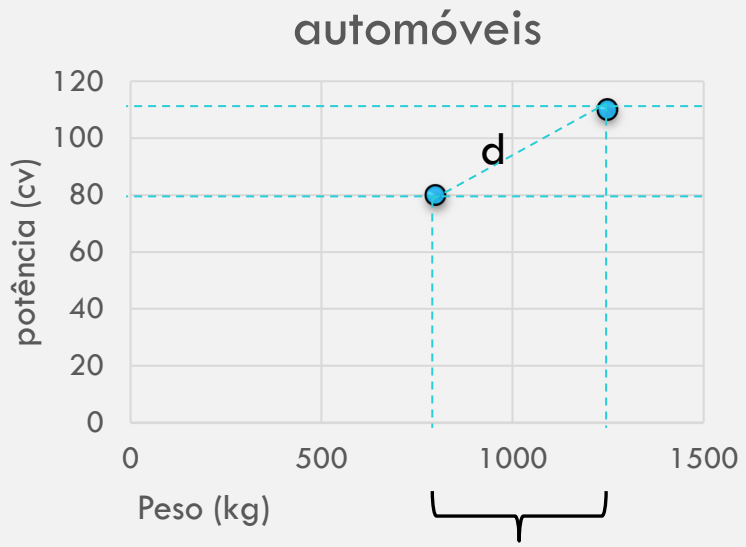
APRENDIZADO EM MÁQUINAS – 1B

Machine Learning – 1B *Prof. Dr. Waldemar Bonventi Jr.* *mar2020*

MAS, HOUSTON, TEMOS UM PROBLEMA...



$\Delta y = 110 - 80 = 30$

$\Delta x = 1250 - 800 = 450$

Pitágoras:

$d^2 = (\Delta x)^2 + (\Delta y)^2$

$= 450^2 + 30^2 = 203400$

$d = \sqrt{203400} = 451$

automóvel	peso(kg)	potencia(cv)	bagagem(litros)
A	1250	110	650
B	800	80	300
C	900	90	450
D	750	100	400
E	1100	90	350
F	1050	90	600

O atributo 'potência' "pesa" pouco na distância
A distância *d* está praticamente determinada pelo atributo 'peso'

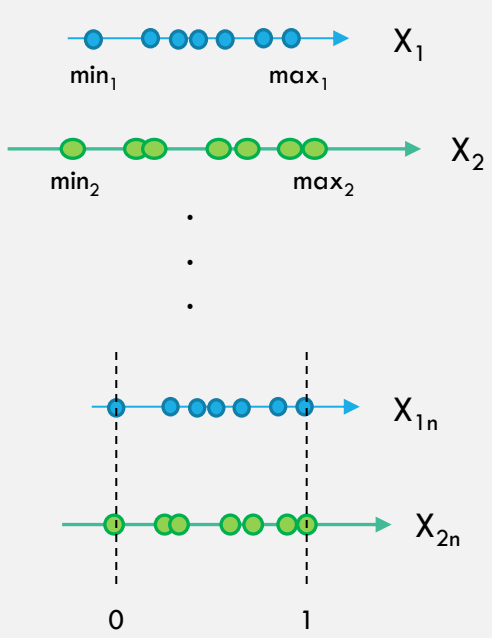
SOLUÇÃO: NORMALIZAÇÃO

automóvel	peso(kg)	potencia(cv)	bagagem(litros)
A	1250	110	650
B	800	80	300
C	900	90	450
D	750	100	400
E	1100	90	350
F	1050	90	600

Deixar cada atributo na mesma escala

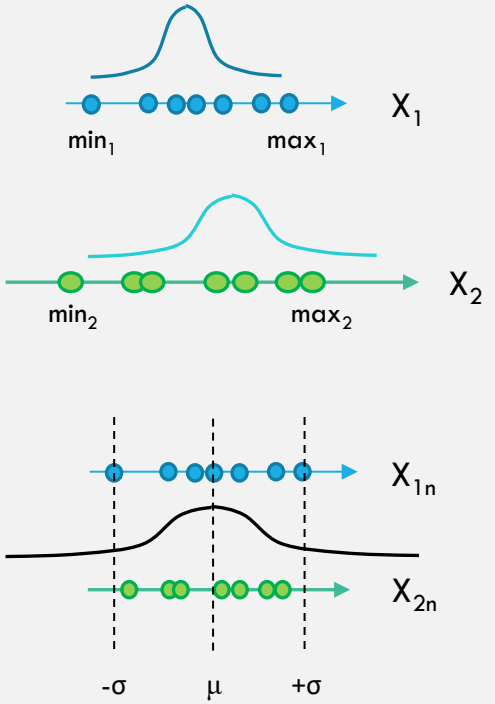
Duas maneiras: 1) escala [0 1] ou média e desvio padrão

