Inteligência Artificial

Aprendizagem não supervisionada
Agrupamento (*Clustering*)

Prof. Fabio Augusto Faria

Material adaptado do Prof. Edirlei Soares de Lima (PUC-RJ)

1º semestre 2015

Tópicos

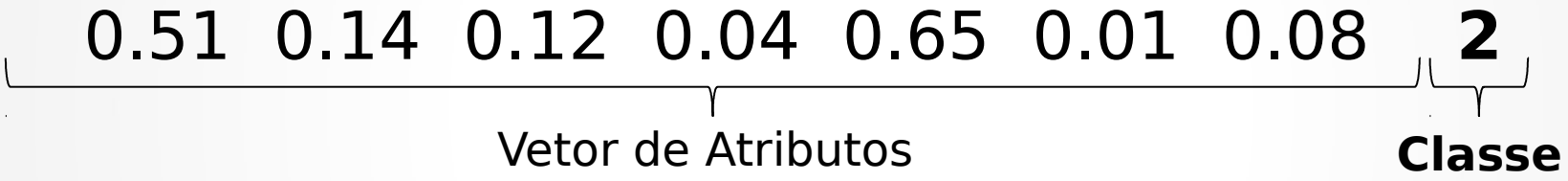
- Formas de Aprendizagem
- Tipos de algoritmos de agrupamento (*clustering*)
- Algoritmos de Agrupamento (k-means e k-medoid)
- Problemas
- Conclusões

Formas de Aprendizagem

- Aprendizagem Supervisionado
- **Aprendizagem Não-Supervisionado**
-
- Aprendizagem Por Reforço

Introdução

- No aprendizado **supervisionado**, todas os exemplos de treinamento eram **rotulados**.



- Estes exemplos são ditos “supervisionados”, pois, contém tanto a **entrada** (atributos), quanto a **saída** (classe).

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.

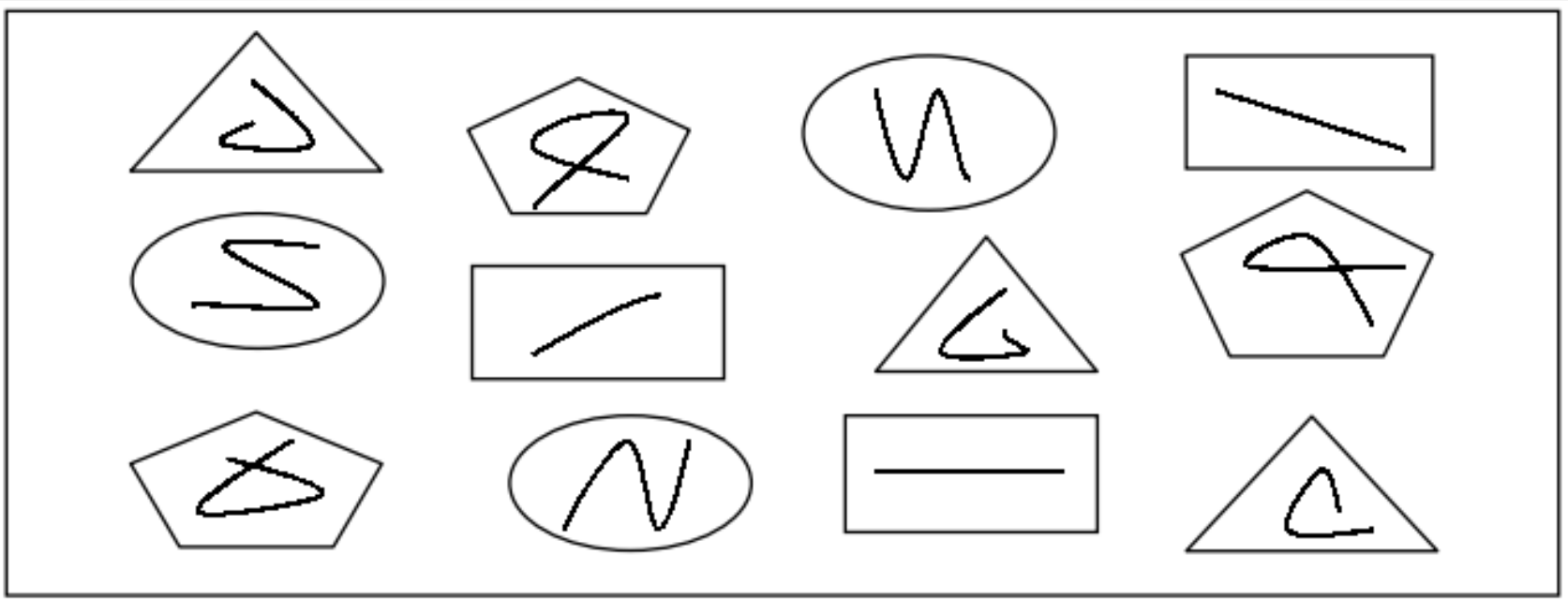
- Esta abordagem é bastante utilizada em aplicações de **mineração de dados** (*data mining*), onde o conteúdo de grandes bases de dados não é conhecido antecipadamente.

Introdução

- O principal interesse do **aprendizado não-supervisionado** é desvendar a organização dos padrões existentes nos dados através de **clusters** (agrupamentos) encontrados;

- Com isso, é possível descobrir **similaridades e diferenças** entre os padrões existentes, assim como derivar conclusões úteis a respeito deles.

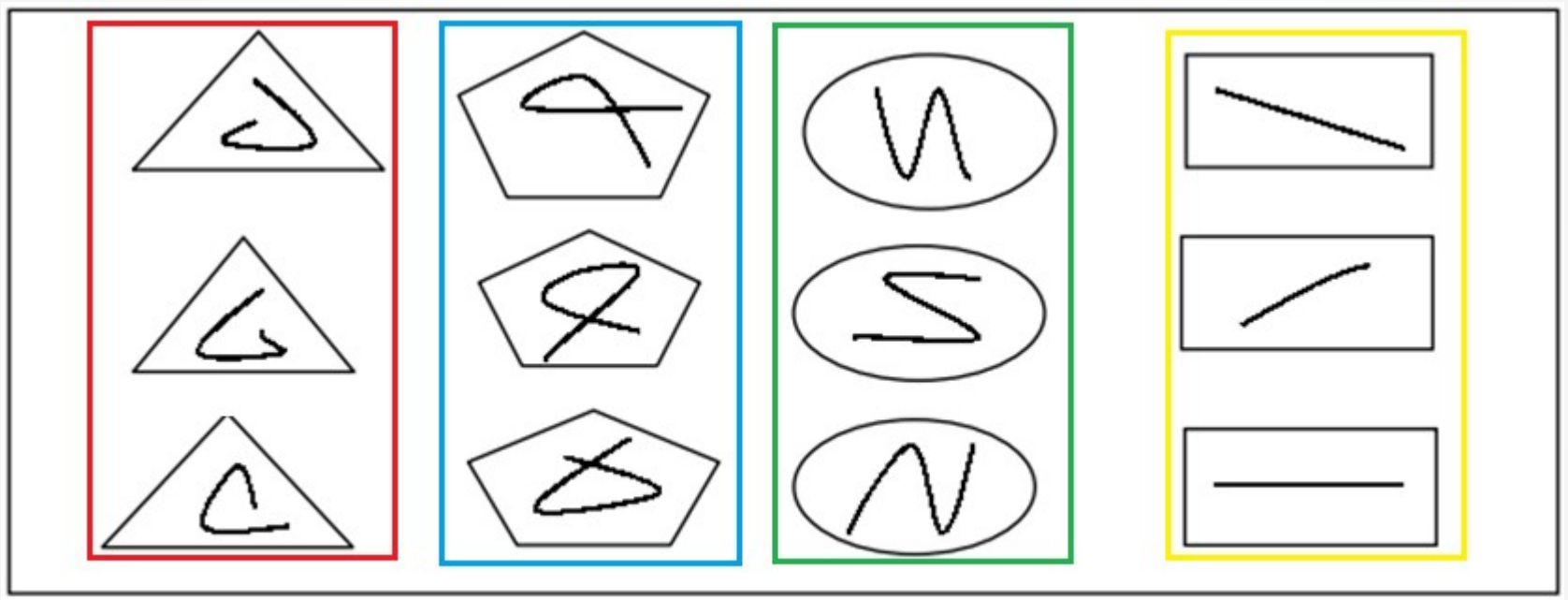
Introdução



Conjunto de objetos

Extraído de [3].

Introdução

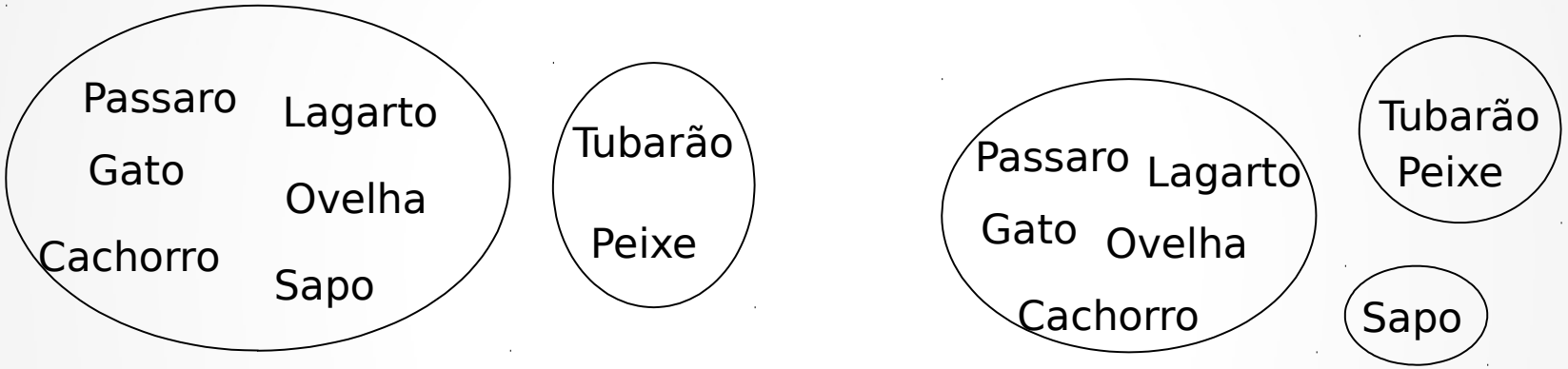


Conjunto de objetos

Extraído de [3].

Introdução

- Exemplos de agrupamentos (clusters):



Existencia de pulmões

Ambiente onde vivem

*Depende do atributo escolhido;

Critério de Similaridade

- A similaridade é difícil de ser definida...

*Depende do critério de similaridade;

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

- Diferentes técnicas de normalização[2]

Min-Max

$$n_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Z-Score

$$n_i = \frac{x_i - mean(x)}{std(x)}$$

Tanh

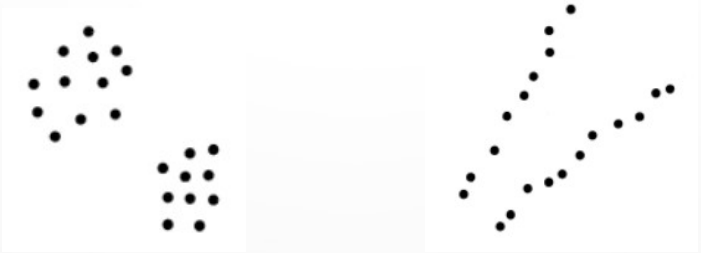
$$n_i = \frac{1}{2}\left[\tanh\left(001\frac{x_i - mean(x)}{std(x)}\right) + 1\right]$$

Soma

$$n_i = \frac{x_i}{\sum x}$$

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 agrupamento** 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 do agrupamento 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 agrupamento levam a **resultados totalmente diferentes**.

- Qual resultado é o correto?

Processo de Aprendizado Não-Supervisionado

- o que caracteriza bons e maus processos de agrupamento/clusterização?

- Para **validar** a saída produzida por um processo de clusterização, geralmente se recorre a critérios de **otimalidade**, muitas vezes definidos de forma subjetiva [1].

Tarefa de Agrupamento

- Dado um conjunto de dados X:

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

- Definimos como um *m*–agrupamento de X, a partição de *X* em *m* grupos (**clusters**) $C_1, C_2, \ldots, C_m$ tal que as três condições seguintes sejam satisfeitas:

  - Nenhum cluster pode ser vazio ($C_i \neq \emptyset$).
  - A **união de todos** os clusters 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** ($C_i \cap C_j = \emptyset$).

Agrupamento

- Os vetores contidos em um cluster C_i devem ser **mais similares** uns aos outros (intra) e **menos similares** aos vetores presentes nos outros clusters (inter).

- Tipos de Clusters:

Clusters compactos

Clusters alongados elipsoidals

Clusters esféricos e

Medidas de Proximidade

- **Medidas de Dissimilaridade*:**
  - Métrica l_p ponderada;
  - Métrica Norma l_∞ ponderada;
  - Métrica l_2 ponderada (Mahalanobis);
  - Métrica l_p especial (Manhattan);
  - Distância de Hamming;

- **Medidas de Similaridade**:**
  - Produto interno (inner);
  - Medida de Tanimoto;

* Maior o valor, menor semelhança;
** Maior o valor, maior semelhança.

Algoritmos de Agrupamento (Clustering)

- Os **algoritmos de agrupamento** buscam identificar padrões existentes em conjuntos de dados.

- Os algoritmos de agrupamento podem ser divididos em varias categorias:
  - Particionais ou Sequenciais;
  - Hierárquicos;
  - Baseados na otimização de funções custo;
  - Outros: Fuzzy, SOM, LVQ...

