

# Inteligência Artificial

## Aula 13 – Aprendizado Não-Supervisionado

Tiago Alves de Oliveira

# Formas de Aprendizado

- Aprendizado Supervisionado
  - Árvores de Decisão.
  - K-Nearest Neighbor (KNN).
  - Support Vector Machines (SVM).
  - Redes Neurais.
- **Aprendizado Não-Supervisionado**
- Aprendizado Por Reforço



# Introdução

- Porém, muitas vezes temos que lidar com exemplos “**não-supervisionados**”, isto é, exemplos **não rotulados**.
- **Por que?**
  - Coletar e rotular um grande conjunto de exemplos pode custar muito tempo, esforço, dinheiro...

# Introdução

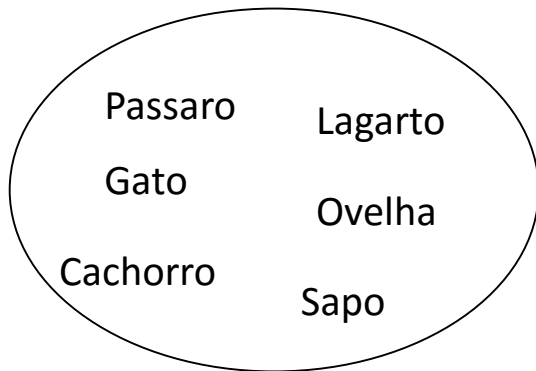
- Entretanto, podemos utilizar grandes quantidades de dados **não rotulados** para encontrar padrões existentes nestes dados. E somente depois supervisionar a rotulação dos agrupamentos encontrados.
- Esta abordagem é bastante utilizada em aplicações de **mineração de dados** (datamining), onde o conteúdo de grandes bases de dados não é conhecido antecipadamente.

# Introdução

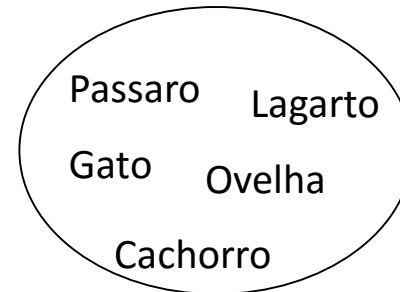
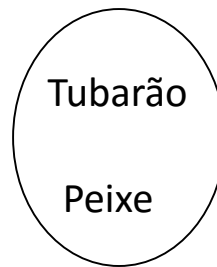
- O principal interesse do aprendizado não-supervisionado é desvendar a organização dos padrões existentes nos dados através de **clusters** (agrupamentos) consistentes.
- Com isso, é possível descobrir **similaridades e diferenças** entre os padrões existentes, assim como derivar conclusões úteis a respeito deles.

# Introdução

- Exemplos de agrupamentos (clusters):



Existencia de pulmões



Ambiente onde vivem



# Clusterização

- A **clusterização** é o processo de **agrupar** um conjunto de objetos físicos ou abstratos em classes de objetos **similares**.
- Um cluster é uma coleção de objetos que são similares uns aos outros (de acordo com algum **critério de similaridade** pré-definido) e dissimilares a objetos pertencentes a outros clusters.



# Critério de Similaridade

- A similaridade é difícil de ser definida...



# Processo de Aprendizado Não-Supervisionado

- As **etapas do processo** de aprendizagem não supervisionada são:
  - (1) Seleção de atributos
  - (2) Medida de proximidade
  - (3) Critério de agrupamento
  - (4) Algoritmo de agrupamento
  - (5) Verificação dos resultados
  - (6) Interpretação dos resultados

# Processo de Aprendizado Não-Supervisionado

- **(1) Seleção de Atributos:**
  - Atributos devem ser adequadamente selecionados de forma a codificar a **maior quantidade possível de informações** relacionada a tarefa de interesse.
  - Os atributos devem ter também uma **redundância mínima** entre eles.

# Processo de Aprendizado Não-Supervisionado

- **(2) Medida de Proximidade:**

- Medida para quantificar quão **similar** ou **dissimilar** são dois vetores de atributos.
- É ideal que todos os atributos **contribuam de maneira igual** no cálculo da medida de proximidade.
  - Um atributo não pode ser dominante sobre o outro, ou seja, é importante normalizar os dados.

# Processo de Aprendizado Não-Supervisionado

- **(3) Critério de Agrupamento:**

- Depende da interpretação que o especialista dá ao termo **sensível** com base no tipo de cluster que são esperados.
- Por exemplo, um cluster compacto de vetores de atributos pode ser sensível de acordo com um critério enquanto outro cluster alongado, pode ser sensível de acordo com outro critério.



# Processo de Aprendizado Não-Supervisionado

- **(4) Algoritmo de Agrupamento:**
  - Tendo adotado uma medida de proximidade e um critério de agrupamento devemos escolher um **algoritmo de clusterização** que revele a estrutura agrupada do conjunto de dados.

# Processo de Aprendizado Não-Supervisionado

- **(5) Validação dos Resultados:**
  - Uma vez obtidos os resultados do algoritmo de agrupamento, devemos verificar se o **resultado esta correto**.
  - Isto geralmente é feito através de testes apropriados.

# Processo de Aprendizado Não-Supervisionado

- **(6) Interpretação dos Resultados:**
  - Em geral, os resultados da clusterização devem ser integrados com outras **evidências experimentais** e análises para chegar as conclusões corretas.



# Processo de Aprendizado Não-Supervisionado

- Diferentes escolhas de atributos, medidas de proximidade, critérios de agrupamento e algoritmos de clusterização levam a **resultados totalmente diferentes**.
- Qual resultado é o correto?

# Clusterização

- Dado um conjunto de dados  $X$ :

$$X = \{x_1, x_2, \dots, x_n\}$$

- Definimos como um  $m$ -agrupamento de  $X$  a partição de  $X$  em  $m$  conjuntos (clusters ou grupos)  $C_1, C_2, \dots, C_m$  tal que as três condições seguintes sejam satisfeitas:
  - Nenhum cluster pode ser vazio ( $C_i \neq \emptyset$ ).
  - A união de todos os cluster deve ser igual ao conjunto de dados que gerou os clusters, ou seja,  $X$ .
  - A interseção de dois clusters deve ser vazio, ou seja, dois cluster não podem conter vetores em comum ( $C_i \cap C_j = \emptyset$ ).

# Clusterização

- Os vetores contidos em um cluster  $C_i$  devem ser mais similares uns aos outros e menos similares aos vetores presentes nos outros clusters.
- Tipos de Clusters:



Clusters compactos



Clusters alongados



Clusters esféricos e elipsoidais

# Medidas de Proximidade

- **Medidas de Dissimilaridade:**
  - Métrica  $l_p$  ponderada;
  - Métrica Norma  $l_\infty$  ponderada;
  - Métrica  $l_2$  ponderada (Mahalanobis);
  - Métrica  $l_p$  especial (Manhattan);
  - Distância de Hamming;
- **Medidas de Similaridade:**
  - Produto interno (inner);
  - Medida de Tanimoto;

# Algoritmos de Clustering

- Os **algoritmos de clusterização** buscam identificar padrões existentes em conjuntos de dados.
- Os algoritmos de clusterização podem ser divididos em varias categorias:
  - Sequenciais;
  - Hierárquicos;
  - Baseados na otimização de funções custo;
  - Outros: Fuzzy, SOM, LVQ...

# Algoritmos Sequenciais

- São algoritmos diretos e rápidos.
- Geralmente, todos os vetores de características são apresentados ao algoritmo uma ou várias vezes.
- O resultado final geralmente depende da ordem de apresentação dos vetores de características.

# Algoritmos Sequenciais

- Basic Sequential Algorithmic Scheme (BSAS)
  - Todos os vetores são apresentados uma única vez ao algoritmo.
  - Número de clusters não é conhecido inicialmente.
  - Novos clusters são criados enquanto o algoritmo evolui.

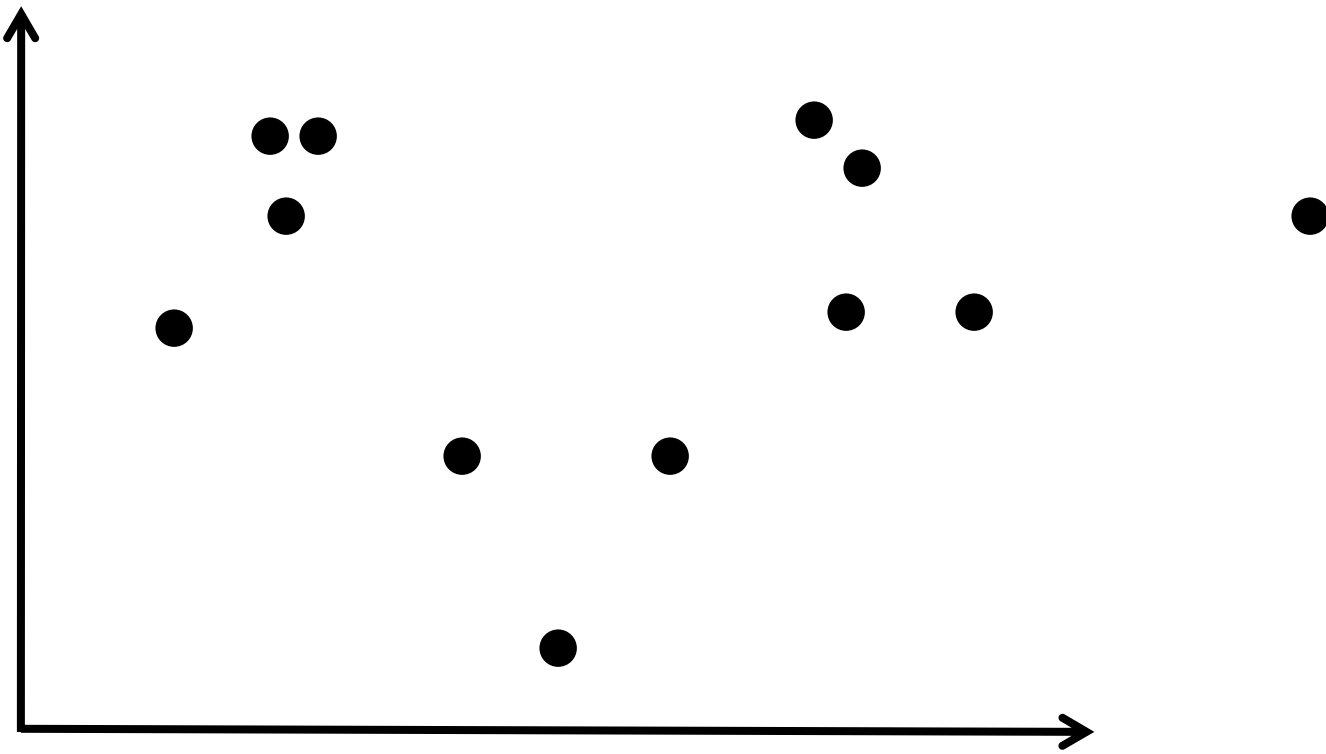
# Basic Sequential Algorithmic Scheme (BSAS)

- **Parâmetros do BSAS:**
  - $d(\mathbf{x}, C)$ : métrica de distância entre um vetor de características  $\mathbf{x}$  e um cluster  $C$ .
  - $\Theta$ : limiar de dissimilaridade.
  - $q$ : número máximo de clusters.
- **Ideia Geral do Algoritmo:**
  - Para um dado vetor de características, designá-lo para um cluster existente ou criar um novo cluster (depende da distância entre o vetor e os clusters já formados).



