Inteligência Artificial

k-means clustering

Clarimar J. Coelho

Escola Politécnica
Pontifícia Universidade Católica de Goiás

Sumário

Como as máquinas podem pensar como os humanos?

- Entender como os humanos pensam

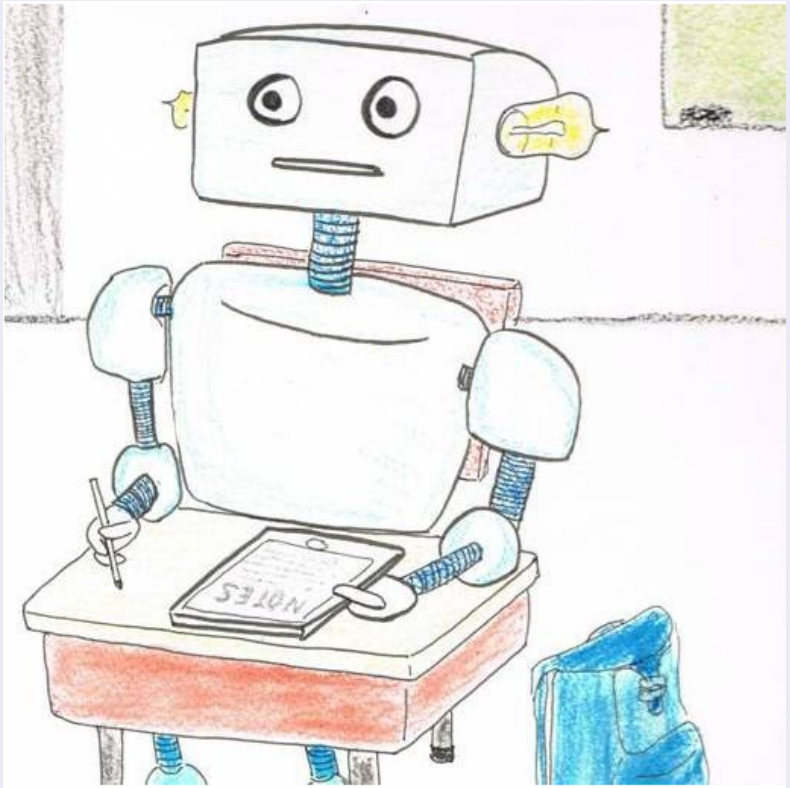
Modelo da natureza do pensamento humano

- Buscar dados massivos/informações
- Senso
- Processo de pensamento no cérebro
- Fazer previsão

Como este modelo pode ser usado em máquina?

O que é aprendizagem não supervisionada?

- Fazendo máquinas que aprendem



Modelo de pensamento de aprendizado de máquina

- A partir de dados massivos
- Aplica algoritmo para ensinar a si mesmo
- Fazer previsões

Como usar o aprendizado de máquina?

- Entenda o problema primeiro
- Tipo de aprendizado de máquina depende do caso

Quantos tipos existem no aprendizado de máquina?

Existem três tipos de aprendizado de máquina

- Aprendizagem supervisionada
- Aprendizagem não supervisionada
- Aprendizagem por reforço