Algoritmos de Agrupamento (Clustering)

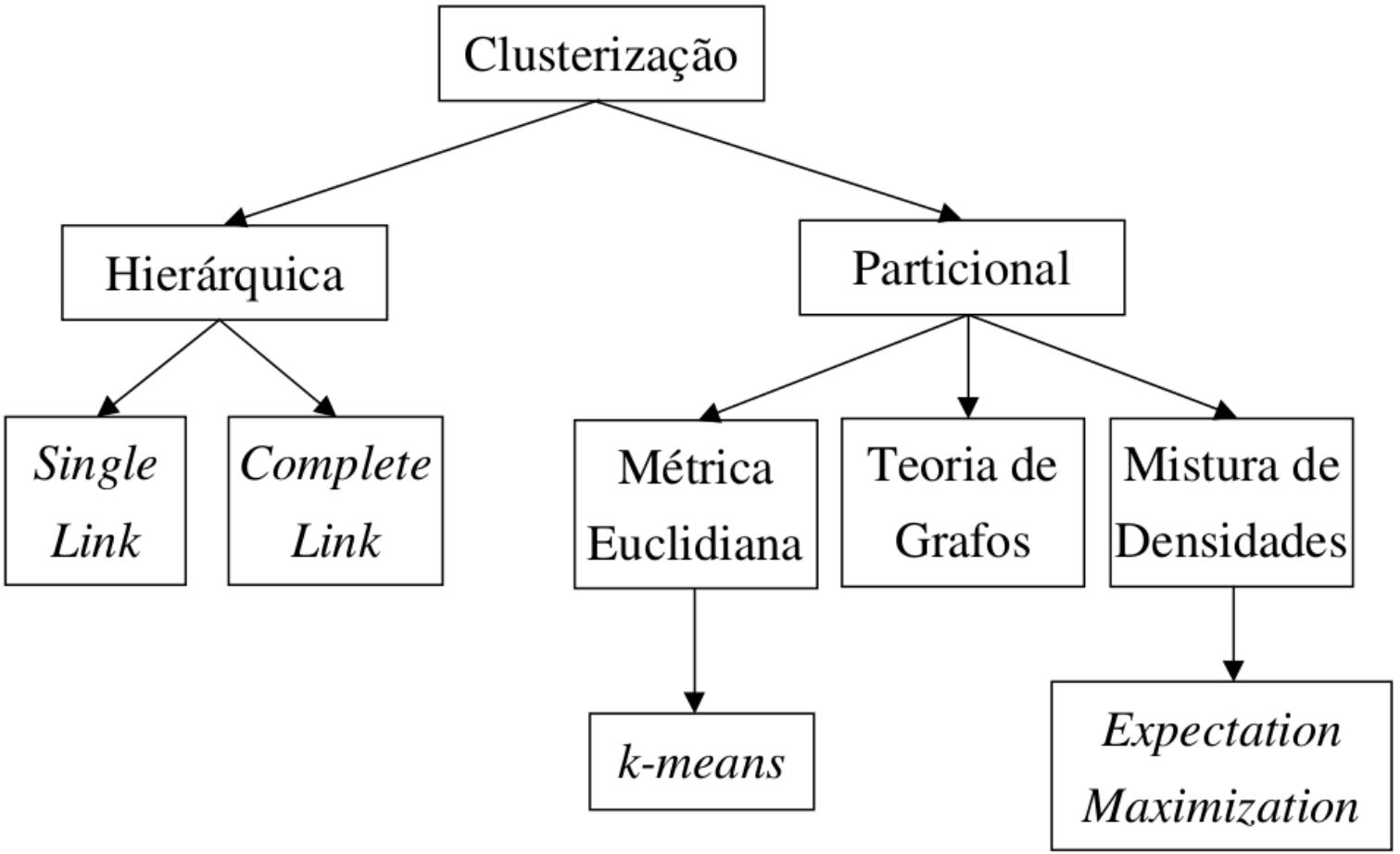


Figura 3 – Taxonomia de Técnicas de Clusterização de Dados

Extraído de [1]

Algoritmos Particionais

- 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 Particionais

- 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.

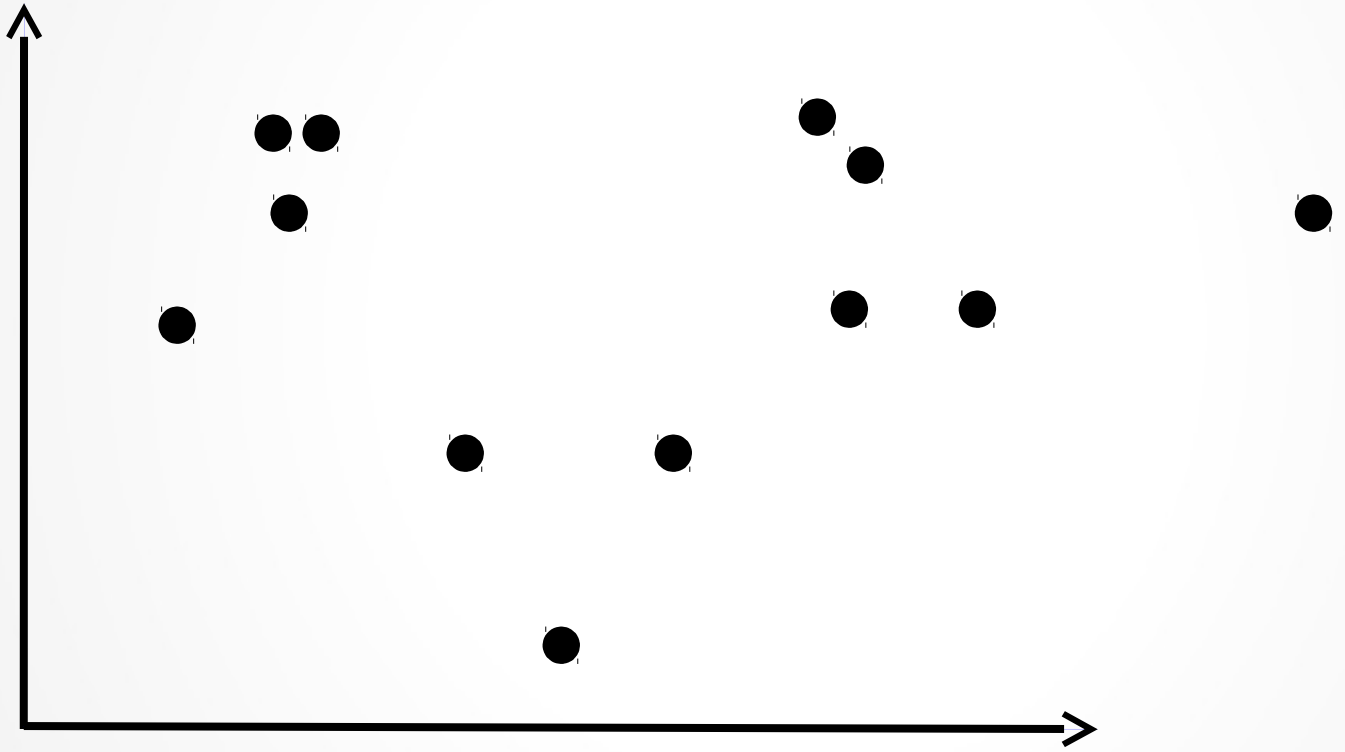
Basic Sequential Algorithmic Scheme (BSAS)

- **Parâmetros do BSAS:**
  - **d(x, C):** métrica de distância entre um vetor de características x e um cluster C.
  - **Θ:** 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).

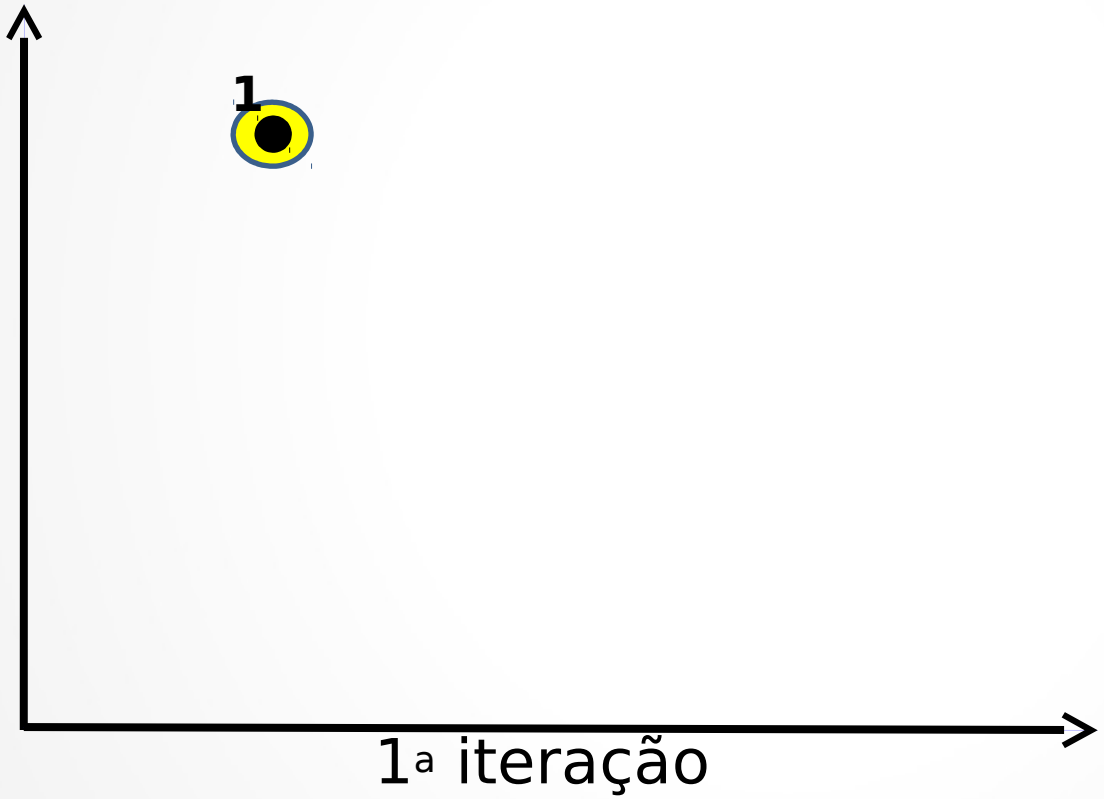
Basic Sequential Algorithmic Scheme (BSAS)

- Exemplo 1:



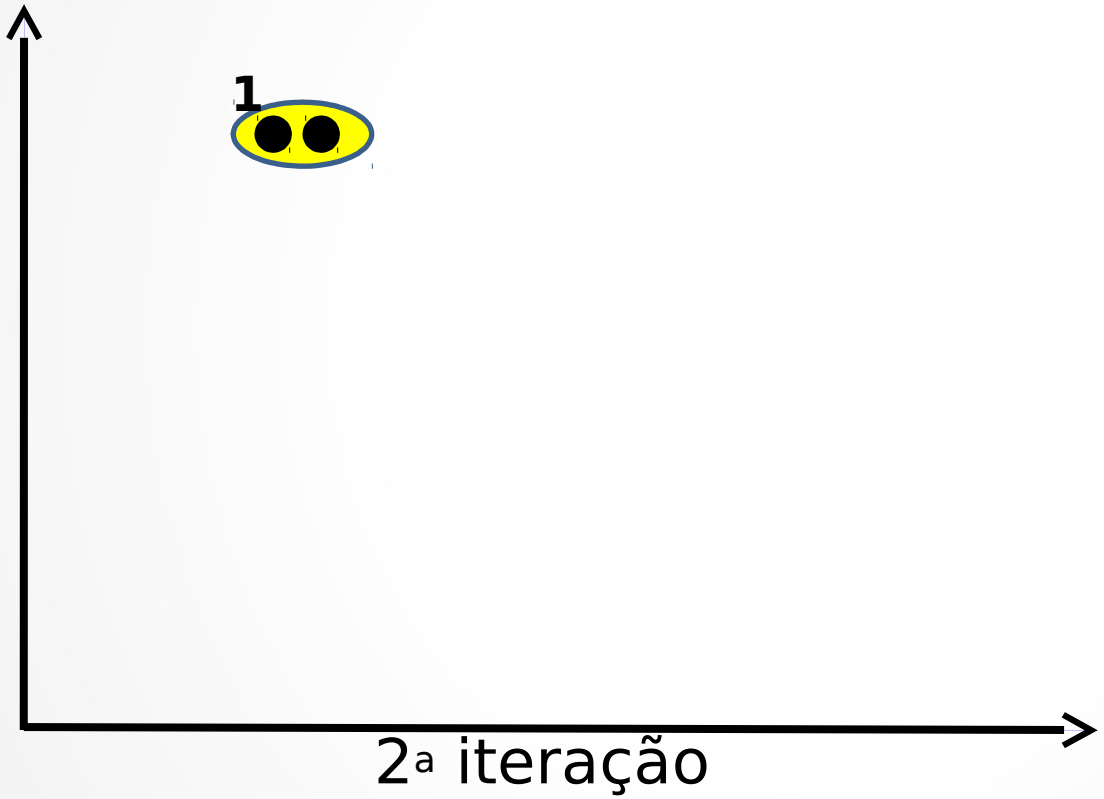
Basic Sequential Algorithmic Scheme (BSAS)

- Exemplo 1:



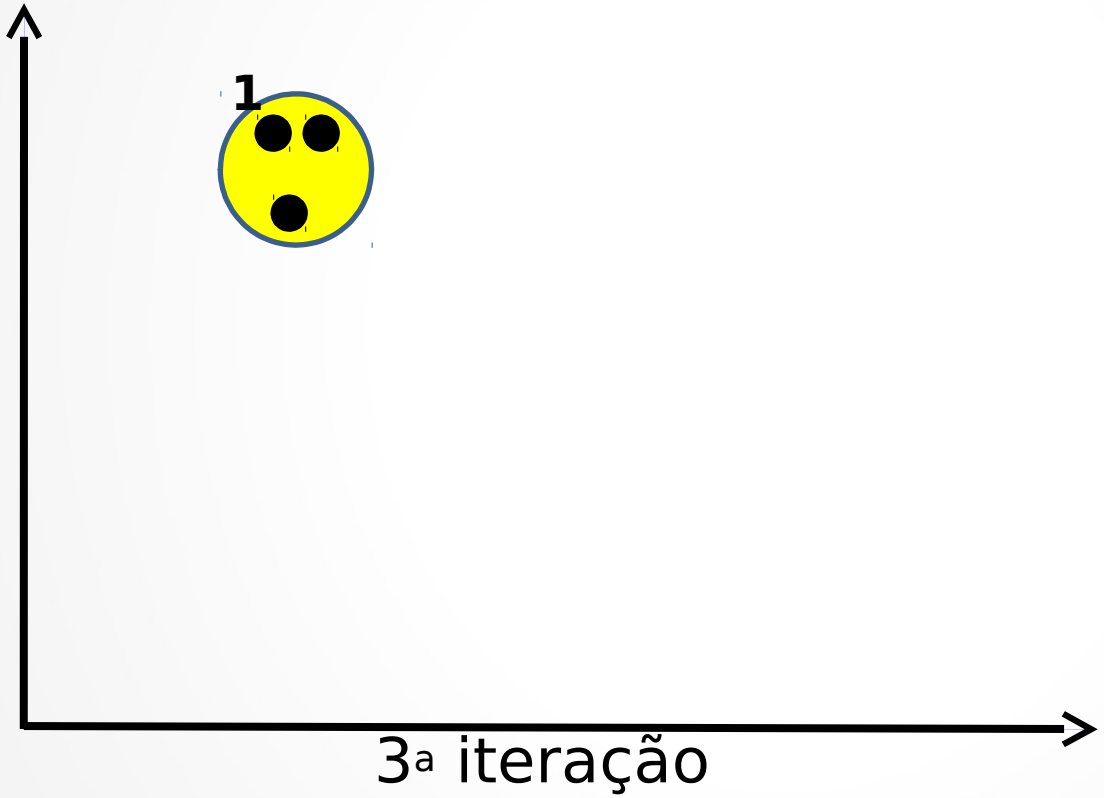
Basic Sequential Algorithmic Scheme (BSAS)

- Exemplo 1:



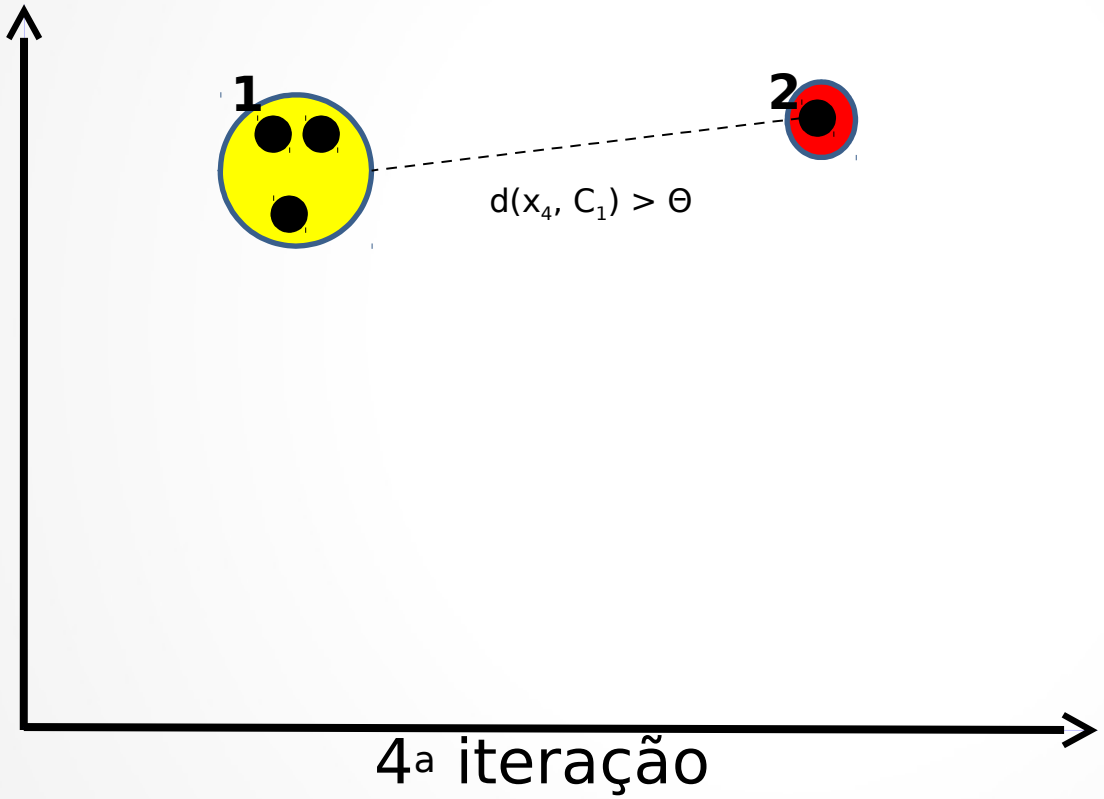
Basic Sequential Algorithmic Scheme (BSAS)

- Exemplo 1:



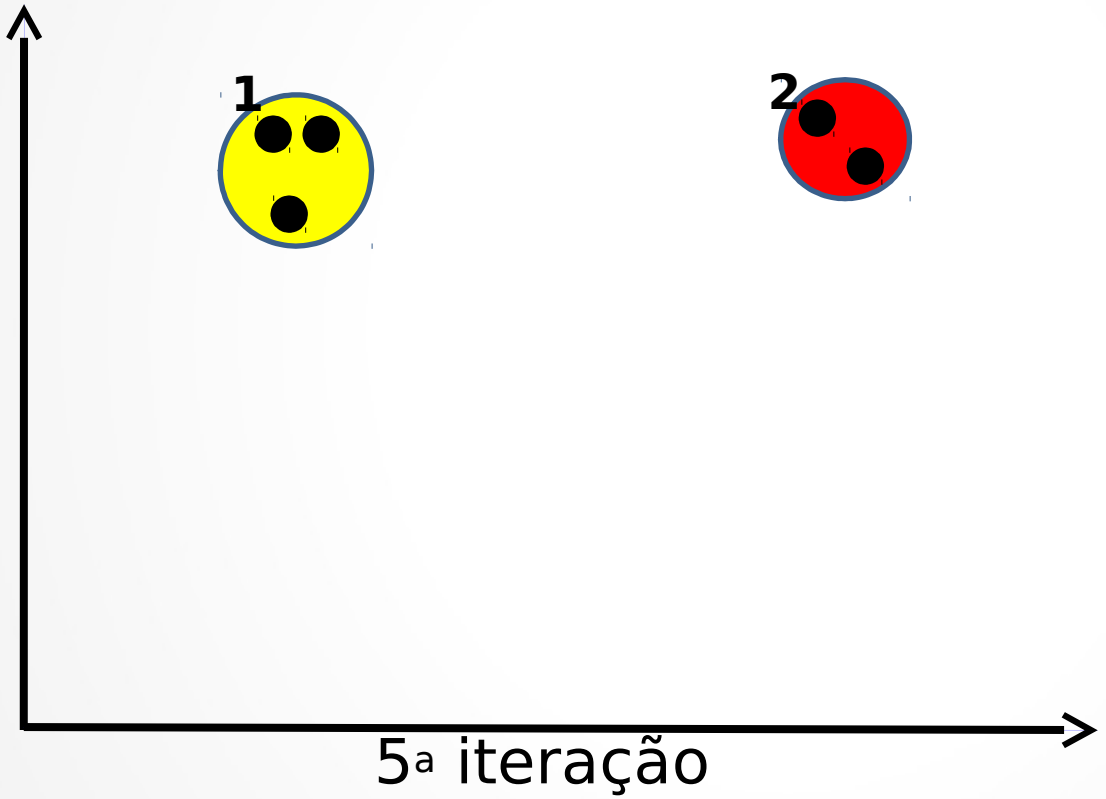
Basic Sequential Algorithmic Scheme (BSAS)

- Exemplo 1:



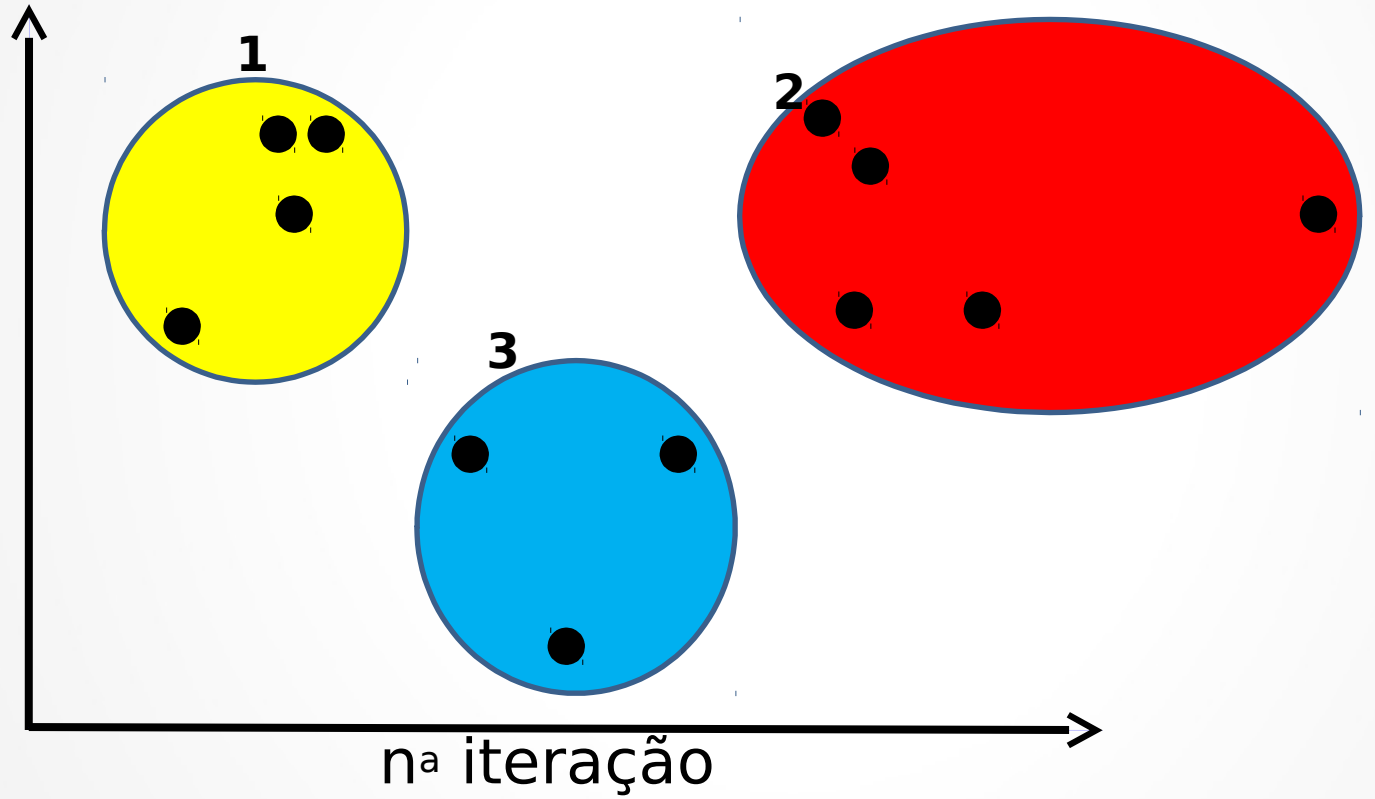
Basic Sequential Algorithmic Scheme (BSAS)

- Exemplo 1:



Basic Sequential Algorithmic Scheme (BSAS)

- Exemplo 1:



Qual pode ser um desafio nessa abordagem?

