Aprendizado de Máquina (machine learning) parte II

por: Rafael Stoffalette João

Data: 14/03/2020

UNIversidade Paulista (UNIP) – Araçatuba

Materiais em: encurtador.com.br/bePY9

Na nossa última aula

O que é um padrão;

O que é ML;

Pré-processamento dos dados;

Ferramenta Weka;

- Construção de base de dados para o weka.

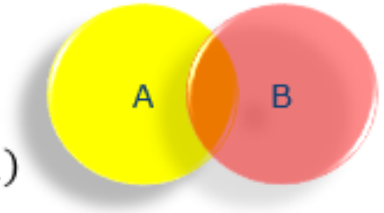
Aprendizado Supervisionado – Tarefa de classificação;

- Algoritmo probabilístico (Naive Bayes);
- Árvores de decisão.

Na nossa última aula

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$



Likelihood of evidence B if A is true

Prior probability

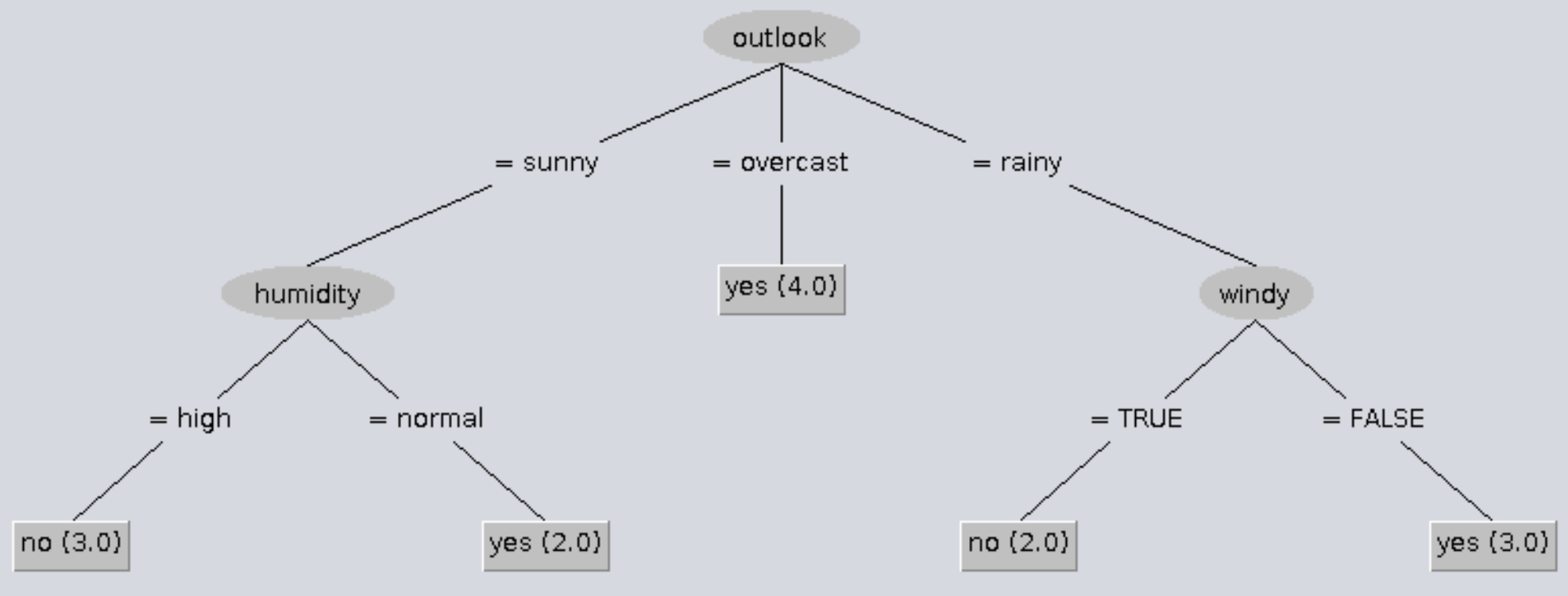
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Posterior probability of A given the evidence B

Prior probability that evidence B is true

P(SIM | sol) = P(sol | SIM) * P(SIM) / P(sol)

Na nossa última aula



Entropia:

$$\mathrm{H}(S) = \sum_{x \in X} -p(x) \log_2 p(x)$$

Ganho de informação:

$$IG(S, A) = \mathrm{H}(S) - \sum_{t \in T} p(t)\mathrm{H}(t)$$

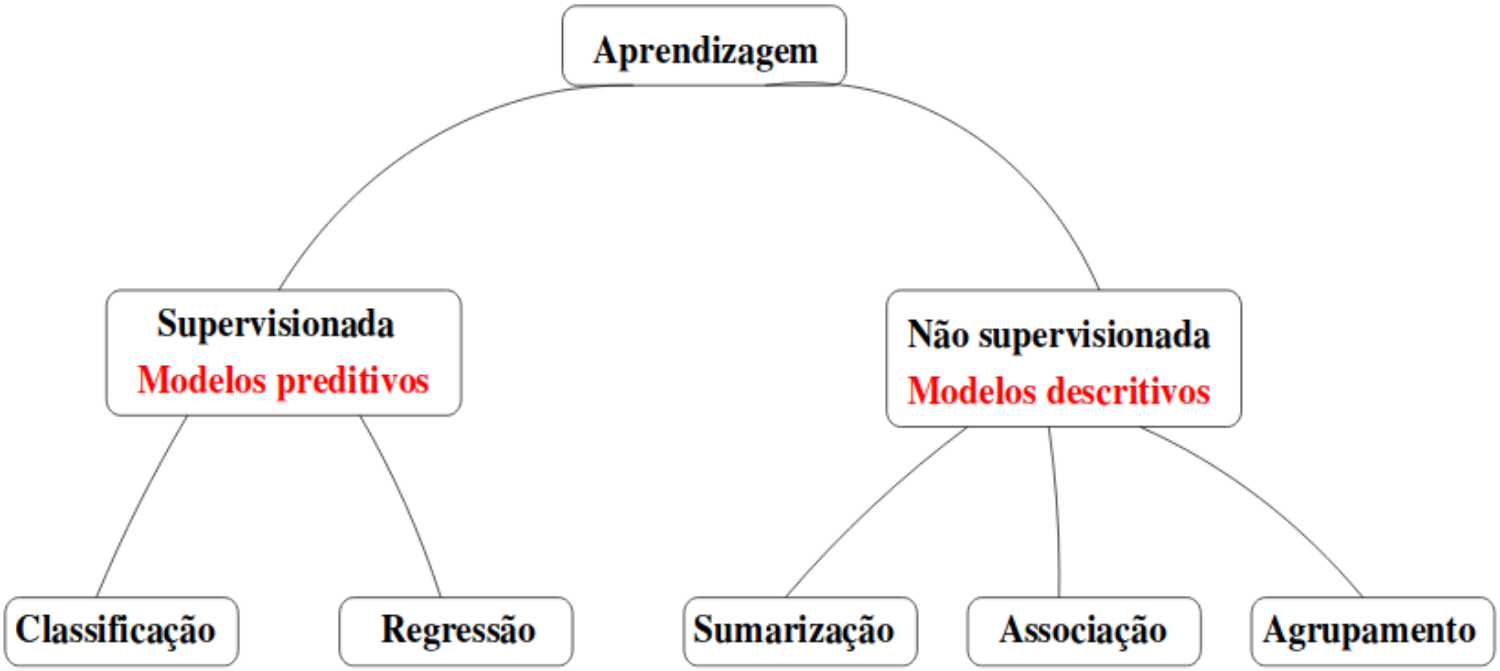
Agenda da aula

- Aprendizado não supervisionado

- Tipos de tarefas

- Agrupamento de dados
  - Hierárquico
  - Plano

Hierarquia



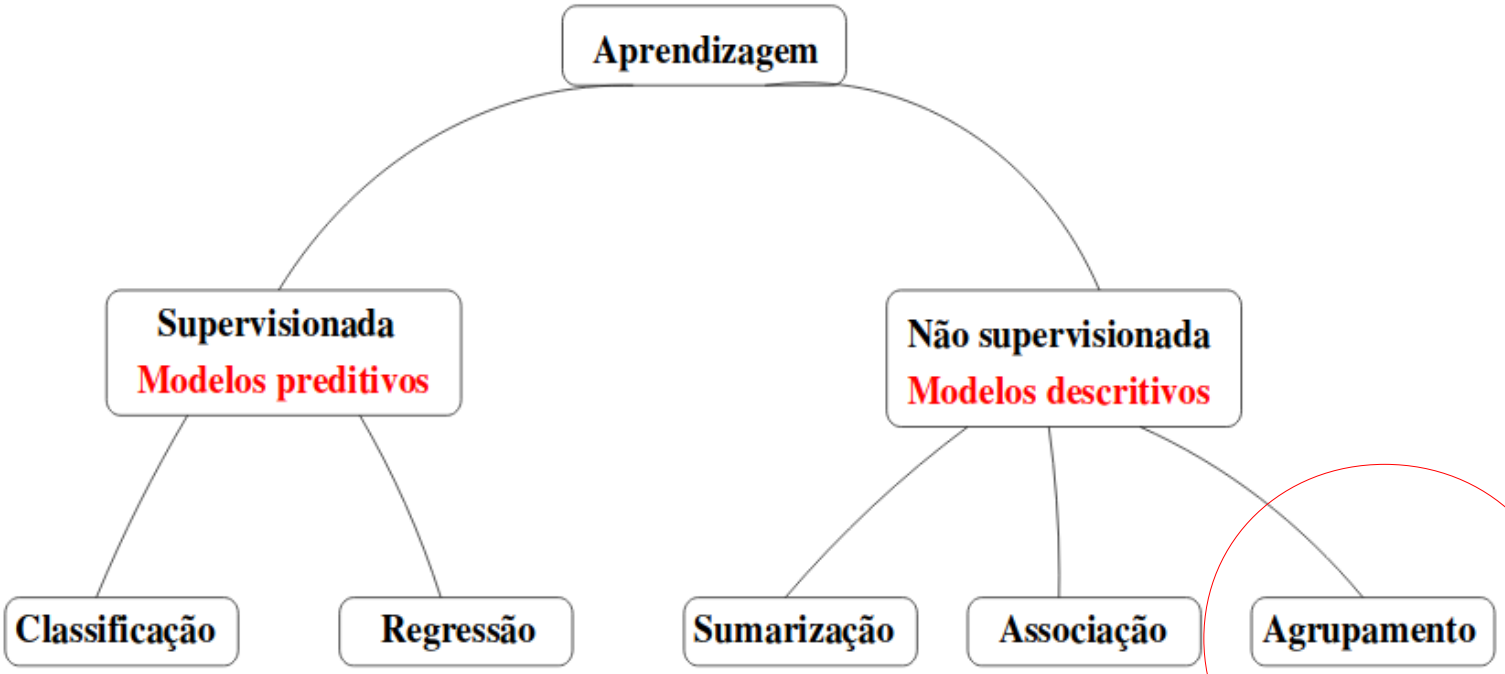
Aprendizado supervisionado

- TODOS os exemplos do TREINAMENTO são **rotulados**

Aprendizado não supervisionado

- O conhecimento é oriundo da similaridade

Hierarquia



Aprendizado supervisionado

- TODOS os exemplos do TREINAMENTO são **rotulados**

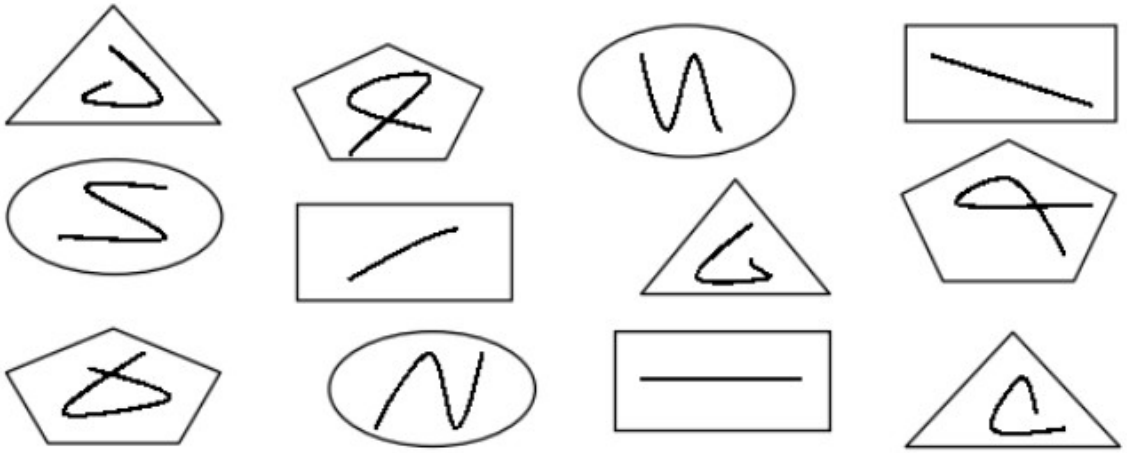
Aprendizado não supervisionado

- O conhecimento é oriundo da similaridade

Aprendizado não supervisionado

Identificar a organização dos **padrões** existentes nos dados através do agrupamentos dos dados – também chamado *clustering*.