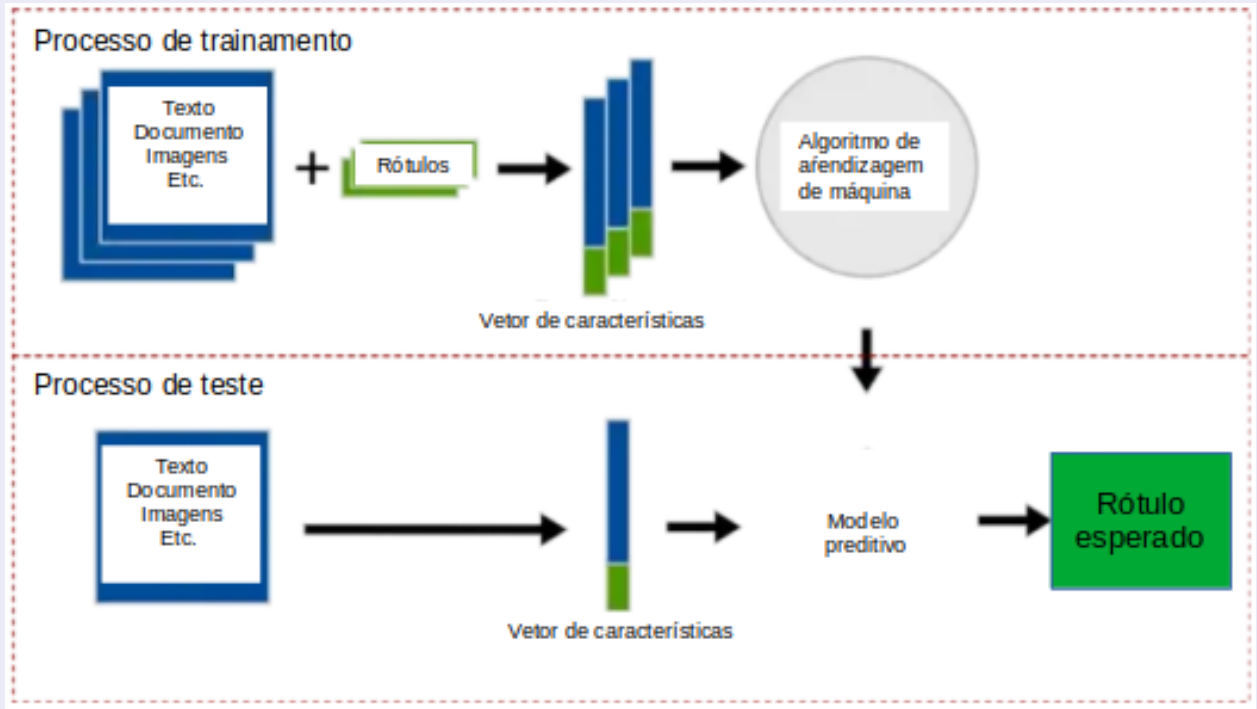
Aprendizagem supervisionada

Tem regra

- Tem dados de entrada e rótulo para treino

Aprendizagem supervisionada

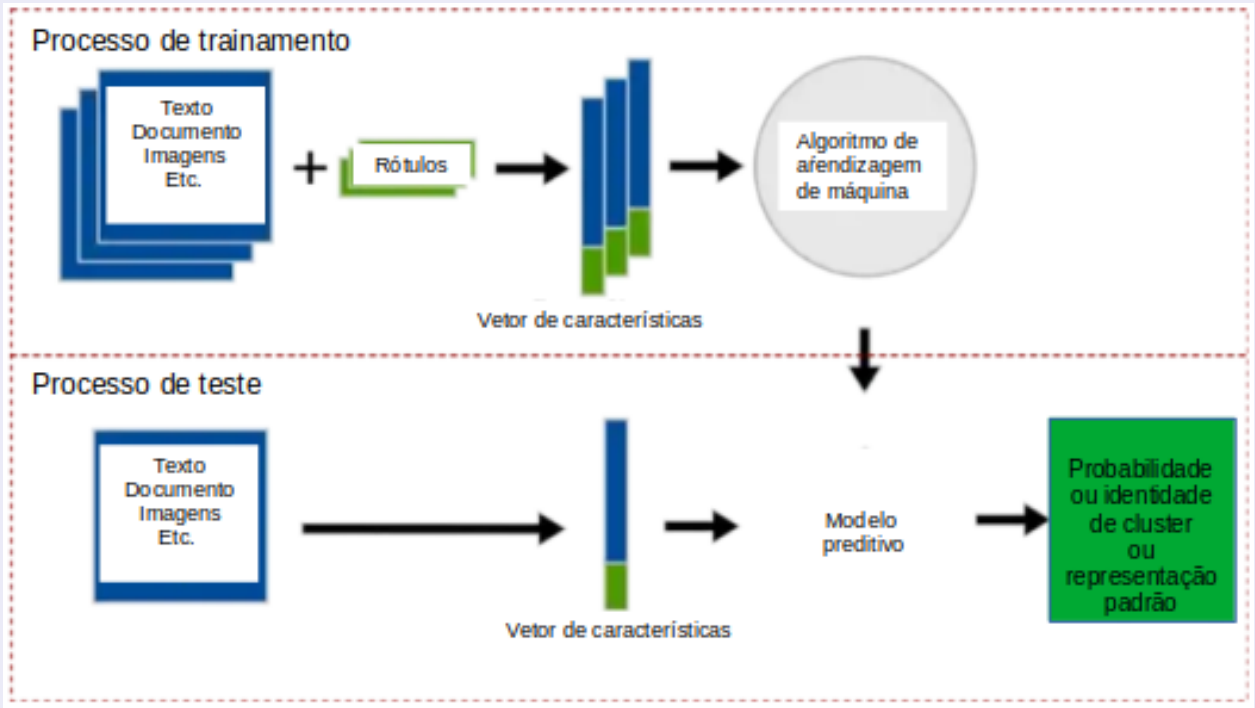
- Tem dados de entrada e rótulo para treino



Aprendizagem não supervisionada

Descoberta de padrões

- Por dados de entrada fornecidos, sem qualquer rótulo



Aprendizagem não supervisionada

Existem três casos

- Agrupamento, redução de dimensionalidade e regra de associação

Aprendizagem não supervisionada

Agrupamento (*clustering*)

- Agrupar dados com base em padrões de similaridade

Aprendizagem não supervisionada

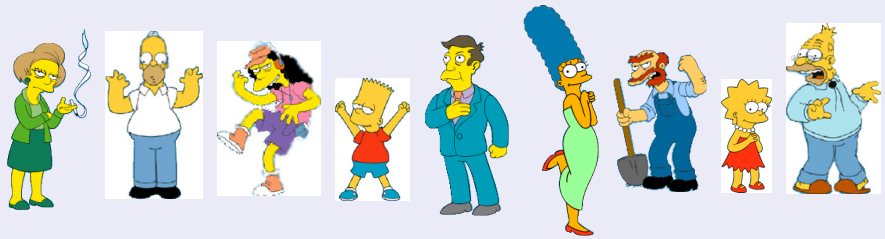
Existem métodos que podem ser usados no (*clustering*)

- *k−means clustering*, *mean shift*, *spectral clustering*, *hierarchical clustering*, DBSCAN, etc.

