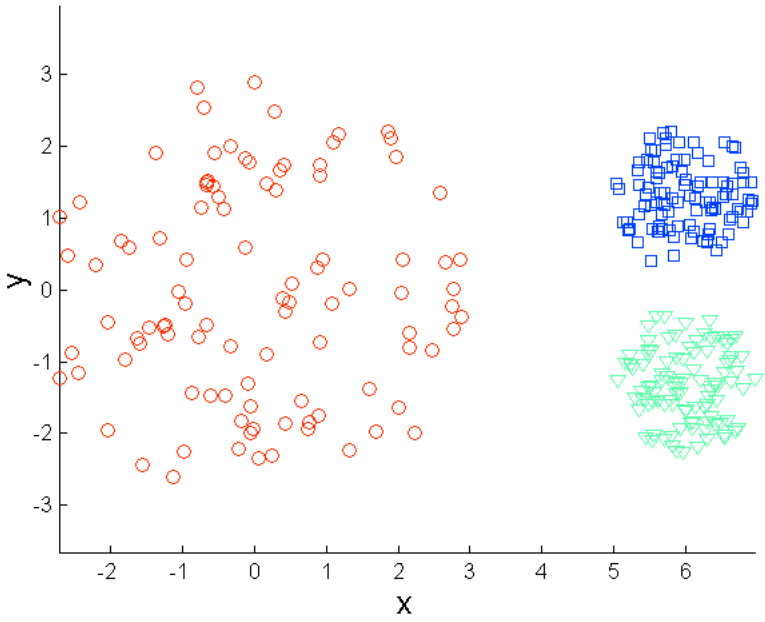


Pontos originais

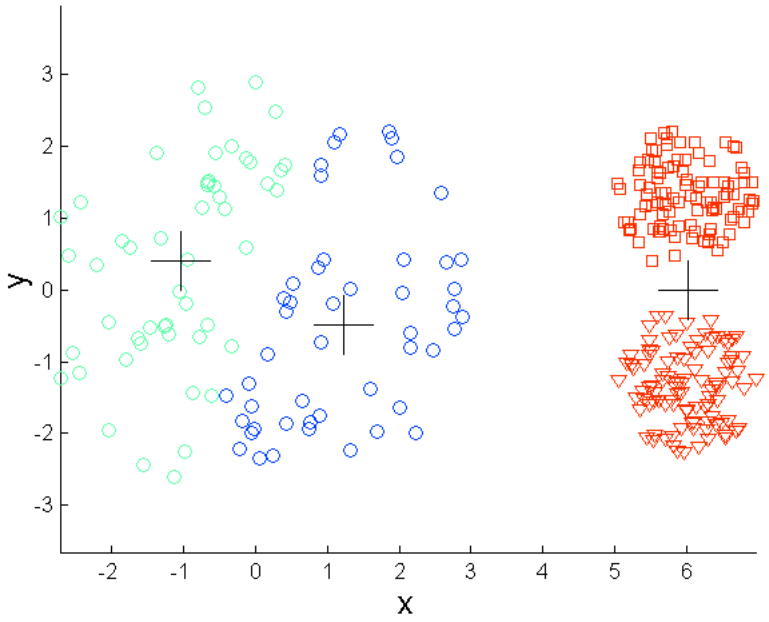


K-means (3 Clusters)

Limitações de K-means: densidades diferentes

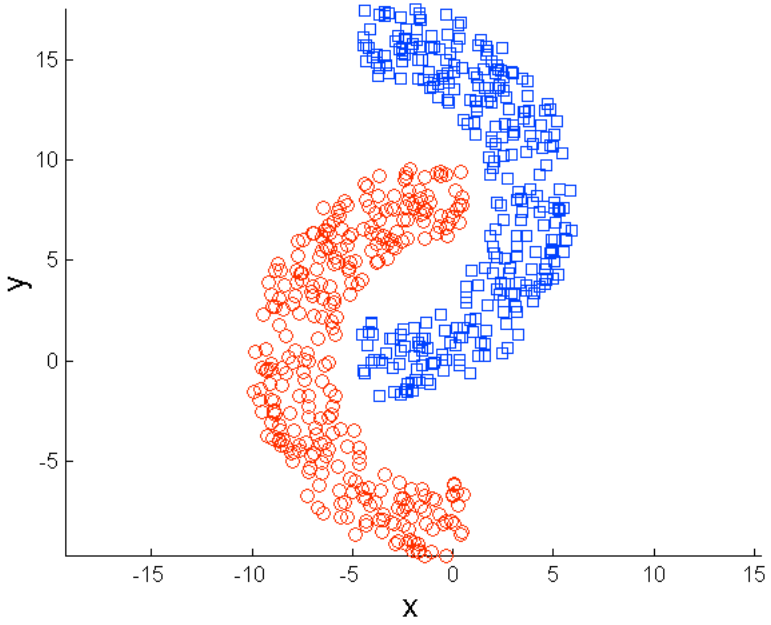


Pontos originais

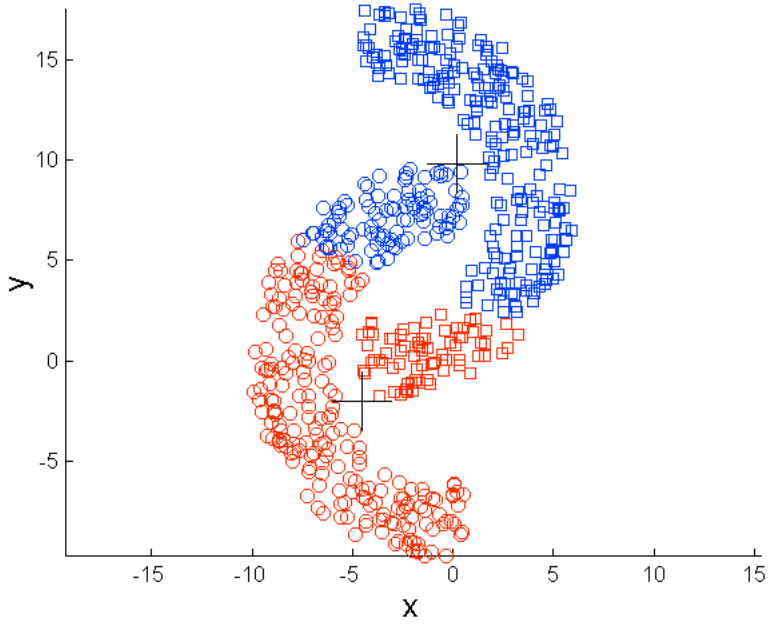


K-means (3 Clusters)

Limitações de K-means: Formas não globulares

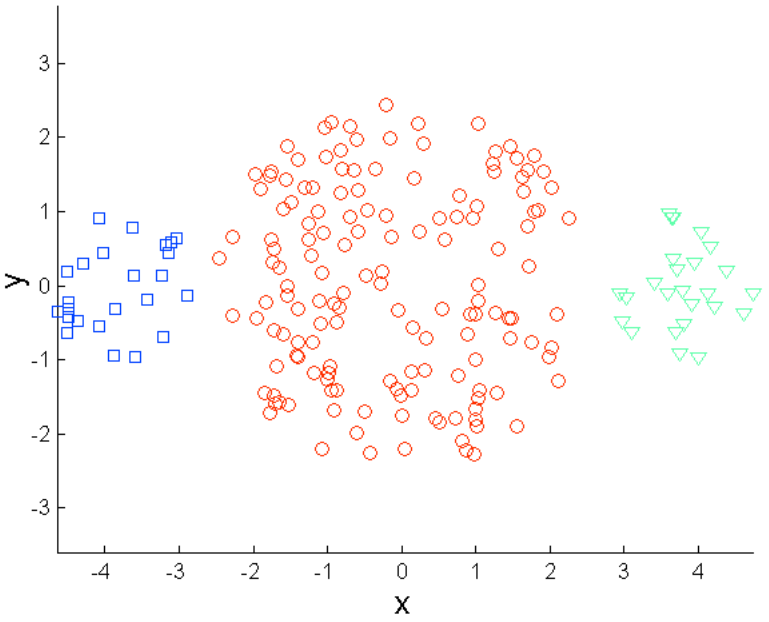


Pontos originais

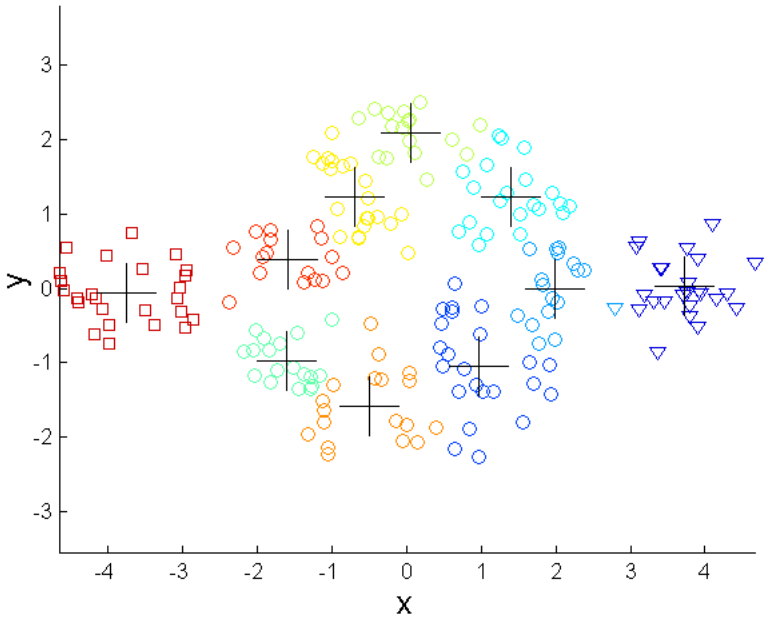


K-means (2 Clusters)

Superando limitações de K-means



Pontos originais

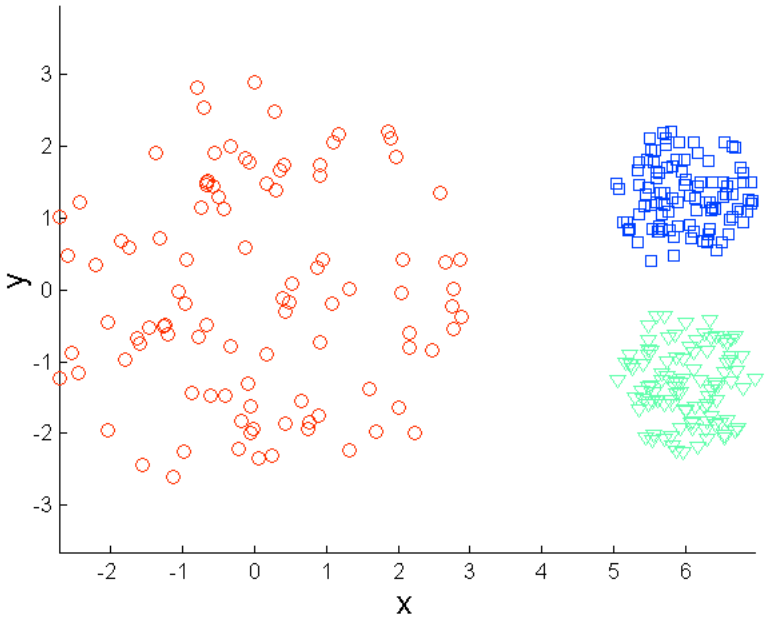


Clusters K-means

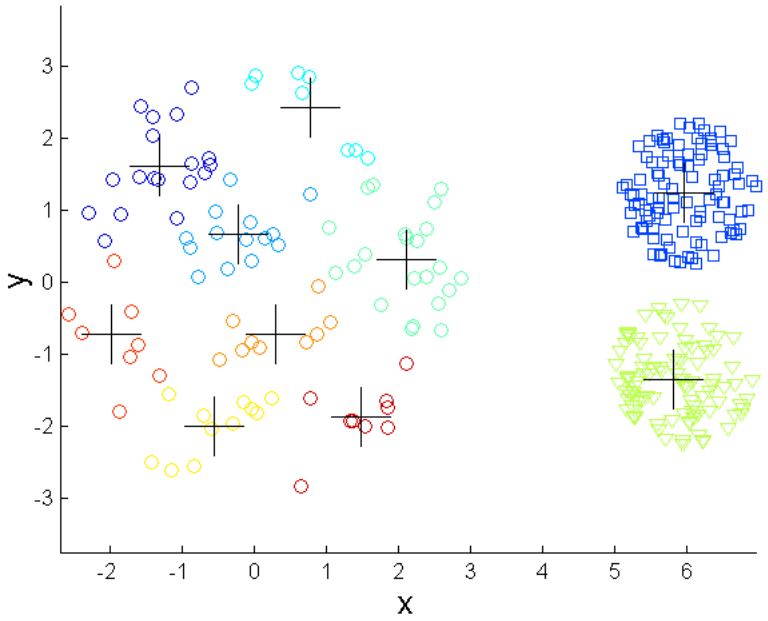
Uma solução é uso de vários clusters.

Encontre partes de clusters com necessidade de juntar depois.