Agrupamento Hierárquica

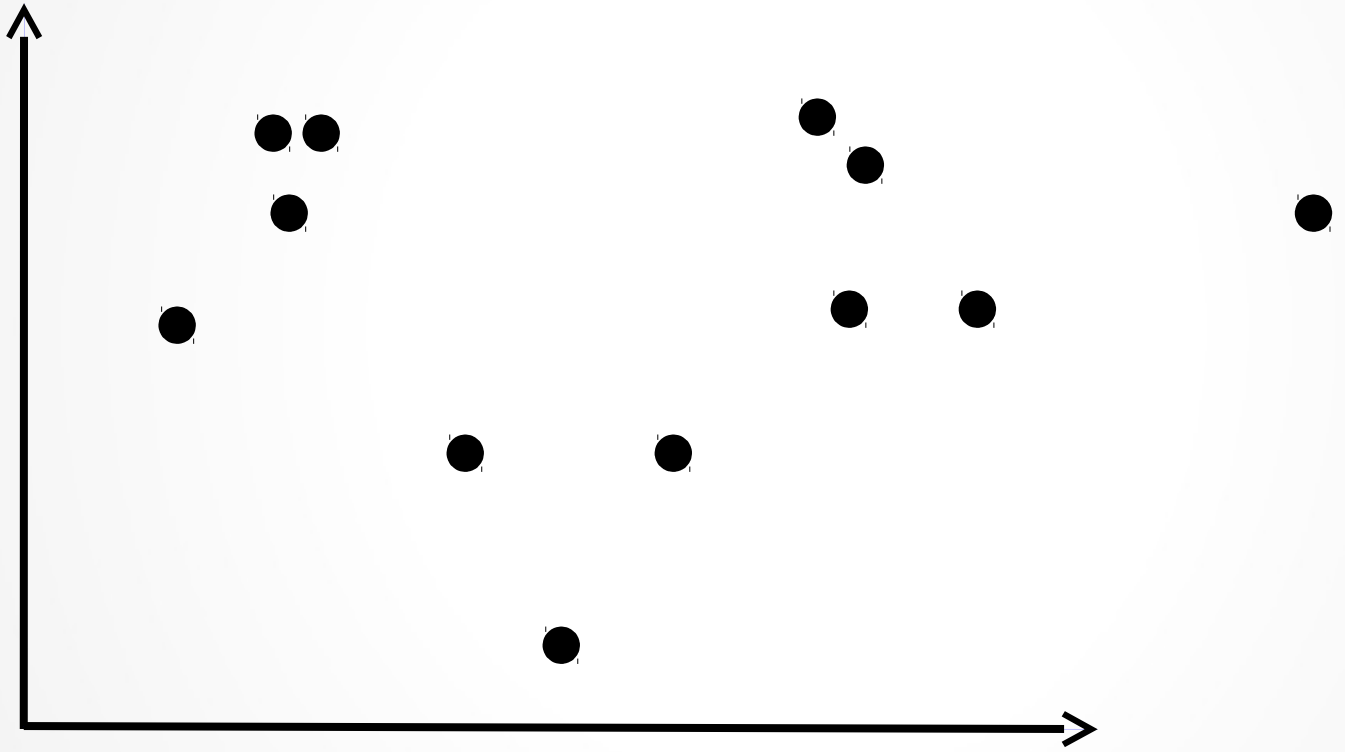
- Os algoritmos de agrupamento **hierárquico** pode ser divididos em 2 subcategorias:

- **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.

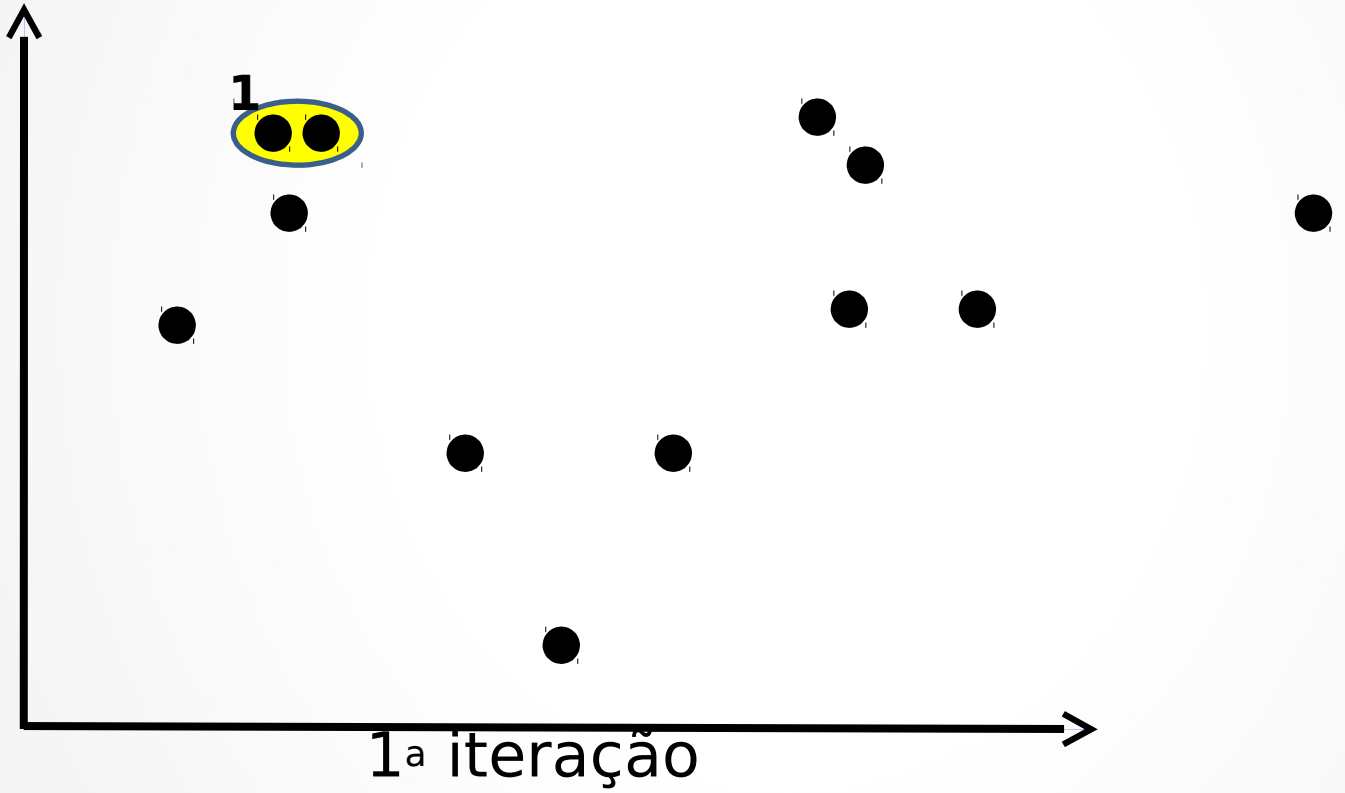
Agrupamento Hierárquica

- Exemplo 1 – Aglomerativo:



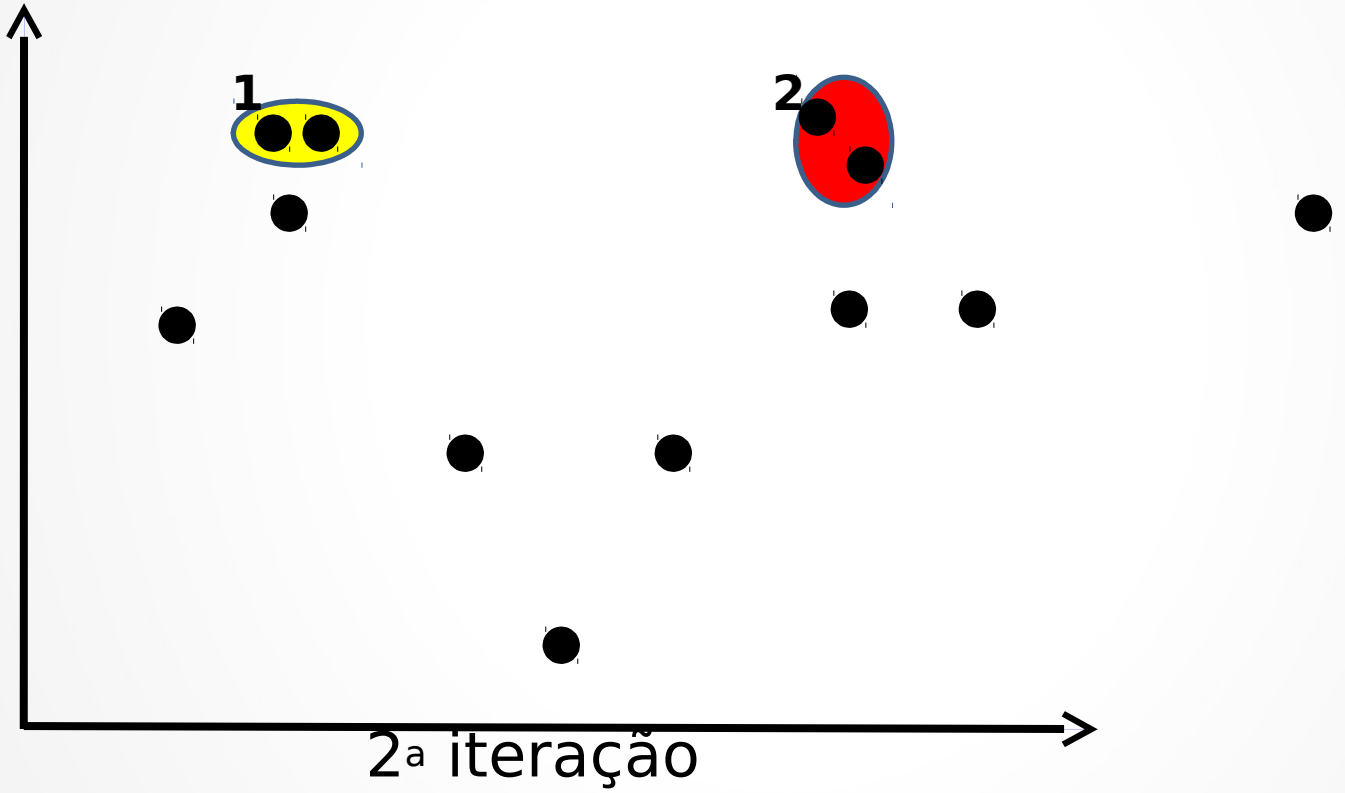
Agrupamento Hierárquica

- Exemplo 1 – Aglomerativo:



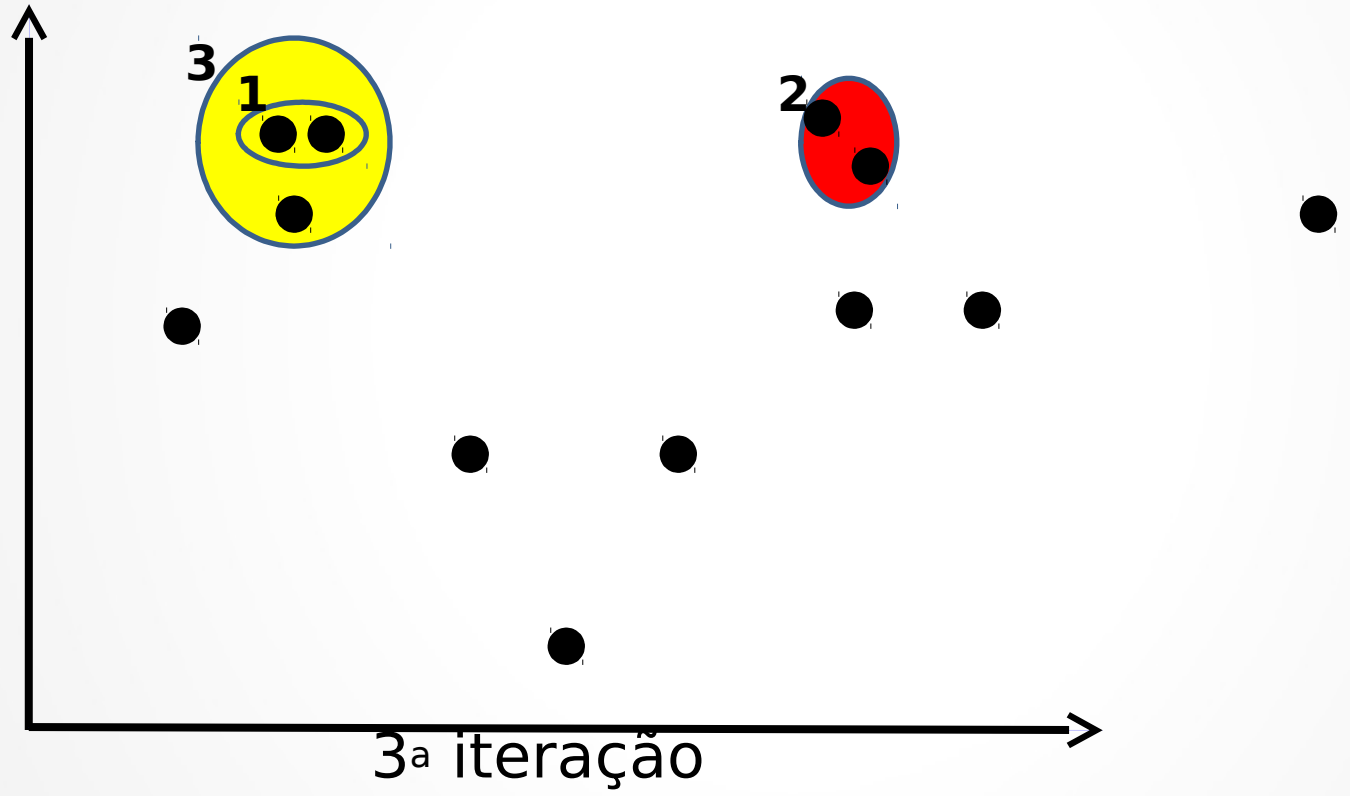
Agrupamento Hierárquica

- Exemplo 1 – Aglomerativo:



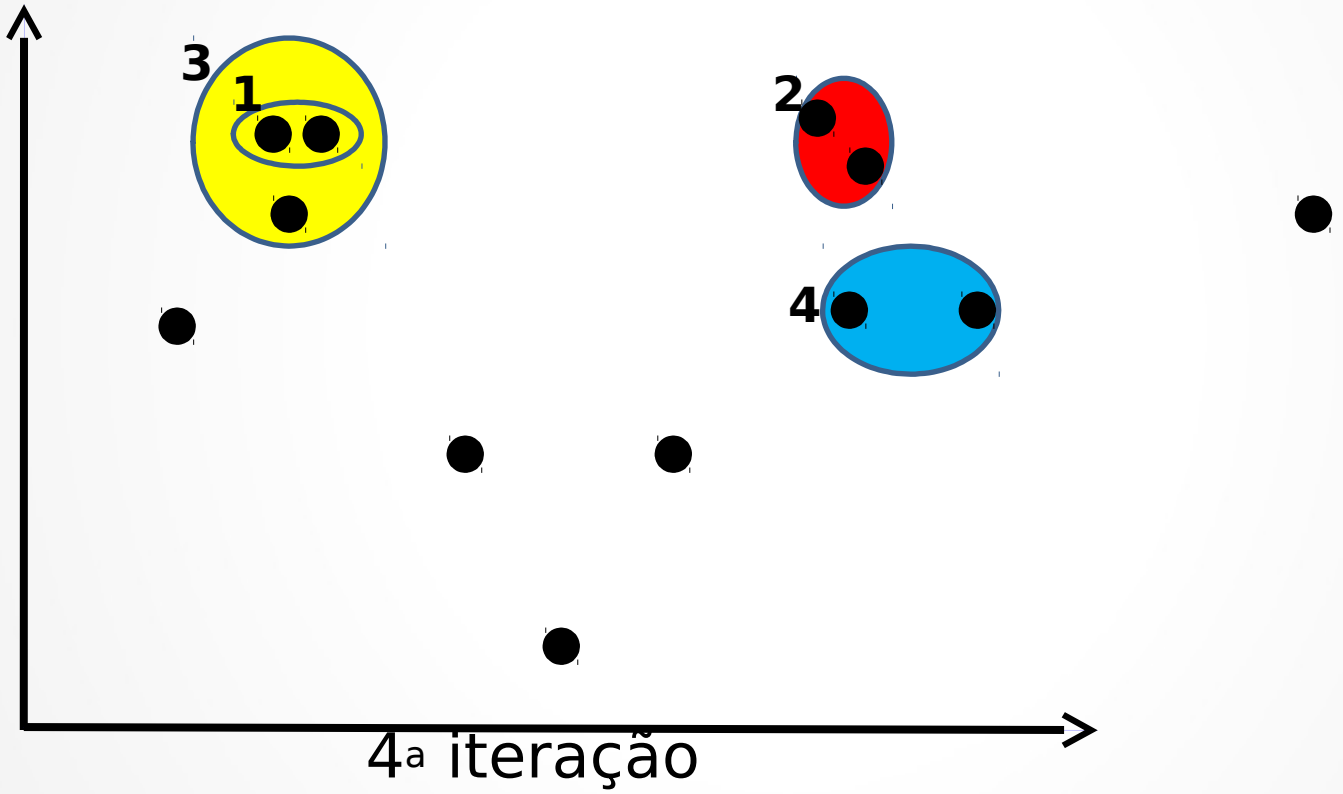
Agrupamento Hierárquica

- Exemplo 1 – Aglomerativo:



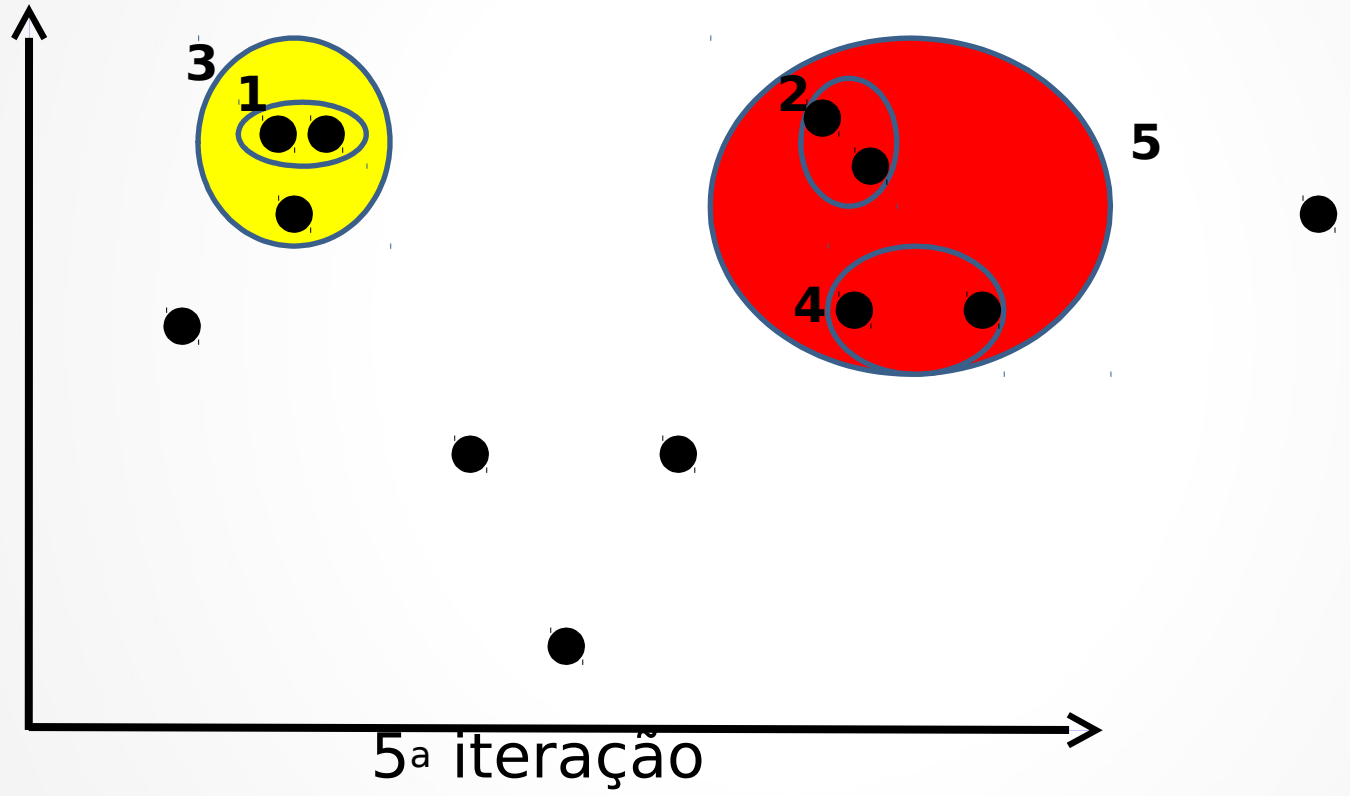
Agrupamento Hierárquica

- Exemplo 1 – Aglomerativo:



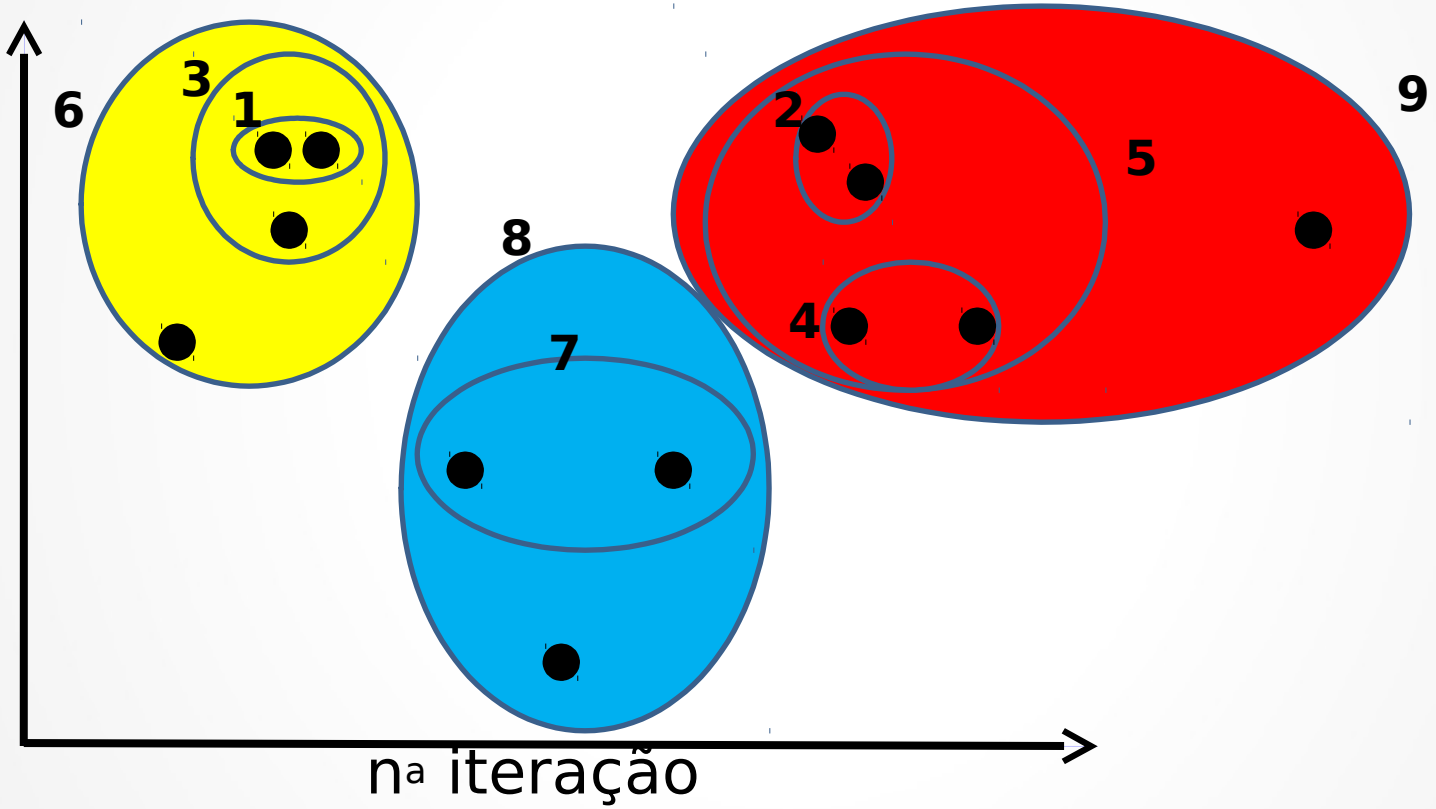
Agrupamento Hierárquica

- Exemplo 1 – Aglomerativo:



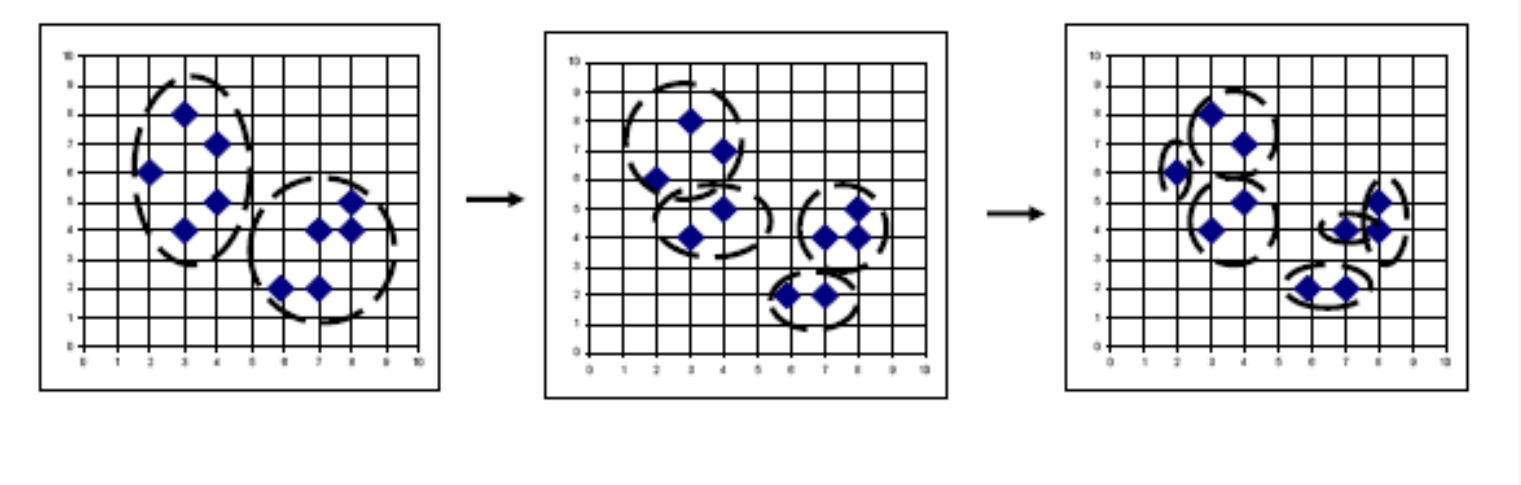
Agrupamento Hierárquica

- Exemplo 1 – Aglomerativo:



Agrupamento Hierárquico

- 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.

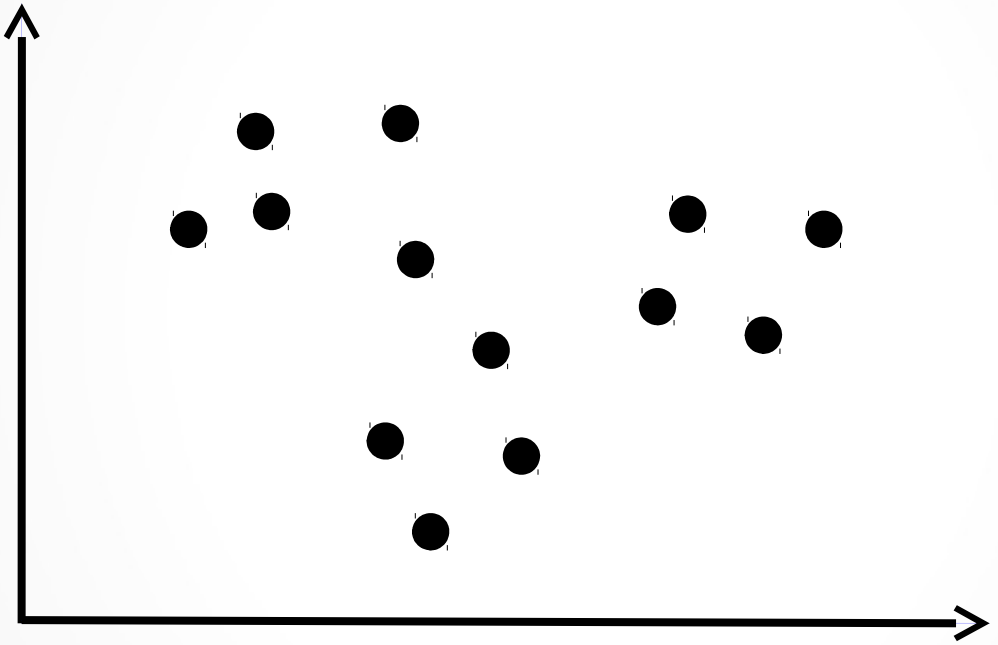
(2) Forme k clusters associando cada exemplo ao seu centróide mais próximo.