1)



$$x_{1n} = \frac{x_1 - min_1}{max_1 - min_1}$$

2)



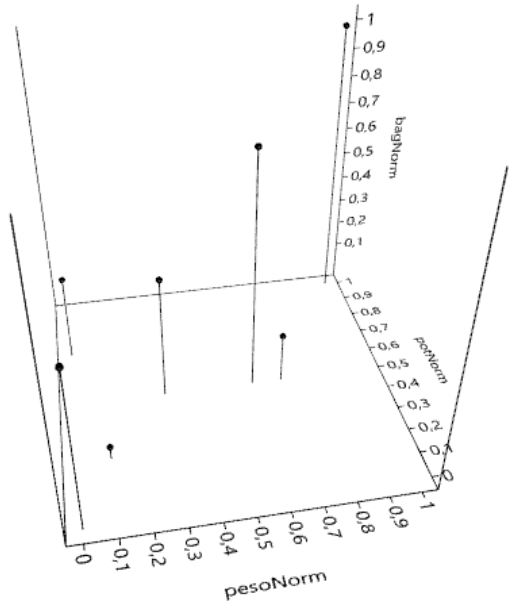
$$x_{1n} = \frac{x_1 - \mu_1}{\sigma_1}$$

NORMALIZAÇÃO [0 1]

automóvel	peso(kg)	potencia(cv)	bagagem(litros)	pesoNorm	potNorm	bagNorm
A	1250	110	650	1,000	1,000	1,000
B	800	80	300	0,100	0,250	0,000
C	900	90	450	0,300	0,500	0,429
D	750	100	400	0,000	0,750	0,286
E	1100	90	350	0,700	0,500	0,143
F	1050	90	600	0,600	0,500	0,857
G	750	70	500	0,000	0,000	0,571

Potência do automóvel C

$$x_n = \frac{90 - 70}{110 - 70} = 0{,}5$$



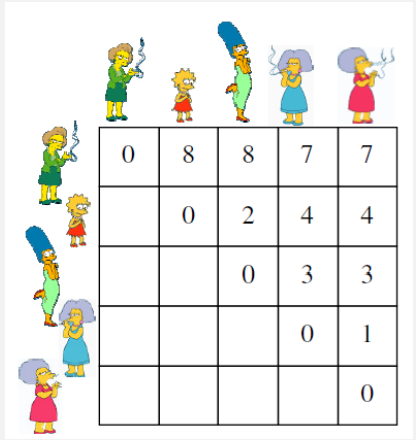
$$d^2_{CE} = (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2 =$$
$$= (0{,}300-0{,}700)^2 + (0{,}500-0{,}500)^2 + (0{,}429-0{,}143)^2$$
$$= 0{,}242$$

$$d = \sqrt{0{,}242} = 0{,}492$$

Matriz de distâncias

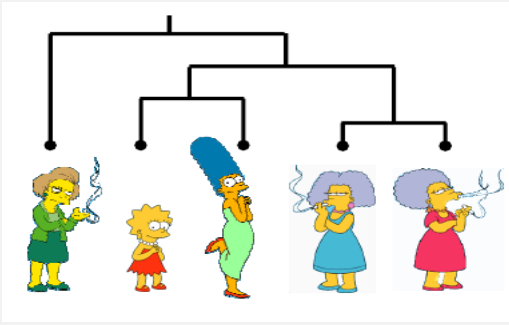
	A	B	C	D	E	F	G
A	0,000	1,540	1,032	1,254	1,037	0,656	1,478
B	1,540	0,000	0,535	0,585	0,666	1,023	0,631
C	1,032	0,535	0,000	0,416	0,492	0,523	0,600
D	1,254	0,585	0,416	0,000	0,757	0,865	0,802
E	1,037	0,666	0,492	0,757	0,000	0,721	0,961
F	0,656	1,023	0,523	0,865	0,721	0,000	0,832
G	1,478	0,631	0,600	0,802	0,961	0,832	0,000