Sumário

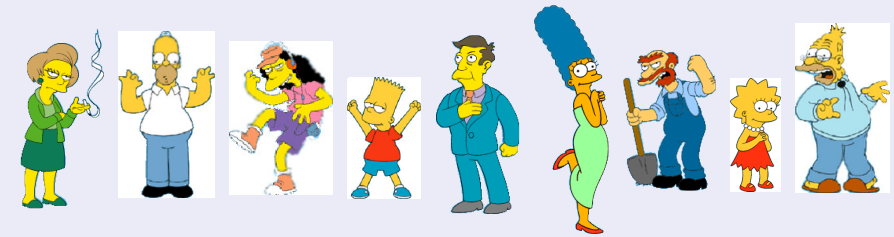
Intuição de *clustering*

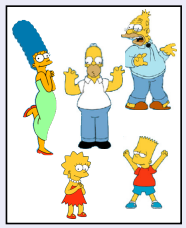
- Qual é o agrupamento natural entre esses objetos?



Intuição de *clustering*

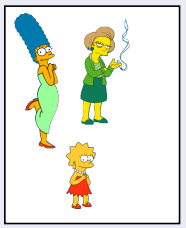
- Qual é o agrupamento natural entre esses objetos?





Familia simpson

Funcionários da escola

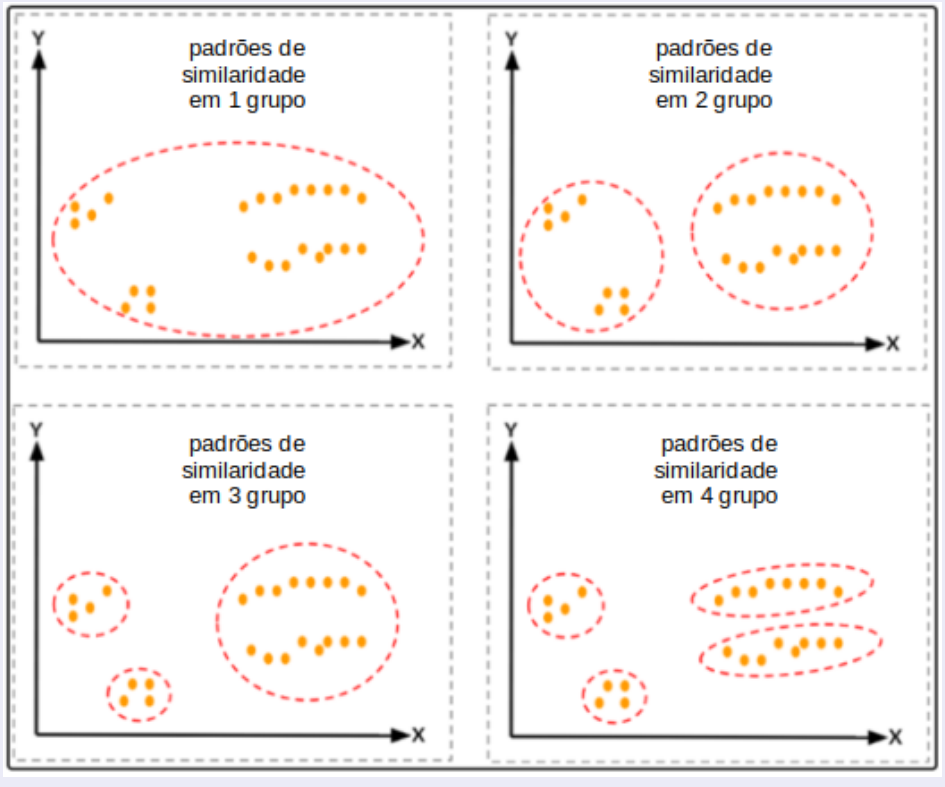


Mulheres

Homens

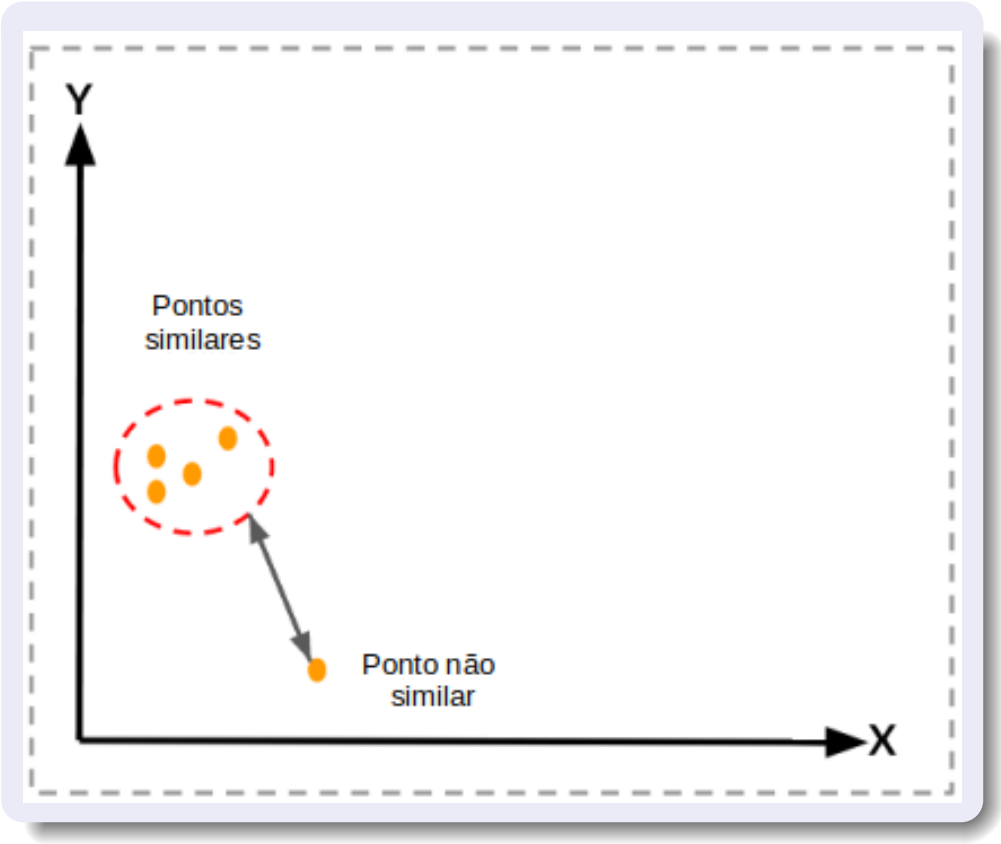
Intuição de *clustering*

- Agrupamento de dados com base em padrões de similaridade
- Com base na distância



Intuição de *clustering*

- Como sabemos que um ponto está no mesmo grupo que outro ponto?
  - Com base em padrões de similaridade

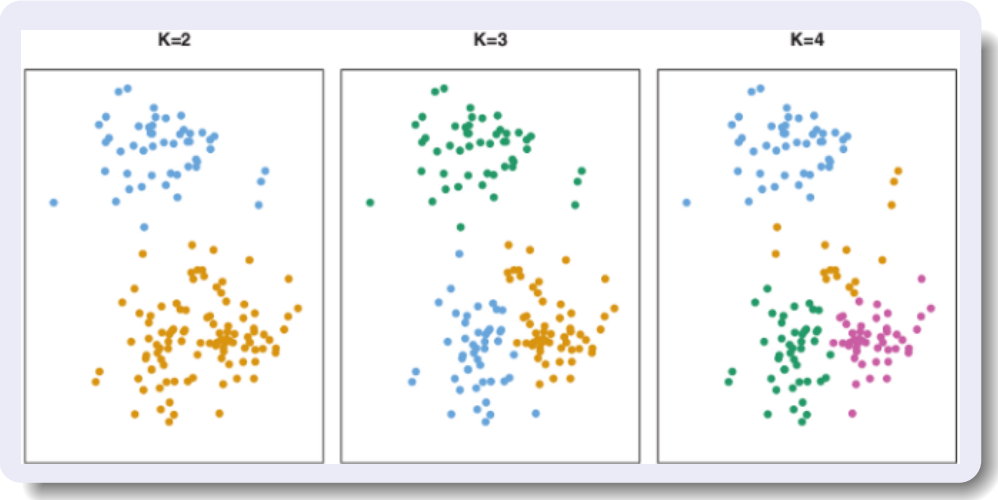


Sumário

O que é $k-means\ clustering$?

Particiona um conjunto de dados

- Em k grupos/*clusters* distintos e não sobrepostos
- Especifique o número desejado de *clusters* k
- O *k-means* atribui cada observação a um dos k agrupamentos