(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.

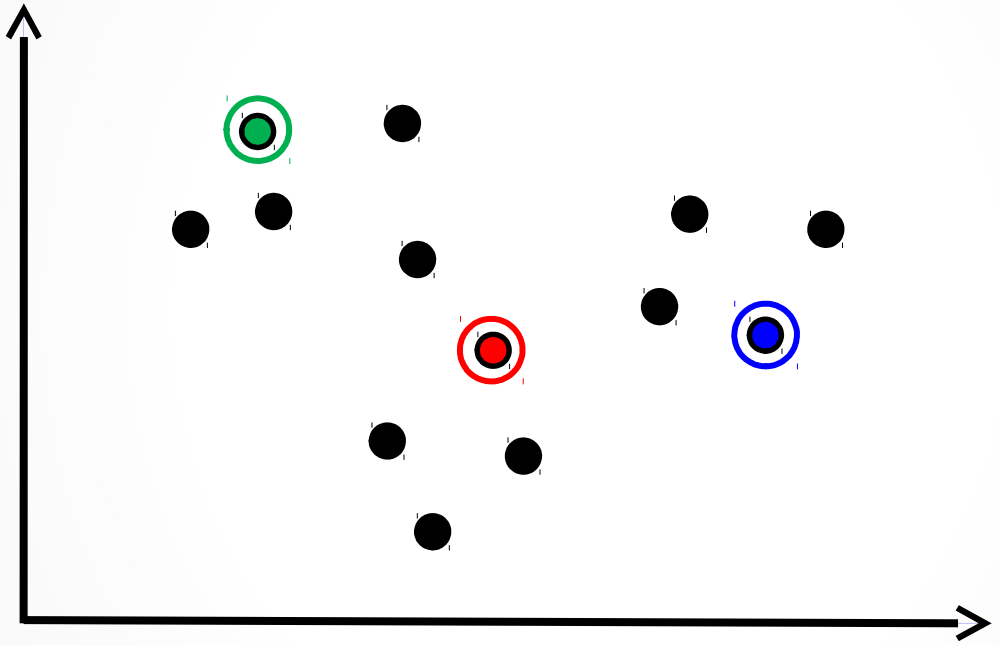
Algoritmo K-Means

- Exemplo:



Algoritmo K-Means

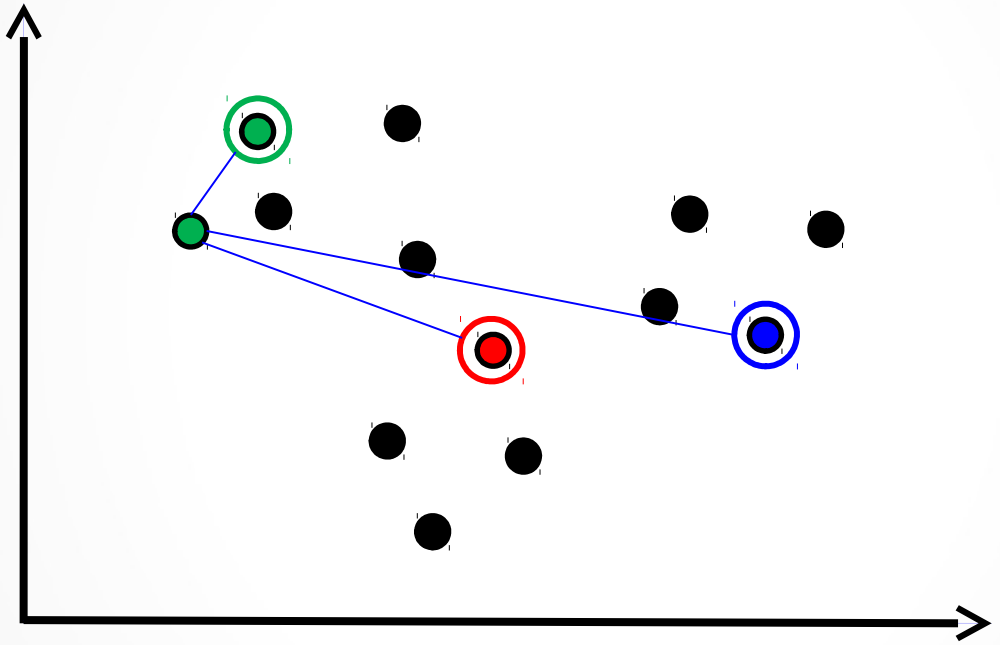
- Exemplo: k = 3



Seleciona-se k centróides iniciais.