COMO DEFINIR SEMELHANÇA? AGRUPAMENTOS

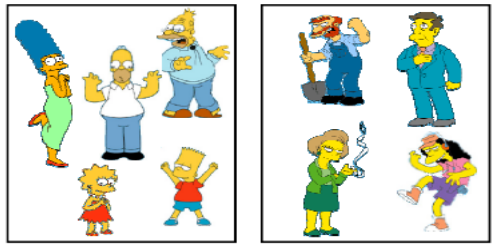


0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0

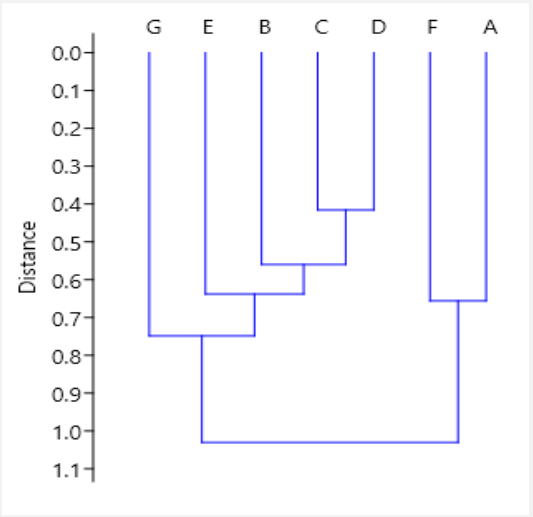
Hierárquico

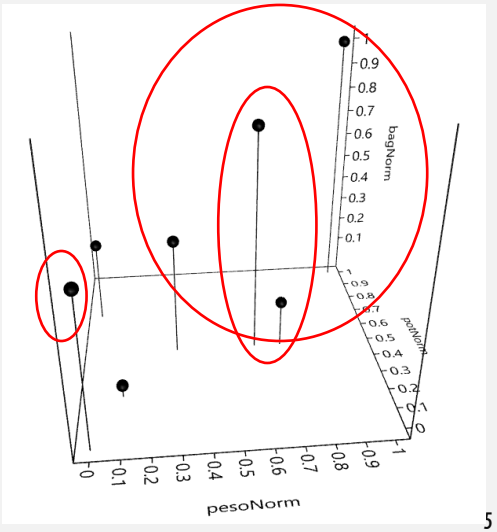


Particional



	A	B	C	D	E	F	G
A	0,000	1,540	1,032	1,254	1,037	0,656	1,478
B	1,540	0,000	0,535	0,585	0,666	1,023	0,631
C	1,032	0,535	0,000	0,416	0,492	0,523	0,600
D	1,254	0,585	0,416	0,000	0,757	0,865	0,802
E	1,037	0,666	0,492	0,757	0,000	0,721	0,961
F	0,656	1,023	0,523	0,865	0,721	0,000	0,832
G	1,478	0,631	0,600	0,802	0,961	0,832	0,000



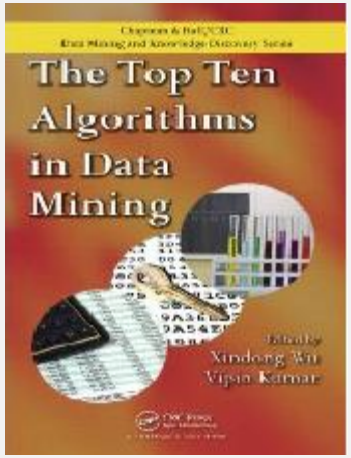


CLUSTERING PARTICIONAL: ALGORITMO K-MEANS

- Um dos algoritmos mais clássicos da área de mineração de dados em geral
- Algoritmo das k-médias ou k-means
- Listado entre os *Top 10 Most Influential Algorithms in DM*

• Wu, X. and Kumar, V. (Editors), The Top Ten Algorithms in Data Mining, CRC Press, 2009