Superando limitações de K-means

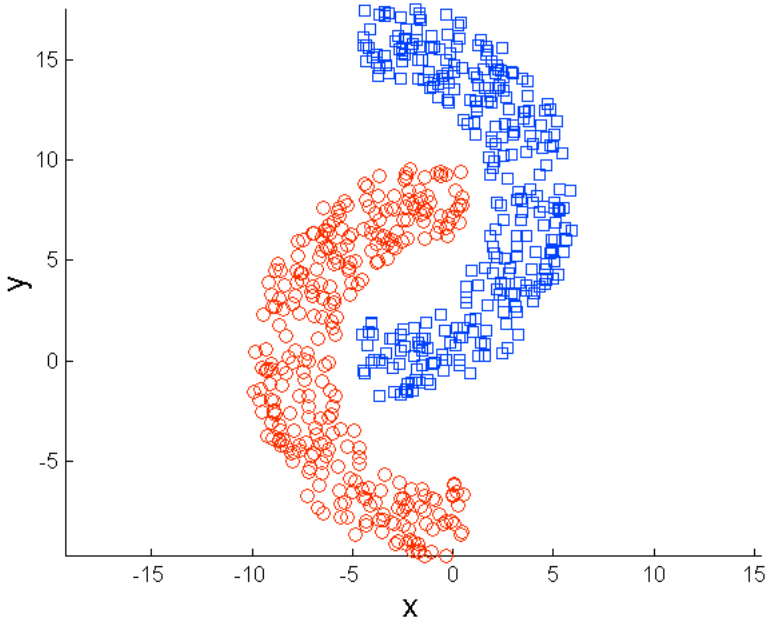


Pontos originais

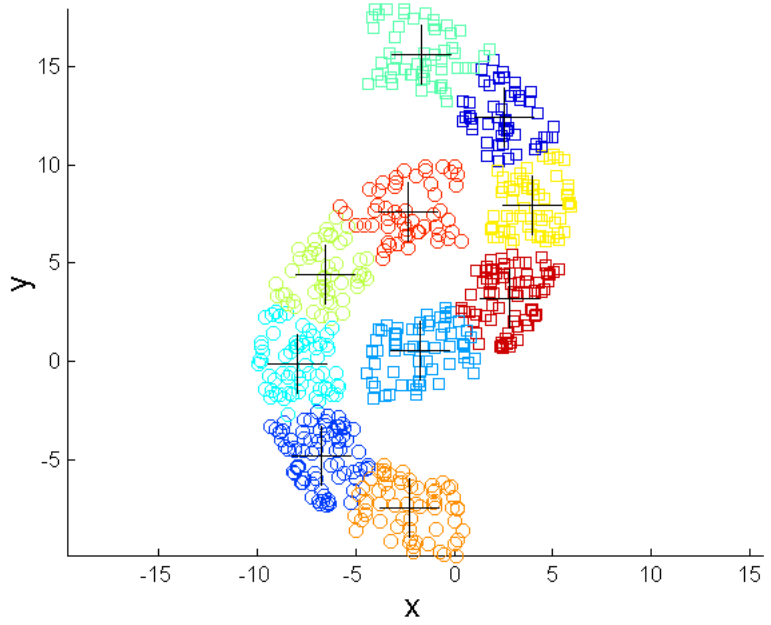


Clusters K-means

Superando limitações de K-means

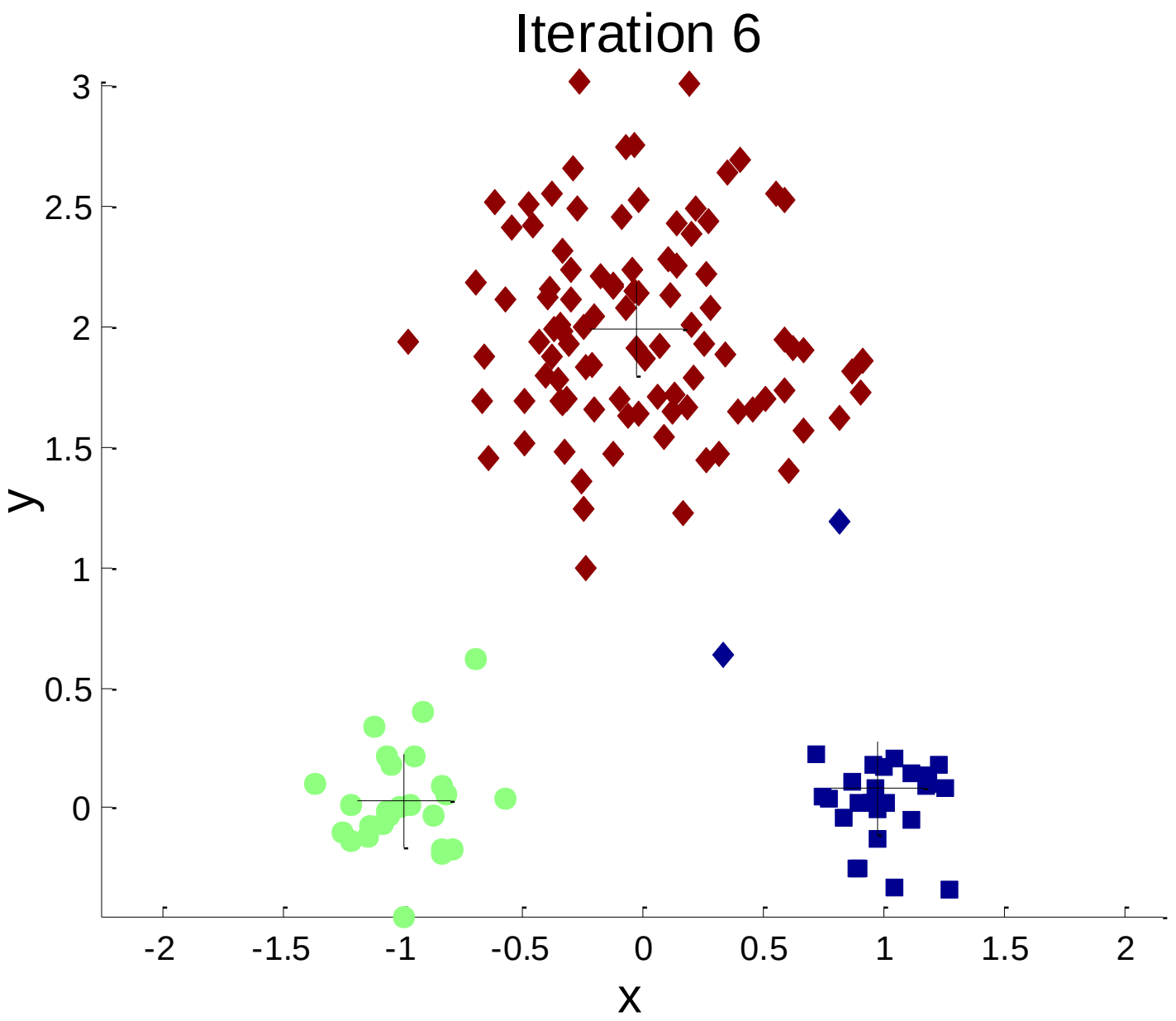


Pontos originais

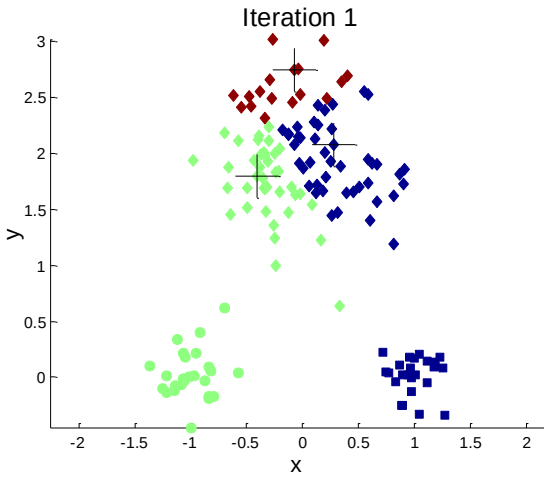


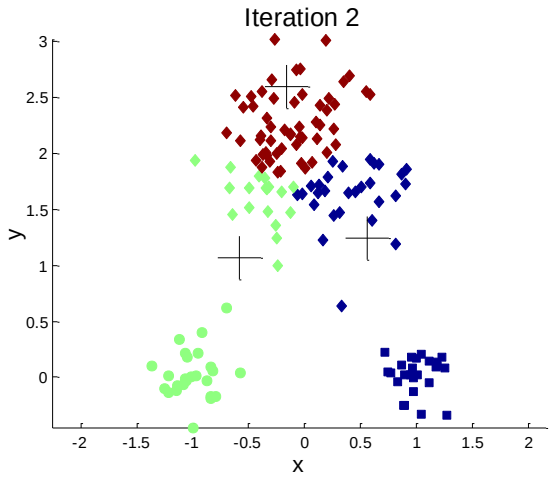
Cluster K-means

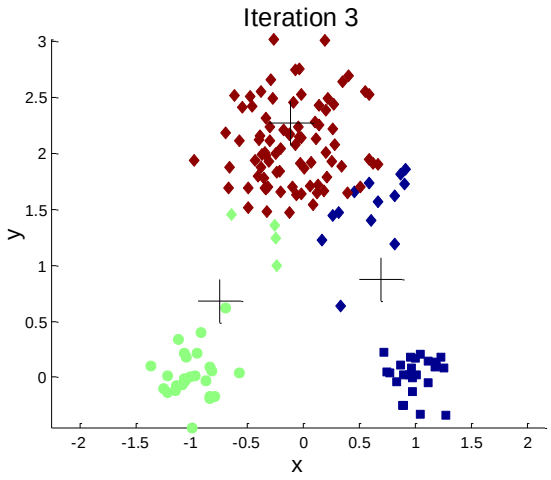
Importância de escolha de centroides iniciais

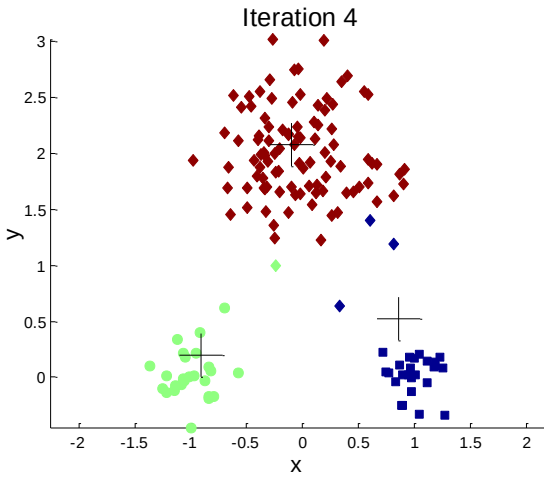


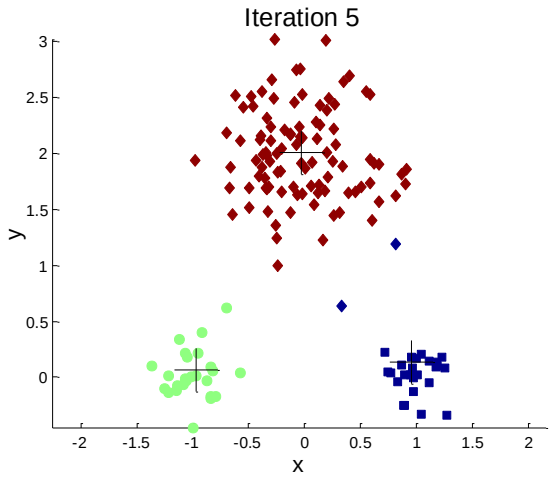
Importância de escolha de centroides iniciais