O que é $k-$means clustering

$k-means$ agrupa os dados

- Tentando separar as amostras em n grupos de variância igual
- Minimiza um critério conhecido como inércia
- Ou soma dos quadrados dentro do *cluster*

O que é $k-$means clustering

$k-means$ agrupa os dados

- Escolhe o centróide que minimiza a inércia
- Ou critério de soma dos quadrados dentro do *cluster*

$$\sum_{i=0}^{n} min(||x_i - \mu_j||^2)$$

O que é $k-$means clustering

Como o $k-means$ funciona

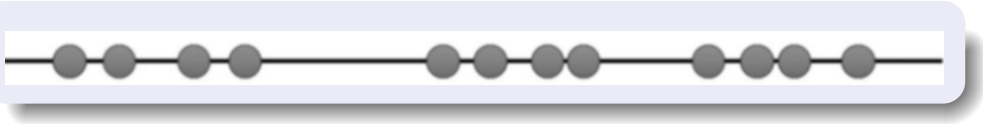
- Para os dados

$$\sum_{i=0}^{n} min(||x_i - \mu_j||^2)$$

O que é $k-means$ *clustering*?

Como o $k-means$ funciona

- Para o conjunto de dados



Como o $k-means$ funciona?

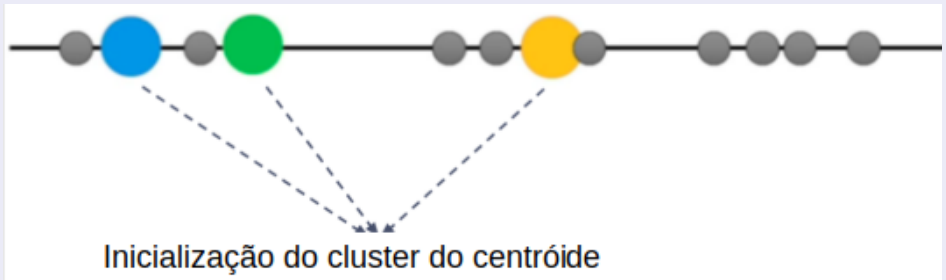
Passo 1 - determine o valor k

- O valor k representa o número de *clusters*
- Selecionamos $k = 3$
- Queremos identificar 3 *clusters*
- Existe alguma maneira de determinar o valor de k?