- Como organizar os desenhos ao lado?

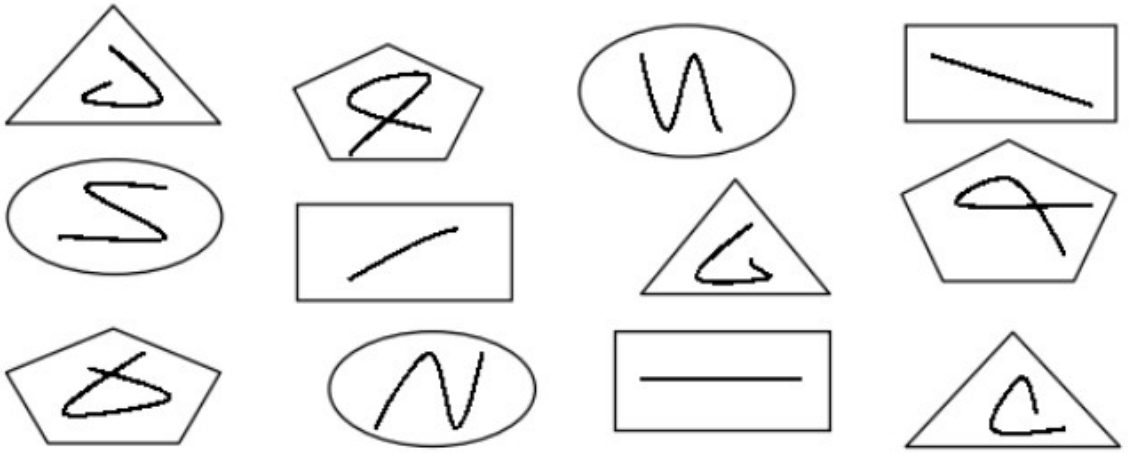


Aprendizado não supervisionado

Identificar a organização dos **padrões** existentes nos dados através do agrupamentos dos dados – também chamado *clustering*.

- Descobrir similaridades e diferenças entre os padrões implícitos.

- Como organizar os desenhos ao lado?

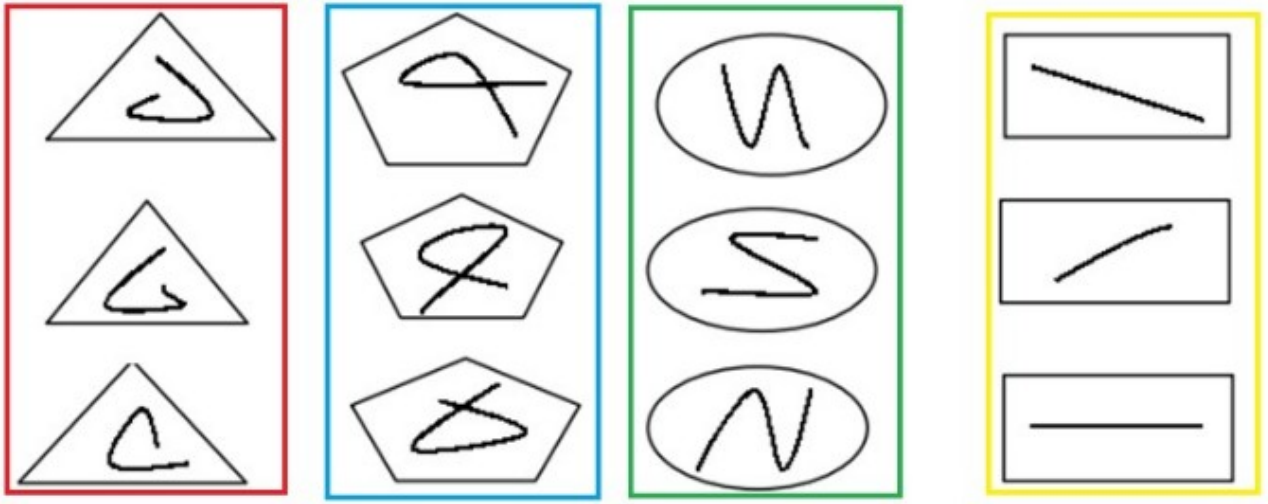


Aprendizado não supervisionado

Identificar a organização dos **padrões** existentes nos dados através do agrupamentos dos dados – também chamado *clustering*.

- Descobrir similaridades e diferenças entre os padrões implícitos.

- Como organizar os desenhos ao lado?



Aprendizado não supervisionado

Entretanto, a similaridade não é fácil de ser identificada na maioria das bases de dados.

Qual atributo utilizar?

Qual critério?

Aprendizado não supervisionado

SEM USAR NENHUM ALGORITMO DE ML

Como Separar os indivíduos em 2 grupos diferentes?

nome	altura	peso	idade	profissão	interesse
Fulano	1,75	85	18	estudante	M
Ciclano	1,54	90	58	pedreiro	M
Fulana	1,45	65	45	programadora	H
Beltrano	1,98	78	17	estudante	H
Anônima	1,65	51	18	estudante	M

TEMPOOOOO...

Aprendizado não supervisionado

SEM USAR NENHUM ALGORITMO DE ML

Como Separar os indivíduos em 2 grupos diferentes?

nome	altura	peso	idade	profissão	interesse
Fulano	1,75	85	18	estudante	M
Ciclano	1,54	90	58	pedreiro	M
Fulana	1,45	65	45	programadora	H
Beltrano	1,98	78	17	estudante	H
Anônima	1,65	51	18	estudante	M

Qual a melhor organização? Se separar por interesse alguém pode ser colocado em grupo errado?

Agrupamento/Clusterização

Algoritmos de agrupamento podem ser classificados em diversos tipos.

Os mais famosos são:

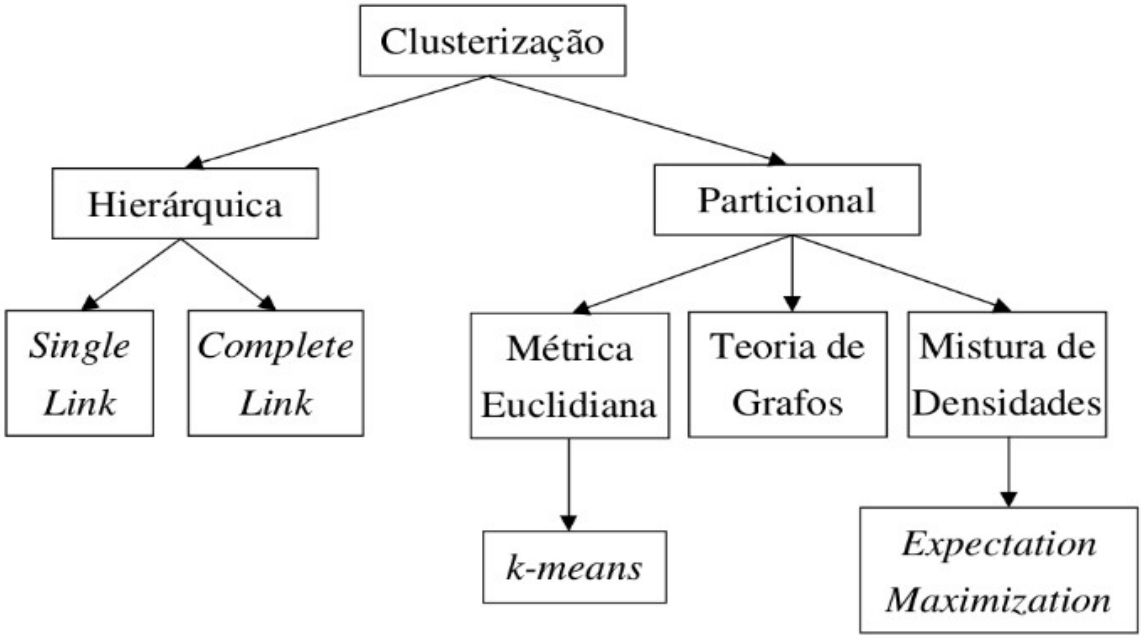
- HIERÁRQUICOS
- NÃO HIERÁRQUCIOS

  (também chamados PARTICIONAIS/SEQUENCIAIS/PLANOS);

Agrupamento/Clusterização

Algoritmos de agrupamento podem ser classificados em diversos tipos.

Os mais famosos são:



- HIERÁRQUICOS
- NÃO HIERÁRQUCIOS

  (também chamados PARTICIONAIS/SEQUENCIAIS/PLANOS);

Agrupamento/Clusterização

Regras para o agrupamento:

Diz-se k-agrupamento, a divisão do conjunto de dados em k grupos.

- Nenhum grupo pode ser vazio ao final da execução;
- Todos os elementos devem participar de pelo menos um grupo;

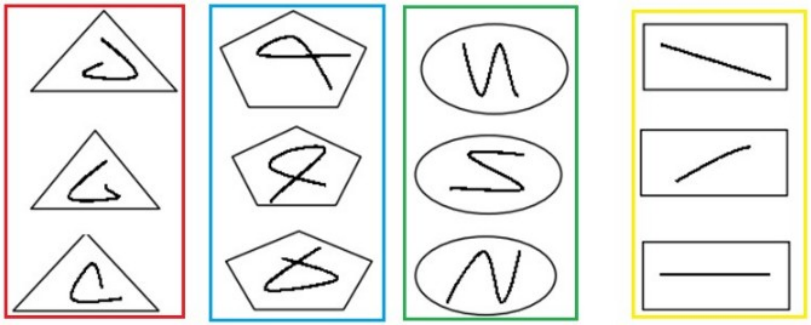
Agrupamento/Clusterização

Regras para o agrupamento:

Diz-se k-agrupamento, a divisão do conjunto de dados em k grupos.

- Nenhum grupo pode ser vazio ao final da execução;
- Todos os elementos devem participar de pelo menos um grupo;

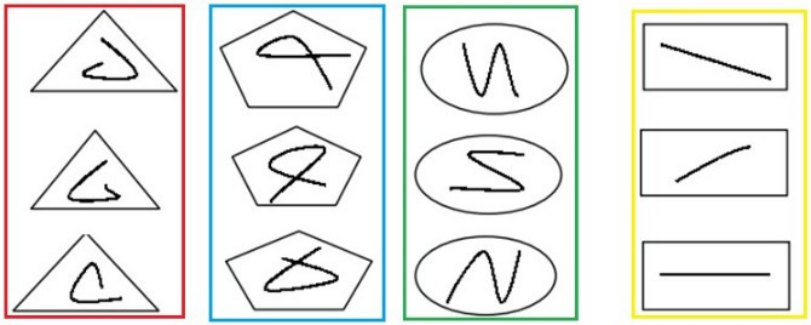
No exemplo anterior: 4-agrupamento



Agrupamento/Clusterização

Mas o que configura a semelhança entre os registros da base de dados?

A semelhança/similaridade de valores em uma base de dados é estabelecida por um cálculo chamado de **distância**.



Algoritmo K-means

É o método mais famoso e simples de todos;

Particiona o conjunto de dados em k grupos mutuamente exclusivos (hard)

– k é definido previamente.

O algoritmo tenta selecionar os registros mais semelhantes e mais distantes dos registros em outros grupos.

É indicado para bases de dados grandes

Algoritmo K-means

A quantidade **K** de clusters deve ser **previamente informada**

Passos:

1. Seleciona-se k elementos da base de dados, onde cada um é um centróide – núcleo do grupo;
2. Para cada registro da base de dados, determina o cluster mais próximo;
3. Recalcula a média de cada cluster.
4. Retoma o processo a partir do passo 2.
5. Resultado: as médias das k partições são os clusters.