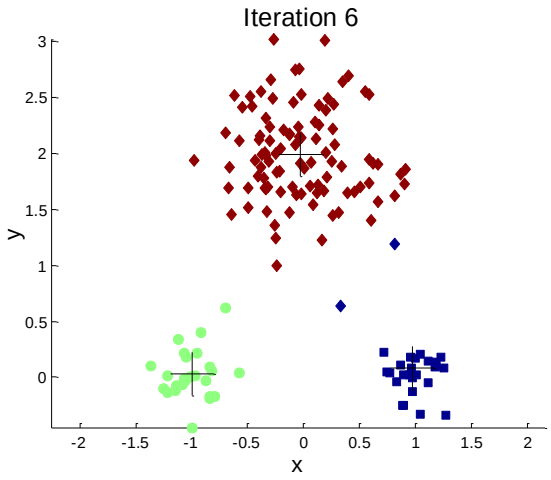




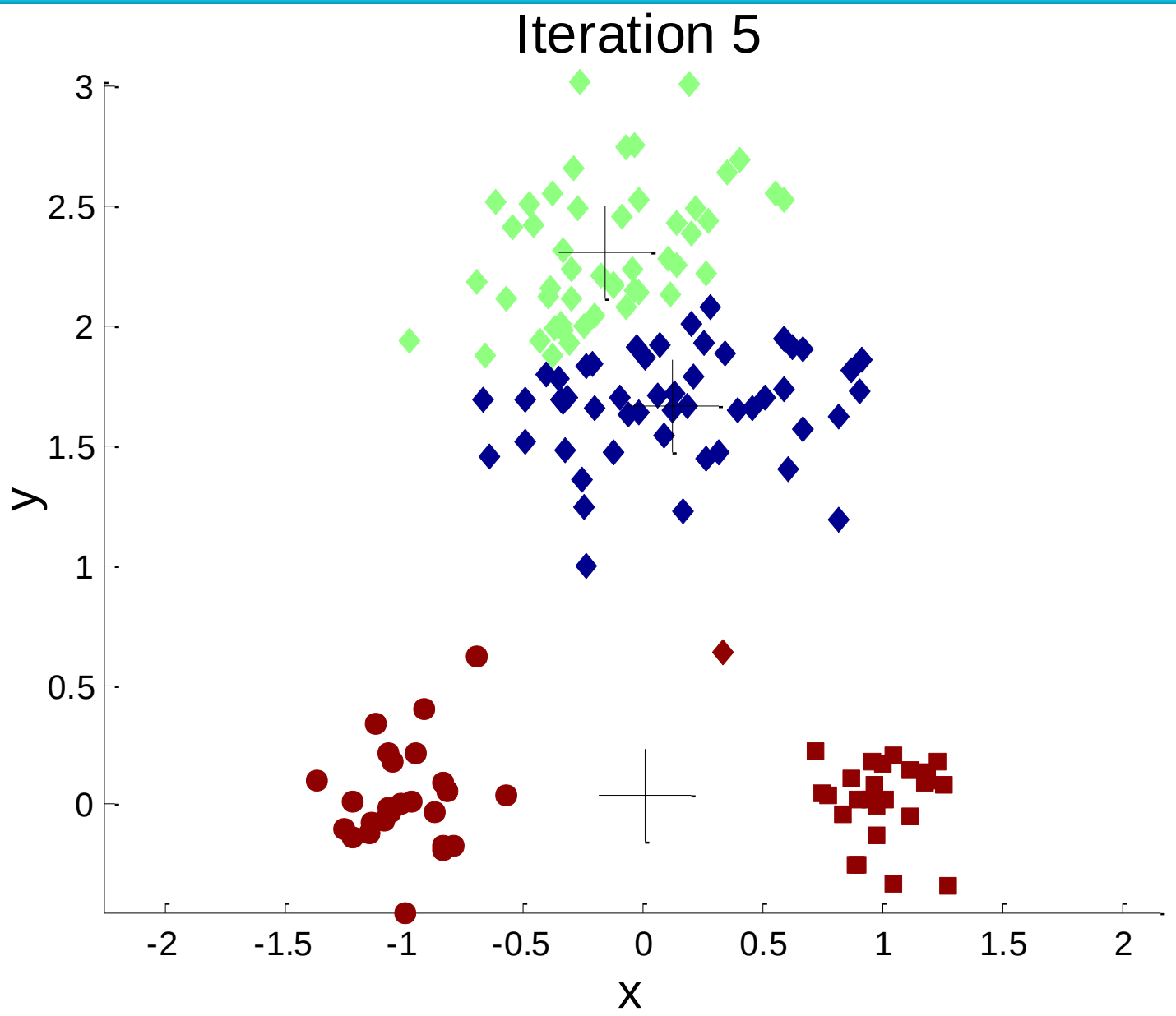




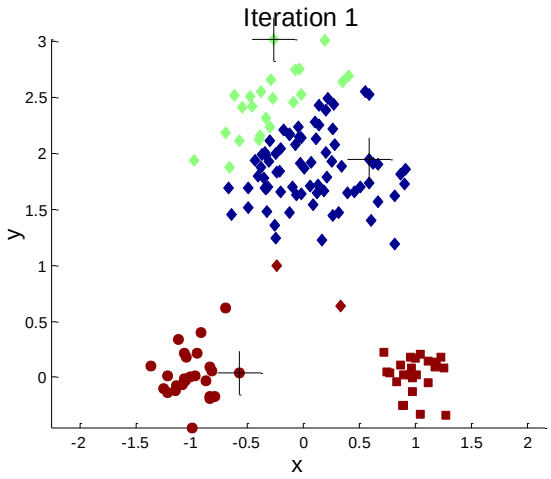


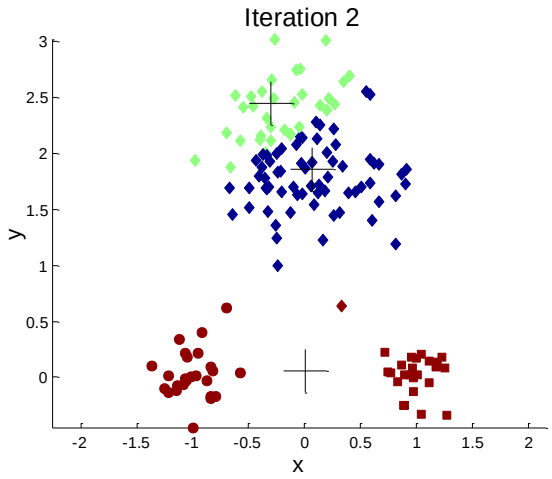


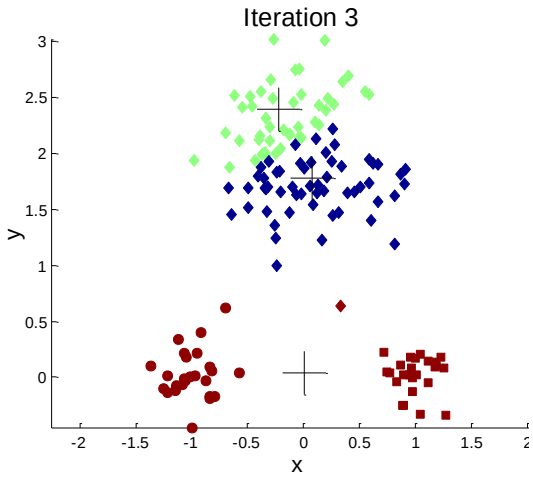
Importância de escolha de centroides iniciais …

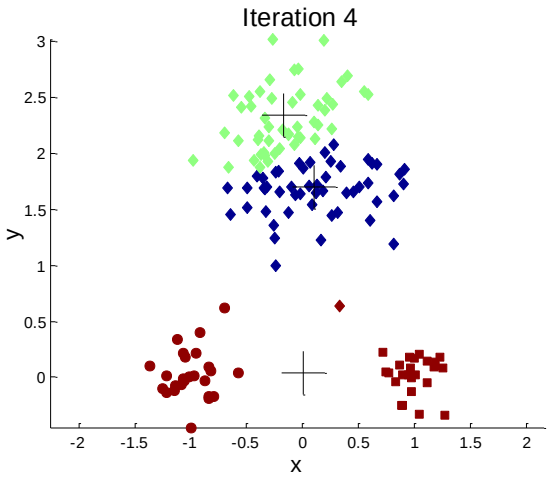


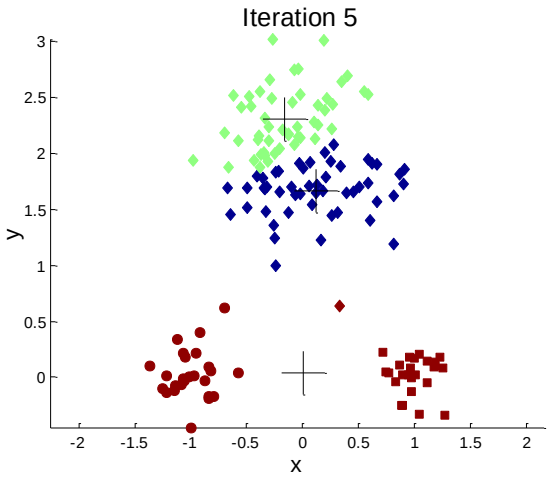
Importância de escolha de centroides iniciais ...











Problemas com escolha de pontos iniciais

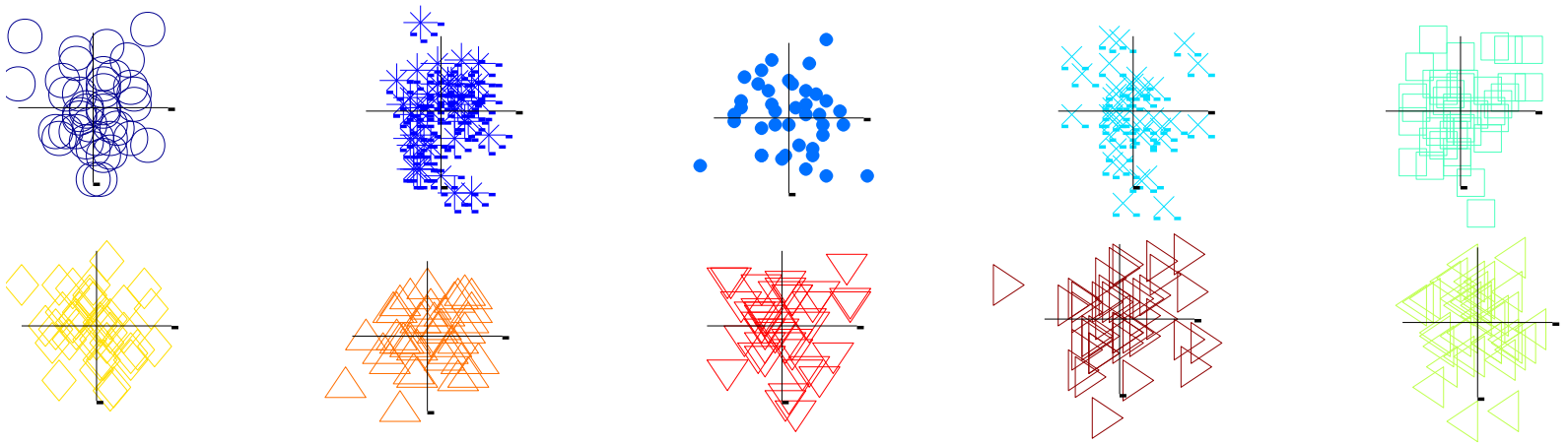
- Se houver K clusters 'reais', a chance de selecionar um centroide de cada cluster é pequena.
  - A chance é relativamente pequena quando K é grande
  - Se os clusters são dos mesmos tamanhos, n, então

$$P = \frac{o\ número\ de\ maneiras\ de\ escolher\ um\ cetroide\ de\ cada\ cluster}{o\ número\ de\ maneiras\ de\ escolher\ K\ centroides}$$

$$P = \frac{K!\, n^K}{(Kn)^K} = \frac{K!}{K^K}$$

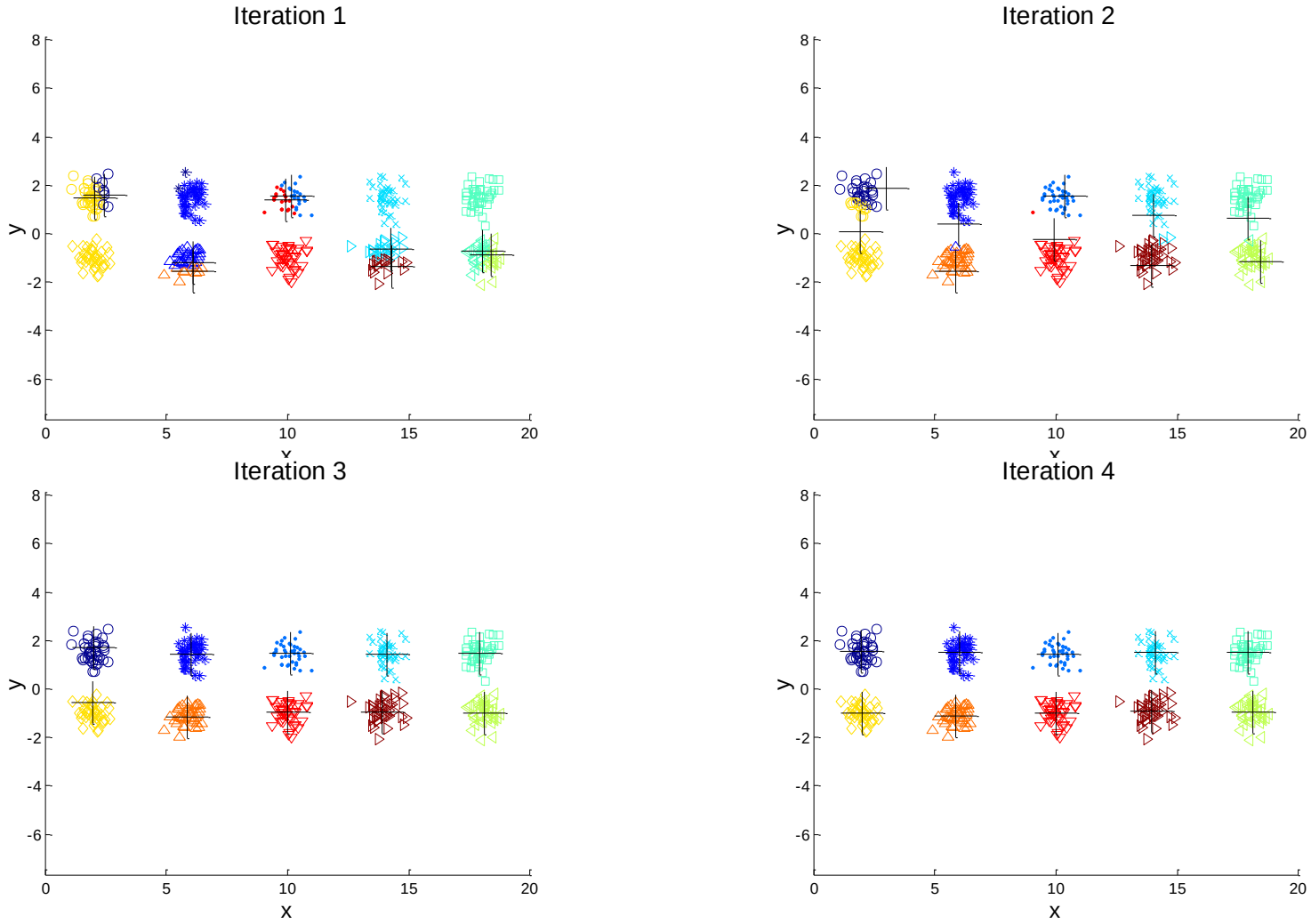
  - Por exemplo, se K = 10, a probabilidade é = $10!/10^{10}$ = 0.00036
  - Às vezes, os centroides iniciais se reajustam da maneira "correta" e, às vezes, não.
  - Considere um exemplo do cinco pares de clusters