Como o $k-means$ funciona?

Passo 2 - Selecione aleatoriamente 3 centroides distintos

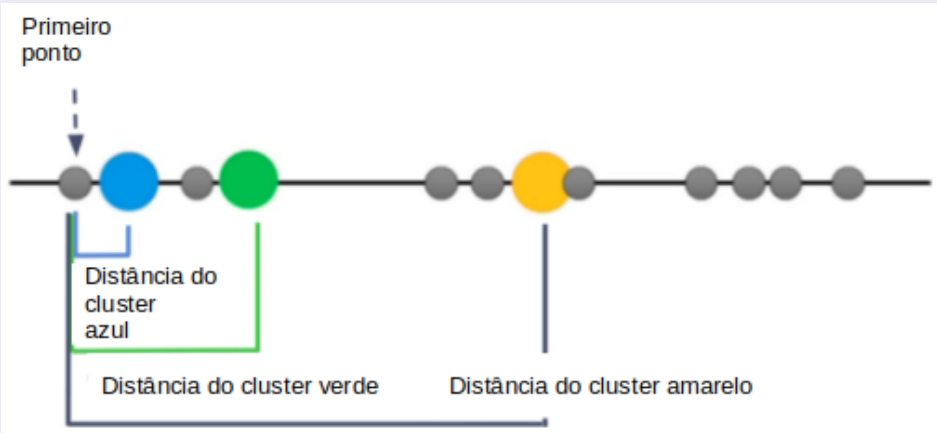
- Novos pontos de dados como inicialização do *cluster*
- Tentativa 1 - k é igual a 3
- Será inicializado com 3 centróides



Como o $k-means$ funciona?

Etapa 3 - calcule a distância (distância euclidiana)

- Entre cada ponto e o centróide
- Calcule a distância entre o primeiro ponto e o centróide



Como o $k-means$ funciona?

Medidas de distância

- Determina como a similaridade de dois elementos é calculada e influencia a forma dos *clusters*
  1. A distância euclidiana (distância de 2 normas) é dada por

$$d(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

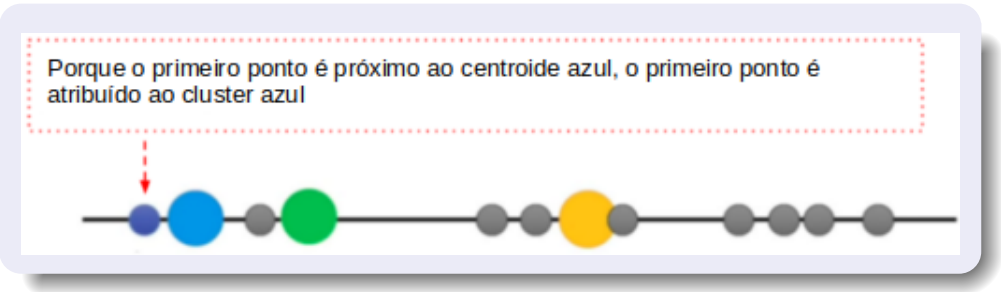
  2. A distância de Manhattan (norma-1) é dada por

$$d(x, y) = \sqrt{\sum_{i=1}^{p} |x_i - y_i|^2}$$

Como o $k-means$ funciona?

Etapa 4 - atribuir cada ponto ao cluster mais próximo

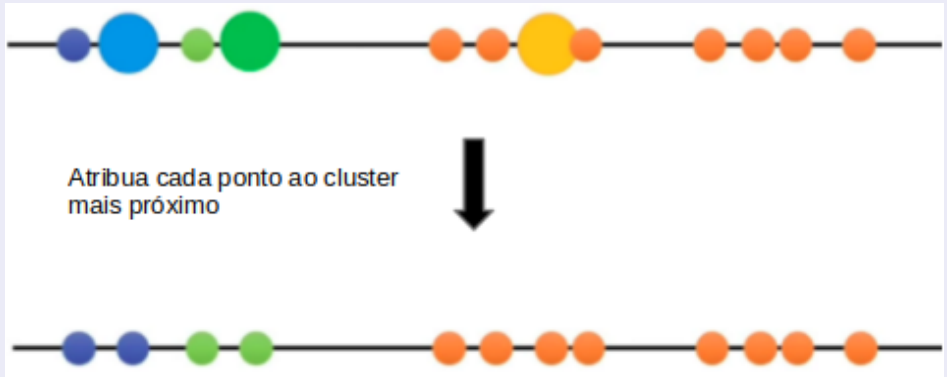
- Calcule a distância entre o primeiro ponto e o centróide



Como o $k-means$ funciona?