Algoritmo K-Means

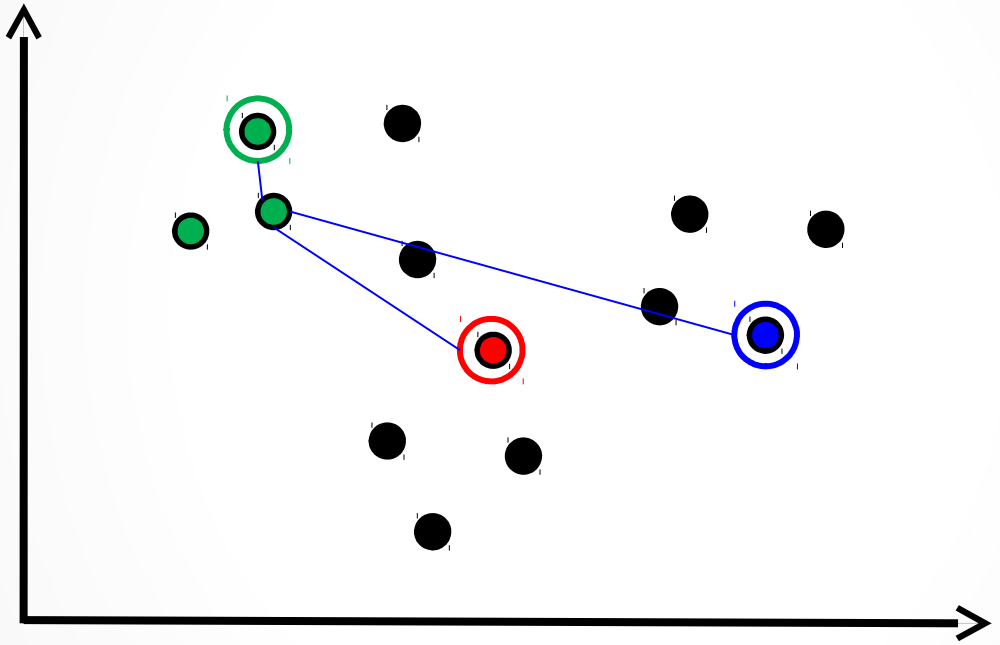
- Exemplo: k = 3



1ª iteração

Algoritmo K-Means

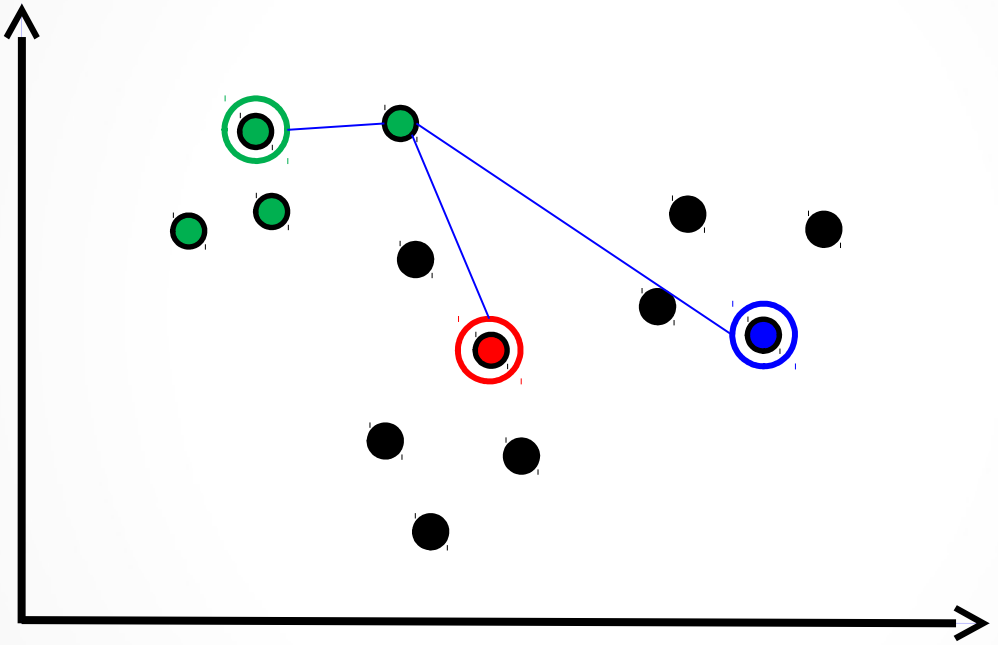
- Exemplo: k = 3



2ª iteração

Algoritmo K-Means

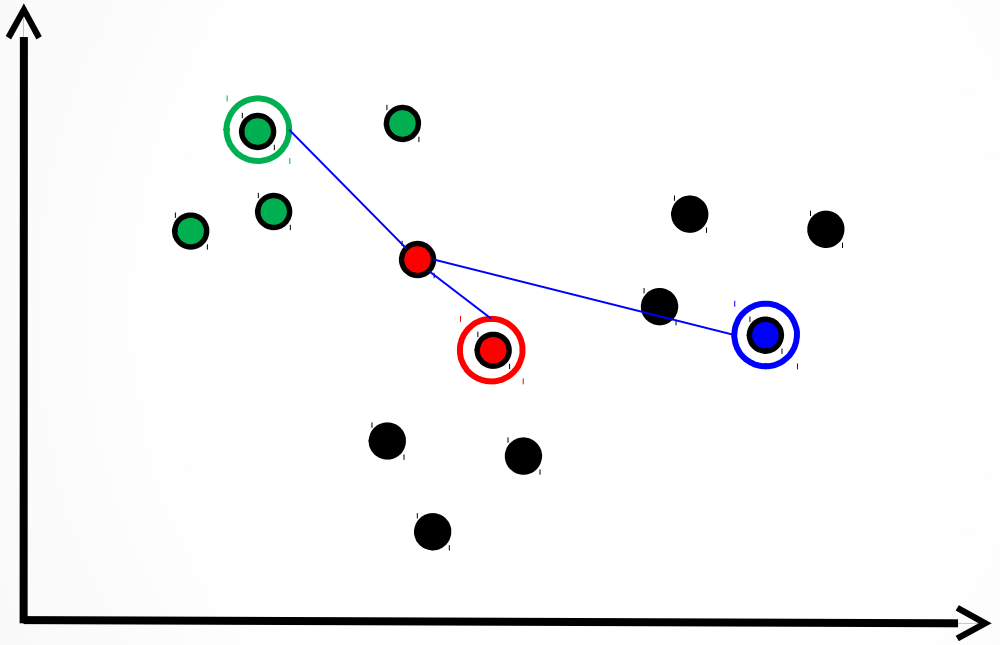
- Exemplo: k = 3



3ª iteração

Algoritmo K-Means

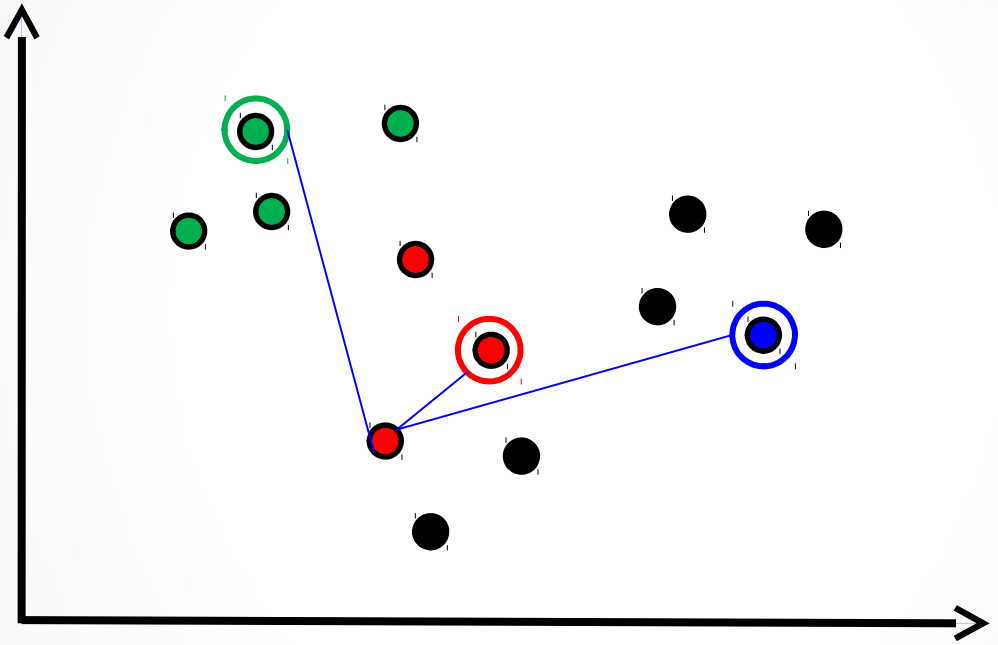
- Exemplo: k = 3



4ª iteração

Algoritmo K-Means

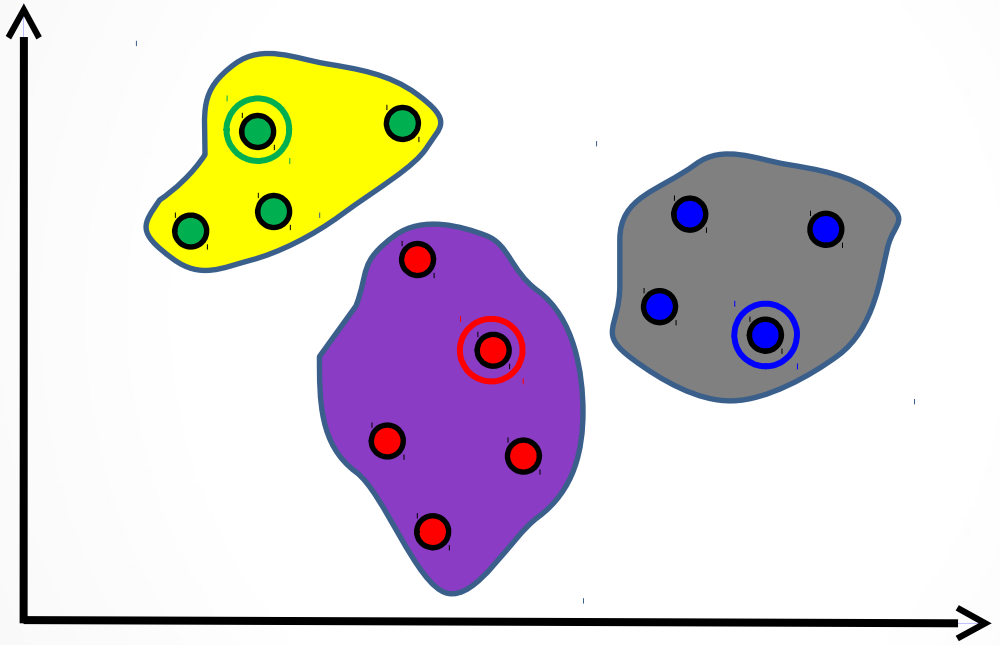
- Exemplo: k = 3



5ª iteração

Algoritmo K-Means

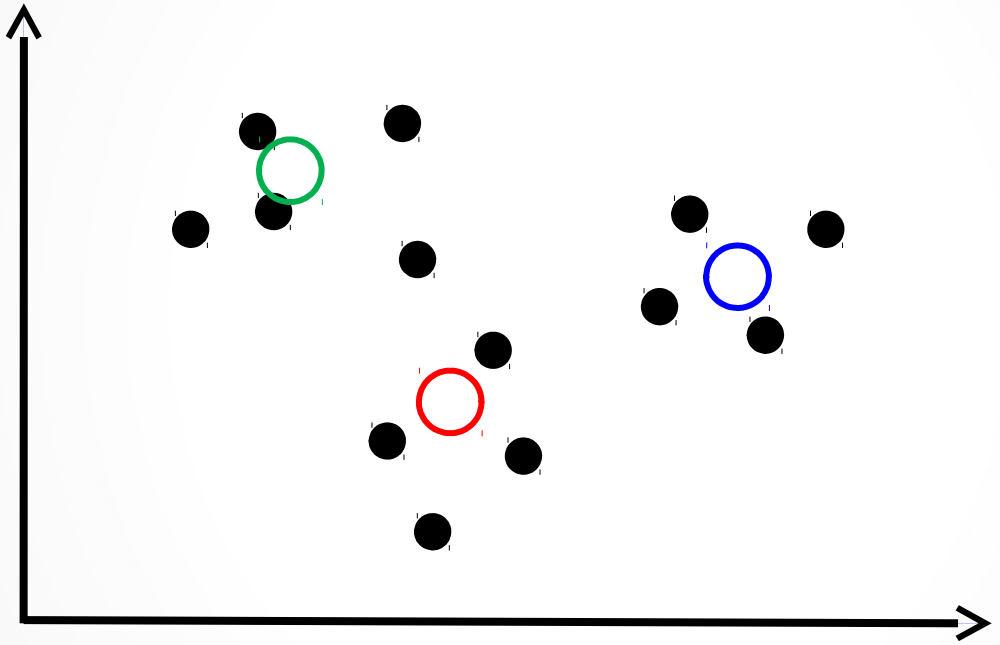
- Exemplo: k = 3



nª iteração

Algoritmo K-Means

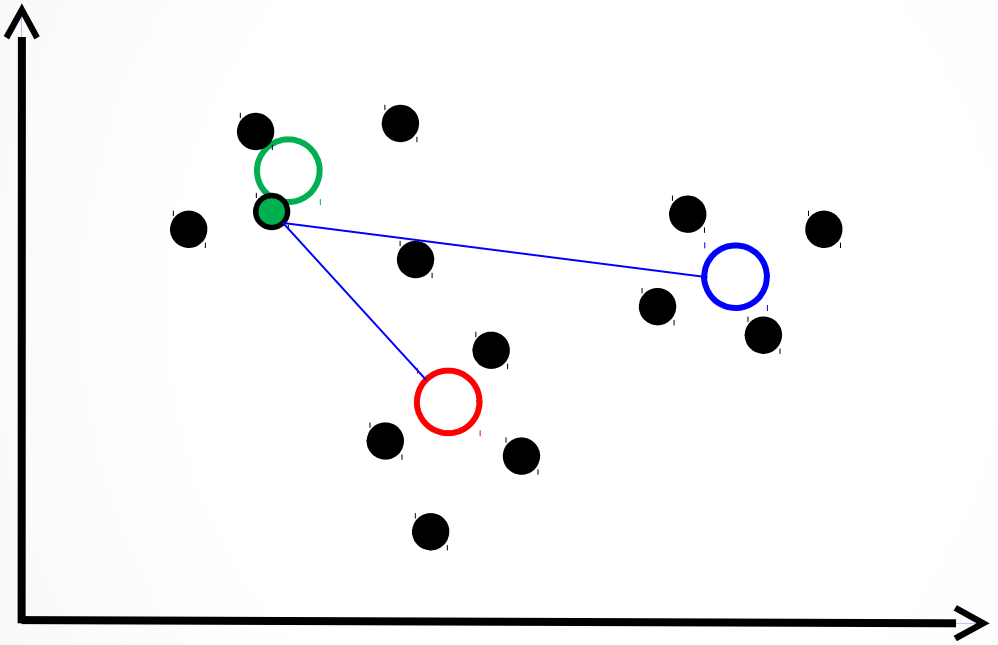
- Exemplo: k = 3



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

Algoritmo K-Means

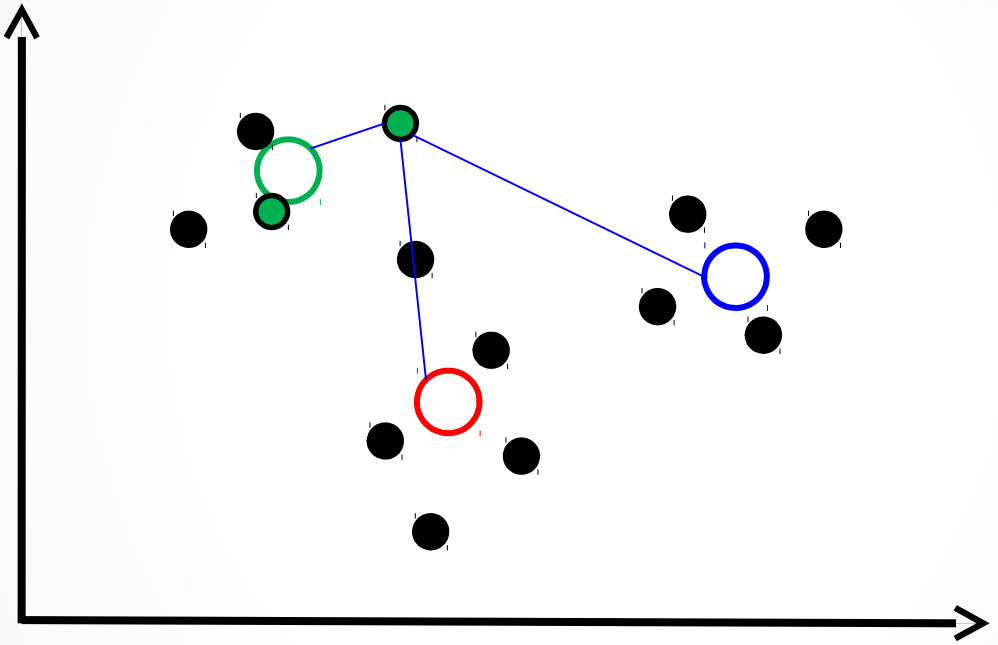
- Exemplo: k = 3



1ª iteração

Algoritmo K-Means

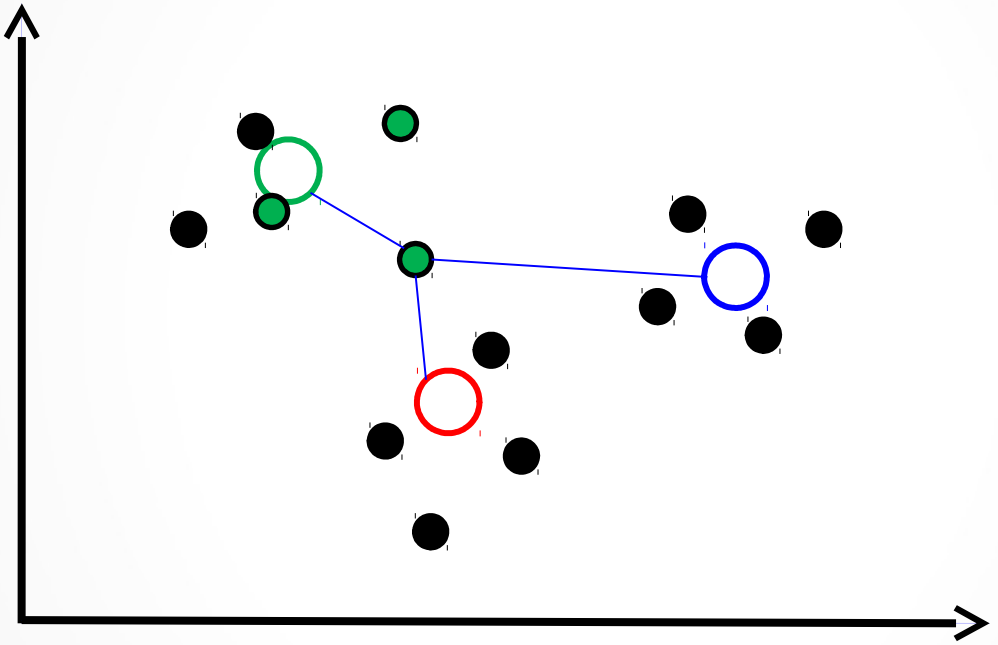
- Exemplo: k = 3



2ª iteração

Algoritmo K-Means

- Exemplo: k = 3



3ª iteração

Problemas do K-Means

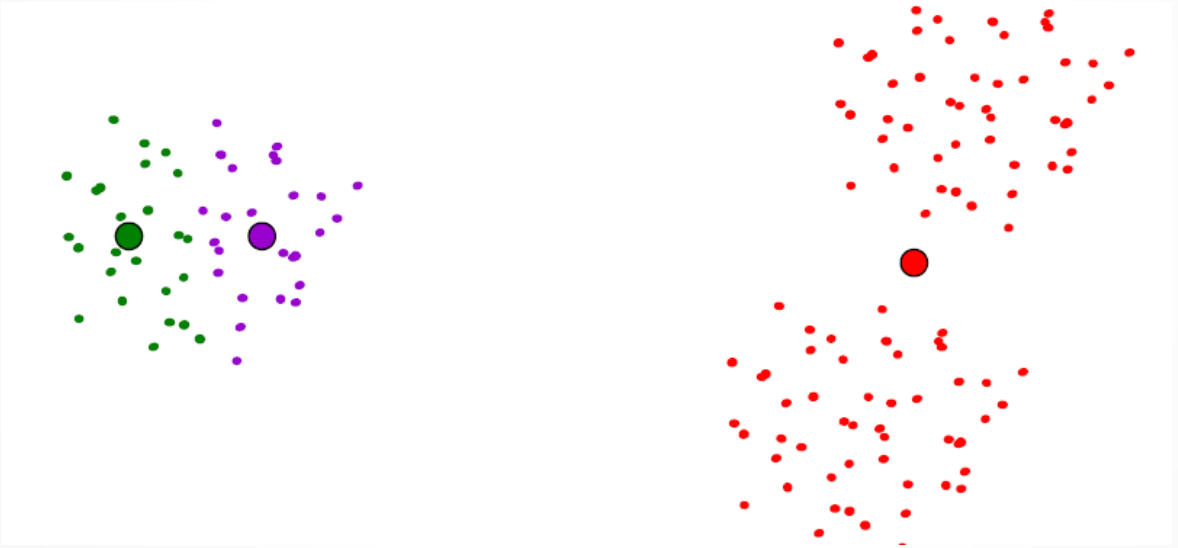
- O principal problema do K-Means é a dependência de uma **boa inicialização**.

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**.