Exemplo de 10 Clusters



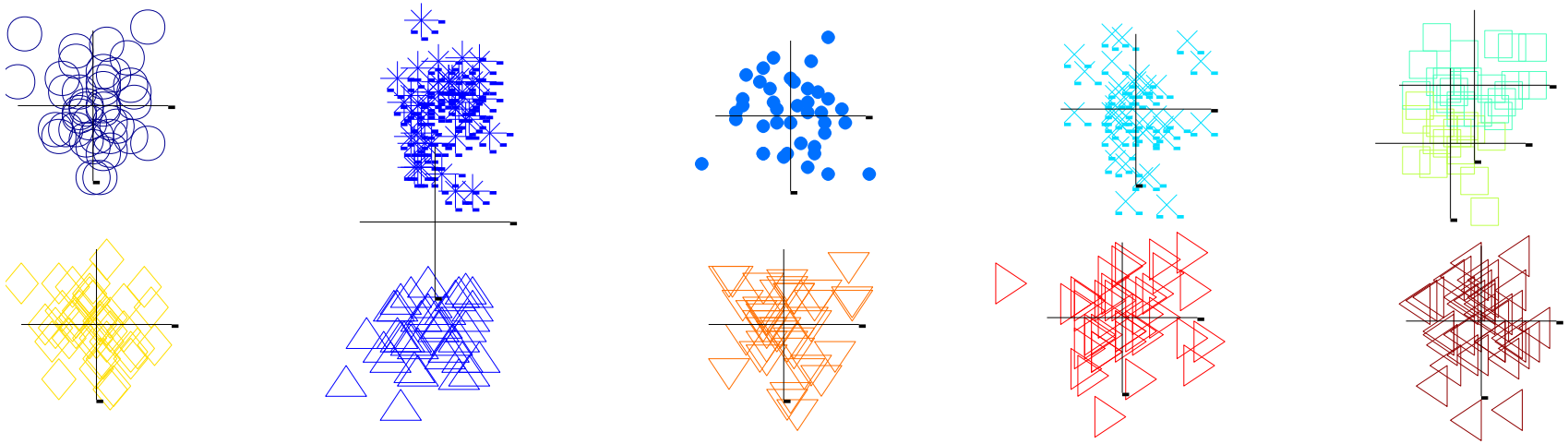
Começando com dois centroides iniciais em um cluster de cada par de clusters

Exemplo de 10 Clusters



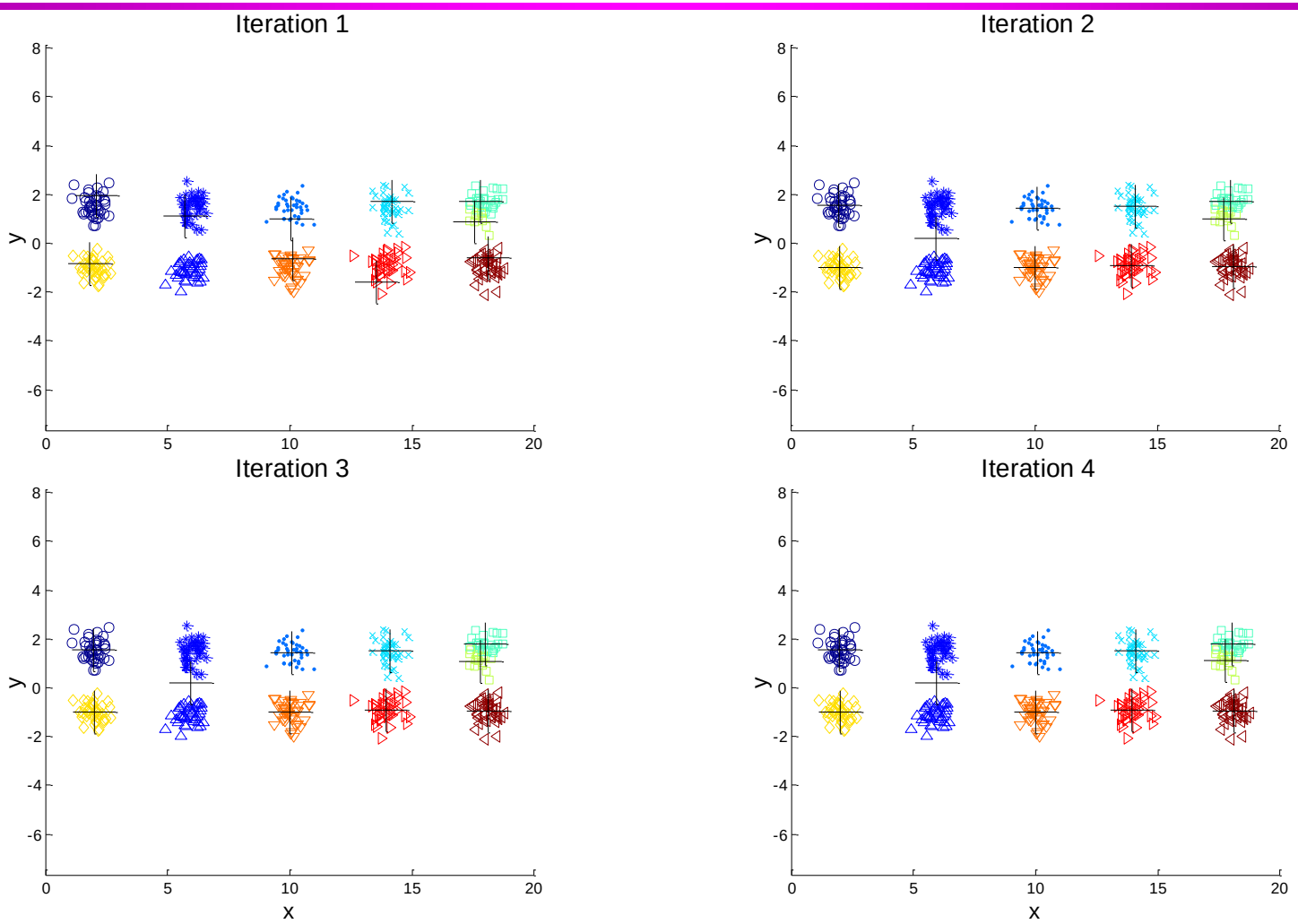
Começando com dois centroides iniciais em um cluster de cada par de clusters

10 Clusters Example



Começando com três centroides iniciais em alguns pares de clusters, enquanto outros possuem apenas um.

Exemplo de 10 Clusters



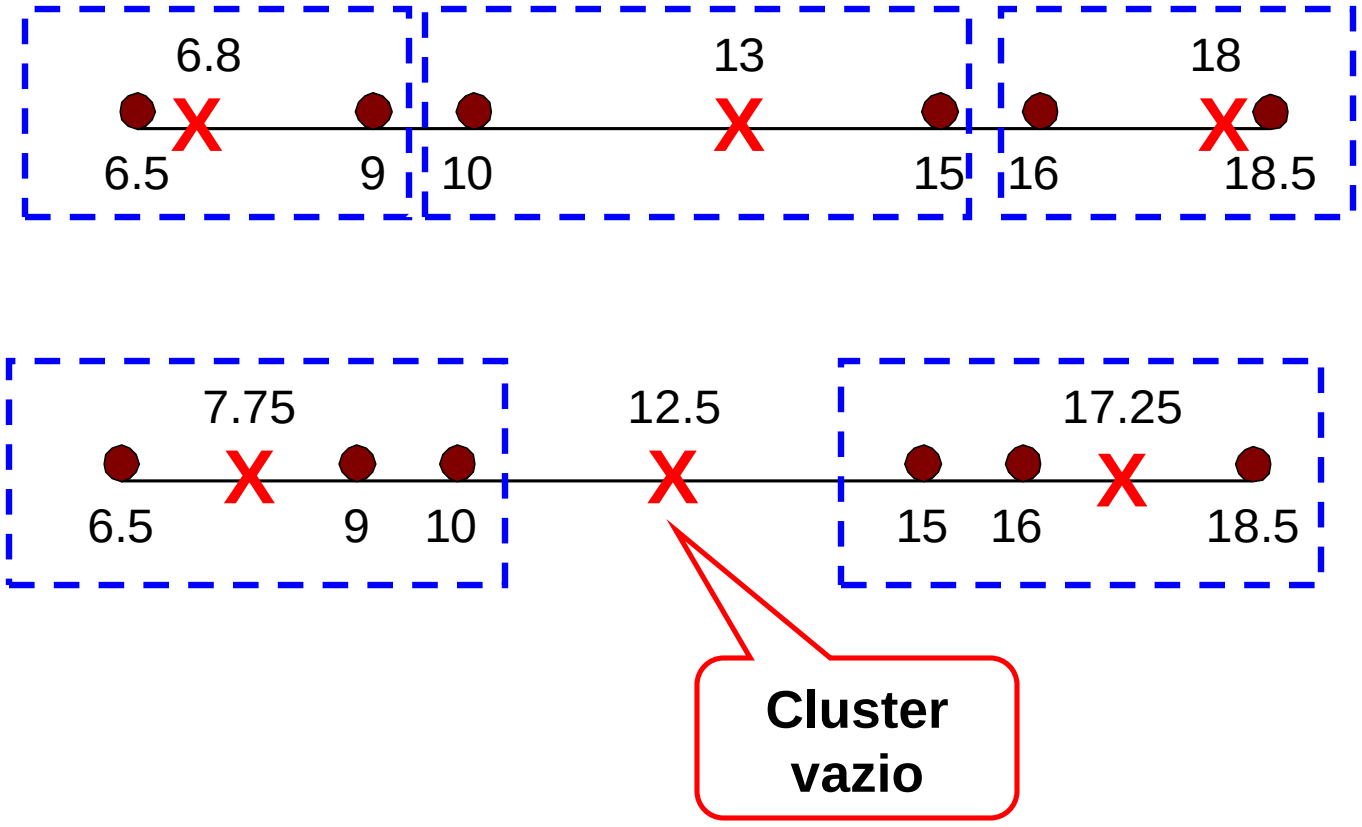
Começando com três centroides iniciais em alguns pares de clusters, enquanto outros possuem apenas um.

Solução de problema de centroides iniciais

- Várias rodadas
  - Ajuda, mas a probabilidade não está do seu lado
- Selecione um amostra e usa um cluster hierárquico para determinar os centroides iniciais
- Selecione mais do que k centroides iniciais e depois escolhe os centroides iniciais entre eles
  - Escolha de centroides mais separados
- Pós-processamento
- Gere um número maior de clusters e, em seguida, execute um cluster hierárquico
- Use bissecção de K-means
  - Não é suscetível a problemas de inicialização

Clusters vazios

- K-means pode produzir agrupamentos vazios



Removendo clusters vazios

- Algoritmos básico de K-means pode produzir clusters vazios

- Algumas estratégias
  - Escolhe o ponto que mais contribui para SSE
  - Escolhe o ponto no cluster com o maior SSE
  - Se houver vários clusters vazios, o acima pode ser repetido várias vezes.

Atualizando centros incrementalmente

- No algoritmo básico de K-means, os centroides são atualizados depois que todos os pontos são atribuídos a um centroide.

- Uma alternativa é atualizar os centroides depois de cada atribuição (abordagem incremental)
  - Cada atribuição atualiza zero ou dois centroides
  - Mais caro
  - Introduz uma dependência de ordem
  - Nunca chega a um cluster vazio
  - Pode usar “pesos” para mudar o impacto

Pré-processamento e Pós-processamento

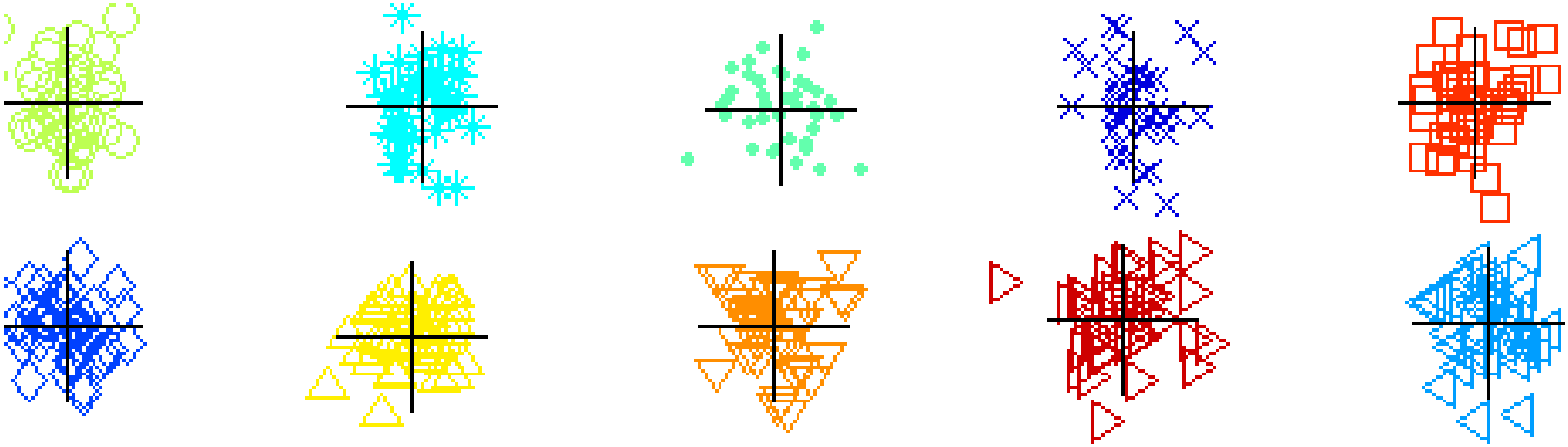
- Pré-processamento
  - Normalização de dados
  - Eliminação de outliers
- Pós-processamento
  - Eliminação de clusters pequenos, que podem representar outliers
  - Divide clusters 'soltos', ou seja, aqueles com SSE relativamente alto
  - Junte clusters ‘próximos’ com SSE relativamente baixo
  - Estes passos podem ser usados no processo de agrupamento

Bisseção de K-means

- Algoritmo de bisseção de K-means
  - Uma versão de K-means que pode produzir um agrupamento em partições hierárquico