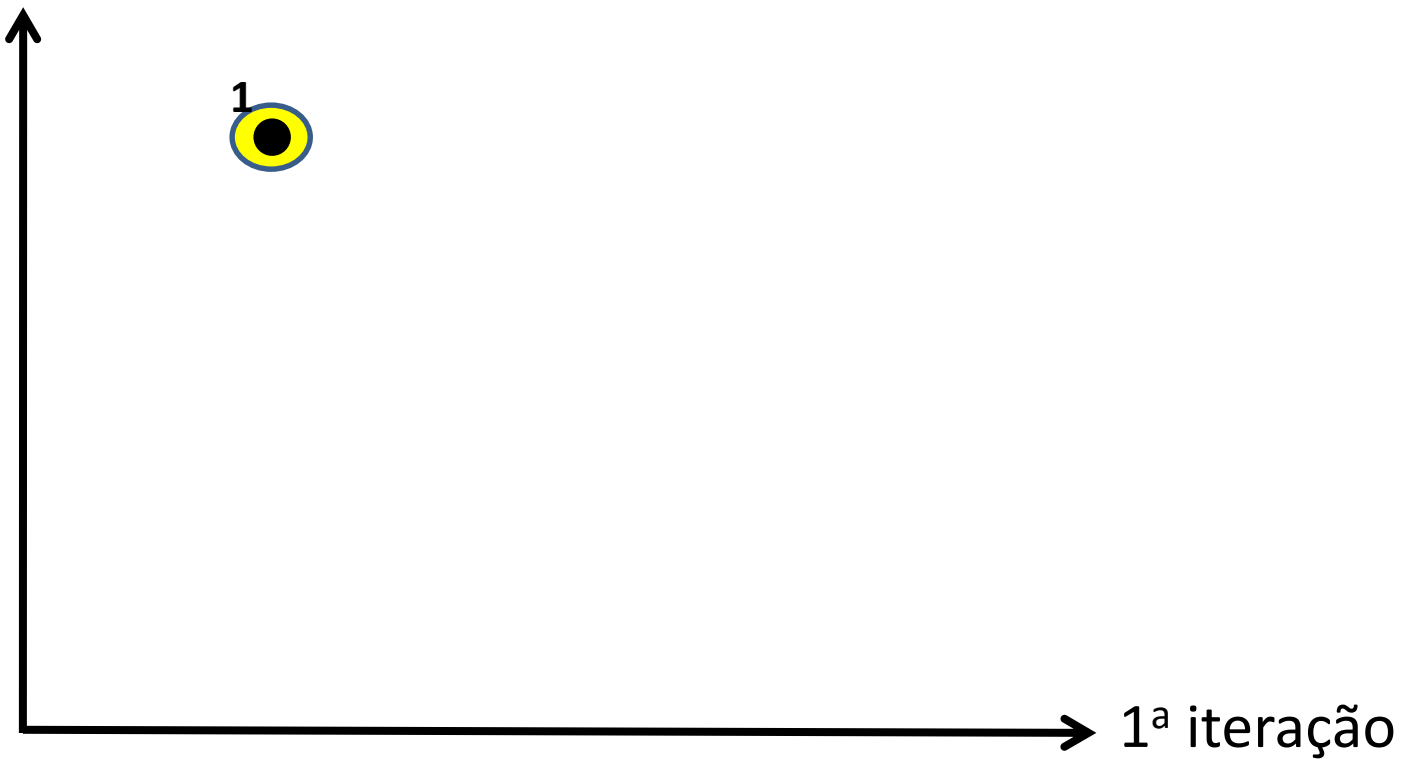
# Basic Sequential Algorithmic Scheme (BSAS)

- Exemplo 1:



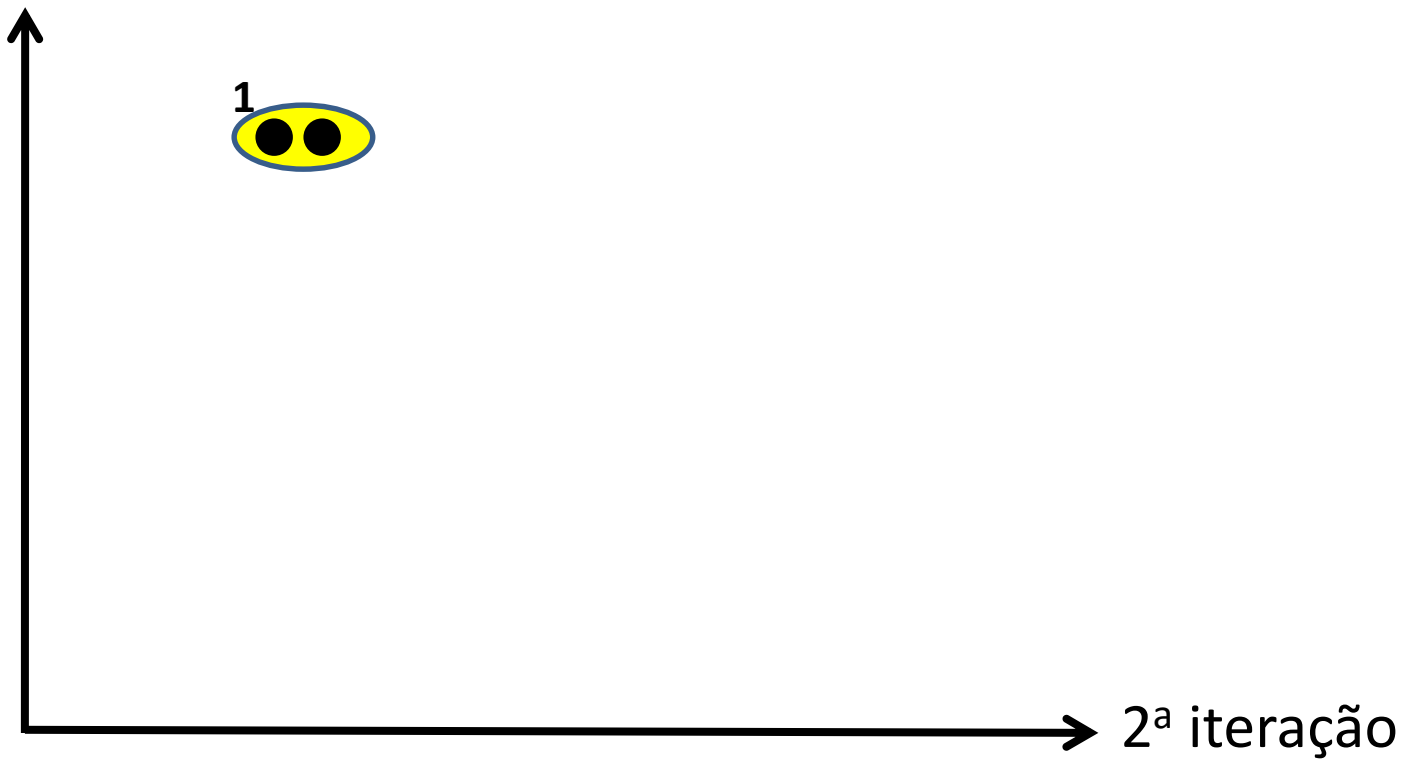
# Basic Sequential Algorithmic Scheme (BSAS)

- Exemplo 1:



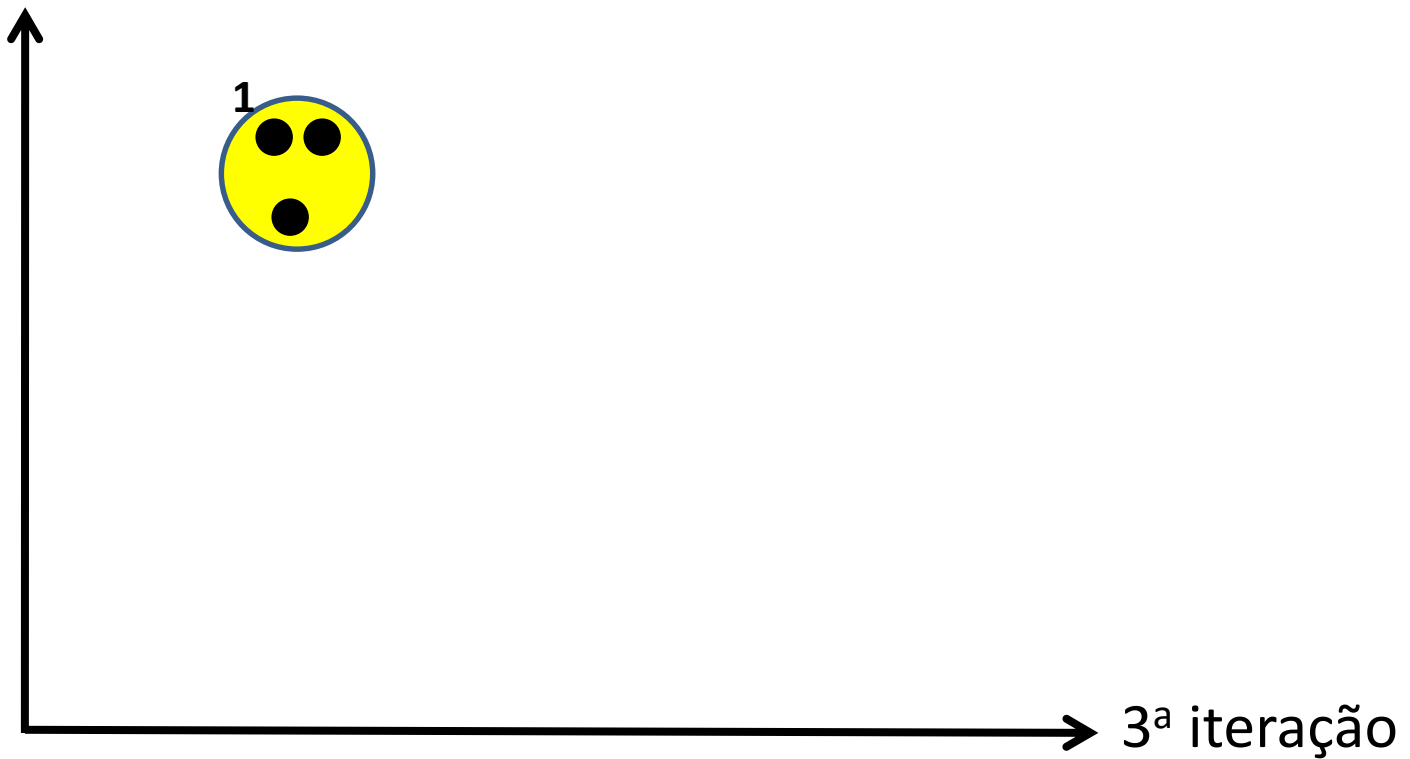
# Basic Sequential Algorithmic Scheme (BSAS)

- Exemplo 1:



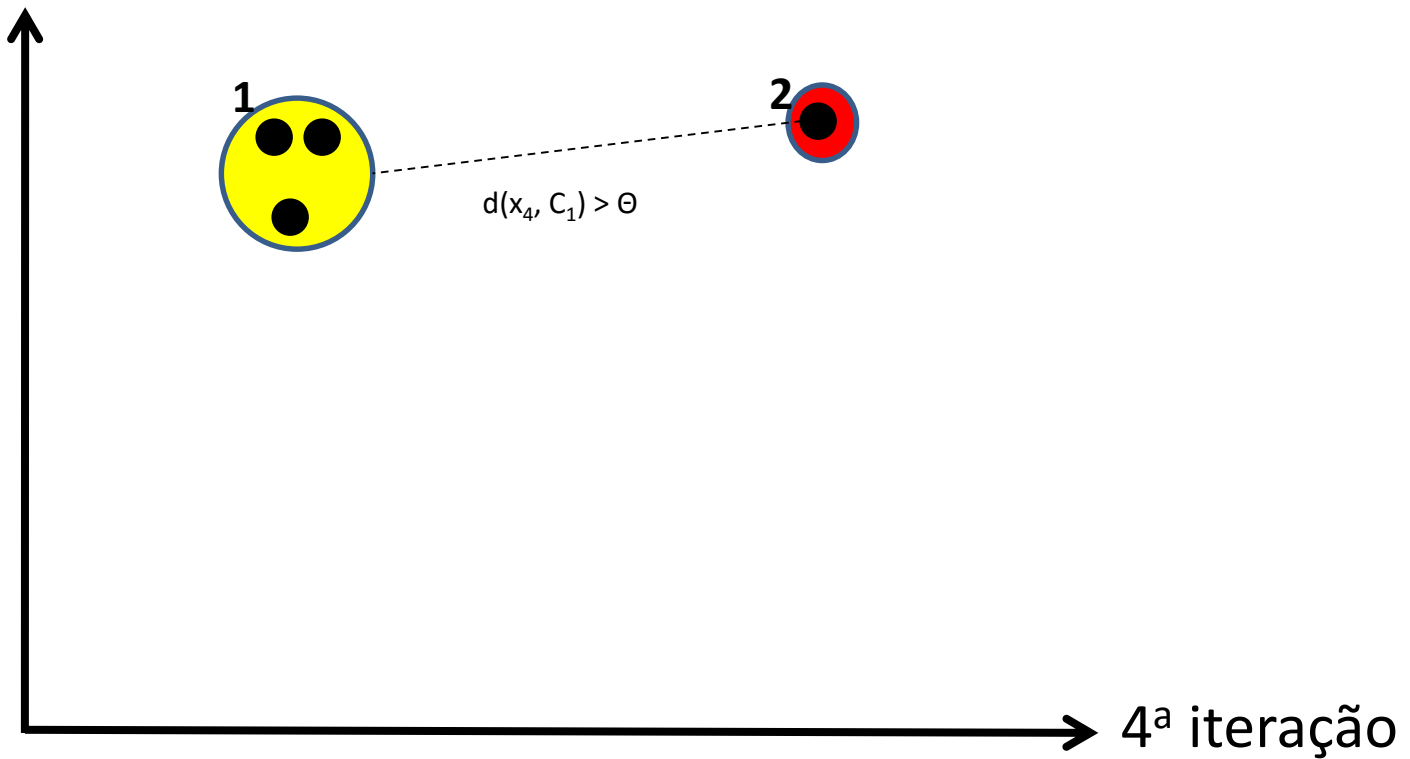
# Basic Sequential Algorithmic Scheme (BSAS)

- Exemplo 1:



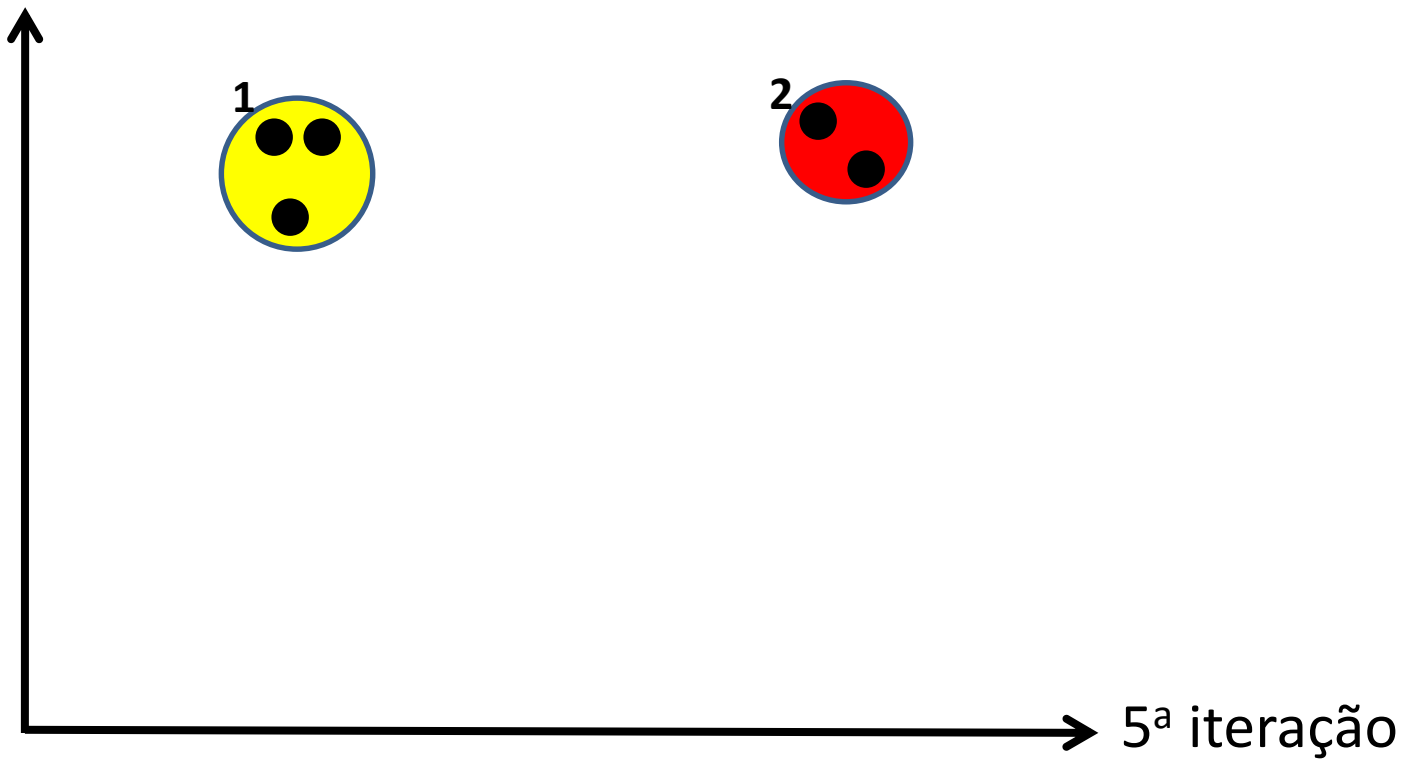
# Basic Sequential Algorithmic Scheme (BSAS)

- Exemplo 1:



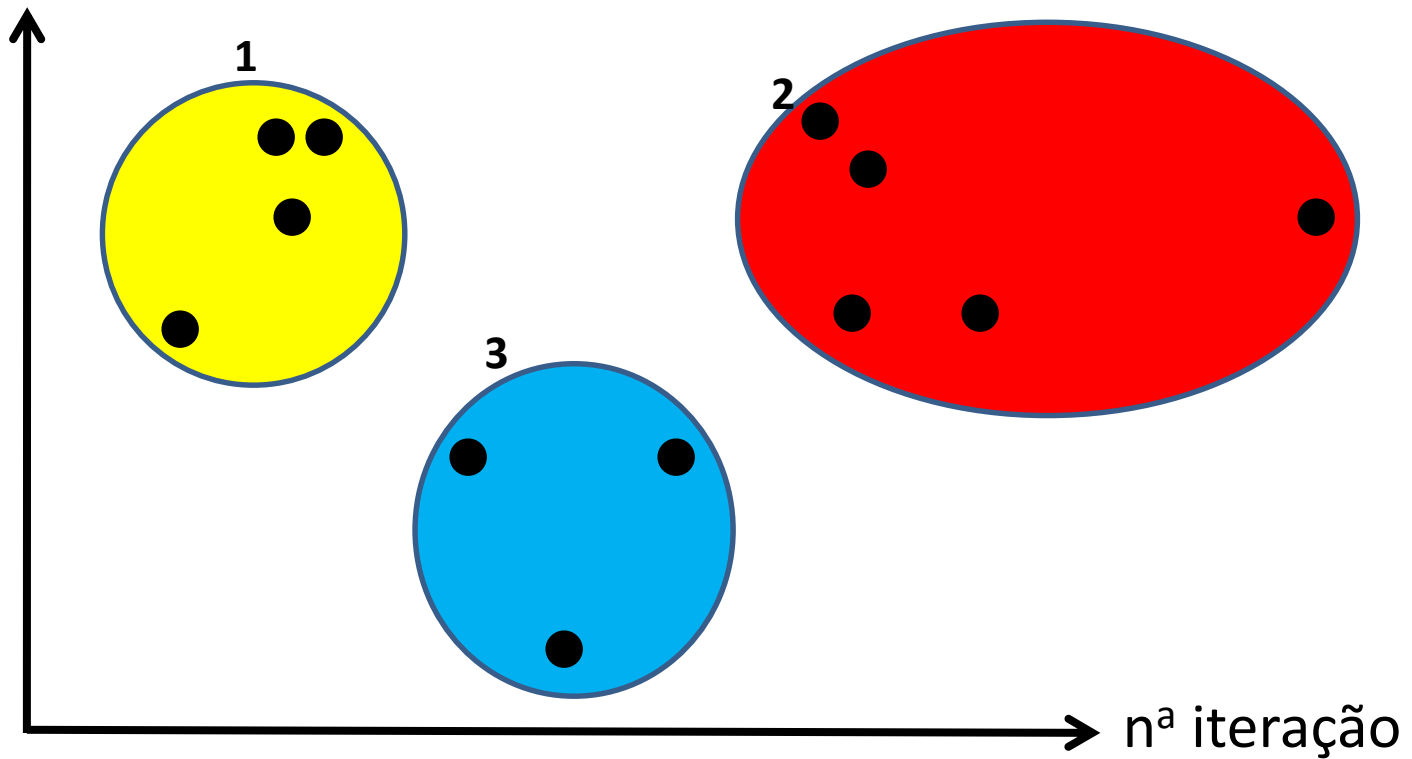
# Basic Sequential Algorithmic Scheme (BSAS)

- Exemplo 1:



# Basic Sequential Algorithmic Scheme (BSAS)

- Exemplo 1:



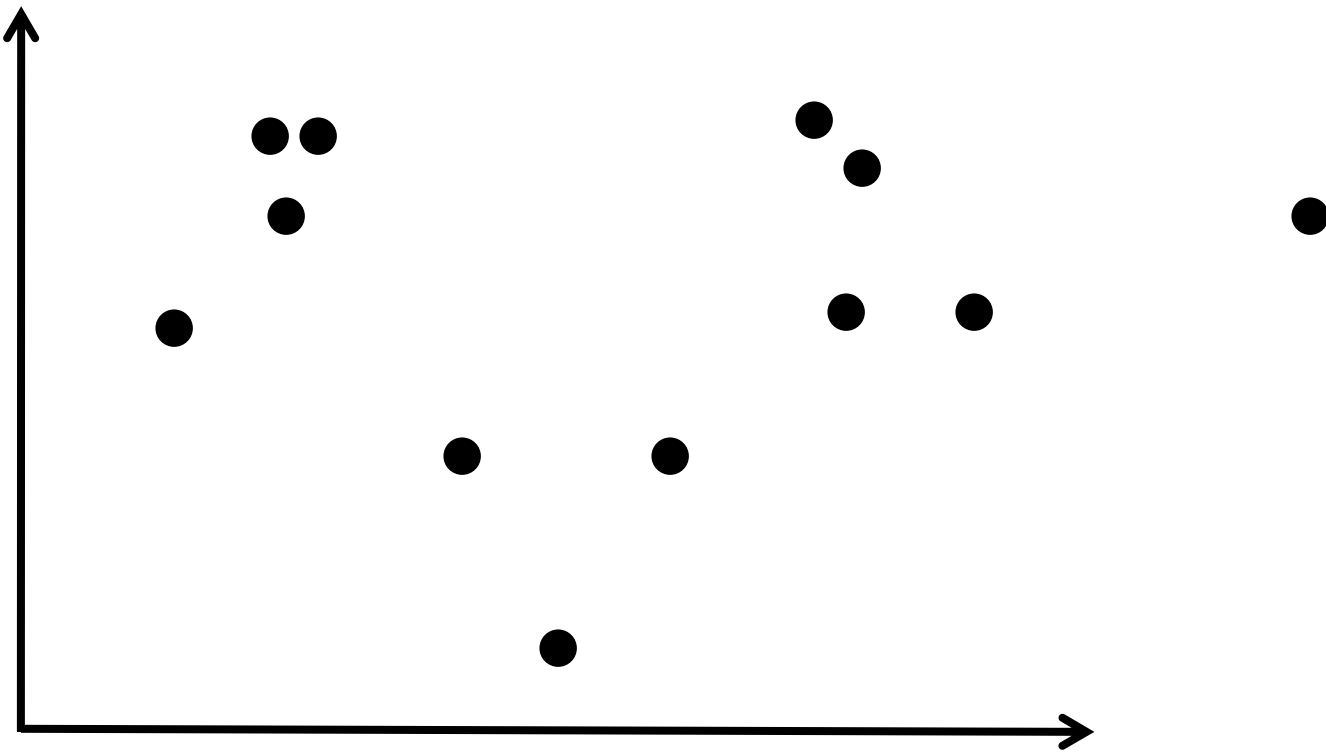
# Clusterização Hierárquica

- Os algoritmos de **clusterização hierárquica** pode ser divididos em 2 subcategorias:
- **Aglomerativos:**
  - Produzem uma sequência de agrupamentos com um número decrescente de clusters a cada passo.
  - Os agrupamentos produzidos em cada passo resultam da fusão de dois clusters em um.
- **Divisivos:**
  - Atuam na direção oposta, isto é, eles produzem uma sequência de agrupamentos com um número crescente de clusters a cada passo.
  - Os agrupamentos produzidos em cada passo resultam da partição de um único cluster em dois.



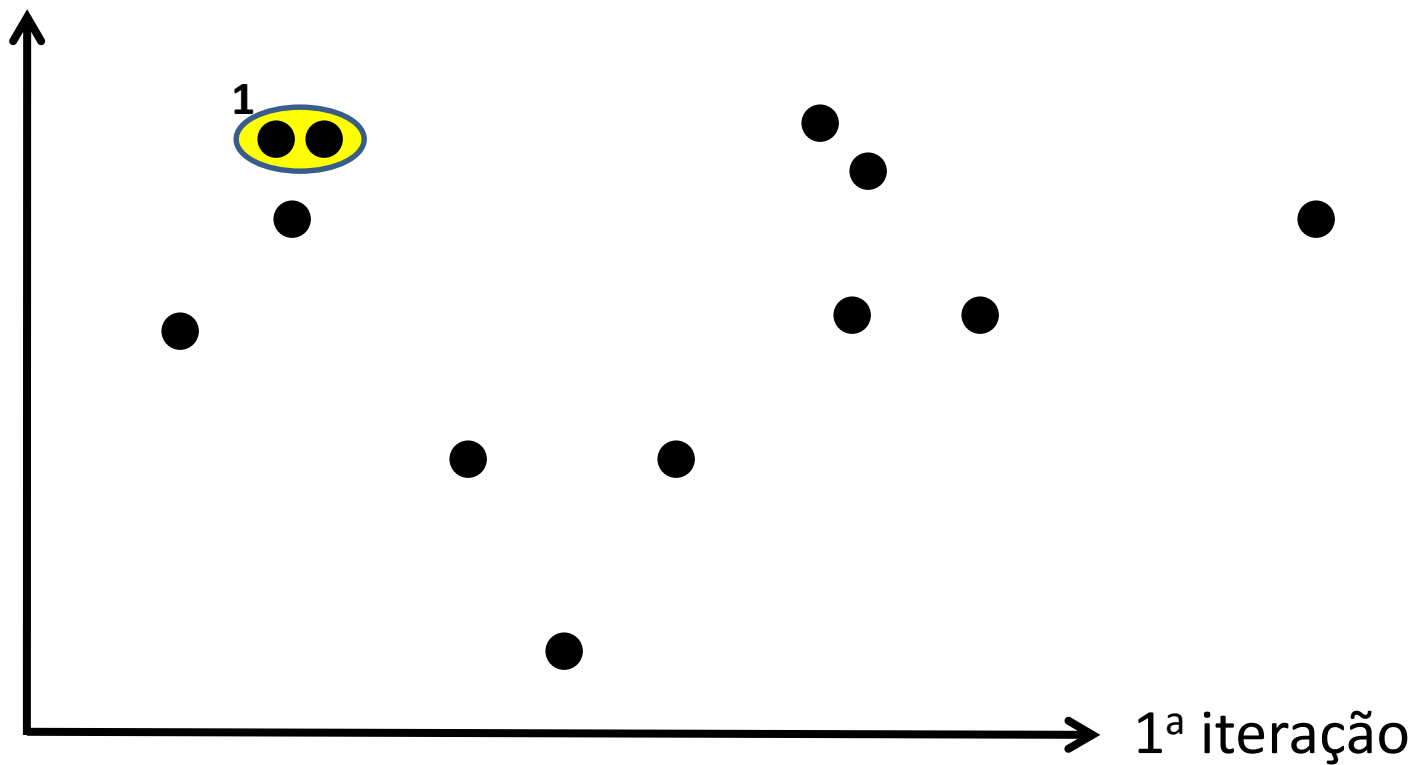
# Clusterização Hierárquica

- Exemplo 1 – Aglomerativo:



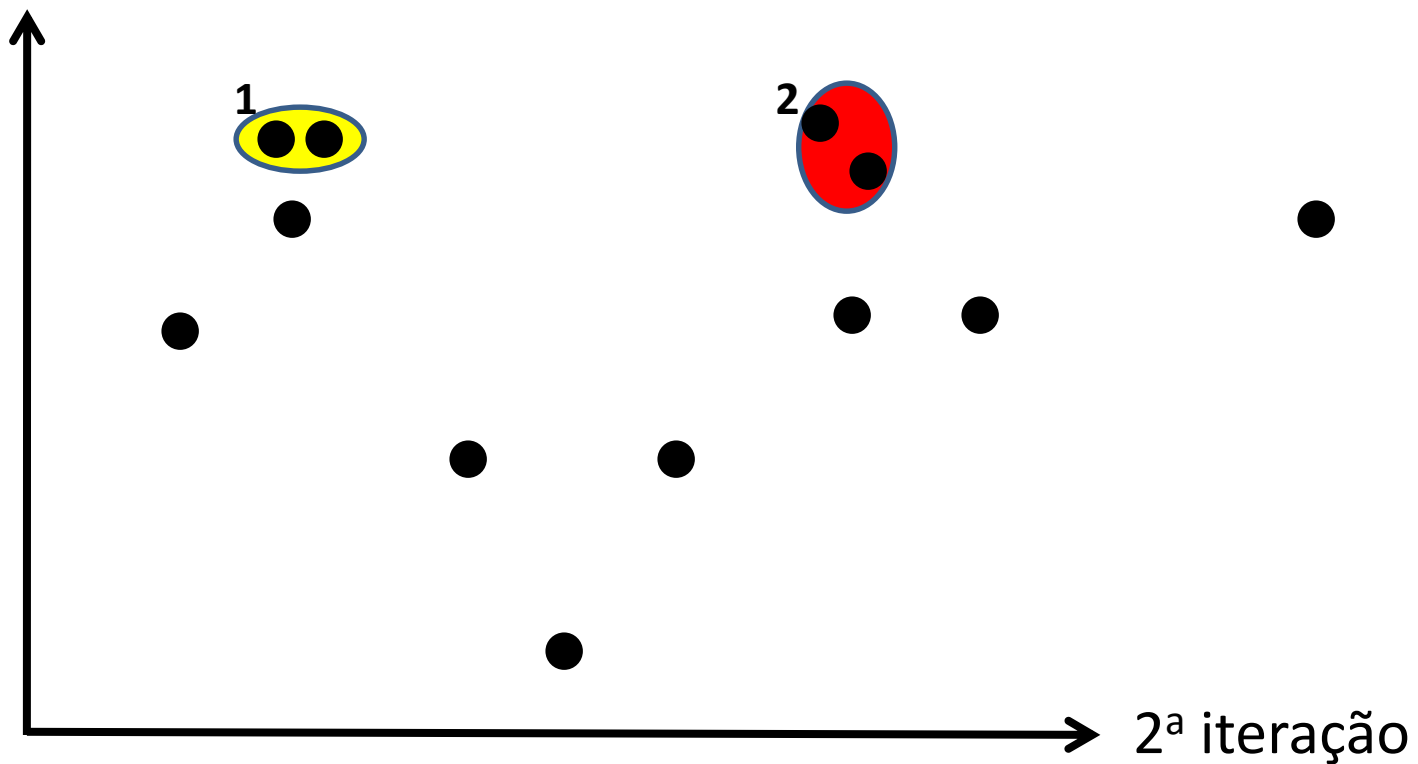
# Clusterização Hierárquica

- Exemplo 1 – Aglomerativo:



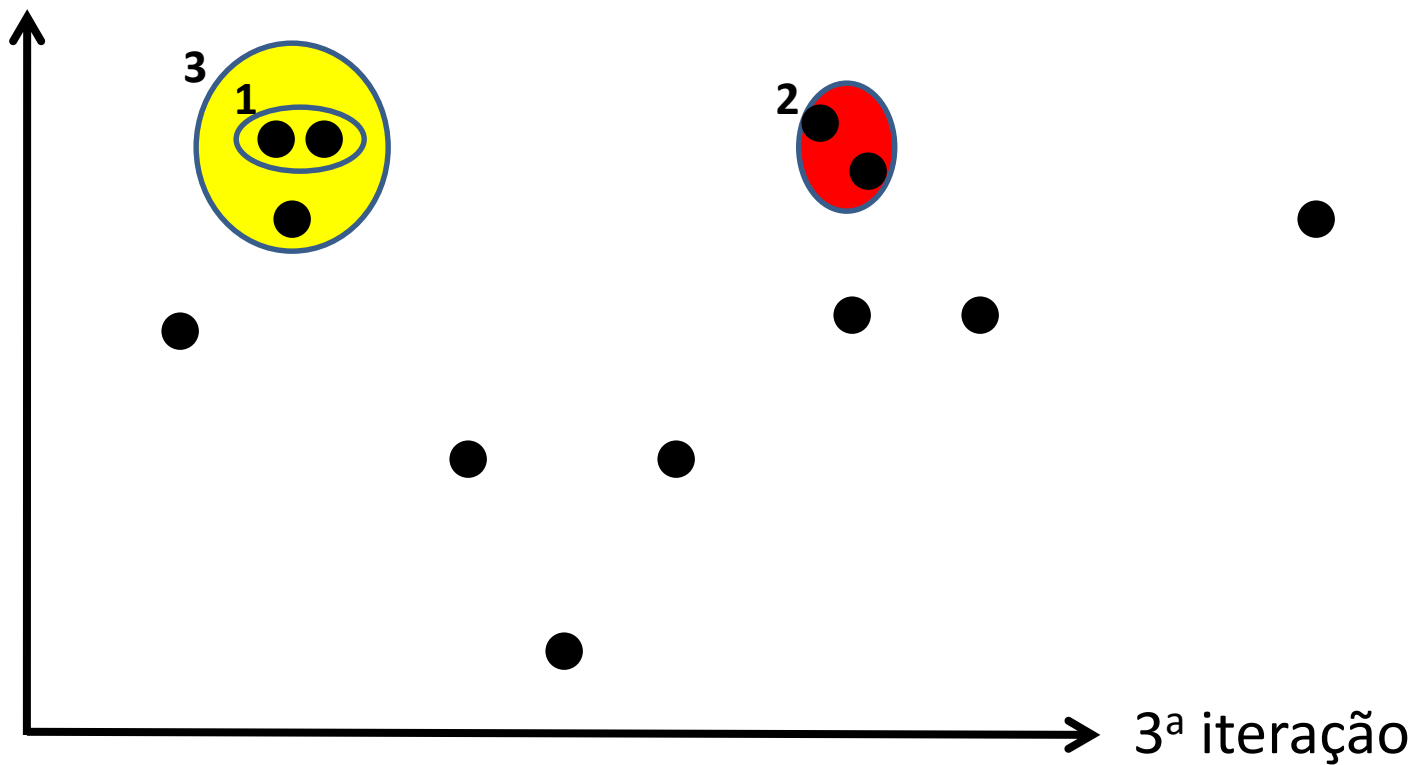
# Clusterização Hierárquica

- Exemplo 1 – Aglomerativo:



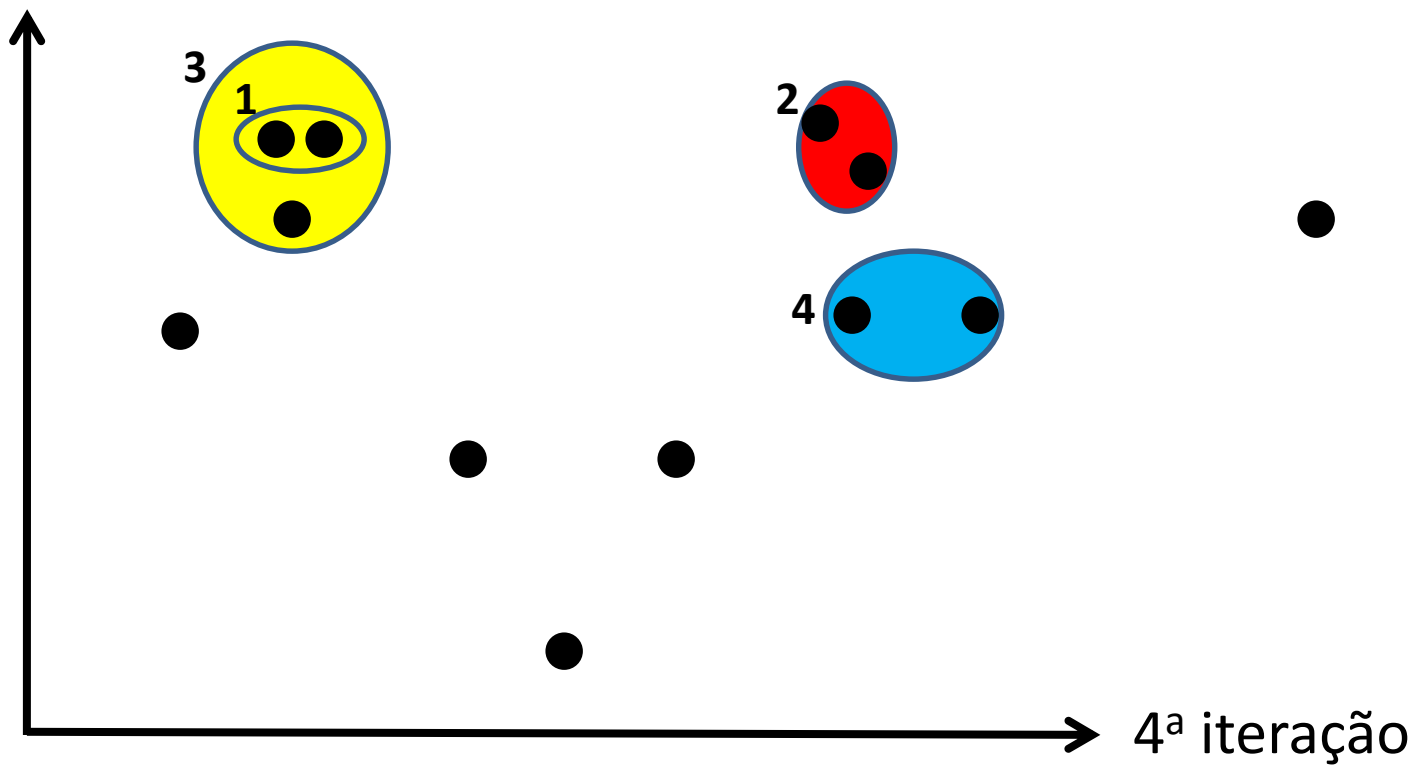
# Clusterização Hierárquica

- Exemplo 1 – Aglomerativo:



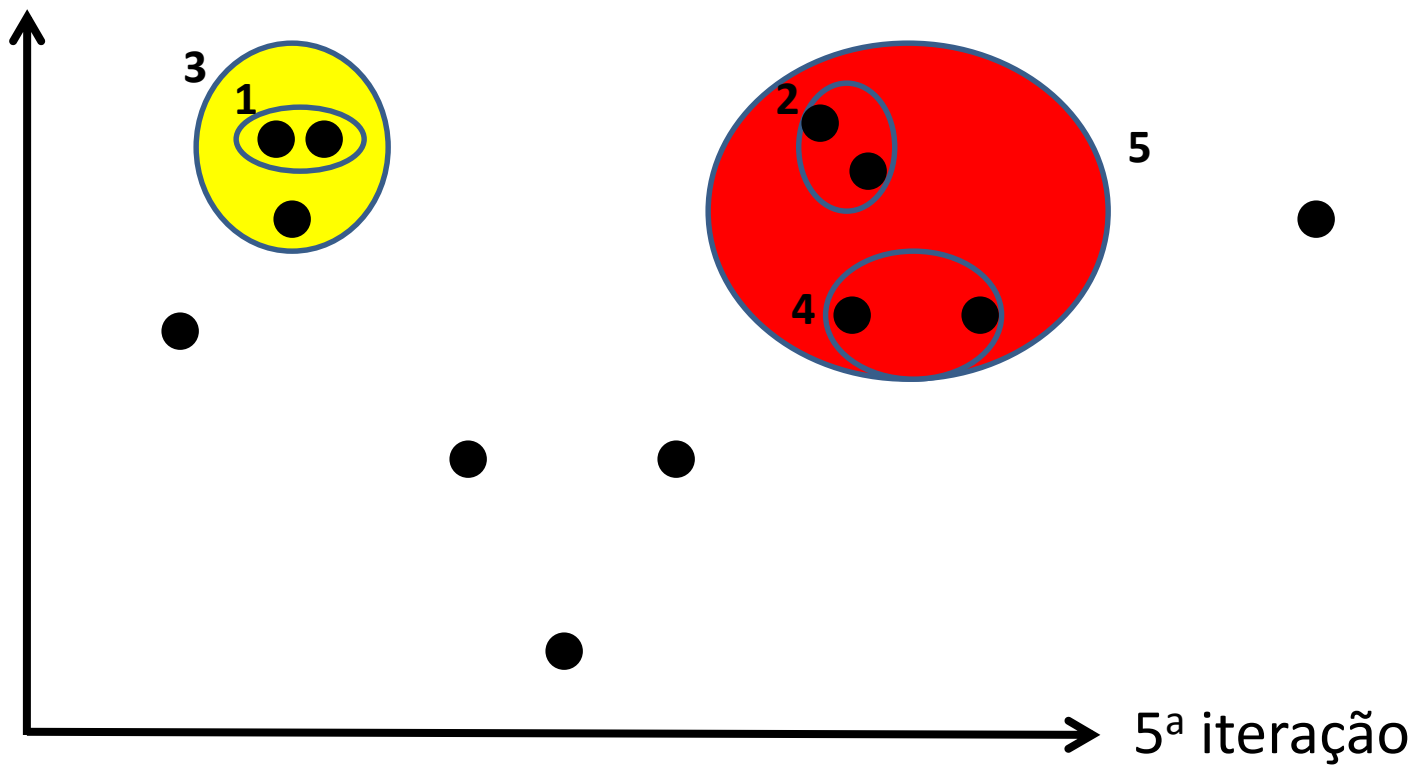
# Clusterização Hierárquica

- Exemplo 1 – Aglomerativo:



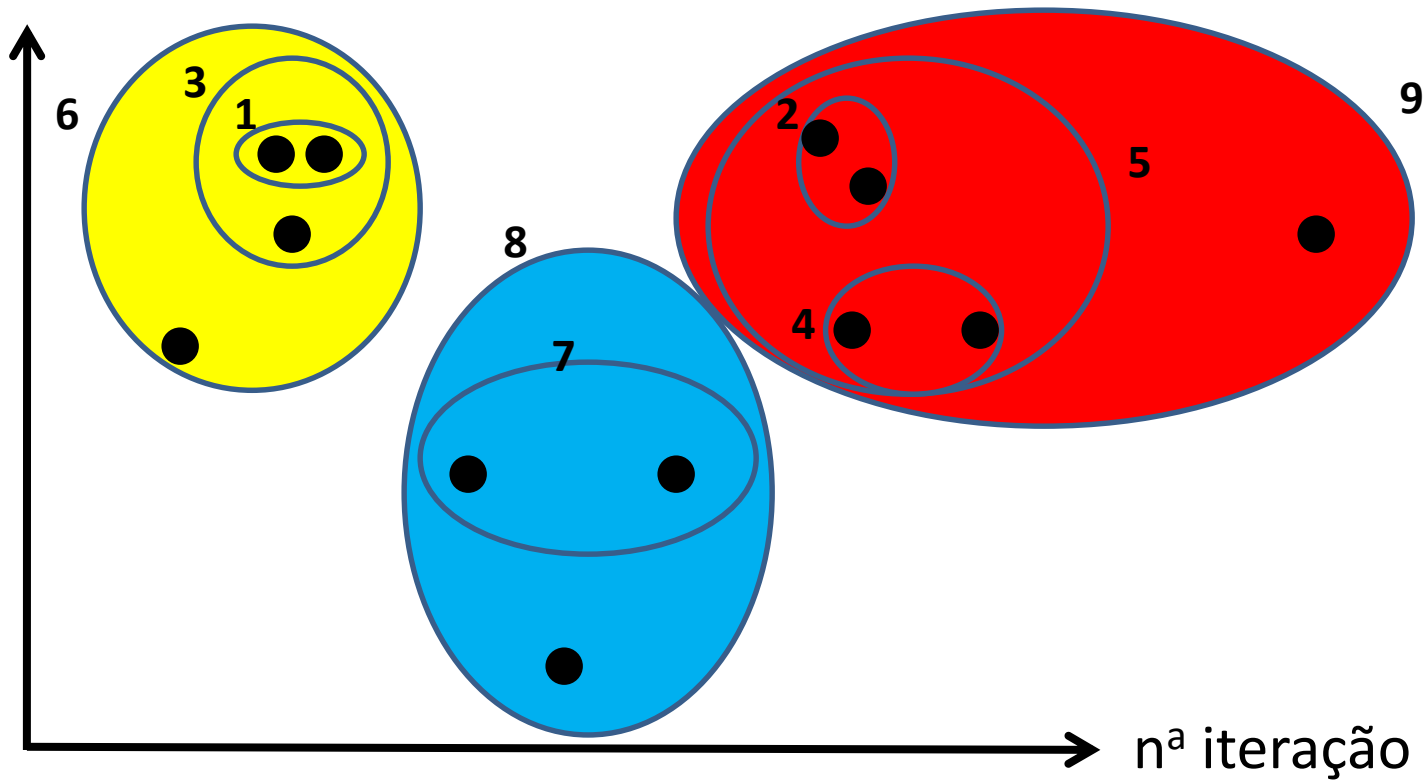
# Clusterização Hierárquica

- Exemplo 1 – Aglomerativo:



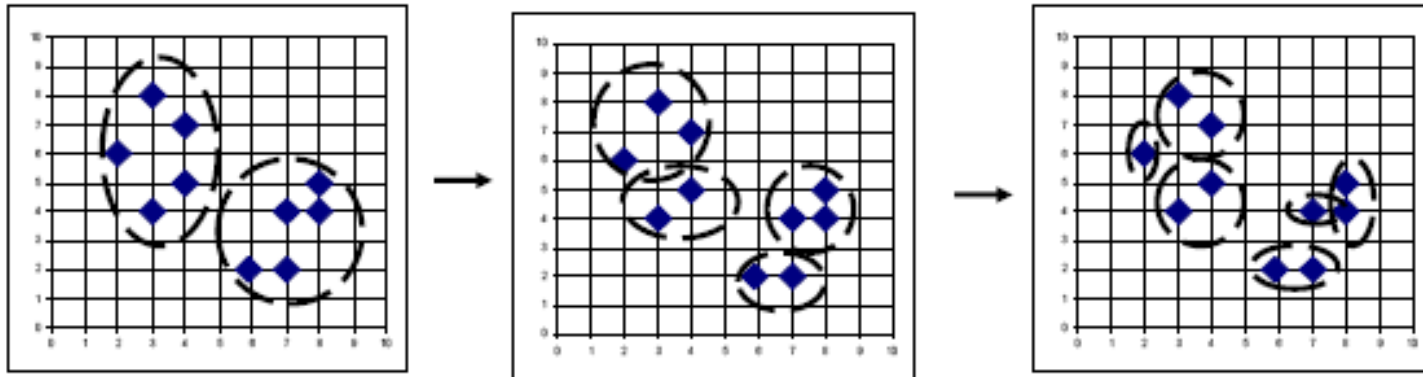
# Clusterização Hierárquica

- Exemplo 1 – Aglomerativo:



# Clusterização Hierárquica

- Exemplo 2 – Divisivo:



- Processo inverso.



# K-Means

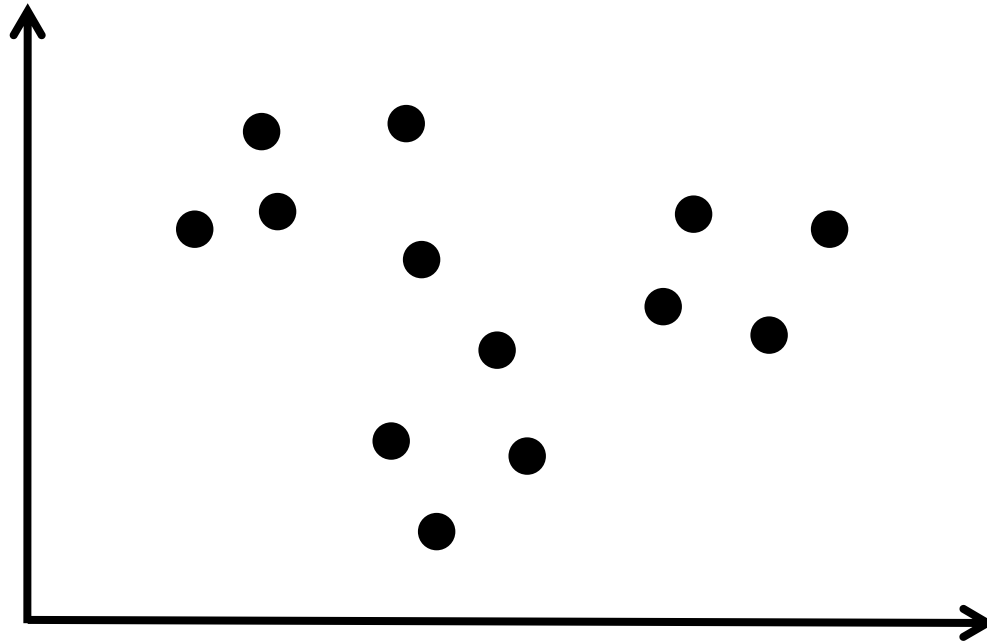
- É a técnica **mais simples** de aprendizagem não supervisionada.
- Consiste em fixar **k centróides** (de maneira aleatória), um para cada grupo (clusters).
- Associar cada indivíduo ao seu **centróide mais próximo**.
- Recalcular os centróides com base nos indivíduos classificados.

# Algoritmo K-Means

- (1)** Selecione  $k$  centróides iniciais.
- (2)** Forme  $k$  clusters associando cada exemplo ao seu centróide mais próximo.
- (3)** Recalcule a posição dos centróides com base no centro de gravidade do cluster.
- (4)** Repita os passos 2 e 3 até que os centróides não sejam mais movimentados.

# Algoritmo K-Means

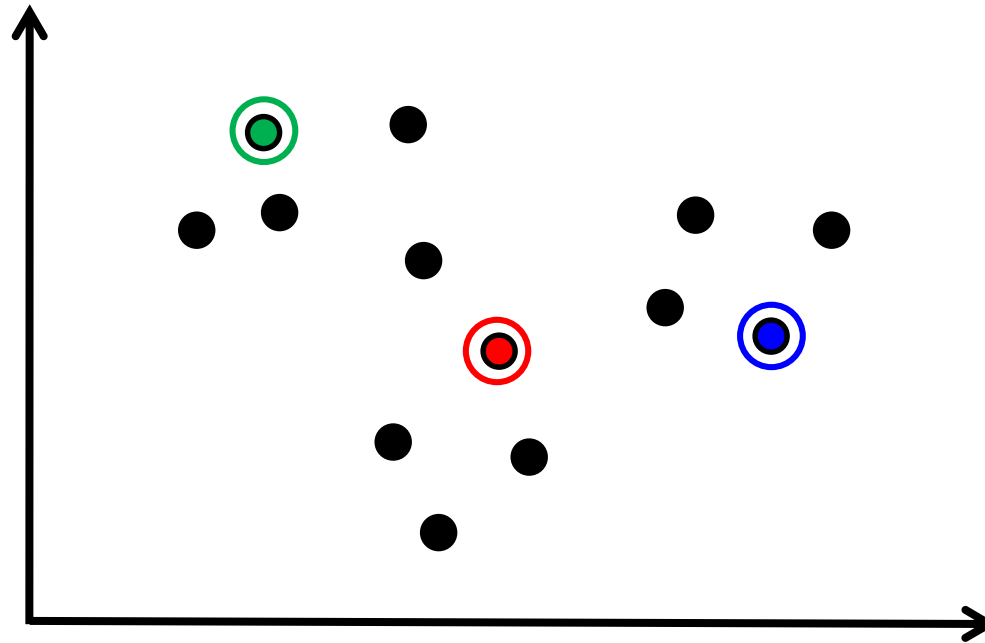
- Exemplo:



# Algoritmo K-Means

- Exemplo:

$k = 3$

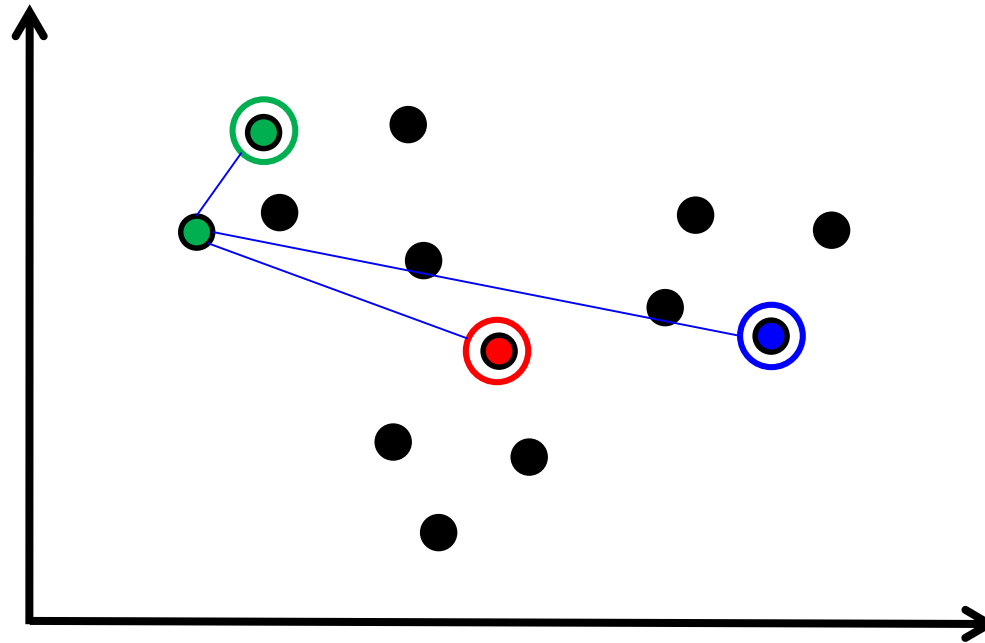


Seleciona-se  $k$  centróides iniciais.