A cada iteração o algoritmo recalcula o centro de massa do cluster.

Algoritmo K-means

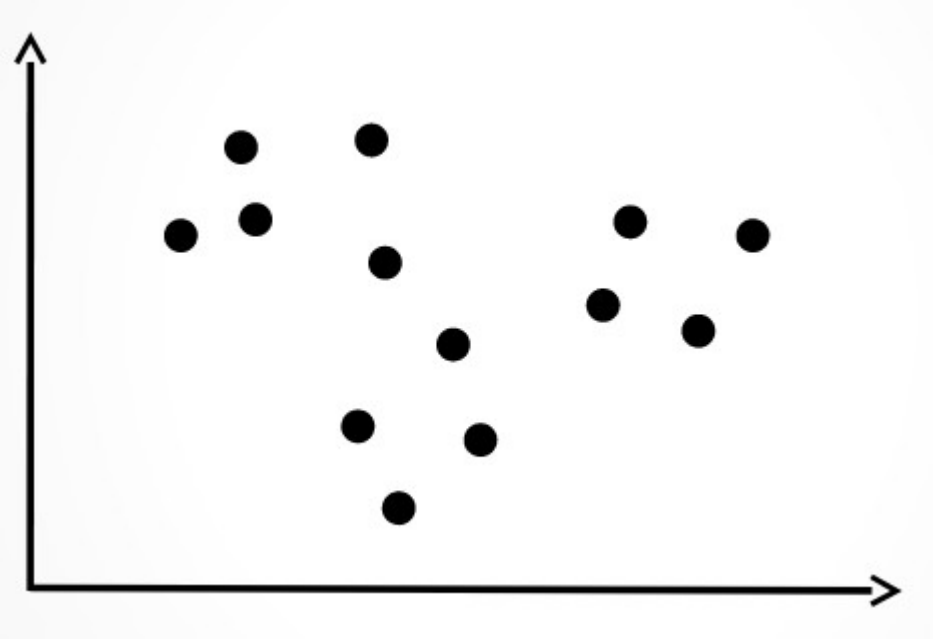
Definição:

Dataset: (x,y)

3 grupos

(clusters)

Iteração: 1



Algoritmo K-means

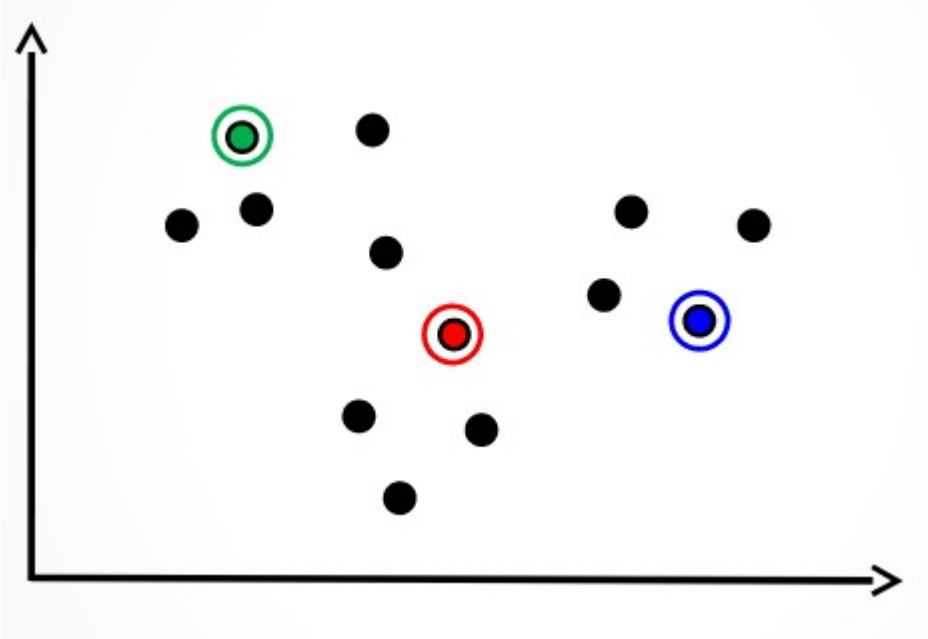
Definição:

Dataset: (x,y)

3 grupos

(clusters)

Iteração: 1



Algoritmo K-means

Definição:

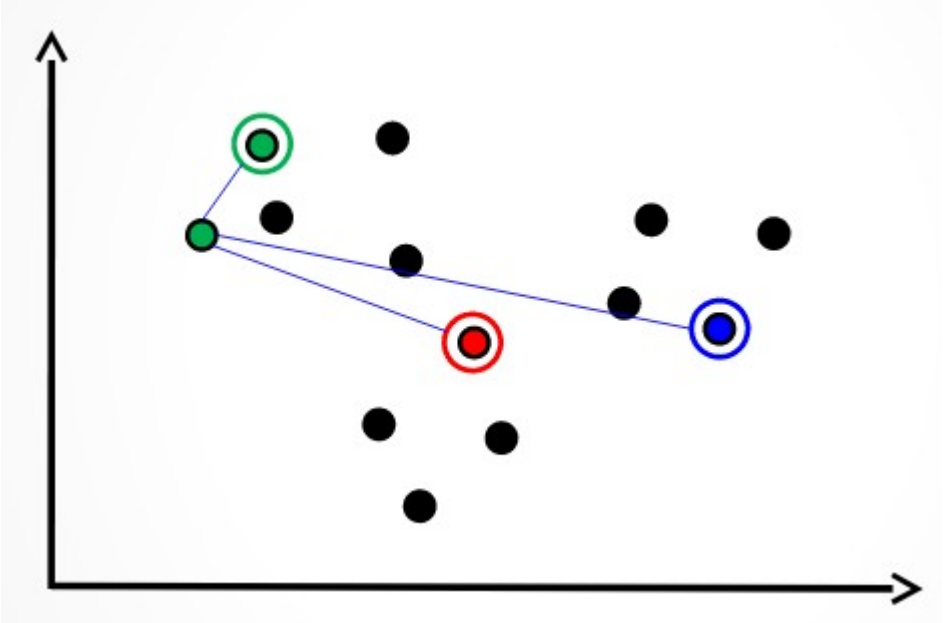
Dataset: (x,y)

Seja:

C1: (2,6)

C2: (5,3)

C3: (8,4)



O primeiro registro selecionado da base de dados é o x_i = (1,5)

Algoritmo K-means

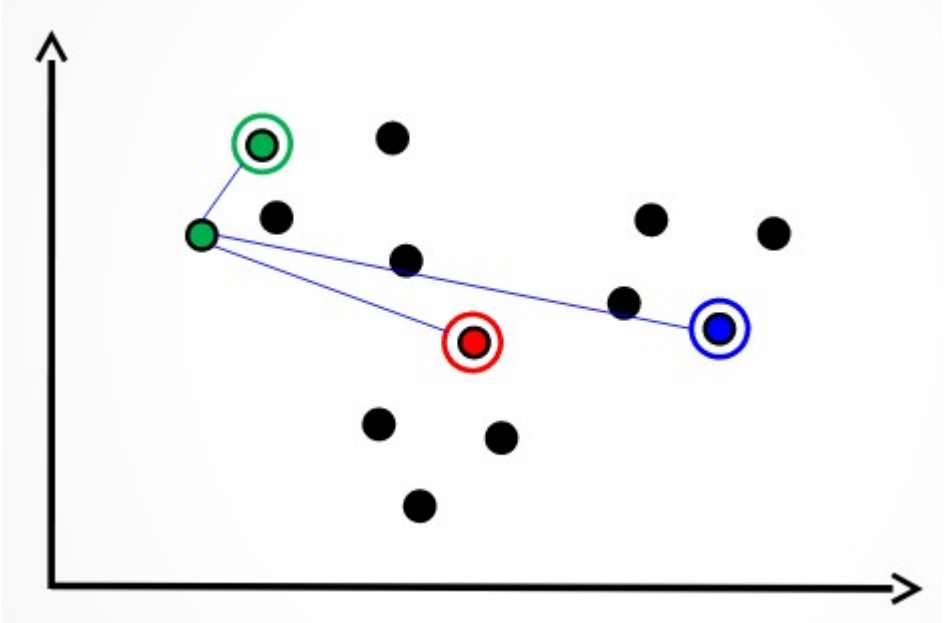
Definição:

Dataset: (x,y)

3 grupos

(clusters)

Iteração: 2



Algoritmo K-means

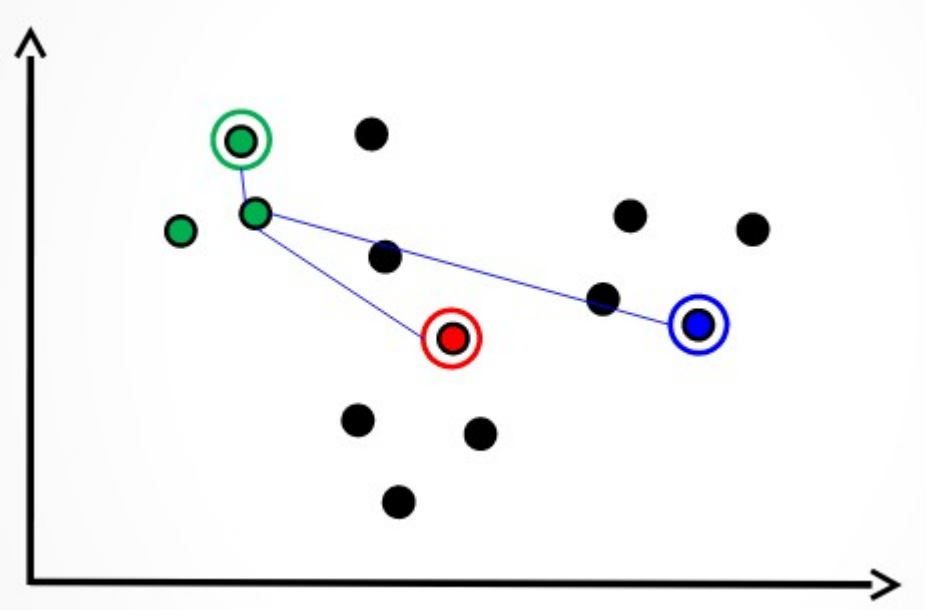
Definição:

Dataset: (x,y)

3 grupos

(clusters)

Iteração: 3



Algoritmo K-means

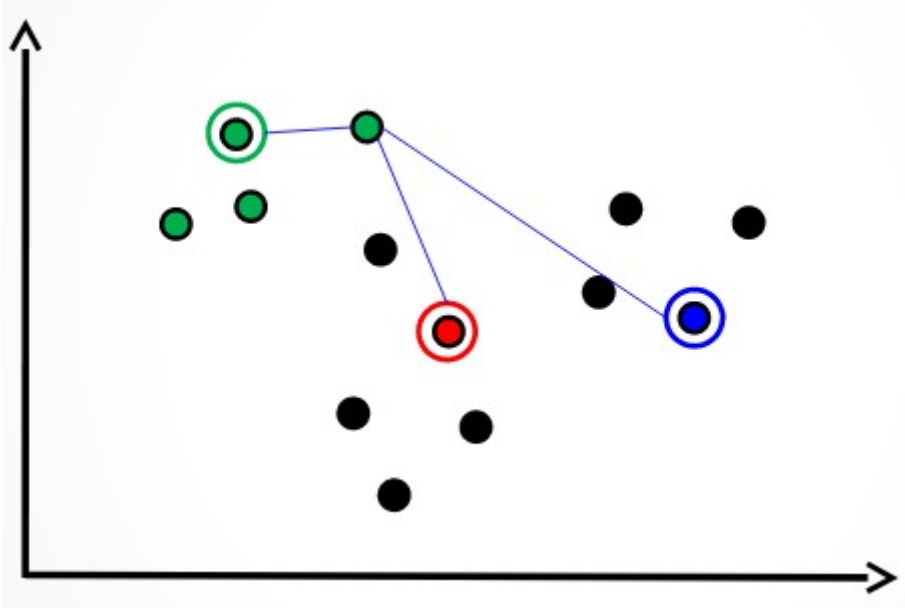
Definição:

Dataset: (x,y)

3 grupos

(clusters)