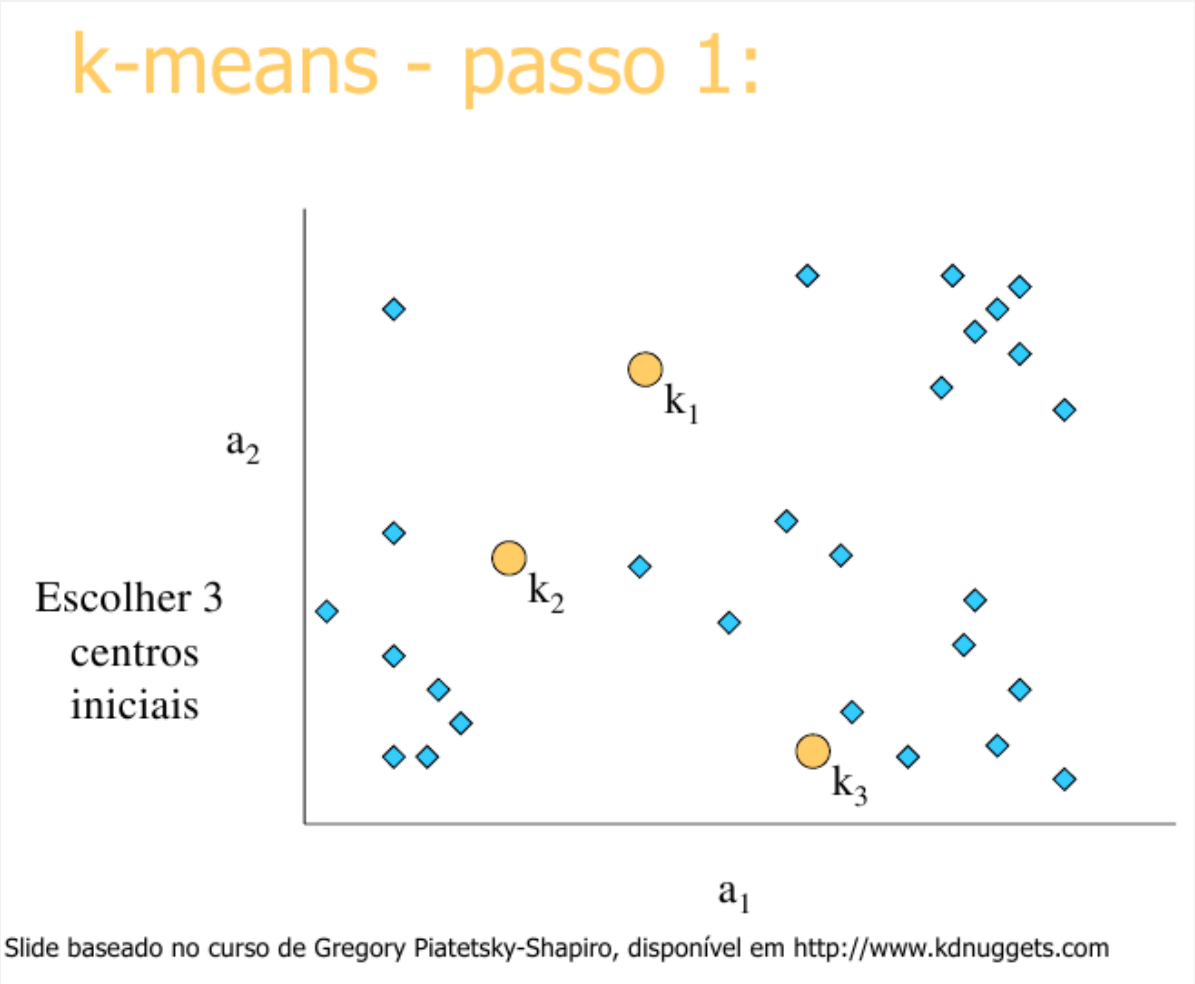
• X. Wu et al., "Top 10 Algorithms in Data Mining", Knowledge and Info. Systems, vol. 14, pp. 1-37, 2008