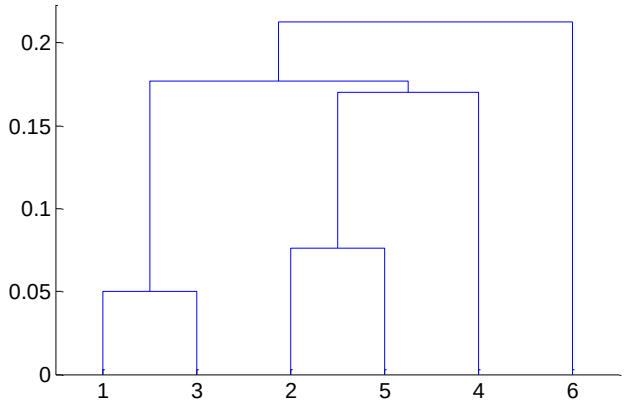
1. Inicialize a lista de clusters para conter o cluster que contém todos os pontos.
2. **Repita**
3. Escolhe um cluster da lista de clusters
4. **Para** i = 1 até *o número de iterações*
5. Aplique bisseção para cluster escolhido usando K-médias.
6. Adicione os dois clusters da bissecção com o menor SSE na lista
7. **Até** que a lista de clusters contem K clusters

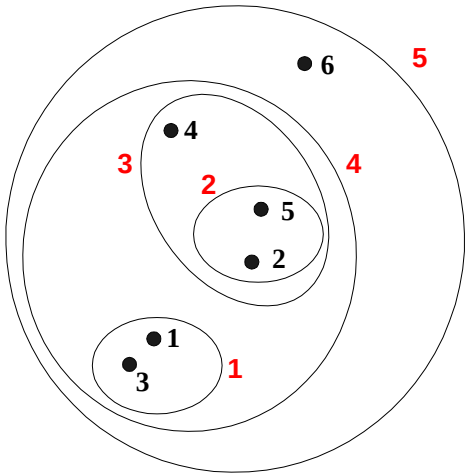
Exemplo de Bisseção de K-means



Agrupamento hierárquico

- Produz um conjunto de clusters aninhados organizados como uma árvore hierárquica
- Pode ser visualizado como um dendrograma
  - Um diagrama como árvore registra a sequência de divisões





Vantagens de agrupamento hierárquico

- Não é necessário assumir qualquer número específico de clusters
  - Qualquer número desejado de clusters pode ser obtido por "cortar" o dendrograma no nível adequado

- Eles podem corresponder a taxonomias significativas
  - Exemplos de aplicação em ciências biológicas: reino animal, reconstrução filogenia,...

Agrupamento hierárquico

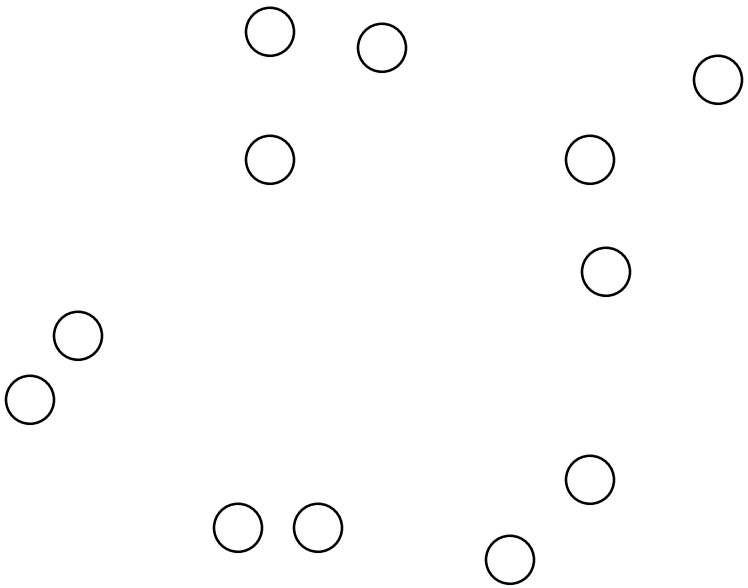
- Duas abordagens principais de agrupamento hierárquico
  - Aglomerativa:
    - Comece com os pontos como clusters individuais
    - Em cada etapa, combina os clusters mais próximos até que apenas um cluster (ou k clusters) sobra.
  - Divisiva:
    - Comece com um cluster com tudo incluído
    - Em cada etapa, divida um cluster até que cada cluster contenha um ponto individual (ou haja clusters k).
- Algoritmos hierárquicos tradicionais usam uma matriz de similaridade ou distância
  - Combine ou divide um cluster de cada vez

Algoritmo de agrupamento aglomerativo

- É a técnica de agrupamento hierárquico mais popular
- O algoritmo básico é simples
  1. Calcule a matriz de proximidade
  2. Deixe que cada ponto de dados seja um cluster
  3. **Repita**
  4. Combine dois cluster mais próximos
  5. Atualize a matriz de proximidade
  6. **Até** que apenas um único cluster permaneça
- Operação-chave é o cálculo da proximidade de dois clusters
  - Diferentes abordagens para definir a distância entre clusters distinguem os diferentes algoritmos

Situação inicial

- Comece pontos como clusters individuais e uma matriz de proximidade

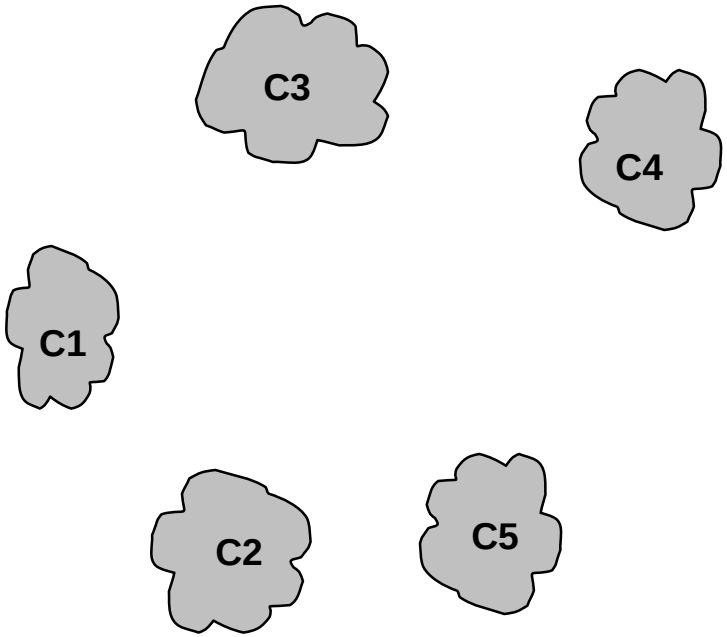


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matriz de proximidade

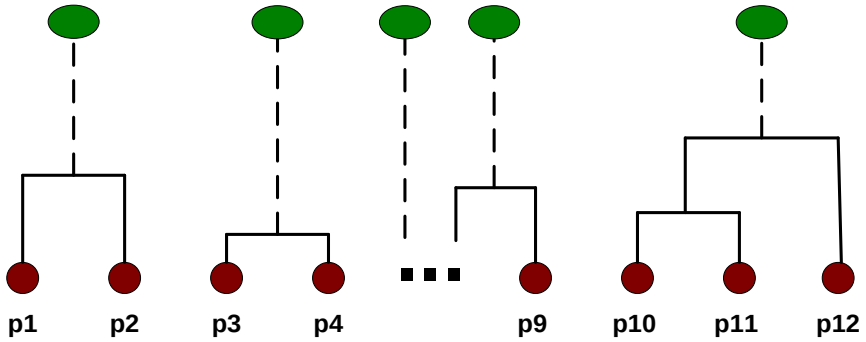
Situação intermediária

- Depois de alguns passos, temos alguns clusters aglomerados



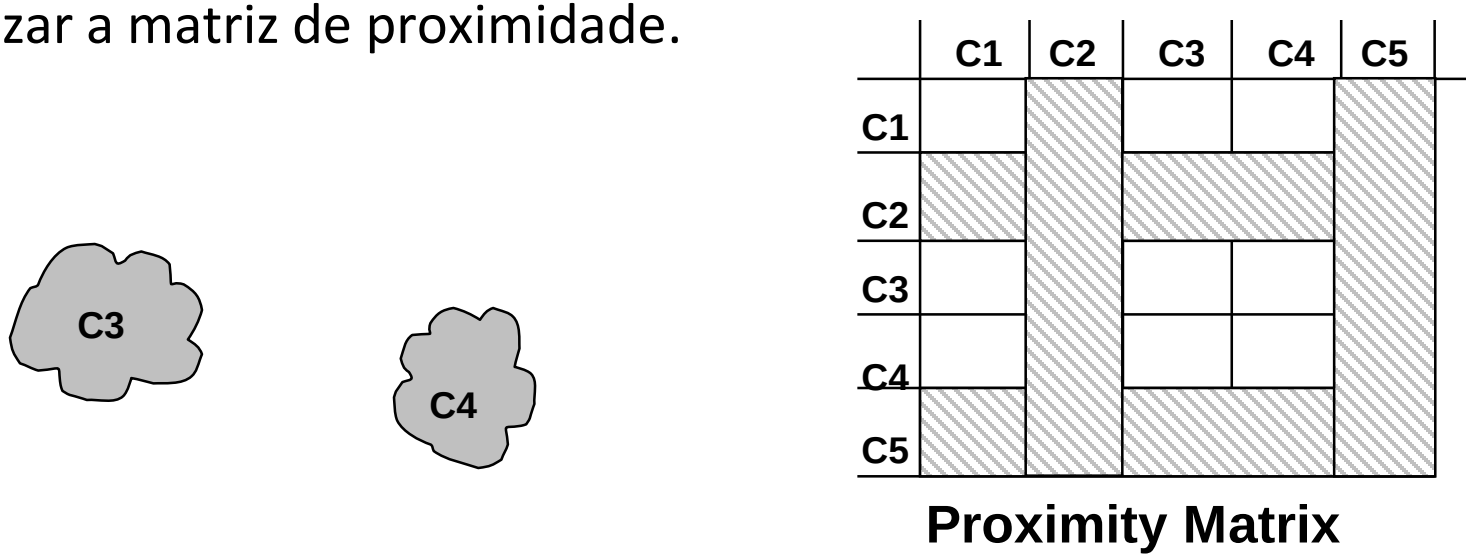
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Matriz de proximidade



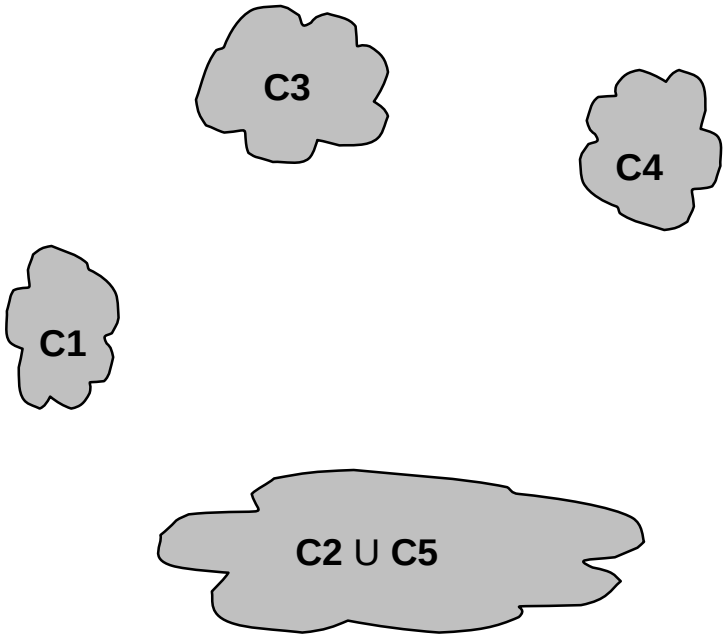
Situação intermediária

- Queremos combinar dois clusters mais próximos (C2 e C5) e atualizar a matriz de proximidade.



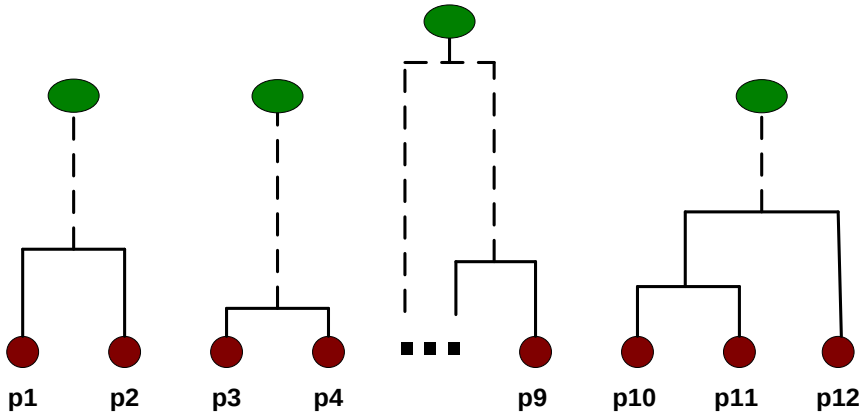
Depois de combinar

- Como atualizar a matriz de proximidade?

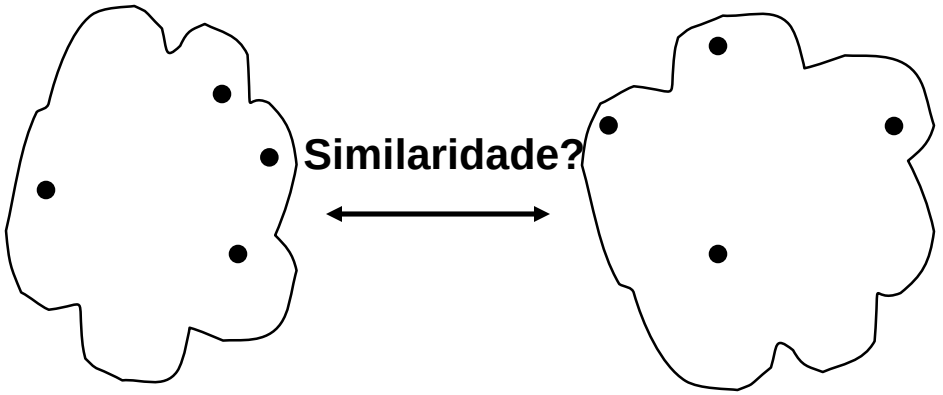


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Matriz de proximidade



Como definir a distância entre clusters?