Iteração: 4



Algoritmo K-means

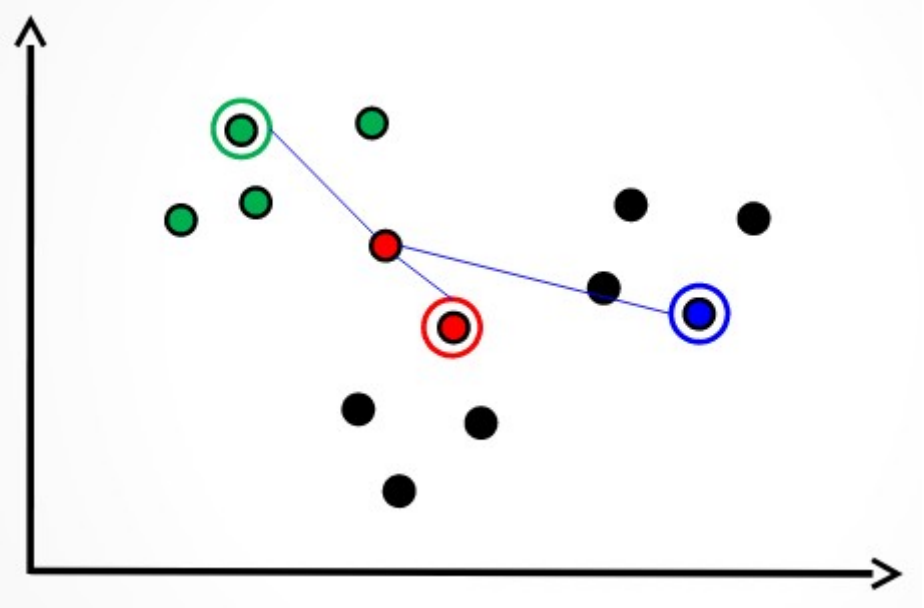
Definição:

Dataset: (x,y)

3 grupos

(clusters)

Iteração: 5



Algoritmo K-means

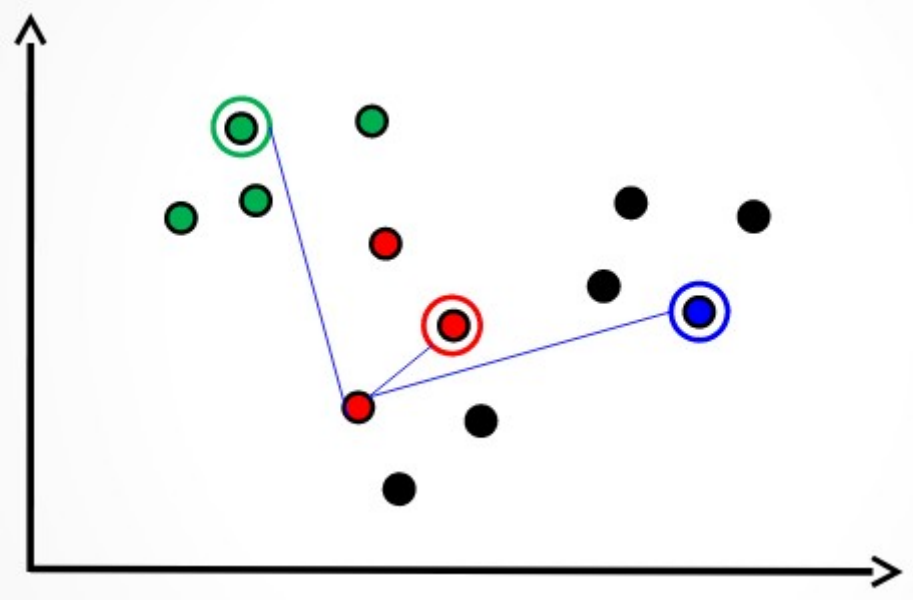
Definição:

Dataset: (x,y)

3 grupos

(clusters)

Iteração: 6



Algoritmo K-means

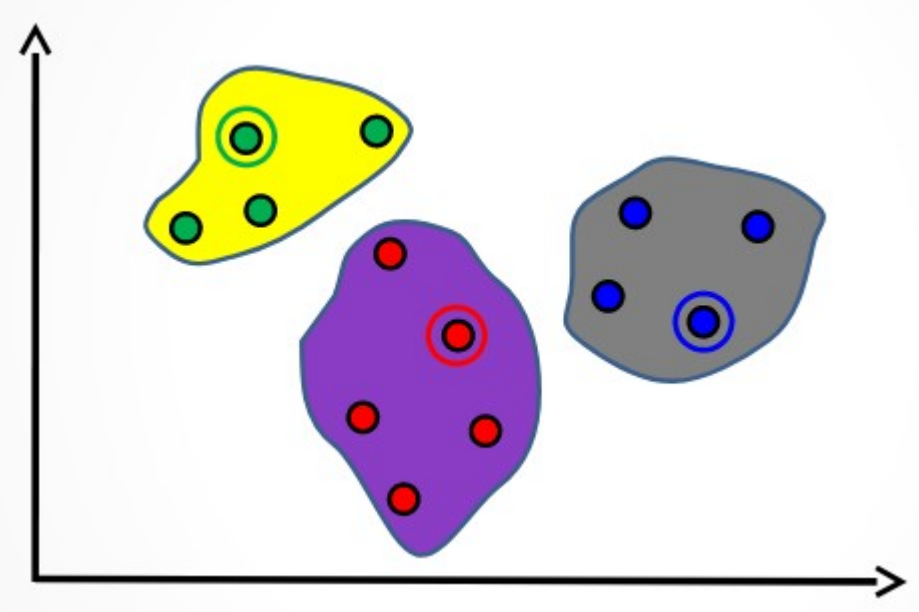
Definição:

Dataset: (x,y)

3 grupos

(clusters)

Iteração: n



Algoritmo K-means

Distância euclidiana: $\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}.$

A distância euclidiana de um ponto à um cluster é a soma das distâncias de cada um dos valores do ponto para os valores do cluster

Algoritmo K-means

Definição:

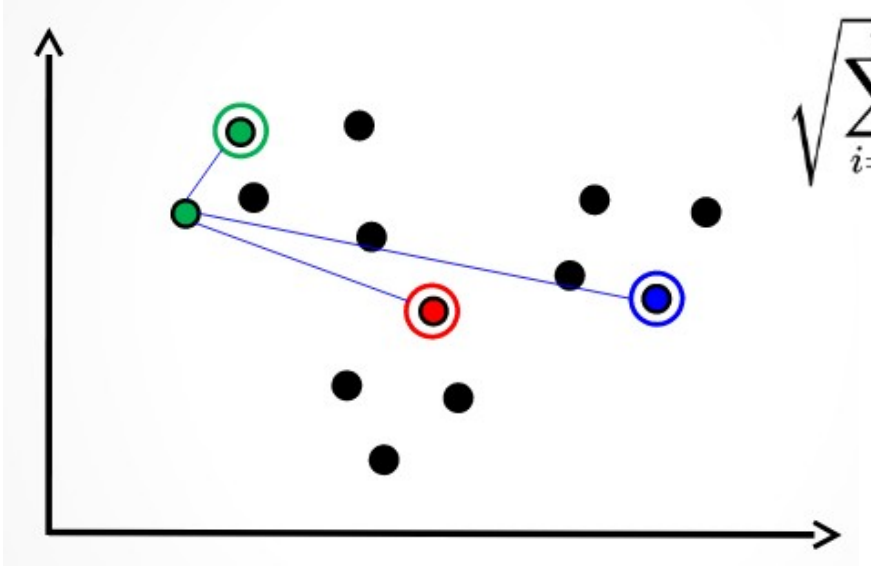
Dataset: (x,y)

Seja:

C1: (2,6)

C2: (5,3)

C3: (8,4)



$$\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}.$$

O primeiro registro selecionado da base de dados é o x_i = (1,5). Pela distância Euclidiana...

$Dx_iC_1 = (1-2)^2 + (5-6)^2$

$Dx_iC_2 = (1-5)^2 + (5-3)^2$

$Dx_iC_3 = (1-8)^2 + (5-4)^2$

Algoritmo K-means

Definição:

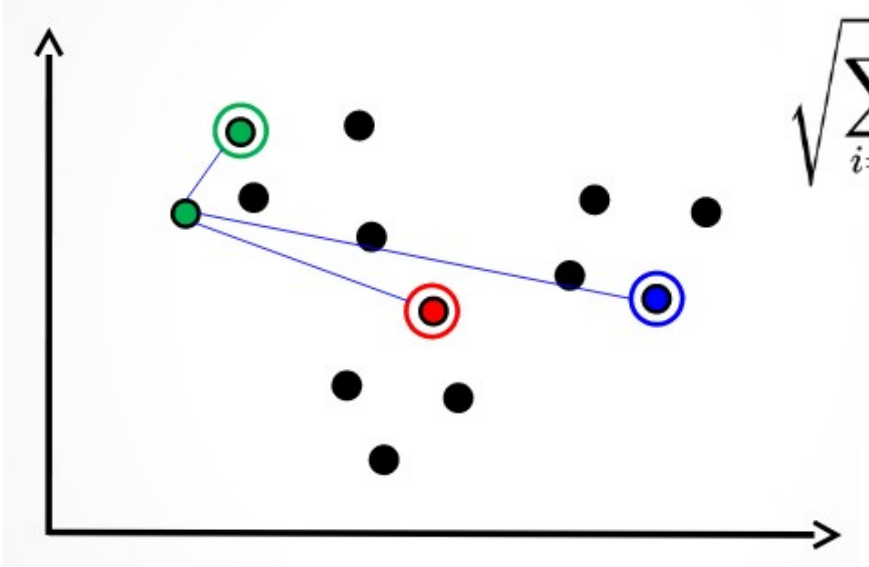
Dataset: (x,y)

Seja:

C1: (2,6)

C2: (5,3)

C3: (8,4)



$$\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}.$$

O primeiro registro selecionado da base de dados é o x_i = (1,5). Pela distância Euclidiana...

Dx_iC_1 = (1-2)² + (5-6)² = -1² + -1² = 2

Dx_iC_2 = (1-5)² + (5-3)² =

Dx_iC_3 = (1-8)² + (5-4)² =

Algoritmo K-means

Definição:

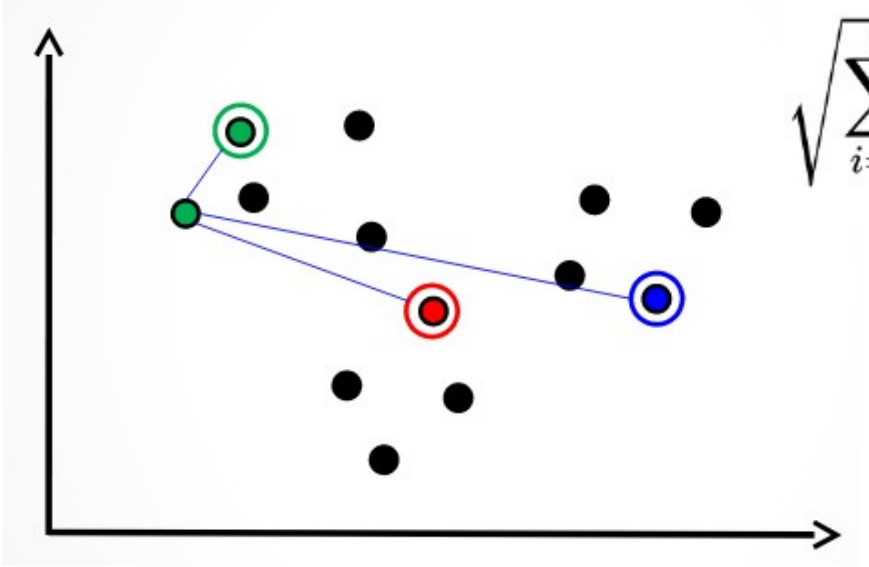
Dataset: (x,y)

Seja:

C1: (2,6)

C2: (5,3)

C3: (8,4)



$$\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}.$$

O primeiro registro selecionado da base de dados é o x_i = (1,5). Pela distância Euclidiana...