# Algoritmo K-Means

- Exemplo:

$k = 3$

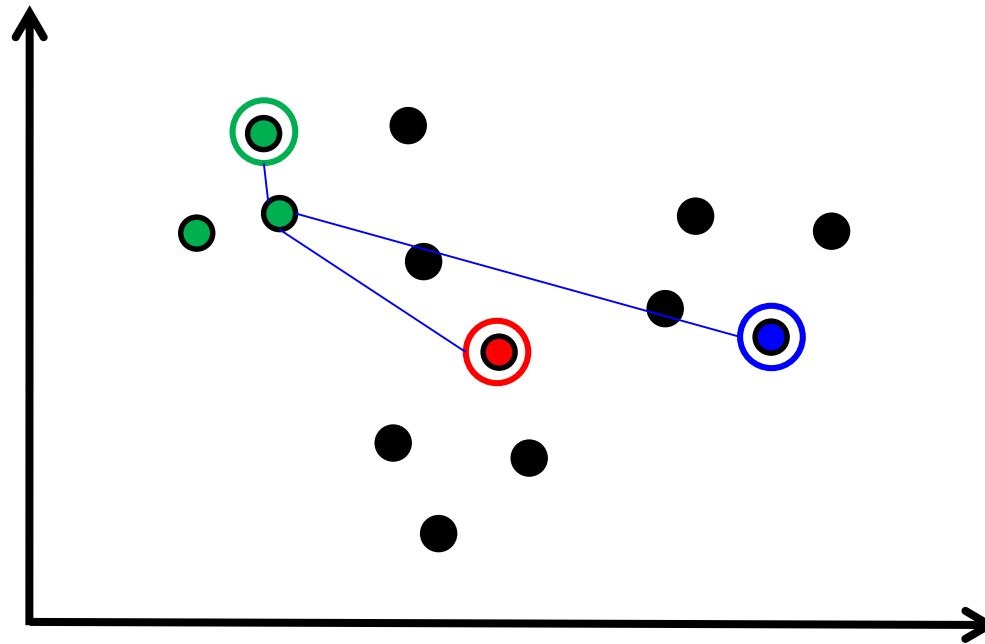


1ª iteração

# Algoritmo K-Means

- Exemplo:

$k = 3$

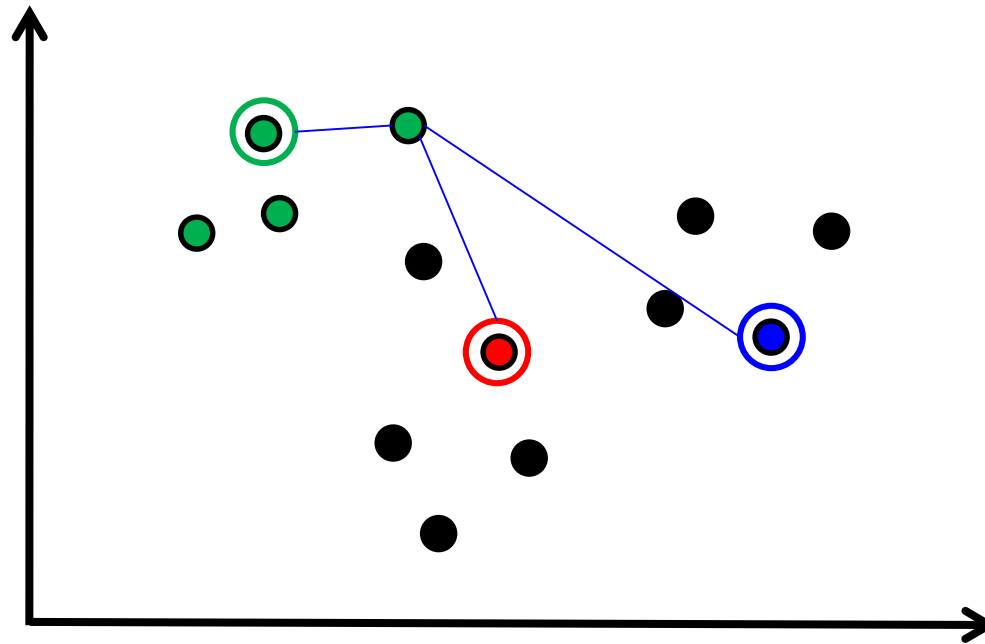


2ª iteração

# Algoritmo K-Means

- Exemplo:

$k = 3$

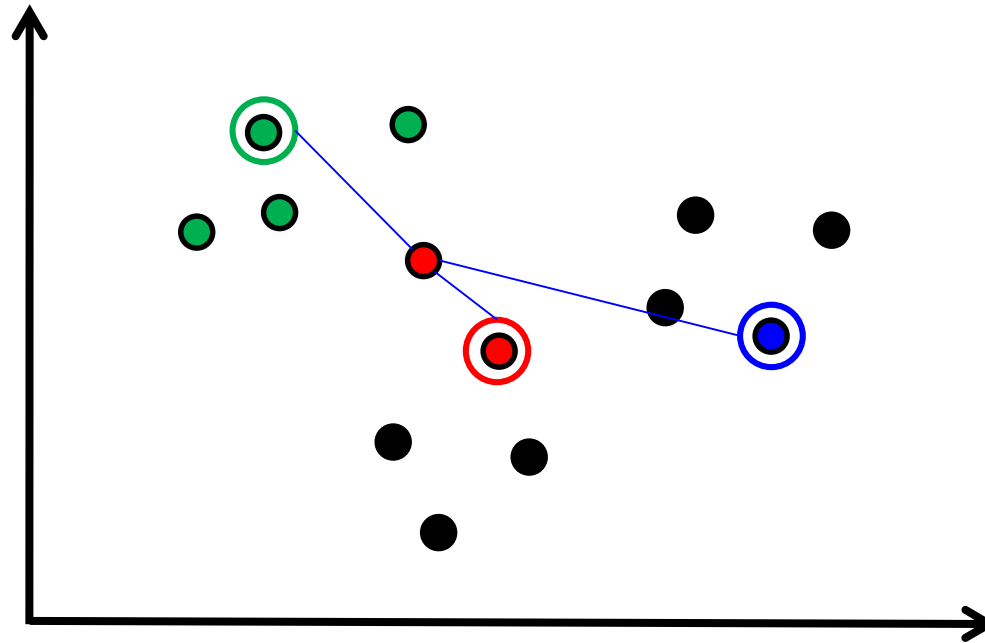


3ª iteração

# Algoritmo K-Means

- Exemplo:

$k = 3$



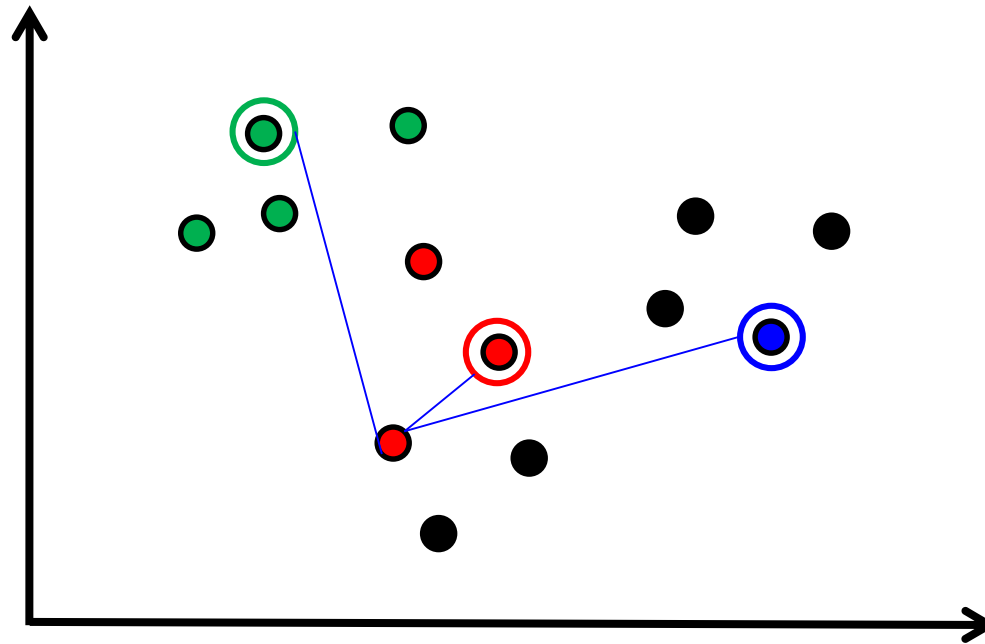
4ª iteração



# Algoritmo K-Means

- Exemplo:

$k = 3$

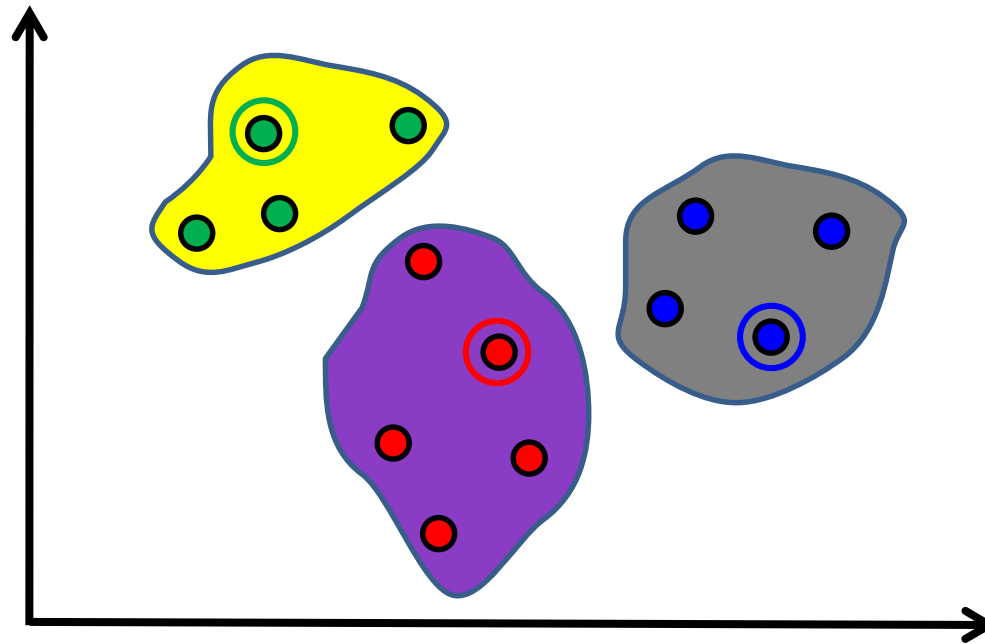


5ª iteração

# Algoritmo K-Means

- Exemplo:

$k = 3$

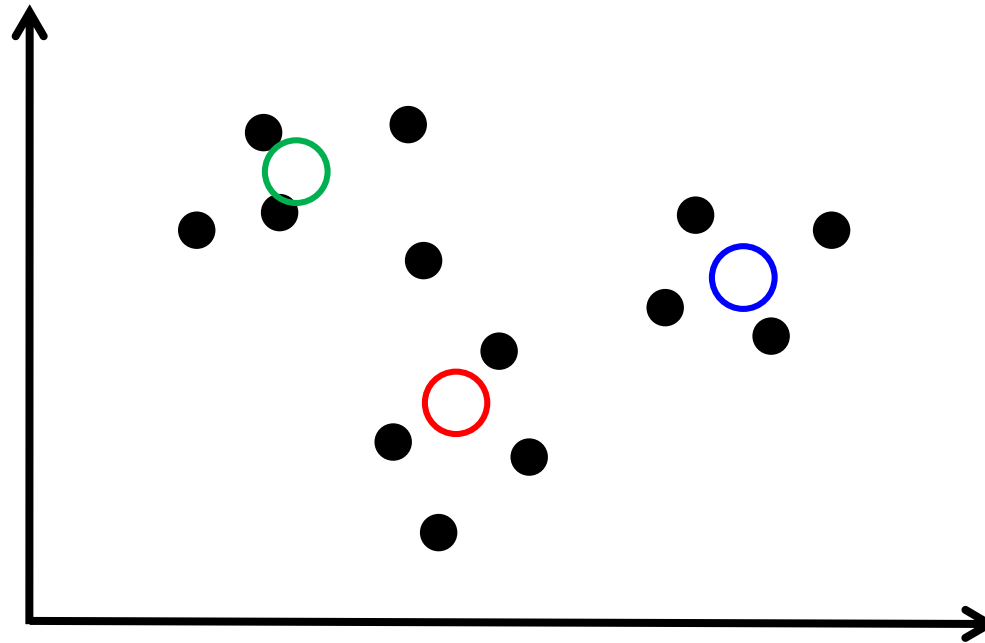


$n^{\text{a}}$  iteração

# Algoritmo K-Means

- Exemplo:

$k = 3$

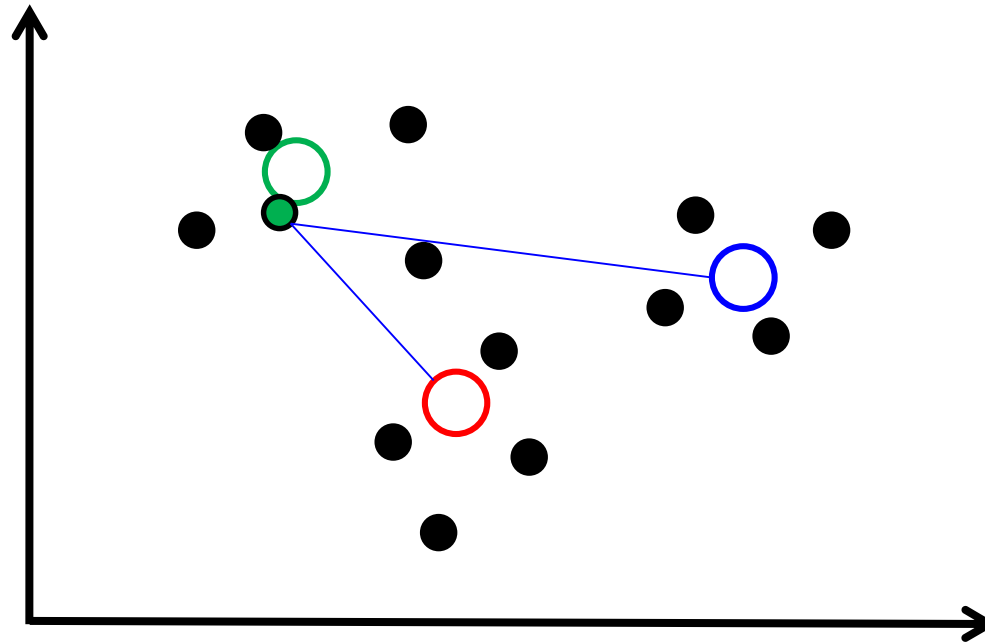


Repete-se os passos anteriores até que os centróides não se movam mais.

# Algoritmo K-Means

- Exemplo:

$k = 3$

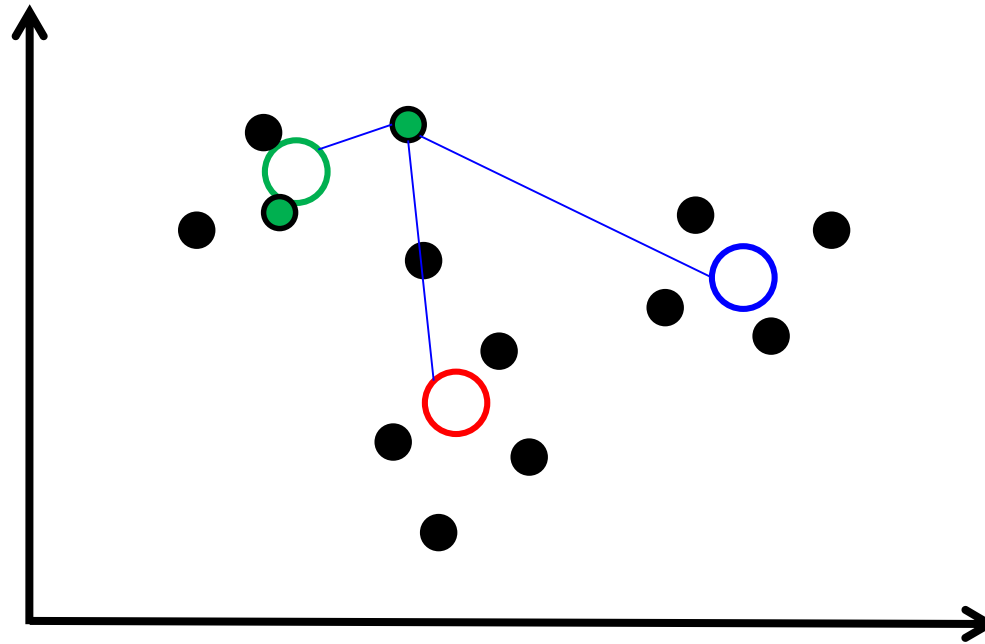


1ª iteração

# Algoritmo K-Means

- Exemplo:

$k = 3$

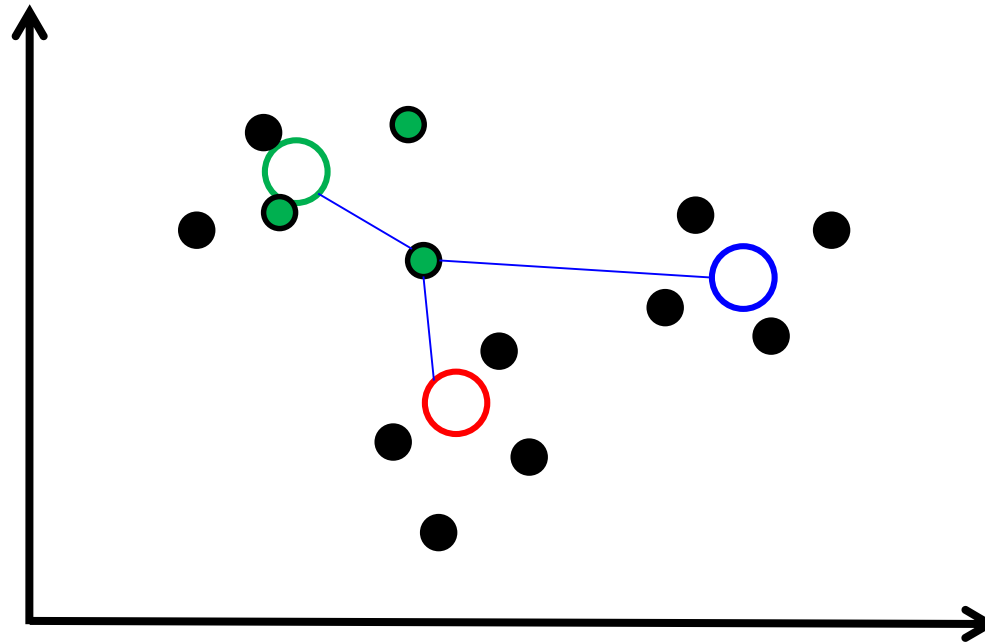


2ª iteração

# Algoritmo K-Means

- Exemplo:

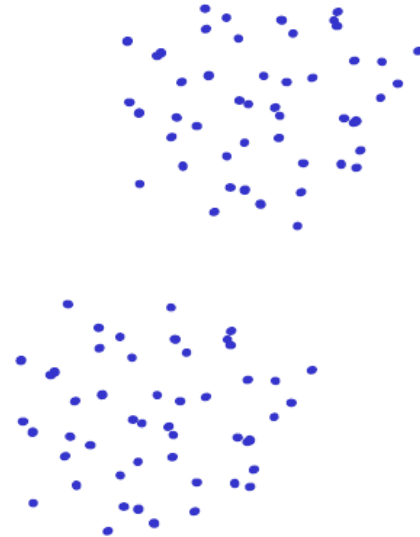
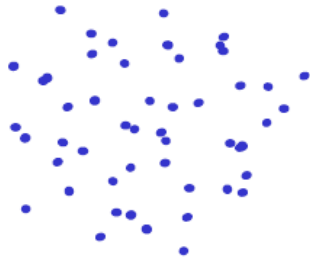
$k = 3$



3ª iteração

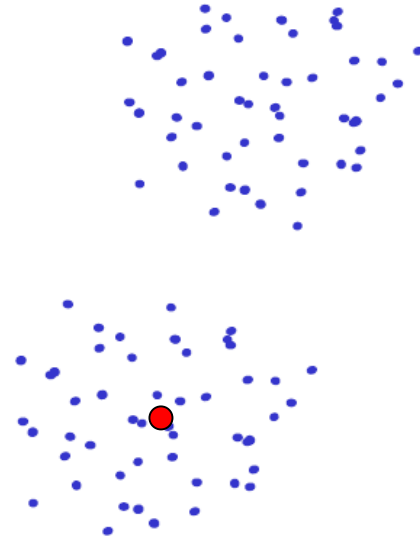
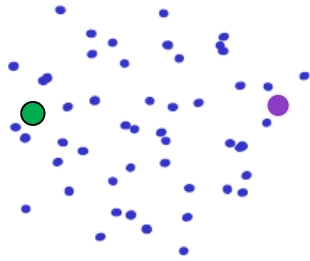
# Problemas do K-Means

- O principal problema do K-Means é a dependência de uma **boa inicialização**.



# Problemas do K-Means

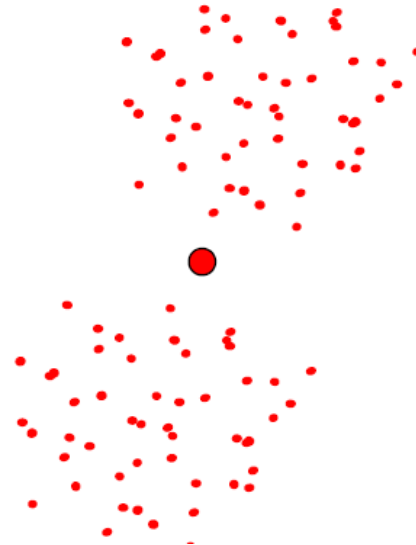
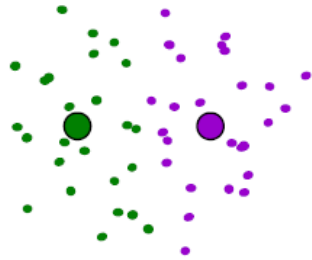
- O principal problema do K-Means é a dependência de uma **boa inicialização**.





# Problemas do K-Means

- O principal problema do K-Means é a dependência de uma **boa inicialização**.



# Aprendizado Não-Supervisionado

- O **aprendizado não-supervisionado** ou clusterização (agrupamento) busca extrair informação relevante de dados **não rotulados**.
- Existem **vários** algoritmos agrupamento de dados.
- Diferentes escolhas de atributos, medidas de proximidade, critérios de agrupamento e algoritmos de clusterização levam a resultados totalmente diferentes.

# Leitura Complementar

- Mitchell, T. **Machine Learning**, McGraw–Hill Science/Engineering/Math, 1997.
- Duda, R., Hart, P., Stork, D., **Pattern Classification**, John Wiley & Sons, 2000