Etapa 4 - atribuir cada ponto ao cluster mais próximo

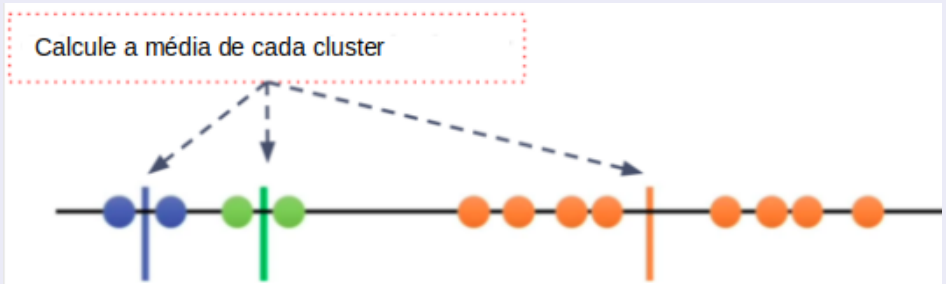
- Faça o mesmo tratamento para o outro ponto não rotulado, até obter a configuração



Como o $k-means$ funciona?

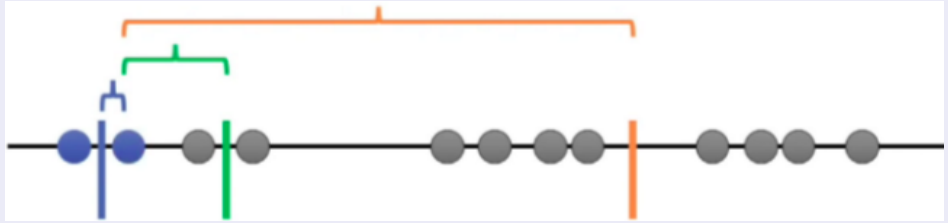
Passo 5. calcule a média de cada cluster como novo centróide

- Atualize o centróide com a média de cada *cluster*



Como o $k-means$ funciona?

Passo 6 - repita as etapas 3 a 5 com o novo centro do *cluster*



Como o $k-means$ funciona?

Passo 6 - repita as etapas 3 a 5 com o novo centro do *cluster*

- Repita até parar:
  - Convergência (sem mais alterações)
  - Número máximo de iterações.
  - Como o agrupamento não mudou durante a última iteração, terminamos

Como o $k-means$ funciona?

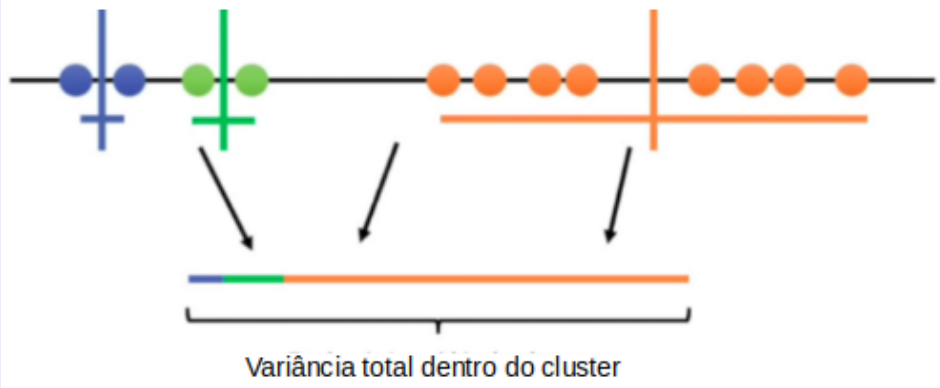
Este processo foi concluído?

- Não...
- O algoritmo *k-means* escolhe o centróide que minimiza a inércia ou o critério de soma dos quadrados dentro do cluster
- Como avaliar os resultados desse agrupamento?

Como o $k-means$ funciona?

Passo 7 - calcule a variância de cada *cluster*

- Repita até parar
  - Convergência (sem mais alterações)
  - Número máximo de iterações.
  - Como o agrupamento não mudou durante a última iteração, terminamos



Como o $k-means$ funciona?

O *k-means clustering*

- Não pode *ver* o melhor agrupamento/*cluster*
- A única opção é acompanhar esses agrupamentos e sua variância total
- E fazer tudo de novo com diferentes pontos de partida

Como o $k-means$ funciona?

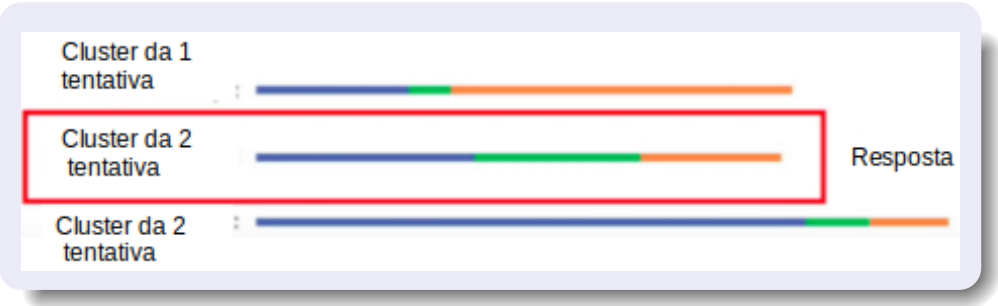
Passo 8 - repita os passos 2 a 7

- Por exemplo - tentativa 3 com diferentes centróides aleatórios

Como o $k-means$ funciona?