Dx_iC_1 = $(1-2)^2 + (5-6)^2$ = $-1^2 + -1^2$ = 2

Dx_iC_2 = $(1-5)^2 + (5-3)^2$ = $-4^2 + -2^2$ = 16 + 4 = 20

Dx_iC_3 = $(1-8)^2 + (5-4)^2$ =

Algoritmo K-means

Definição:

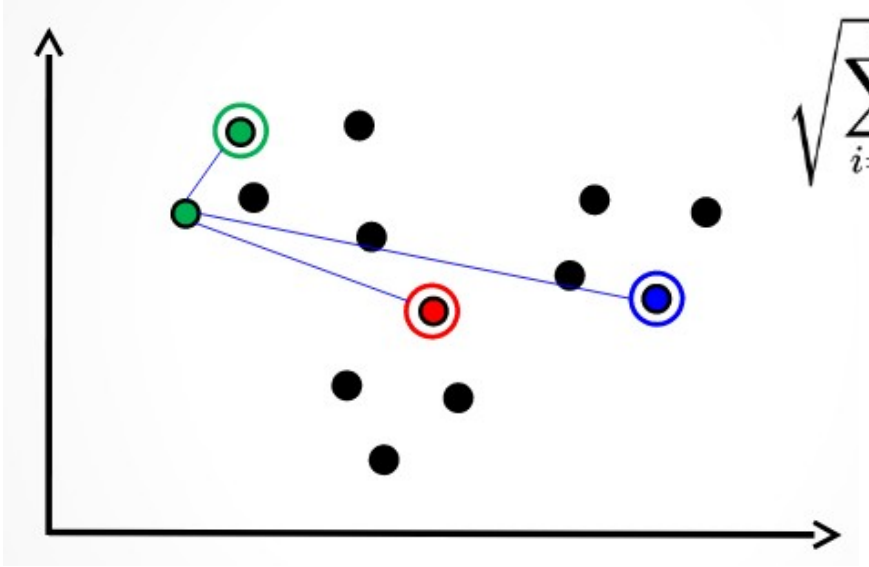
Dataset: (x,y)

Seja:

C1: (2,6)

C2: (5,3)

C3: (8,4)



$$\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}.$$

O primeiro registro selecionado da base de dados é o x_i = (1,5). Pela distância Euclidiana...

$Dx_iC_1 = (1-2)^2 + (5-6)^2 = -1^2 + -1^2 = 2$

$Dx_iC_2 = (1-5)^2 + (5-3)^2 = -4^2 + -2^2 = 16 + 4 = 20$

$Dx_iC_3 = (1-8)^2 + (5-4)^2 = -7^2 + 1^2 = 49 + 1 = 50$

Algoritmo K-means

$$\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}.$$

Dado a base de dados de países emergentes, abaixo:

#	exp_vida	%_estudantes	cresc_PIB	cotacao_dolar
A	45	0.4	0.6	2.34
B	75	0.7	1.3	1.5
C	77	0.5	-0.2	4.87
D	68	0.9	3.2	4.1

Qual a distância euclidiana do ponto

X_4 : 72, 0.80, 1.25 4.5

???

Algoritmo K-means

$$\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}.$$

Dado a base de dados de países emergentes, abaixo:

#	exp_vida	%_estudantes	cresc_PIB	cotacao_dolar
A	45	0.4	0.6	2.34
B	75	0.7	1.3	1.5
C	77	0.5	-0.2	4.87
D	68	0.9	3.2	4.1

Qual a distância euclidiana do ponto

X_4 : 72, 0.80, 1.25 4.5

???

Dist Eucl. = calcular a média de cada atributo e calcular a raiz quadrada da soma quadrada das diferenças do registro para as médias

Algoritmo K-means

```
C   =   [45,0.4,0.6,12.34]
reg =   [72,0.8,1.25,4.5]

distanciaEuclidiana =  sum((attReg - attC)*(attReg - attC) for attReg, attC in
zip( reg, C))
print(distanciaEuclidiana)
```

E se precisássemos calcular a distância euclidiana de cada cluster?

Fazer média dos valores de cada atribudo de cada cluster;

Calcular a distância do reg para o centróide;

Verificar qual é a menor distância.

Algoritmo K-means

C1	exp_vida	%_estudantes	cresc_PIB	cotacao_dolar
K	45	0.4	0.6	12.34
T	55	0.3	0.3	81.5
I	57	0.5	-0.2	4.81
D	58	0.9	-3.2	4.41

$$\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}.$$

C2	exp_vida	%_estudantes	cresc_PIB	cotacao_dolar
A	75	0.74	1.6	2.34
B	75	0.7	1.3	1.5
E	87	0.85	2.2	1.87
G	78	0.9	3.2	2.1

X5 : 72, 0.80, 1.25 4.5

C3	exp_vida	%_estudantes	cresc_PIB	cotacao_dolar
H	65	0.14	0.6	24.8
J	71	0.47	-1.1	51.3
C	68	0.45	-1.2	14.83
F	68	0.19	-5.2	141

Algoritmo K-means

```
Clusters = [
    [
        [45,0.4,0.6,12.34],
        [55,0.3,0.3,81.5],
        [57,0.5,-0.2,4.81],
        [58,0.9,-3.2,4.41]
    ],
    [
        [75,0.74,1.6,2.34],
        [75,0.7,1.3,1.5],
        [87,0.85,2.2,1.87],
        [78,0.9,3.2,2.1]
    ],
    [
        [65,0.14,0.6,24.8],
        [71,0.47,-1.1,51.3],
        [68,0.45,-1.2,14.83],
        [68,0.19,-5.2,141]
      ]
    ]

#C1 = Clusters[0]
#C2 = Clusters[1]
#C3 = Clusters[2]
```

Algoritmo K-means

A partir do momento que um registro é inserido (ou dito pertencer) a um Cluster C_i a os valores do centróide são atualizados

Ci	exp_vida	%_estudantes	cresc_PIB	cotacao_dolar
Média				

Quais os novos valores do Cluster C_i após a identificação que o registro X_5 faz parte dele?

X_5 : 72, 0.80, 1.25 4.5

Algoritmo K-means

A partir do momento que um registro é inserido (ou dito pertencer) a um Cluster C_i a os valores do centróide são atualizados

Ci	exp_vida	%_estudantes	cresc_PIB	cotacao_dolar
x_5	72	0.80	1.25	4.5
Média				

Quais os novos valores do Cluster C_i após a identificação que o registro X_5 faz parte dele?

X_5 : 72, 0.80, 1.25 4.5

Algoritmo K-means

Vamos ver uma implementação do algoritmo K-means...

Base de dados: genérica

Ferramenta Weka/Orange/Scikit-learn.

http://localhost:8888/notebooks/K-mean.ipynb

Otimização

Antes de agrupar dados:

- Atributos devem ser selecionados para evitar redundância

- Para que todos atributos contribuam de forma equalitária é interessante **NORMALIZAR** os atributos.

Outras medidas podem ser consideradas?

Distância Manhattan:

$d(xi,xj) = \sum|xi - xj|;$

menor número de substituições necessárias para transformar uma string na outra

elabore" e "melhore" é 4.

Distância de Hamming

Fuzzy K-means

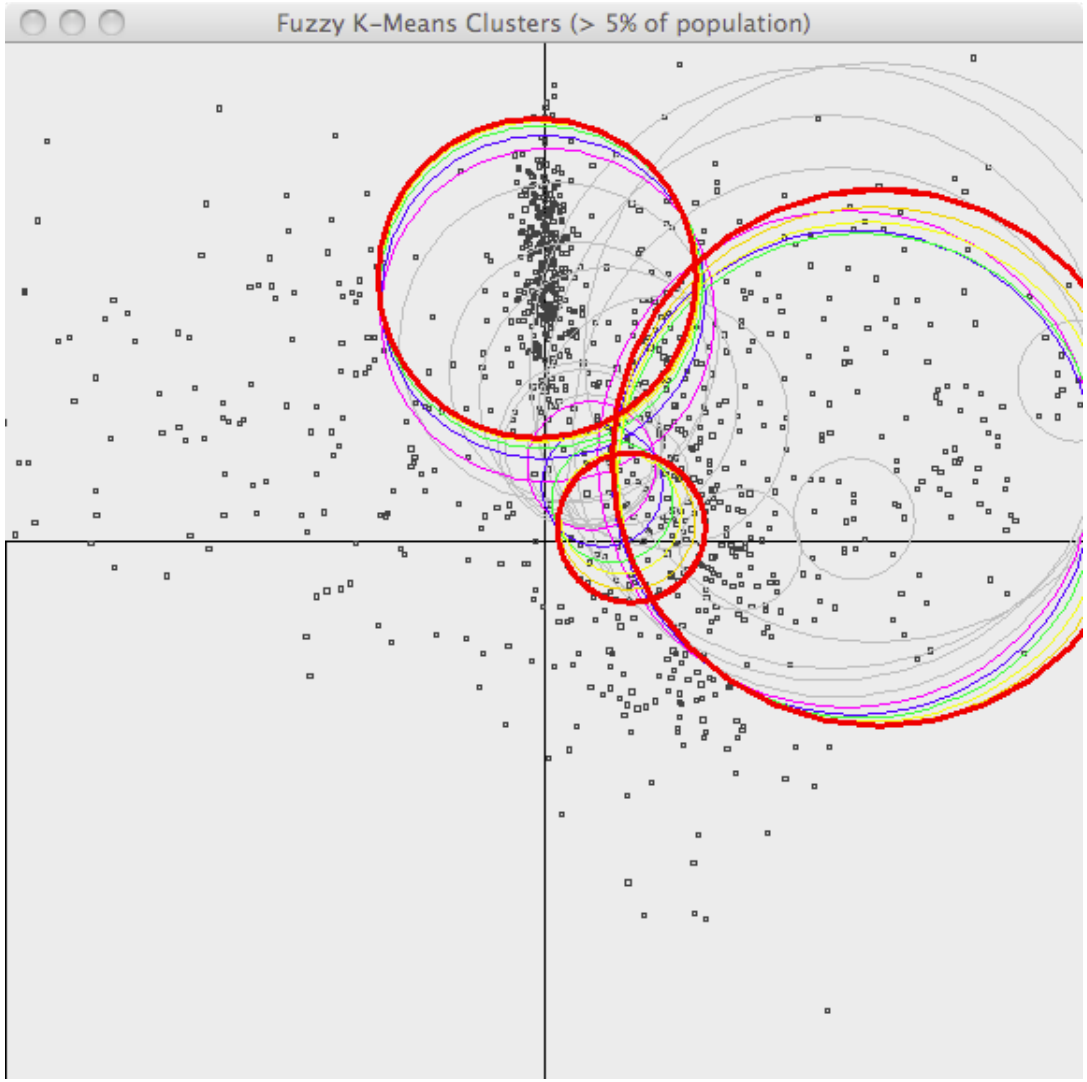
Leva em consideração o grau de pertinência de um registro aos grupos;

Por exemplo, ao calcular a distância euclidiana do registro xi aos clusters A, B e C, o resultado poderia ser:

Xi pertence a A: 80%

Xi pertence a B: 15%

Xi pertence a C: 5%



K-medians

Enquanto o K-means busca a média dos valores de cada atributo, o K-medians assume como valor do centróide do cluster a mediana.

Os valores presentes em um cluster C1 são:

C1: {(0.1,10)(5,-2)(0.2,200)}

Pelo K-means: {(1.76, 69.3)}

Pelo K-medians: {(0.2, 10)}

O k-medians é mais robusto quanto à presença de outliers.

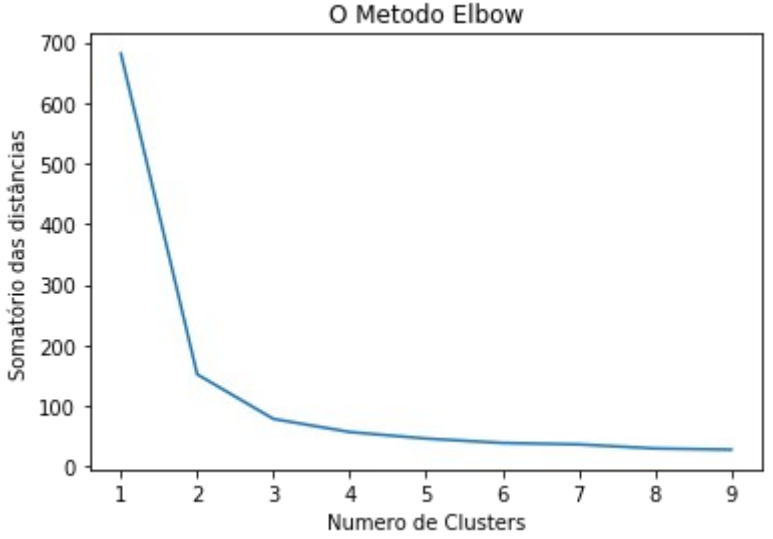
Escolha do número de clusters

Quantos clusters são necessários?