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

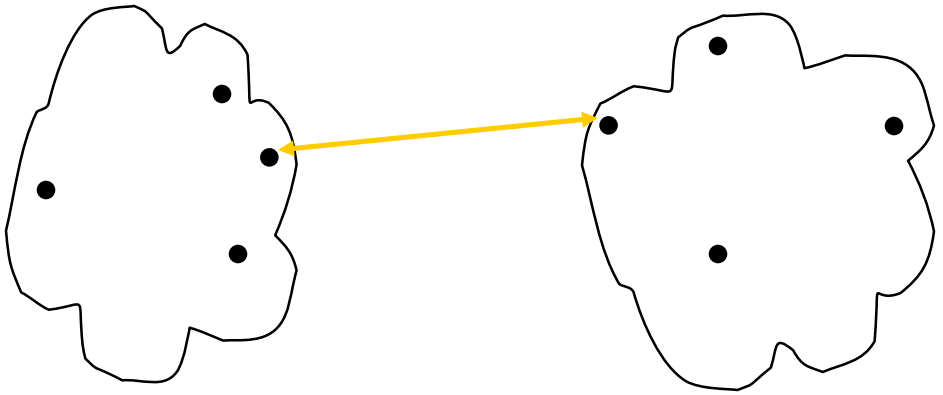
.

.

Matriz de proximidade

- MIN
- MAX
- Média do Grupo
- Distância entre centroides
- Outros métodos por função objetiva
  - O método de Ward usa erro quadrático

Como definir a distância entre clusters?

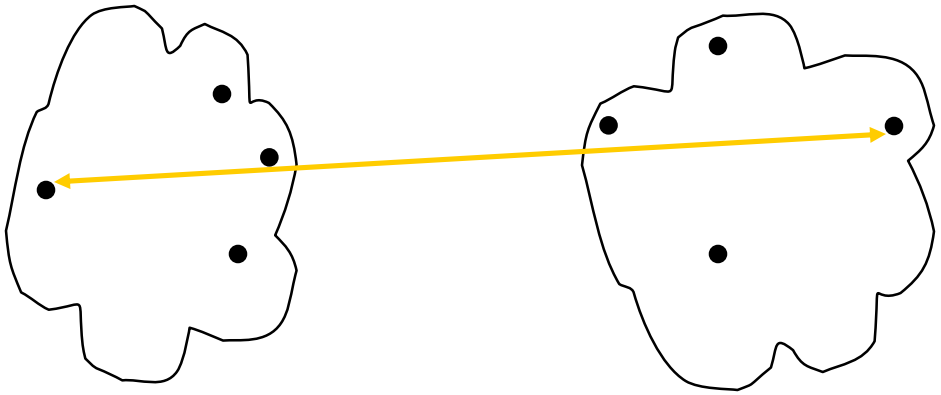


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matriz de proximidade

- MIN
- MAX
- Média do Grupo
- Distância entre centroides
- Outros métodos por função objetiva
  - O método de Ward usa erro quadrático

Como definir a distância entre clusters?

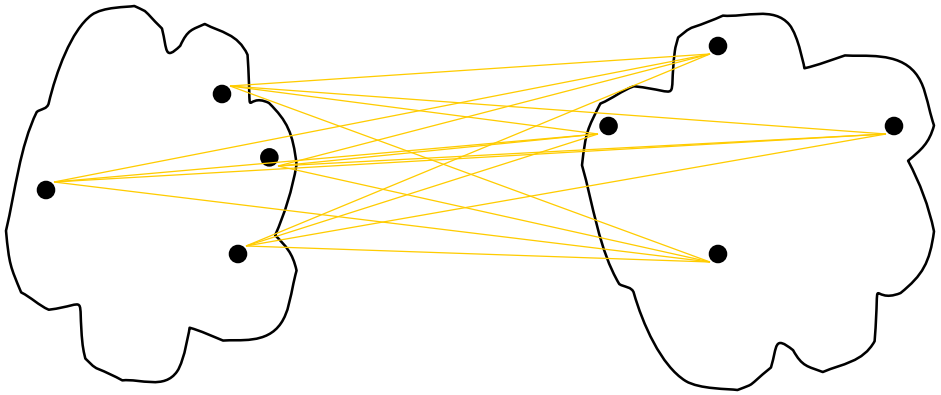


- MIN
- MAX
- Média do Grupo
- Distância entre centroides
- Outros métodos por função objetiva
  - O método de Ward usa erro quadrático

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matriz de proximidade

Como definir a distância entre clusters?

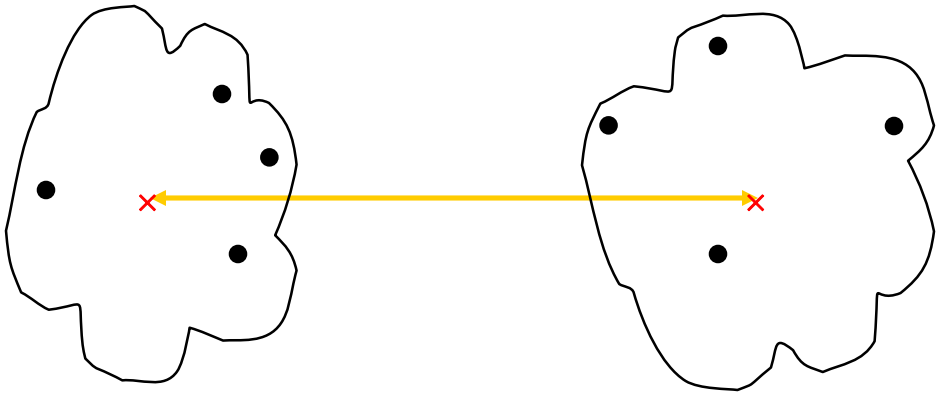


- MIN
- MAX
- Média do Grupo
- Distância entre centroides
- Outros métodos por função objetiva
  - O método de Ward usa erro quadrático

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matriz de proximidade

Como definir a distância entre clusters?



- MIN
- MAX
- Média do Grupo
- Distância entre centroides
- Outros métodos por função objetiva
  - O método de Ward usa erro quadrático

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

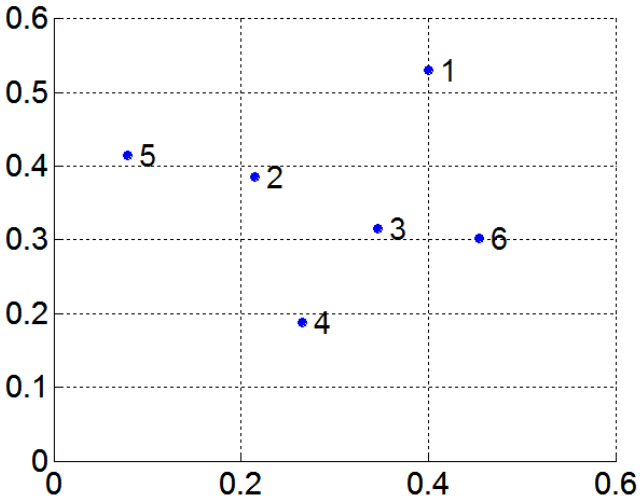
.

.

Matriz de proximidade

MIN ou Single Link

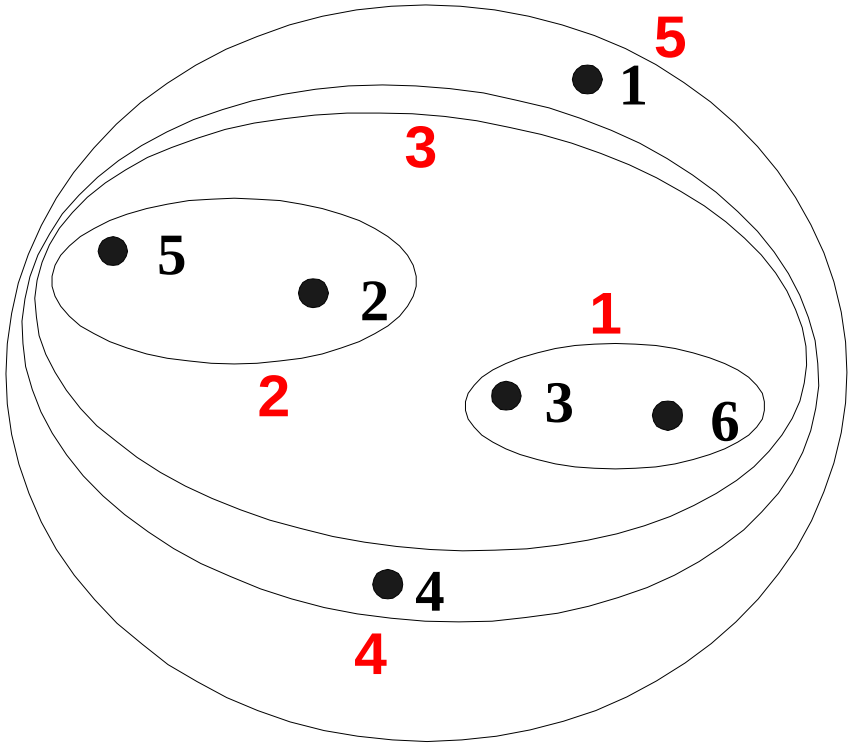
- A proximidade de dois clusters baseia-se nos dois pontos mais próximos dos diferentes clusters
  - Determinado por um par de pontos, ou seja, por um link no gráfico de proximidade
- Exemplo:

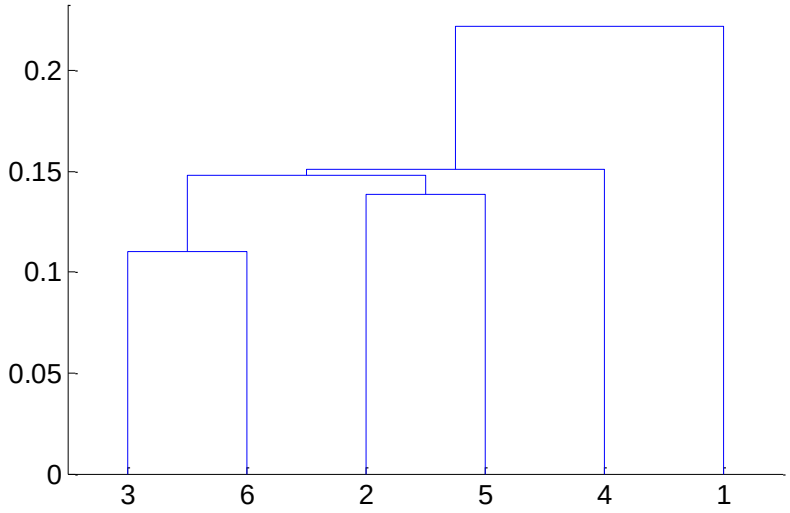


Matriz de distância:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Agrupamento hierárquico: MIN