Repita até parar

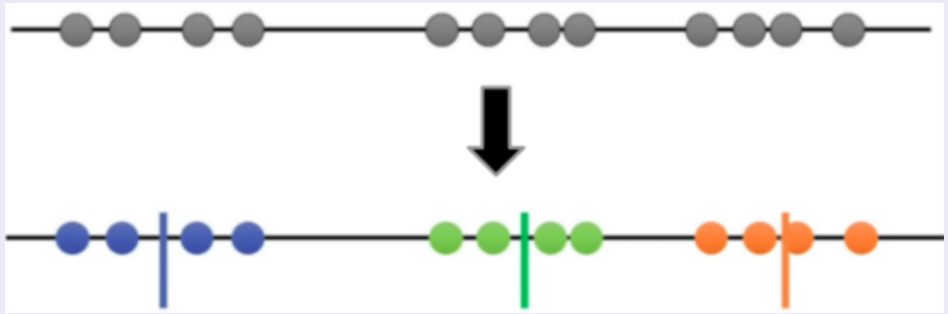
- Até obter a menor soma de variância e escolher os *clusters* como nosso resultado



Como o $k-means$ funciona?

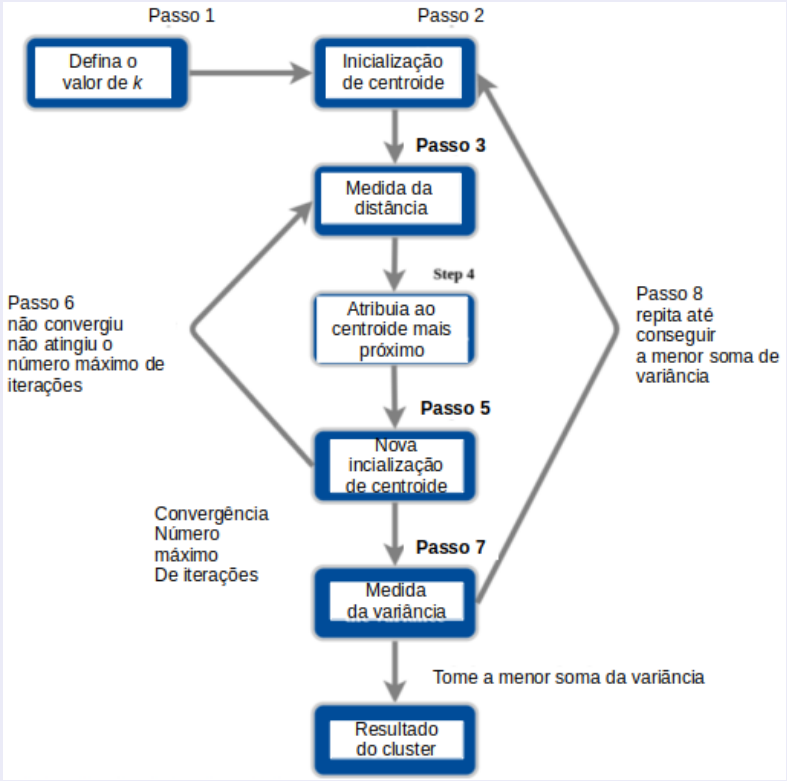
Este processo foi concluído?

- Sim...
- O resultado do agrupamento é



Como o $k-means$ funciona?

Fluxograma do *k-means clustering*



Sumário

Exemplo $k-means$

Problema

- Agrupe os oito pontos de dados (com (x, y) representando localizações) em três grupos
  - A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8 (4, 9)
- Os centros de agrupamento iniciais são: A1(2, 10), A4(5, 8) e A7(1, 2)
- A função de distância entre dois pontos a = (x1, y1) e b = (x2, y2) é definida como

$$P(a, b) = |x2 - x1| + |y2 - y1|$$

- Use o *k-means* para encontrar os três centros de *cluster* após a segunda iteração

Exemplo $k-means$

Passo 1

- Calculamos a distância de cada ponto de cada um dos centros dos três grupos
- A distância é calculada usando a função de distância dada
- A ilustração mostra o cálculo da distância entre o ponto A1(2, 10) e cada centro dos três clusters

Exemplo $k-means$

Passo 1

- Cálculo da distância entre A1(2, 10) e C1(2, 10)

$$\begin{aligned} P(A1, C1) &= |x2 - x1| + |y2 - y1| \\ &= |2 - 2| + |10 - 10| \\ &= 0 \end{aligned}$$

Exemplo $k-means$

Passo 1

- Calculando a distância entre A1(2, 10) e C2(5, 8)

$$\begin{aligned} P(A1, C2) &= |x2 - x1| + |y2 - y1| \\ &= |5 - 2| + |8 - 10| \\ &= 3 + 2 \\ &= 5 \end{aligned}$$

Exemplo $k-means$

Passo 1

- Calculando a distância entre A1(2, 10) e C3(1, 2)

$$\begin{aligned} P(A1, C3) &= |x2 - x1| + |y2 - y1| \\ &= |1 - 2| + |2 - 10| \\ &= 1 + 8 \\ &= 9 \end{aligned}$$

Exemplo $k-means$

De maneira semelhante

- Calculamos a distância de outros pontos de cada um dos centros dos três *clusters*
- Desenhamos uma tabela com todos os resultados
- Usando a tabela, decidimos qual ponto pertence a qual cluster
- O ponto dado pertence aquele *clusters* cujo centro está mais próximo dele