Método elbow (cotovelo)

Basicamente testa a variância dos dados em relação ao número de clusters.

Até o momento que aumentar o número de clusters não representa melhoria significativa.



A medida que a quantidade de clusters aumenta, a soma das distâncias quadráticas tende a zero.

http://localhost:8888/notebooks/elbow.ipynb

Hora de fazer amigos – busca por similaridade

Vamos construir uma base de dados suficientemente grande para agrupar os alunos da sala e descobrir nossos colegas que mais compartilham das nossas características

Atributos que poderíamos considerar:

Nome,
idade,
altura,
cor_pref,
linguagem_programacao_pref,
livro_pref,
cidade_ori,
estilo_mus, ... quais mais?

http://encurtador.com.br/bJQVY

Hora de fazer amigos – busca por similaridade

Passos para um bom agrupamento:

- Implementar a função elbow (cotovelo) para encontrar a quantidade ideal de grupos;
- Construir os grupos pelo algoritmo K-means;
- Visualizar os grupos gerados;
- Verificar em qual grupo cada um de nós se encaixa...

E agora?

Quais estratégias você usaria nos problemas abaixo:

- Implementar um filtro de spam para um cliente que está cansado de fazer isso *na mão*;

- Encontrar clientes que poderíam se interessar por um novo produto que foi testado por um grupo pequeno de clientes.

Avaliação do modelo

Visualizar o modelo é a melhor estratégia.

O espaçamento entre os clusters está bom?

Forma mais conhecida:

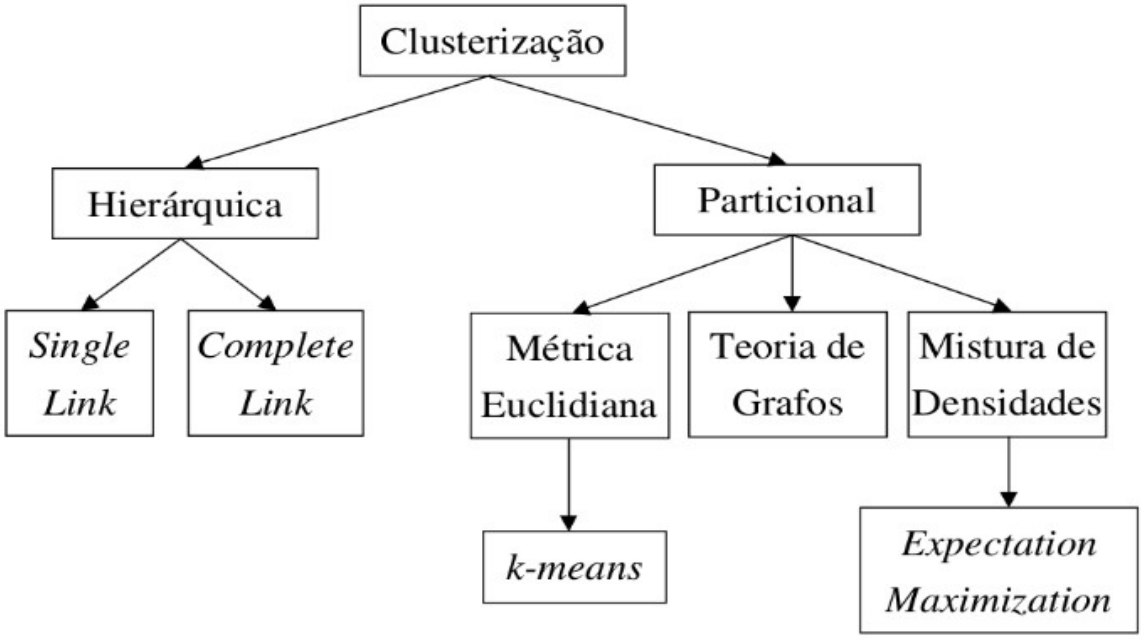
Monte Carlo (roleta)

Um registro qualquer é selecionado e uma classificação quanto ao grupo que deve pertencer é realizada.

Agrupamento/Clusterização

Algoritmos de agrupamento podem ser classificados em diversos tipos.

Os mais famosos são:



- HIERÁRQUICOS
- NÃO HIERÁRQUCIOS

  (também chamados PARTICIONAIS/SEQUENCIAIS/PLANOS);

Métodos hierárquicos

Técnicas simples – os dados são particionados a cada iteração;

- Não requer pré estipulação do número de clusters (k)

Constrói a chamada matriz de similaridade

	G1	G2	G3
G1	0	0,1	0,3
G2	0,1	0	0,4
G3	0,3	0,4	0

G1 e G2 são mais similares

G2 e G3 são menos similares

Métodos hierárquicos

Estratégia **aglomerativa**

- Inicia com cada registro compondo um grupo (N registros = N clusters).

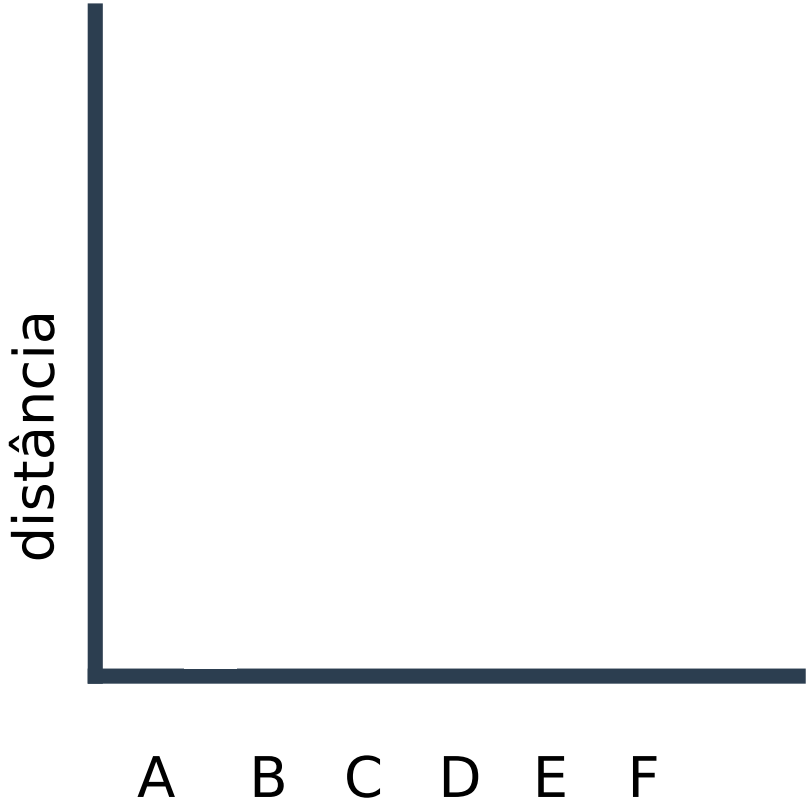
- Calcula-se a matriz de similaridade para todos os registros;

- Os clusters dos dois registros com maior similaridade são unificados em um só - menor valor na matriz de similaridade.

- Repete-se os passos até que todos os registros sejam agrupados em um cluster só.