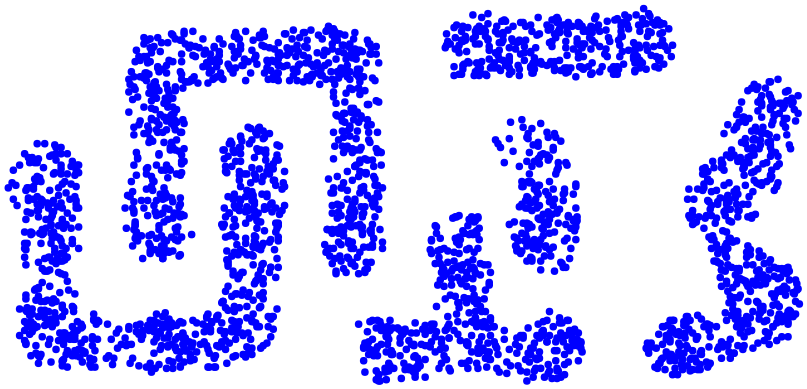




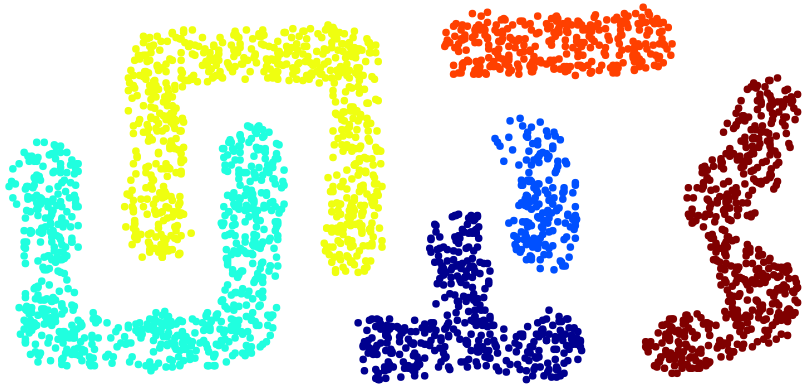
Clusters aninhados

Dendrograma

Pontos fortes de MIN



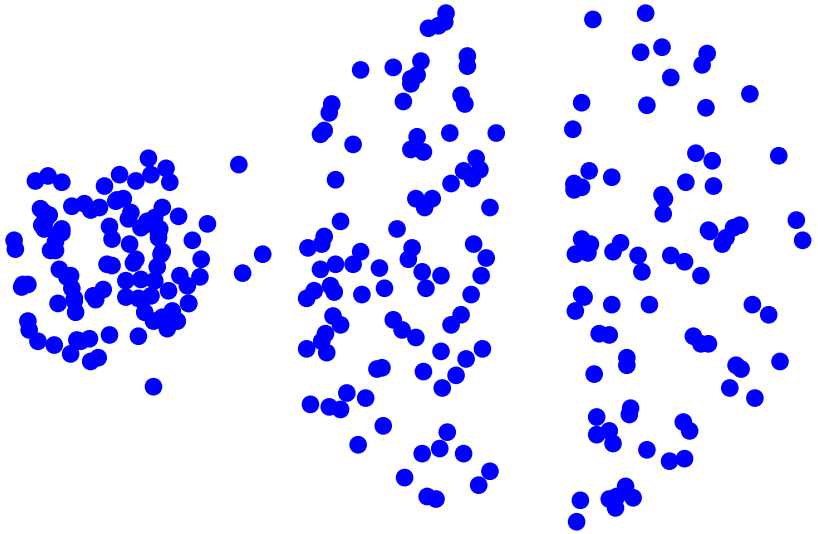
Pontos originais



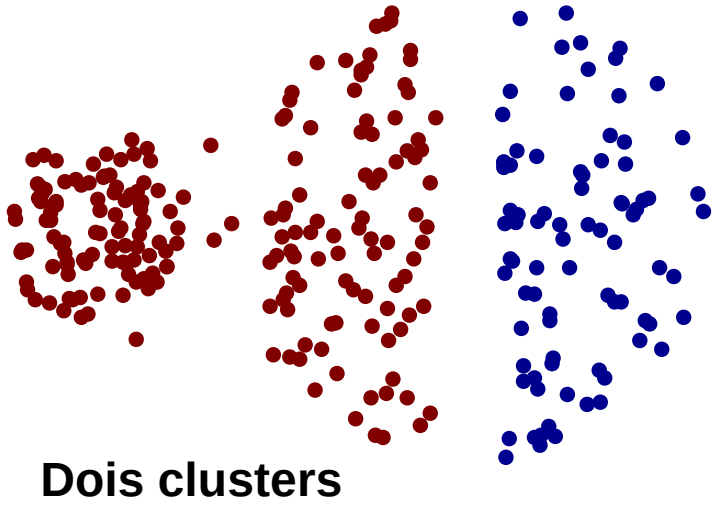
Seis Clusters

- Pode analisar formas não elípticas

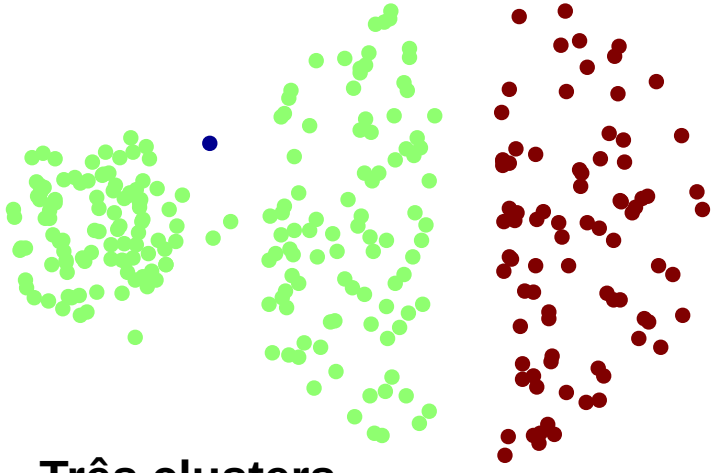
Limitações de MIN



Pontos originais



Dois clusters

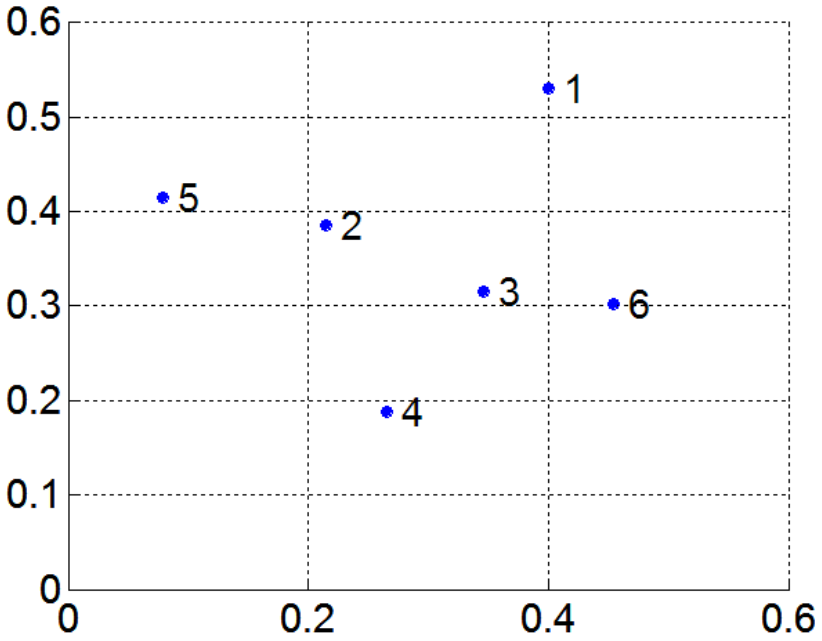


Três clusters

- Sensível ao ruído e aos outliers

MAX ou Complete Linkage

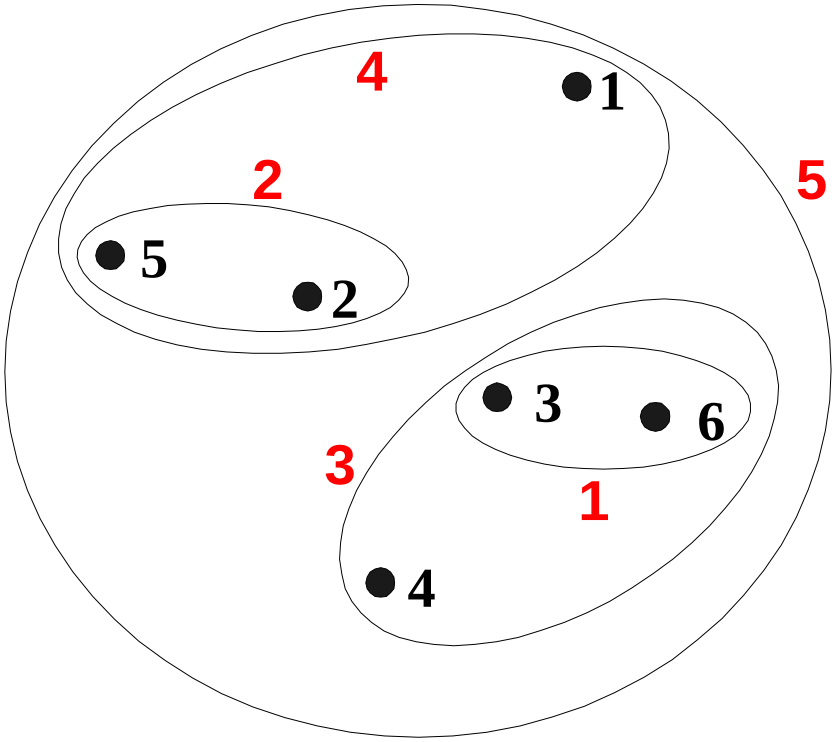
- A proximidade de dois clusters baseia-se nos dois pontos mais distantes dos diferentes clusters
  - Determinado por todos os pares de pontos nos dois clusters



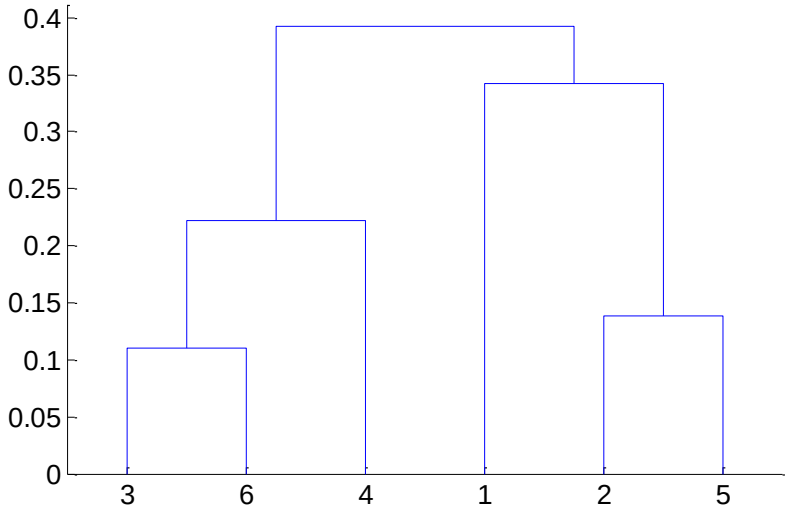
Matriz de distância:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Agrupamento hierárquico: MAX

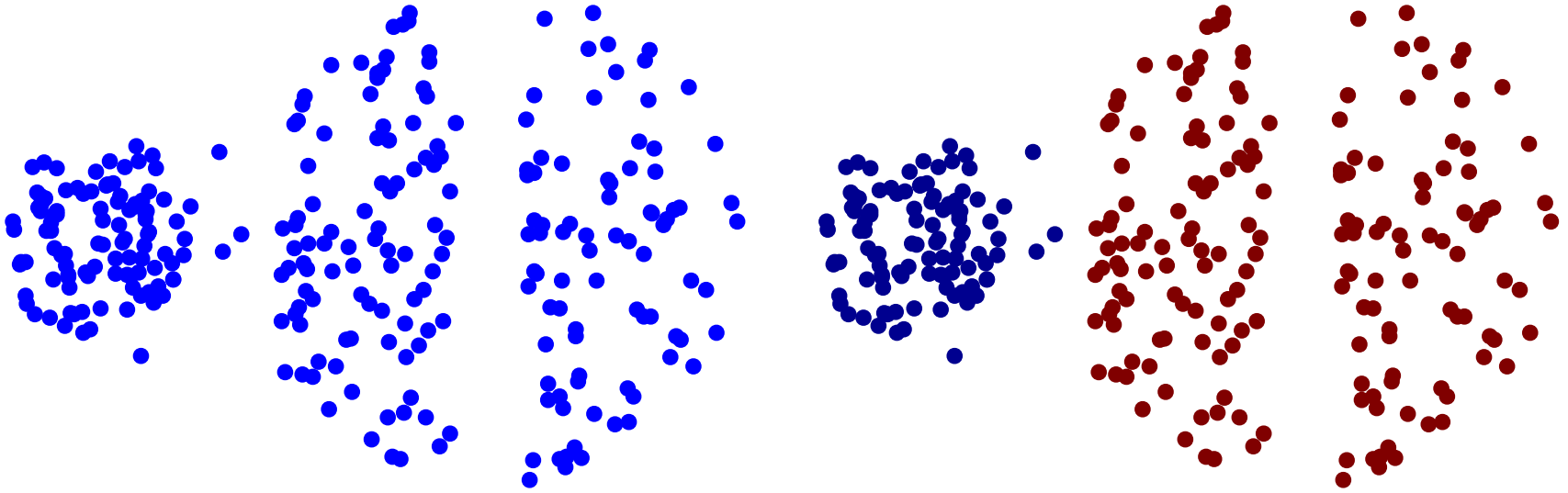


Clusters aninhados



Dendrograma

Pontos fortes de MAX

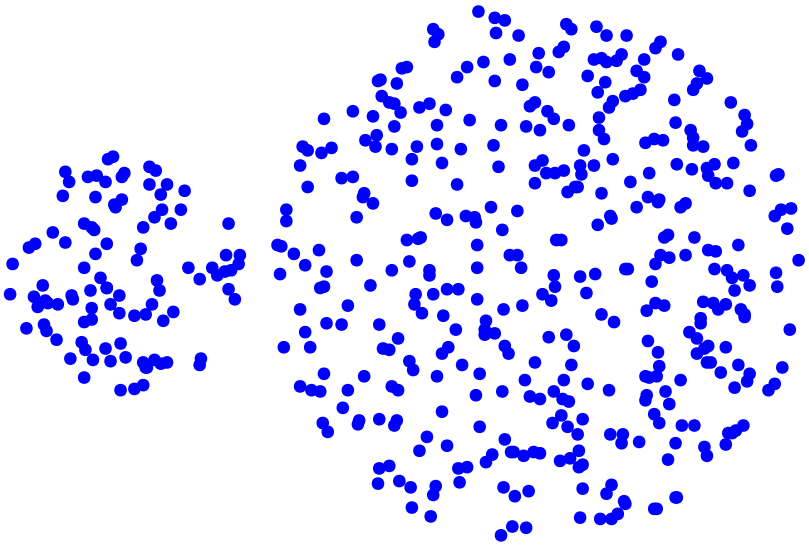


Pontos originais

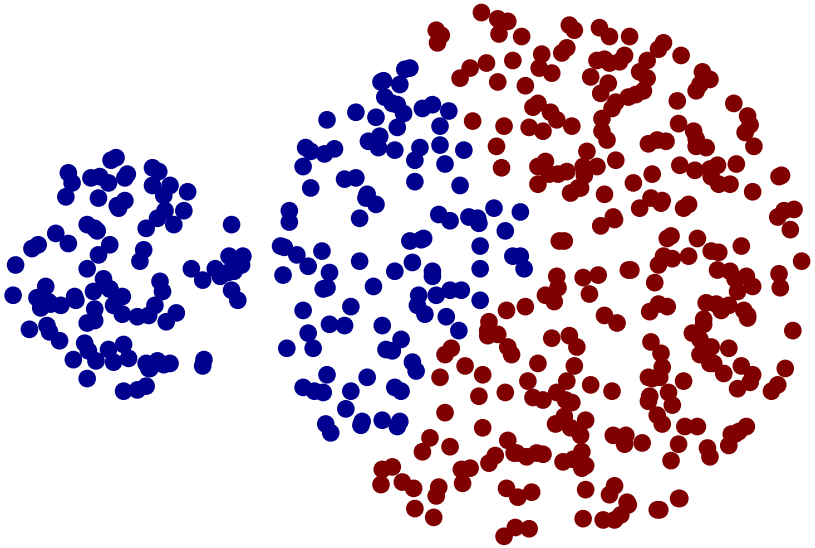
Dois Clusters

- Menos suscetível ao ruído e aos outliers

Limitações de MAX



Pontos originais



Dois Clusters

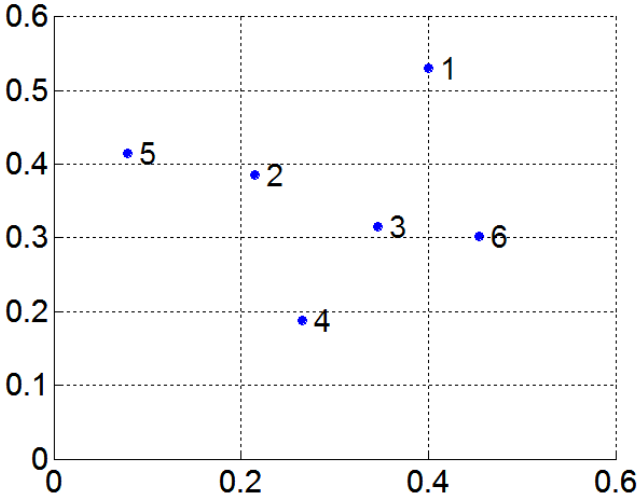
- Tende a quebrar aglomerados grandes
- Viés para clusters globulares

Média do grupo

- A proximidade de dois clusters é a média de proximidade emparelhados entre os pontos nos dois clusters.