Exemplo $k-means$

Usando a tabela

- Decidimos qual ponto pertence a qual cluster

Pontos	Distância do centro (2,10) do cluster 1	Distância do centro (5,8) do cluster 2	Distância do centro (1,2) do cluster 3	Pontos pertencem ao cluster
A1(2, 10)	0	5	9	C1
A2(2, 5)	5	6	4	C3
A3(8, 4)	12	7	9	C2
A4(5, 8)	5	0	10	C2
A5(7, 5)	10	5	9	C2
A6(6, 4)	10	5	7	C2
A7(1, 2)	9	10	0	C3
A8(4, 9)	3	2	10	C2

- O ponto dado pertence àquele *cluster* cujo centro está mais próximo dele

Exemplo $k-means$

Os novos clusters são

- *Clusters*-01: A1(2, 10)
- *Clusters*-02: A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A8(4, 9)
- *Clusters*-03: A2(2, 5), A7(1, 2)

Exemplo $k-means$

Recalculamos os novos *clusters*

- O novo centro do *clusters* é calculado a média de todos os pontos contidos nesse cluster
- Para *cluster*-01
  - Temos apenas um ponto A1(2, 10) no *clusters*-01
  - Portanto, o centro do cluster permanece o mesmo
- Para *cluster*-02
  - Centro do *cluster*-02 = ((8 + 5 + 7 + 6 + 4)/5, (4 + 8 + 5 + 4 + 9)/5) = (6, 6)
- Para *cluster*r-03
  - Centro do *cluster*-03 = ((2 + 1)/2, (5 + 2)/2) = (1,5, 3,5)
- O passo 1 está concluído

Exemplo $k-means$

Passo 2

- Calculamos a distância de cada ponto de um dos centros dos três grupos
- A distância é calculada usando a função de distância fornecida.

Exemplo $k-means$

Passo 2

- Cálculo da distância entre o ponto A1(2, 10) e cada centro dos três *clusters* A1(2, 10) e C1(2, 10)

$$\begin{aligned} P(A1, C1) &= |x2 - x1| + |y2 - y1| \\ &= |2 - 2| + |10 - 10| \\ &= 0 \end{aligned}$$

Exemplo $k-means$

Passo 2

- Calculando a distância entre A1(2, 10) e C2(6, 6)

$$\begin{aligned} P(A1, C2) &= |x2 - x1| + |y2 - y1| \\ &= |6 - 2| + |6 - 10| \\ &= 4 + 4 \\ &= 8 \end{aligned}$$

Exemplo $k-means$

Passo 2

- Calculando a distância entre A1(2, 10) e C3(1,5, 3,5)

$$\begin{aligned} P(A1, C3) &= |x2 - x1| + |y2 - y1| \\ &= |1,5 - 2| + |3,5 - 10| \\ &= 0,5 + 6,5 \\ &= 7 \end{aligned}$$

Exemplo $k-means$

Passo 2

- De maneira semelhante, calculamos a distância de outros pontos de cada um dos centros dos três clusters
- Desenhamos uma tabela com todos os resultados

Pontos	Distância do centro (2,10) do cluster 1	Distância do centro (5,8) do cluster 2	Distância do centro (1,2) do cluster 3	Pontos pertencem ao cluster
A1(2, 10)	0	5	9	C1
A2(2, 5)	5	6	4	C3
A3(8, 4)	12	7	9	C2
A4(5, 8)	5	0	10	C2
A5(7, 5)	10	5	9	C2
A6(6, 4)	10	5	7	C2
A7(1, 2)	9	10	0	C3
A8(4, 9)	3	2	10	C2

Exemplo $k-means$

Passo 2

- Usando a tabela, decidimos qual ponto pertence a qual *cluster*
- O ponto dado pertence àquele cluster cujo centro está mais próximo dele
- Os novos clusters são
  - Cluster-01: A1(2, 10), A8(4, 9)
  - Cluster-02: A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4)
  - Cluster-03: A2(2, 5), A7(1, 2)

Exemplo $k-means$

Passo 2

- Recalculamos os novos centros dos clusters
- É calculado tomando a média de todos os pontos contidos nesse *cluster*
  - *Cluster*-01 $= ((2 + 4)/2, (10 + 9)/2) = (3, 9.5)$
  - *Cluster*-02 $= ((8 + 5 + 7 + 6)/4, (4 + 8 + 5 + 4)/4) = (6.5, 5.25)$
  - *Cluster*-03 $= ((2 + 1)/2, (5 + 2)/2) = (1.5, 3.5)$
- Concluimos o Passo-02

Exemplo $k-means$

Passo 2

- Recalculamos os novos centros dos clusters
- É calculado tomando a média de todos os pontos contidos nesse *cluster*
  - *Cluster*-01 $= ((2 + 4)/2, (10 + 9)/2) = (3, 9.5)$
  - *Cluster*-02 $= ((8 + 5 + 7 + 6)/4, (4 + 8 + 5 + 4)/4) = (6.5, 5.25)$
  - *Cluster*-03 $= ((2 + 1)/2, (5 + 2)/2) = (1.5, 3.5)$
- Concluimos o Passo-02

Exemplo $k-means$

Passo 2

- O centro dos três clusters são C1(3, 9.5), C2(6.5, 5.25), C3(1.5, 3.5)