Métodos hierárquicos

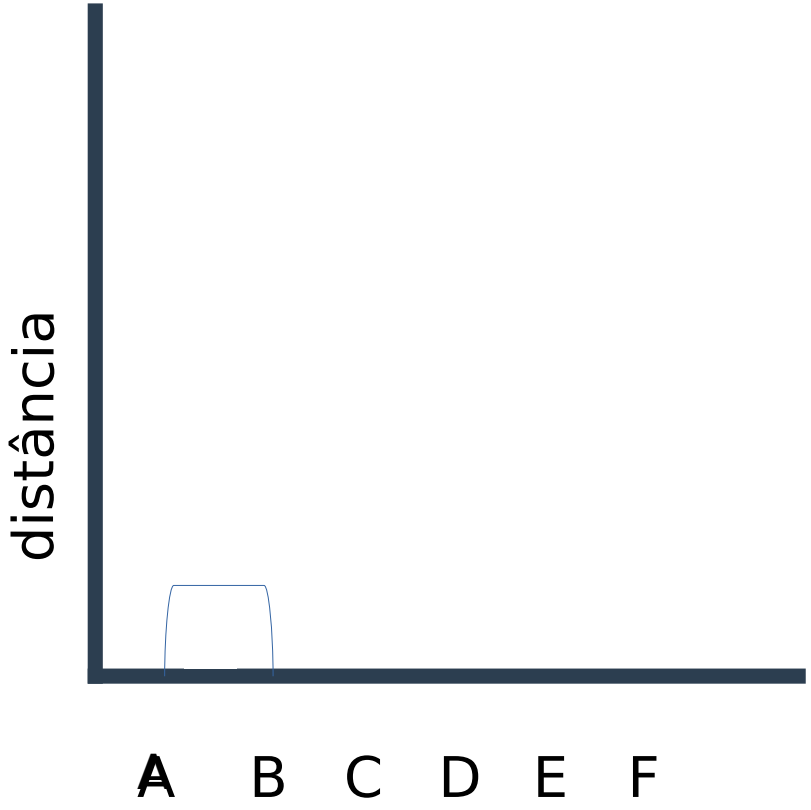
Aglomerativo



	A	B	C	D	E	F
A	0	-	-	-	-	-
B	14	0	-	-	-	-
C	20	74	0	-	-	-
D	33	45	41	0	-	-
E	48	51	56	25	0	-
F	33	85	34	95	31	0

Métodos hierárquicos

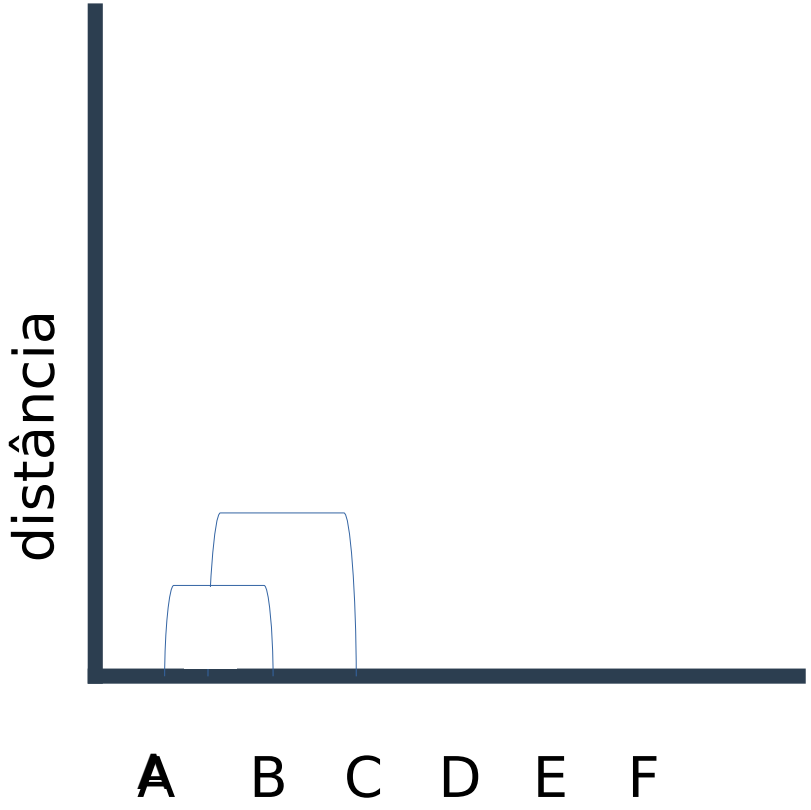
Aglomerativo



	A	B	C	D	E	F
A	0	-	-	-	-	-
B	14	0	-	-	-	-
C	20	74	0	-	-	-
D	33	45	41	0	-	-
E	48	51	56	25	0	-
F	33	85	34	95	31	0

Métodos hierárquicos

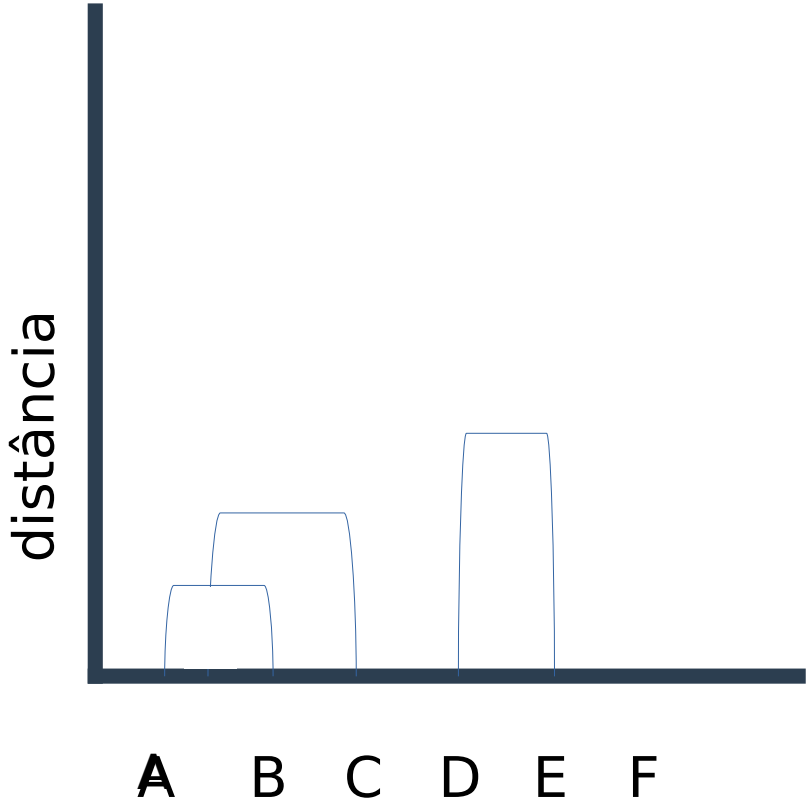
Aglomerativo



	A	B	C	D	E	F
A	0	-	-	-	-	-
B	14	0	-	-	-	-
C	20	74	0	-	-	-
D	33	45	41	0	-	-
E	48	51	56	25	0	-
F	33	85	34	95	31	0

Métodos hierárquicos

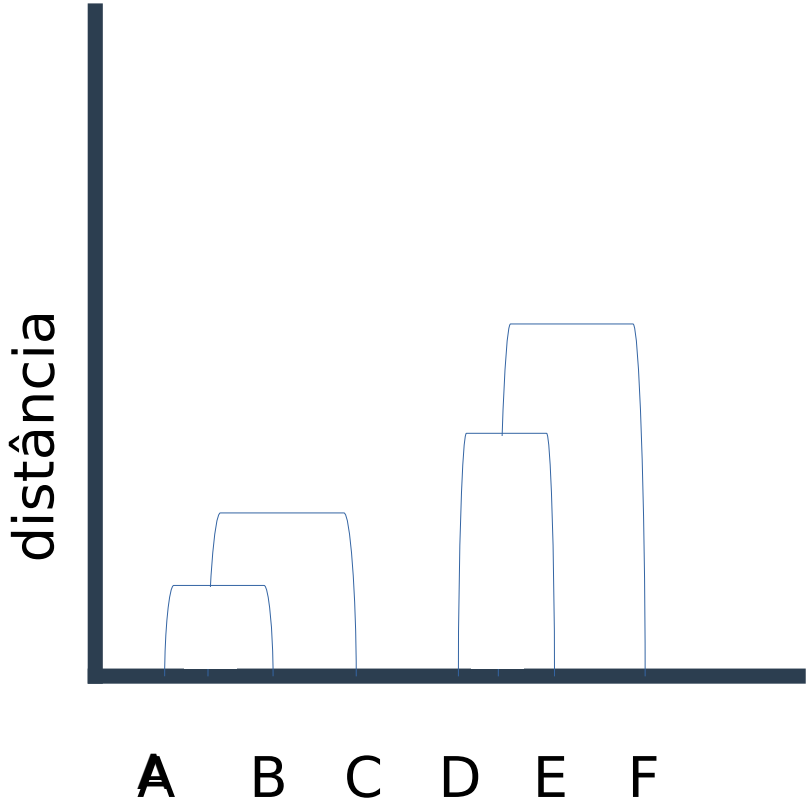
Aglomerativo



	A	B	C	D	E	F
A	0	-	-	-	-	-
B	14	0	-	-	-	-
C	20	74	0	-	-	-
D	33	45	41	0	-	-
E	48	51	56	25	0	-
F	33	85	34	95	31	0

Métodos hierárquicos

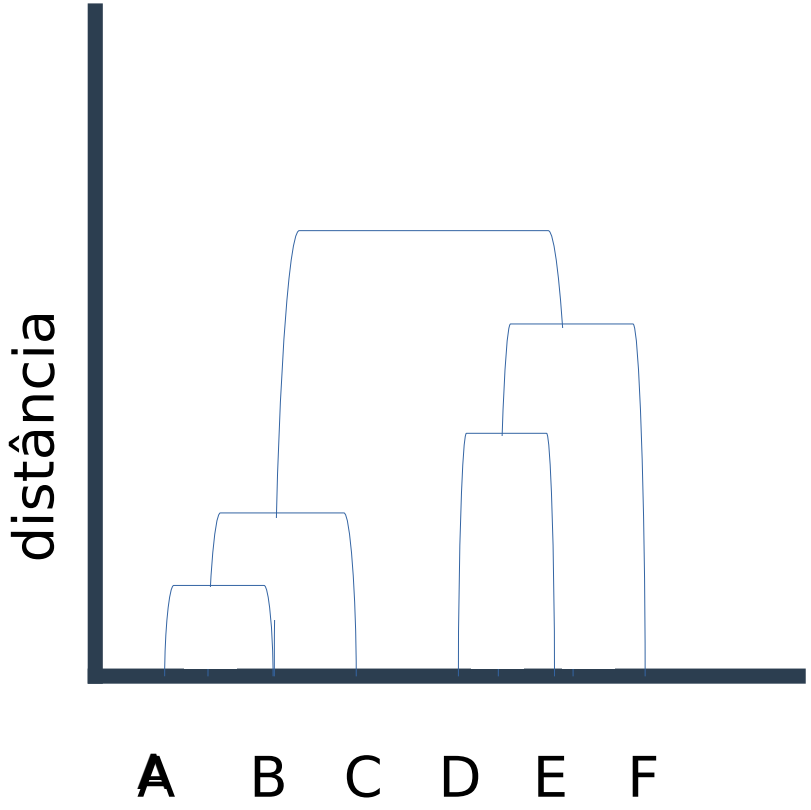
Aglomerativo



	A	B	C	D	E	F
A	0	-	-	-	-	-
B	14	0	-	-	-	-
C	20	74	0	-	-	-
D	33	45	41	0	-	-
E	48	51	56	25	0	-
F	33	85	34	95	31	0

Métodos hierárquicos

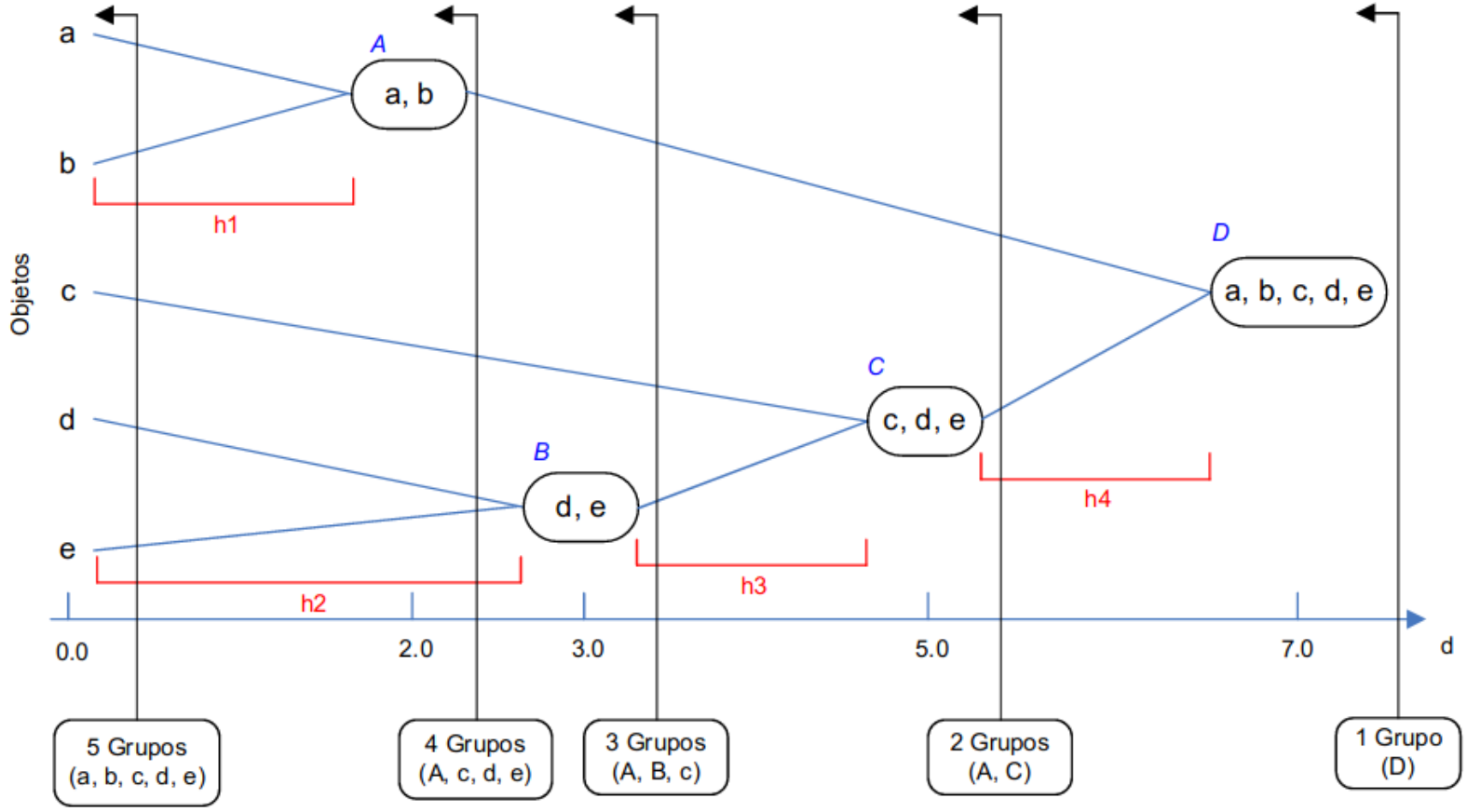
Aglomerativo



	A	B	C	D	E	F
A	0	-	-	-	-	-
B	14	0	-	-	-	-
C	20	74	0	-	-	-
D	33	45	41	0	-	-
E	48	51	56	25	0	-
F	33	85	34	95	31	0

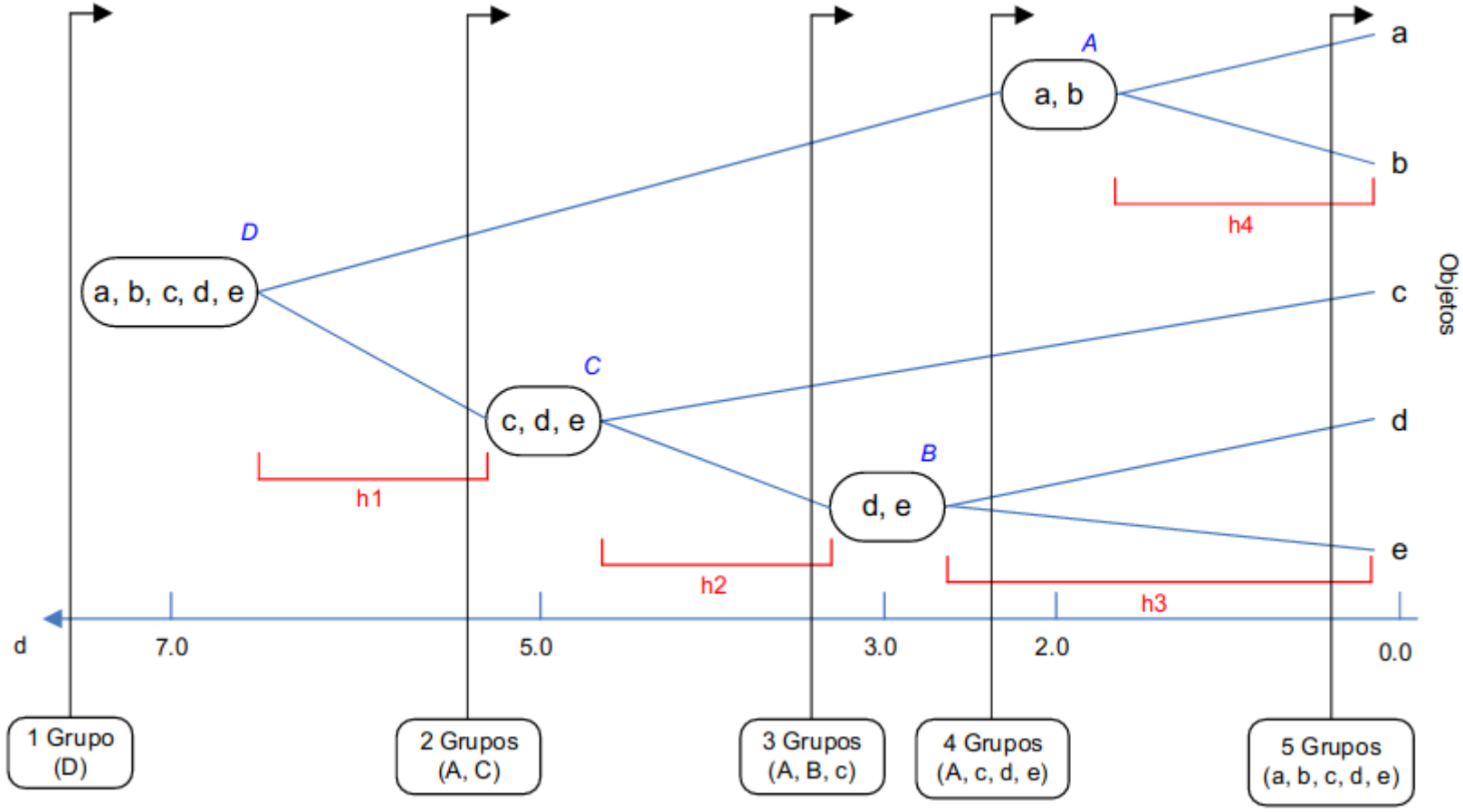
Métodos hierárquicos - dendograma

Aglomerativo



Métodos hierárquicos - dendograma

Divisivo



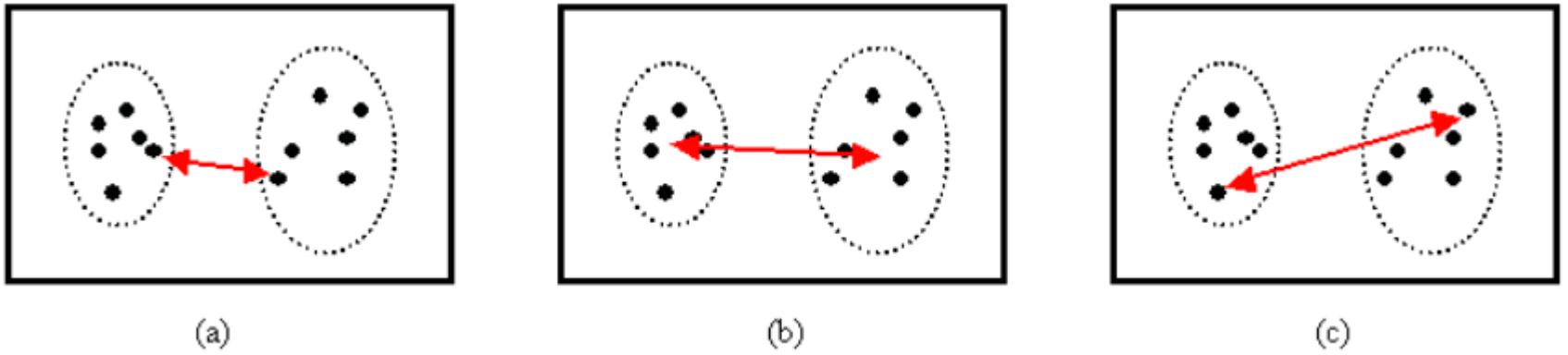
Ligação de Clusters

Distância entre dois clusters:

- **Single Link**: A distância entre dois clusters é dada pela distância entre os seus pontos mais próximos;

- **Average Link**: Distância entre seus centróides – média do grupo.

- **Complete Link**: A distância entre clusters é a distância entre seus pontos mais distantes.



Tipos de agrupamento

Métodos Hierárquicos

- Algoritmos Aglomerativos
- Algoritmos Divisivos

Métodos Particionais

- Algoritmos Exclusivos
- Algoritmos Não-exclusivos