Critérios de Otimização [2]

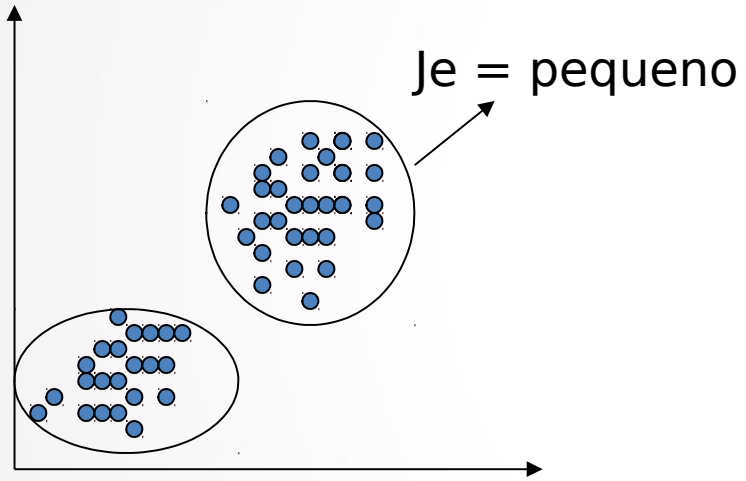
- O problema consiste em encontrar os *clusters* que minimizam/maximizam um dado critério.

- Alguns critérios de otimização:
  - Soma dos Erros Quadrados.
  - Critérios de Dispersão

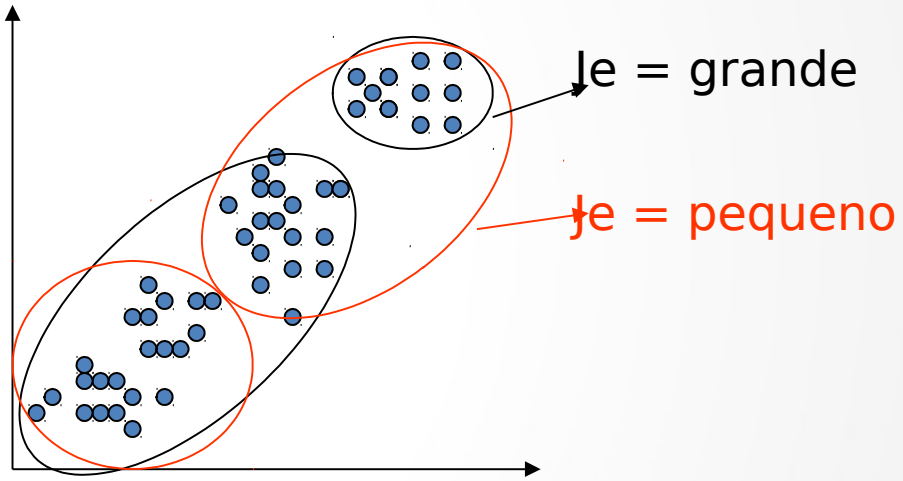
Soma dos Erros Quadrados

- É o mais simples e usado critério de otimização em *clustering.*
- Seja n_i o número de exemplos no cluster D_i e seja $\mathbf{m_i}$ a média desse exemplos $$m_i = \frac{1}{n_i} \sum_{x \in D_i} x$$
- A soma dos erros quadrados é definida $$J_e = \sum_{i=1}^{c} \sum_{x \in D_i} \left\| x - m_i \right\|^2$$

Soma dos Erros Quadrados



Adequado nesses casos
- Separação natural



Não é muito adequado para dados mais dispersos.
Outliers podem afetar bastante os vetores médios **m**

Critérios de Dispersão

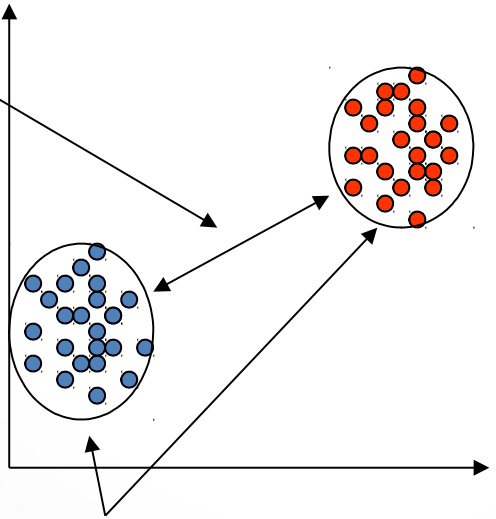
- Vetor médio do cluster i $m_i = \frac{1}{n_i} \sum_{x \in D_i} x$
- Vetor médio total $\mathrm{m} = \frac{1}{n} \sum_{D} x$
- Dispersão do cluster i $S_i = \sum_{x \in D_i} (x - m_i)(x - m_i)^t$
- Within-cluster $S_w = \sum_{i=1}^{c} S_i$
- Between-cluster $S_B = \sum_{i=1}^{c} n_i (m_i - \mathrm{m})(m_i - \mathrm{m})^t$

Critérios de Dispersão

- Relação Within-Between

Caso ideal

Alto between (S_b)
Clusters distantes um do outro.

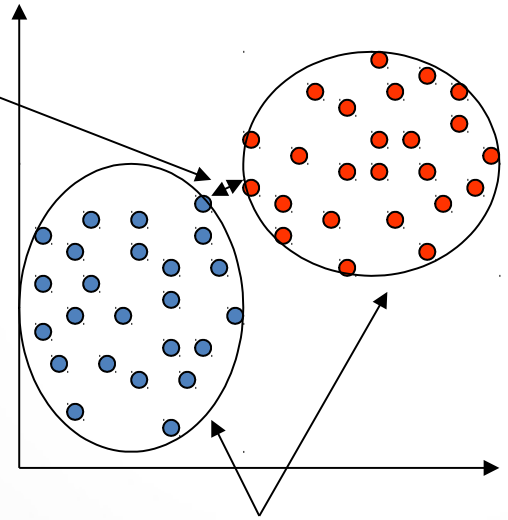


Baixo within (S_w)
(boa compactação)

Critérios de Dispersão

Caso não ideal

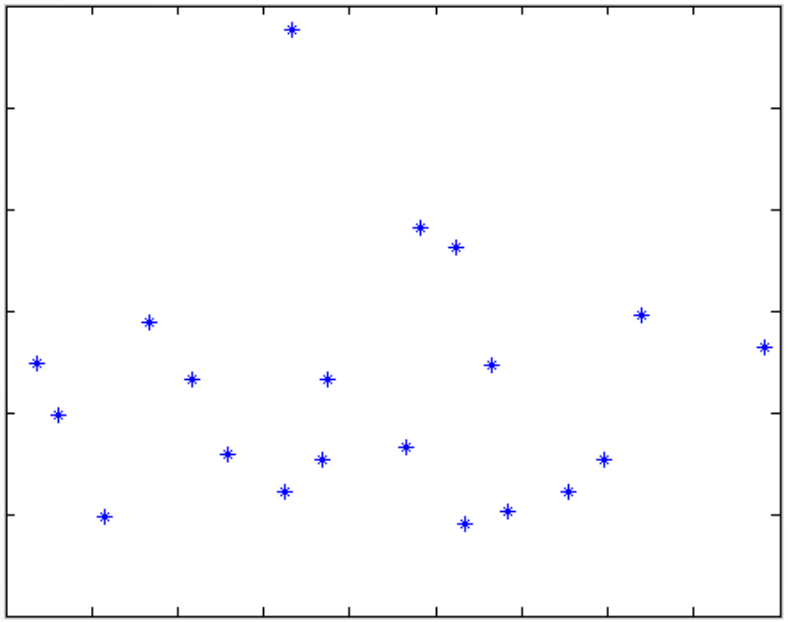
Baixo between (S_b)
Baixa distância entre os clusters.



Clusters dispersos
Alto within

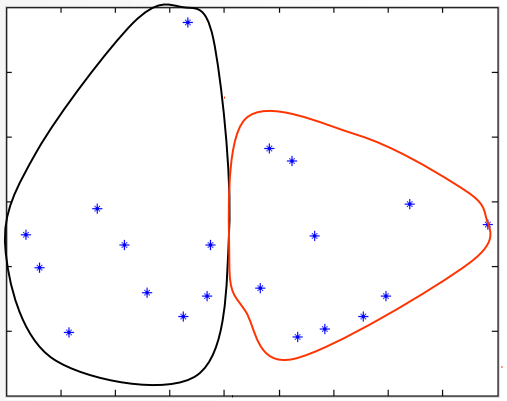
Critérios de Dispersão

- Podemos entender melhor os critérios de dispersão analisando o seguinte exemplo:

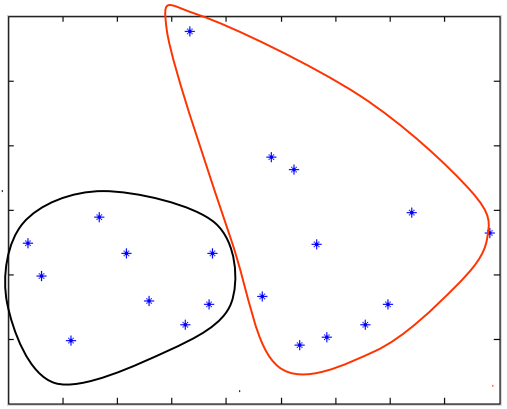


Diferentes clusters para c=2 usando diferentes critérios de otimização

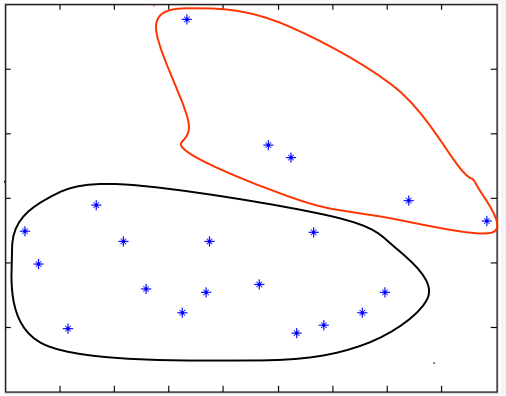
Erro Quadrado



S_w

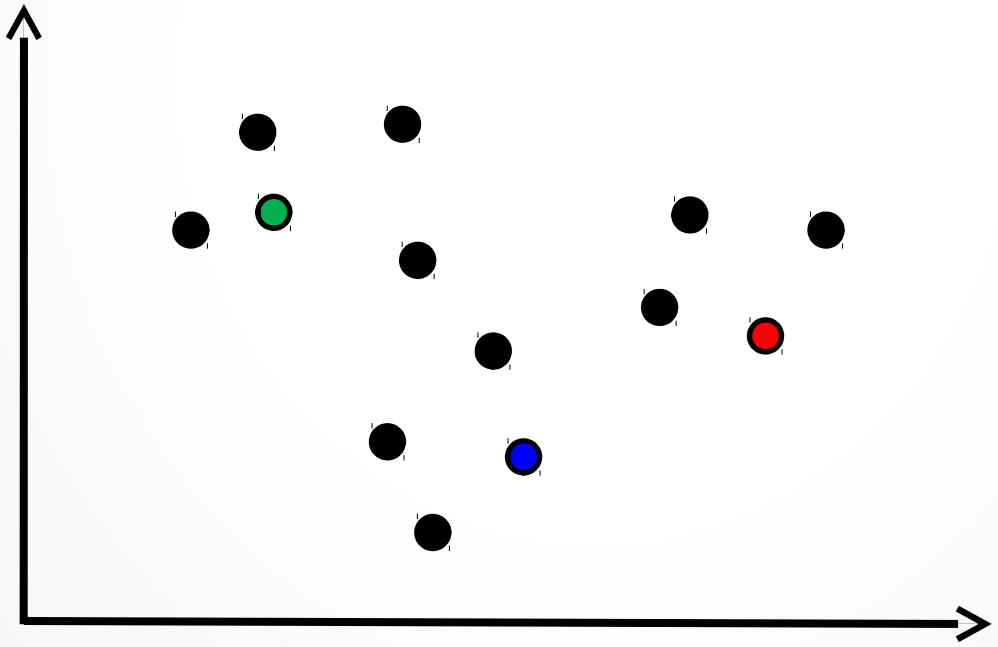


Relação S_w/S_b



Algoritmo K-Medoids

- Diferença para o k-means é que o **representante** do grupo é uma **instância** do próprio grupo e não mais um **centróide** (ponto médio);



Características Desejáveis

- descobrir clusters com forma arbitrária;
- identificar clusters de tamanhos variados;

trabalhar com objetos com qualquer número de atributos (dimensões)[4];

Características Desejáveis

- ser escalável para lidar com qualquer quantidade de objetos;
- exigir o mínimo de conhecimento para determinar os parâmetros de entrada;
- encontrar o número adequado de clusters[4].[3]

Considerações Finais

- O **aprendizado não-supervisionado** ou agrupamento (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 agrupamento levam a resultados totalmente diferentes.

Considerações Finais

- O problema de clusterização é NP-Completo;
- Para um conjunto com 10 elementos:
  - com 2 clusters são 511 grupos possíveis;
  - na clusterização automática serão 115.975[4].

Considerações Finais

- Na área de negócios, Clustering pode ajudar a descobrir grupos distintos nas bases de clientes;

- E caracterizar os grupos de clientes baseado nos padrões de compras[4].

Considerações Finais

- Etapa de pré-processamento para outros algoritmos, tais como caracterização e classificação, que iriam então operar nos clusters detectados[4].

Referência

1- Análise de Dados em Bioinformática – Profs. Moscato & Von Zuben DCA/FEEC/Unicamp.
2- Aula de *Clustering* de *pixels* por Kmeans. A. Falcão & D. Menotti (UNICAMP e UFOP)
3- Aula de Análise de Agrupamento. C. A. A. Varella (UFRRJ).
4- Aula de Clustering. S. Tinôco (UFOP).