$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| \times |\text{Cluster}_j|}$$

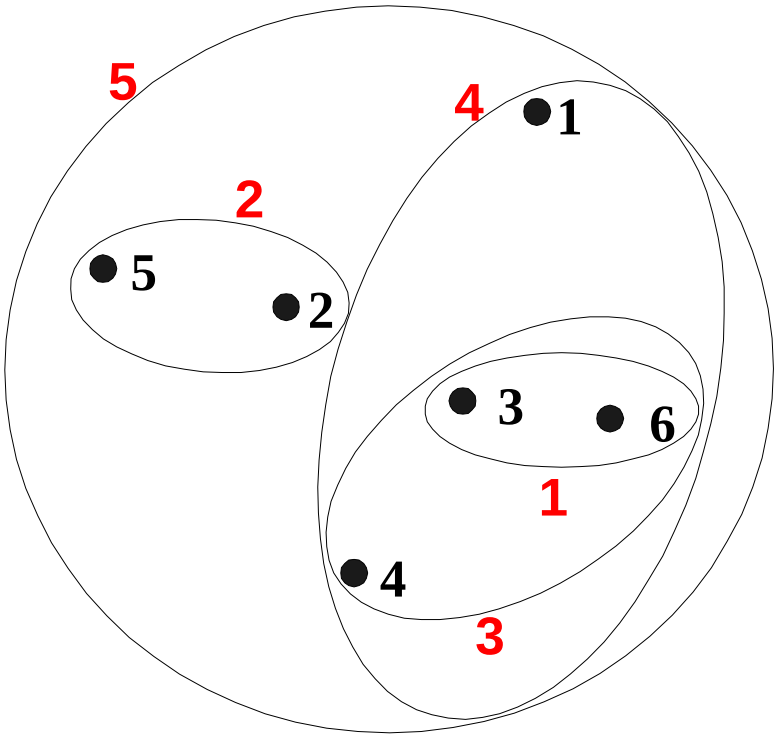
- Necessidade de usar a conectividade média para escalabilidade, uma vez que a proximidade total favorece grandes clusters.



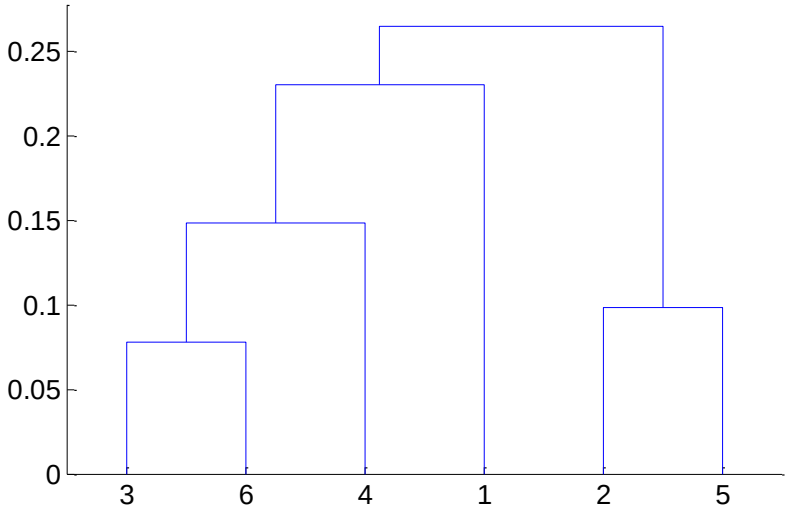
Matriz de distância:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Agrupamento hierárquico: média do grupo



Clusters aninhados



Dendrograma

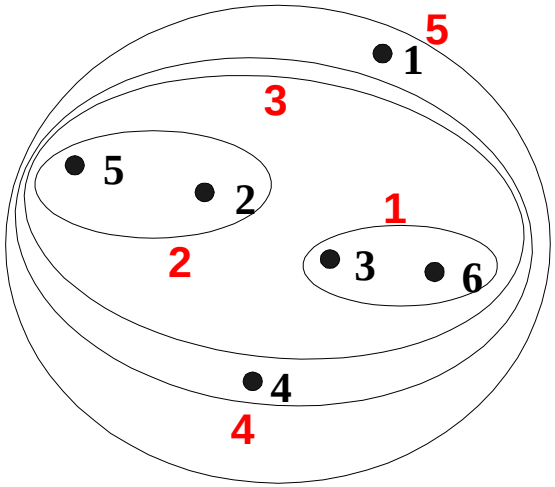
Agrupamento hierárquico: média do grupo

- Balanço entre Single Link e Complete Link
- Vantagens
  - Menos suscetível ao ruído e aos outliers
- Limitações
  - Viés para clusters globulares

Similaridade de cluster: método de Ward

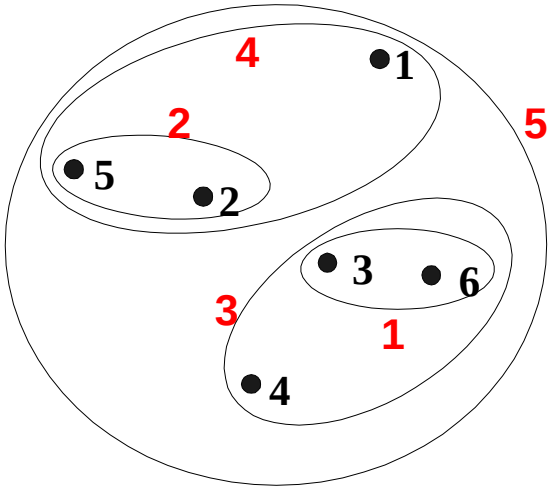
- A similaridade de dois clusters baseia-se no aumento do erro quadrado quando dois clusters são combinados
  - Semelhante à média do grupo se a distância entre pontos é a distância ao quadrado
- Menos suscetível a ruído e outliers
- Viés para clusters globulares
- Análogo hierárquico de K-means
  - Pode ser usado para inicializar K-means

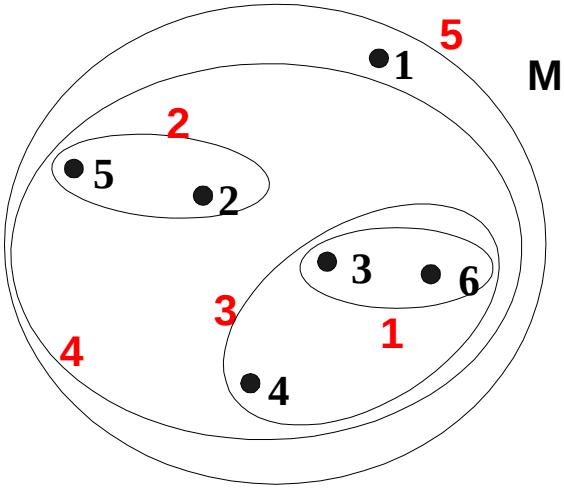
Agrupamento hierárquico: comparação



MIN

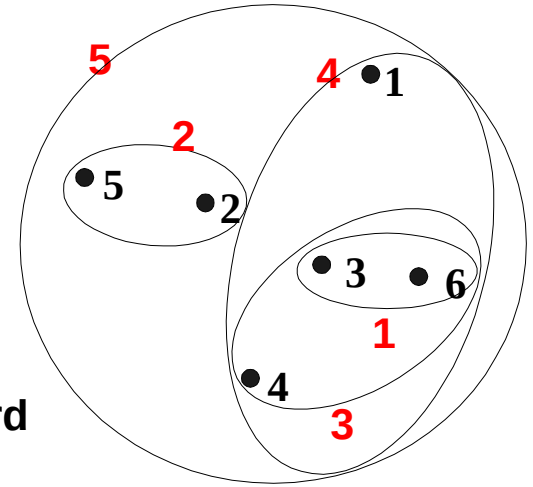
MAX





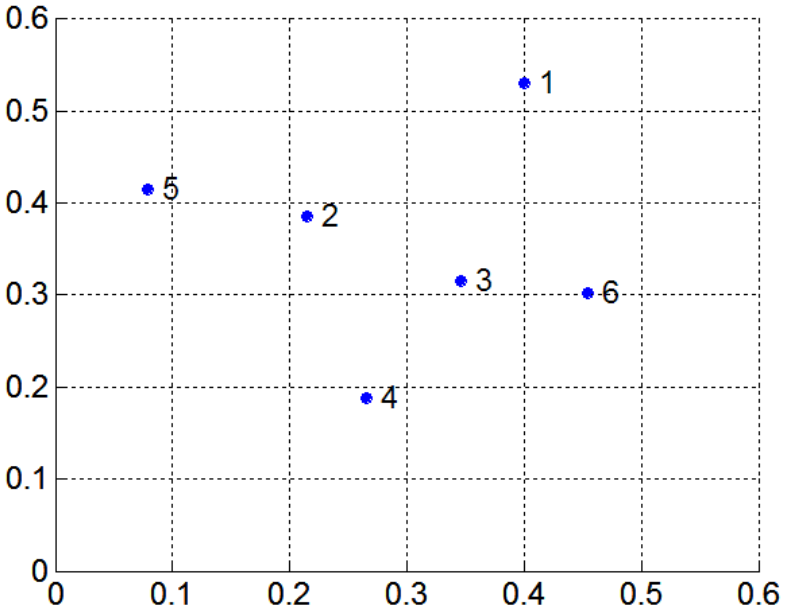
Média do grupo

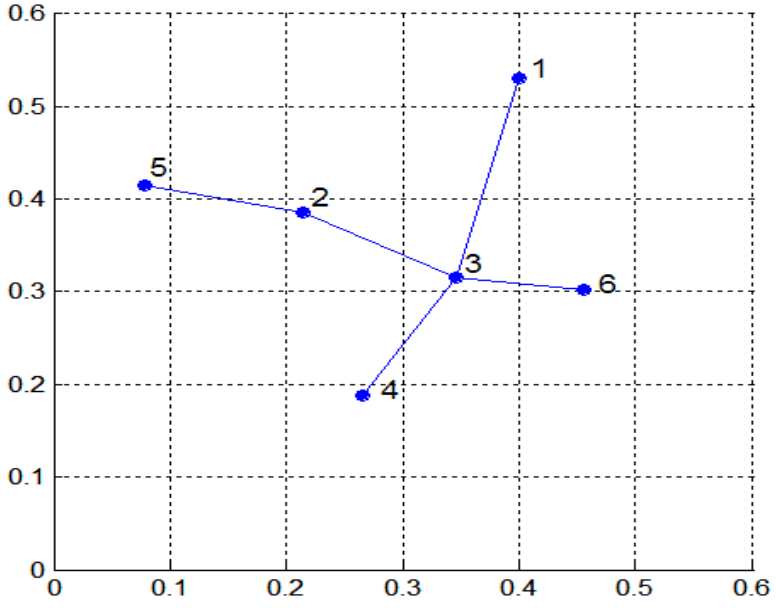
Método de Ward



MST: Agrupamento hierárquico divisivo

- MST (Minimum Spanning Tree)
  - Comece com uma árvore que consista em qualquer ponto
  - Em etapas sucessivas, procure o par mais próximo de pontos (p, q) de tal forma que um ponto (p) está na árvore atual, mas o outro (q) não é
  - Insere q à árvore e coloca uma borda entre p e q





MST: Agrupamento hierárquico divisivo

- Como usar o MST para construir a hierarquia de clusters?

1. Calcule a MST para o gráfico de proximidade
2. **Repita**
3. Crie um novo cluster quebrando o link correspondente à maior distância (menor semelhança).
4. **Até** que apenas clusters com únicos pontos permanecem

Agrupamento hierárquico: requisitos de tempo e espaço

- $O(N^2)$ em espaço, pois usa a matriz de proximidade.
  - N é o número de pontos.

- $O(N^3)$ em tempo em muitos casos
  - Há N passos e em cada passo precisamos atualizar e pesquisar a matriz de proximidade de tamanho N^2
  - A complexidade pode ser reduzida para $O(N^2 \log(N))$ com algumas técnicas (FFT – Fast Fourier Transform)

Agrupamento hierárquico: problemas e limitações

- Uma vez que uma decisão é tomada para combinar dois clusters, não pode ser desfeita
- Nenhuma função objetiva global é minimizada diretamente
- Diferentes esquemas têm problemas com um ou mais dos seguintes:
  - Sensibilidade a ruído e outliers
  - Dificuldade com clusters de tamanhos diferentes e formas não globulares
  - Quebra de clusters grandes