Métodos hierárquicos - dendograma

Para dados de difícil interpretação é a melhor estratégia - análise exploratória de dados;

Menos eficiente em tempo de execução e consumo de memória do que o método plano (K-means)

Atividade Hamming

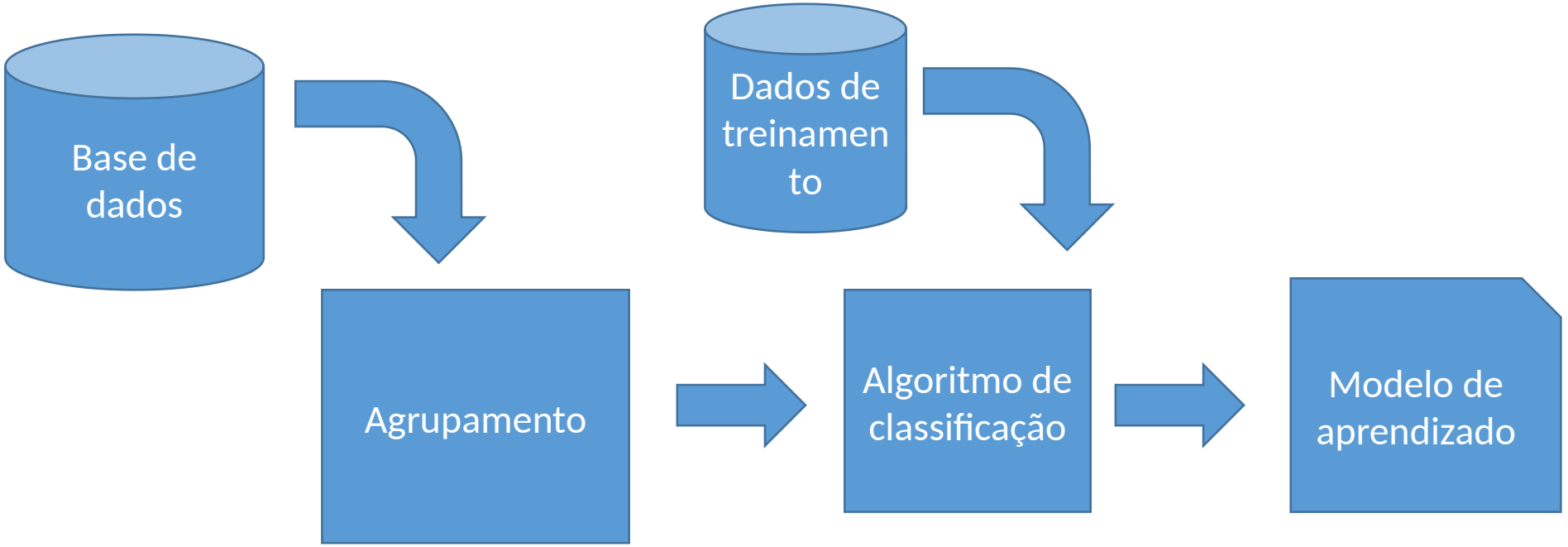
Levando em conta a função de distância Hamming, abaixo:

```
def hamming_distance(str1, str2):
    if len(str1) == len(str2):
        return sum(char1 != char2 for char1, char2 in zip(str1, str2))
```

Implemente um agrupador de Pessoas da sala (ignora todos atributos numéricos)

O que é machine learning

O agrupamento pode ser uma eficiente forma de seleção de características.



Fonte: Elaborada pelo autor

Combinação de modelos

Um agrupamento pode colaborar com a tarefa de classificação? COMO?

E o contrário, é possível?

Como?

Atividade dengue

Base de dados dengue:

Selecionar TODOS classificados como Dengue

Selecionar TODOS classificados como não dengue

Realizar agrupamento hierárquico para cada caso. Por que hierárquico?

→ os atributos dos agrupamentos podem ser utilizados para classificar novos pacientes com dengue.

Sumarizando

- Aprendizado não supervisionado

- Tipos de tarefas

- Agrupamento de dados

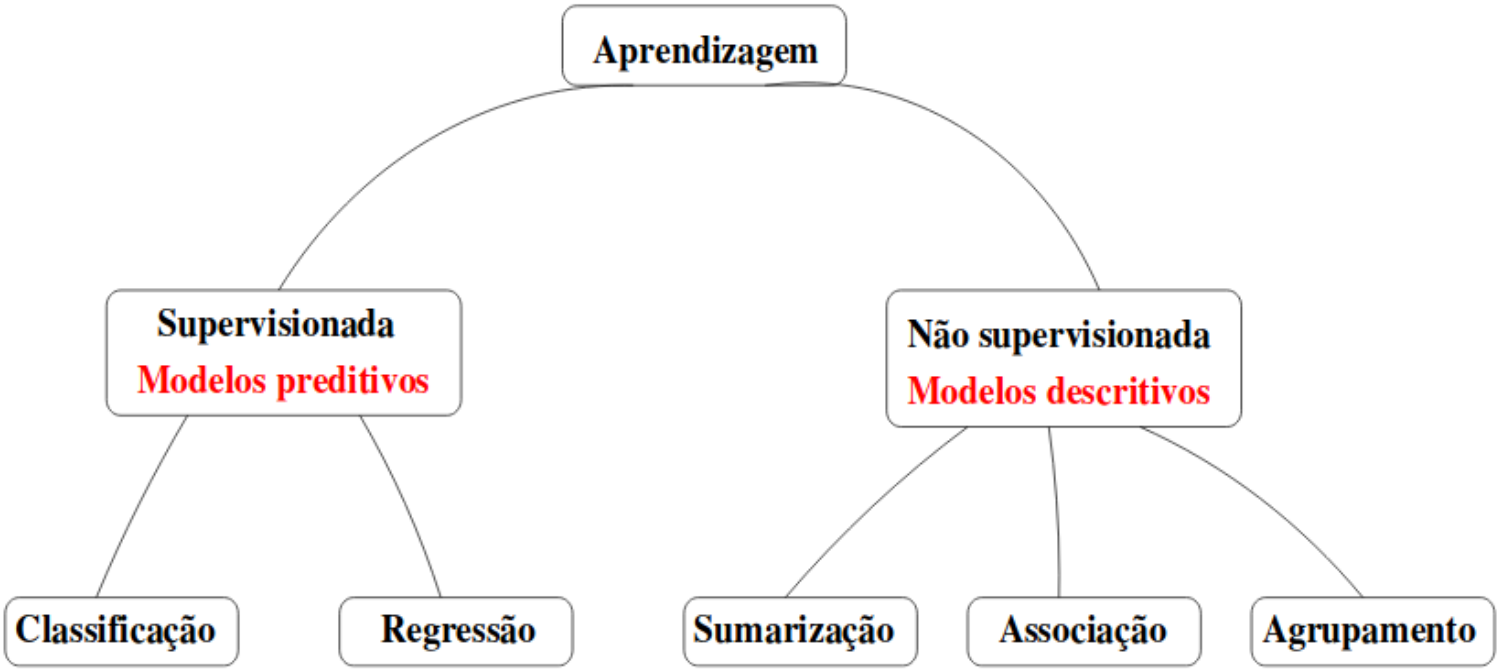
Hierárquico
Plano

Agrupar é:
Maior entendimento dos dados;

Evidenciar correlações entre os atributos;

Pré-processar dados para outros algoritmos.

Hierarquia



Aprendizado supervisionado

- TODOS os exemplos do TREINAMENTO são **rotulados**

Aprendizado não supervisionado

- O conhecimento é oriundo da similaridade

Sumarização de dados

Busca realizar uma descrição simples e compacta dos dados.

- Nuvens de tags;
- Métricas estatísticas (média, mediana, desvio padrão,...);;
- valores mais frequentes para cada atributo;
- ... mais detalhes no módulo de processamento de linguagem natural