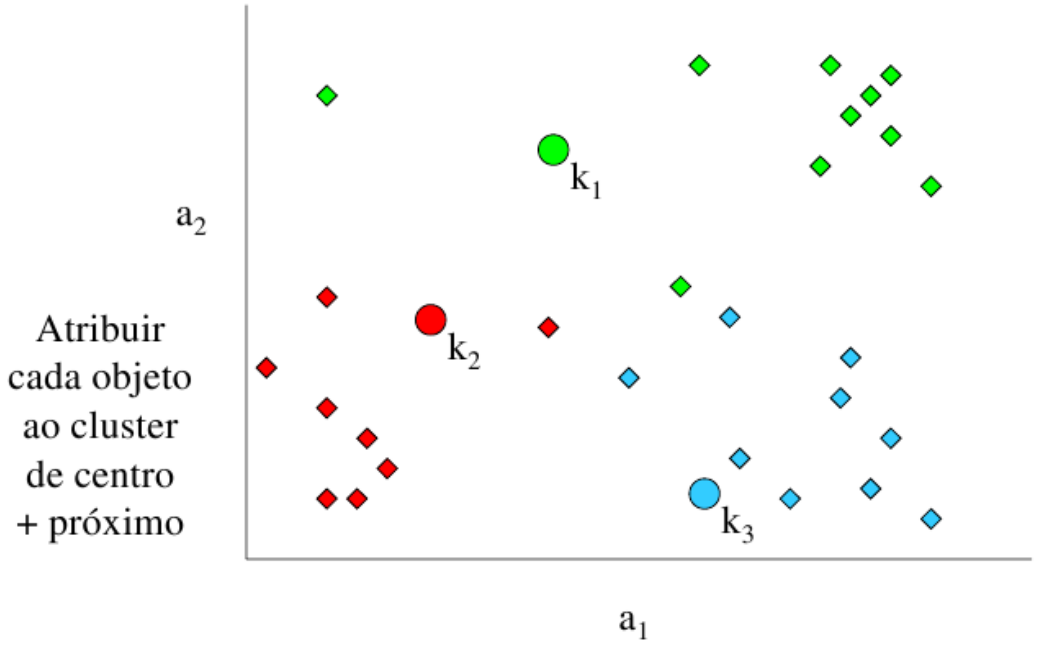
COMO FUNCIONA?

Necessário definir “em quantos grupos se deseja particionar o conjunto de dados” --> valor de *k*

No exemplo a seguir, escolhido *k*=3



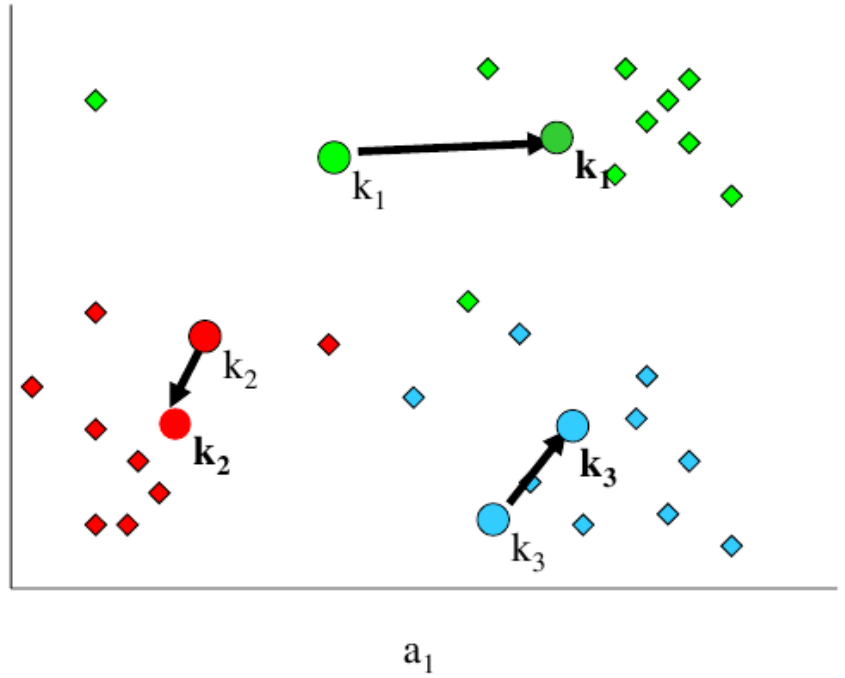
k-means - passo 2:



Atribuir cada objeto ao cluster de centro + próximo

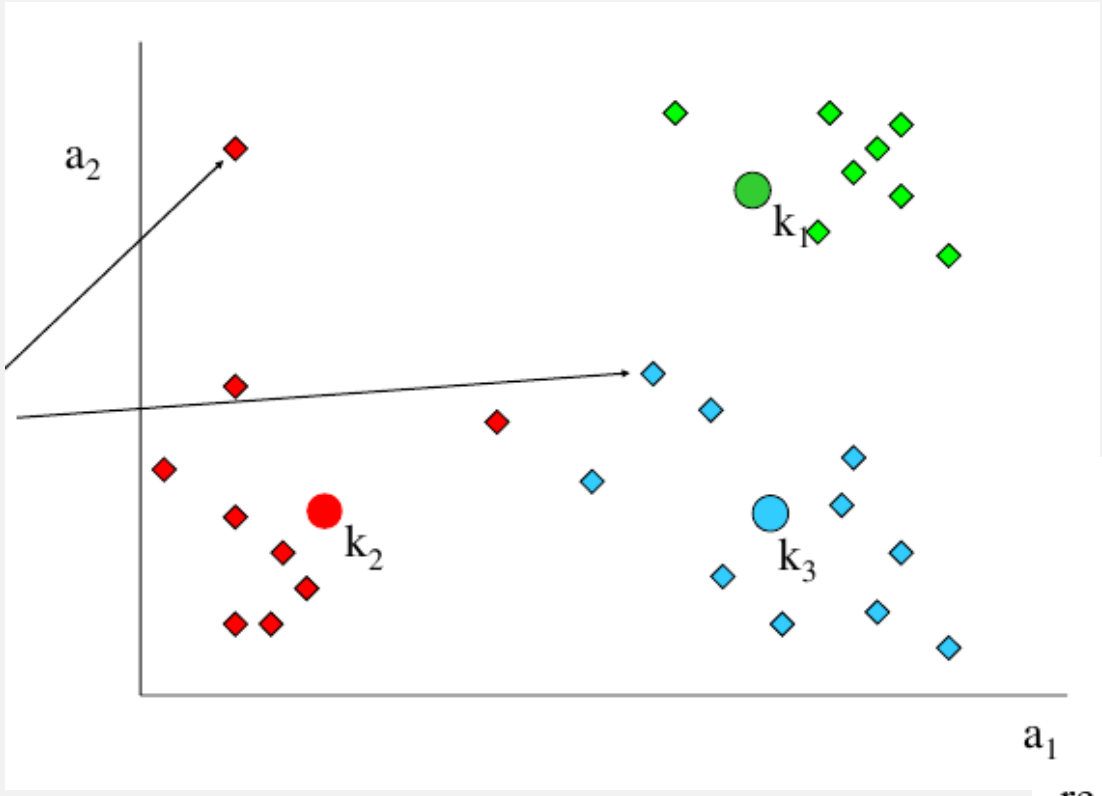
Slide baseado no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com

k-means - passo 3:

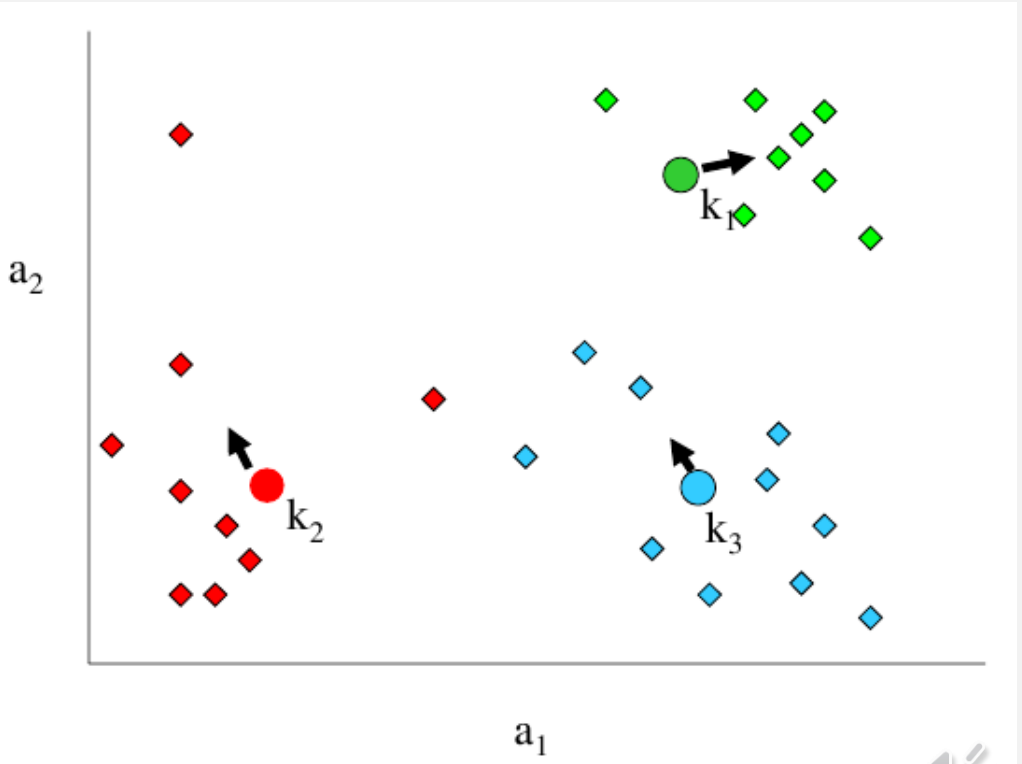


Mover cada centro para o vetor médio do cluster (centróide)

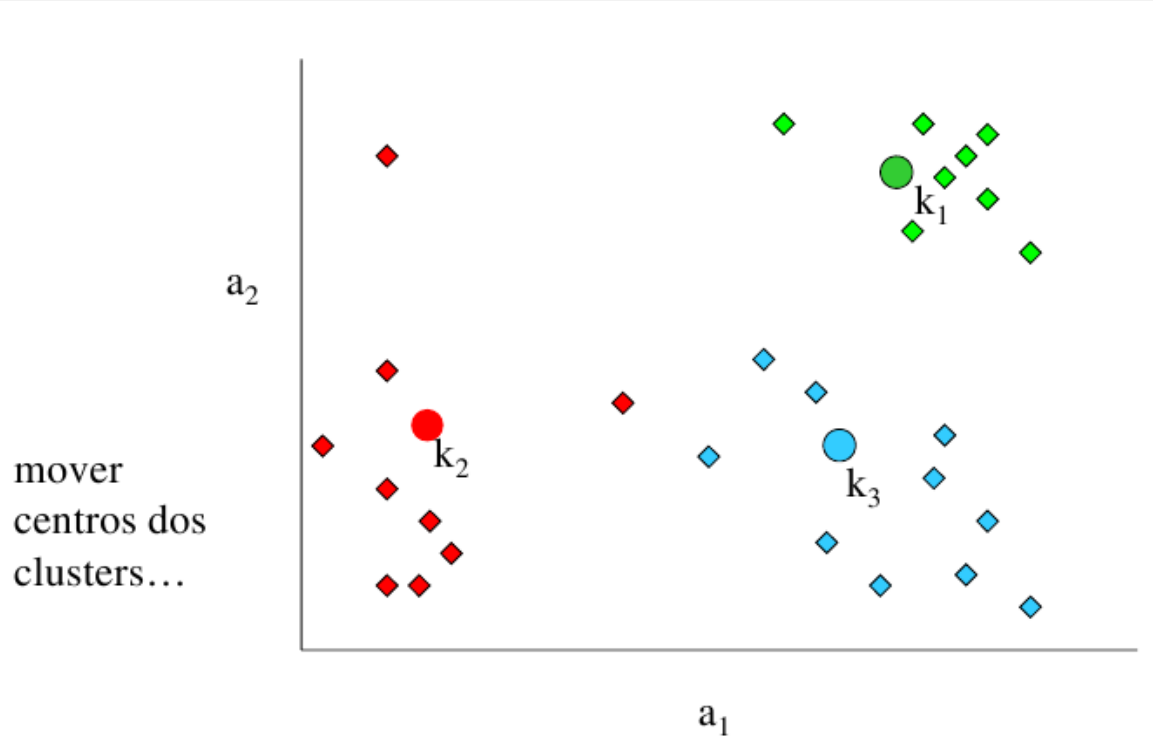
reatribuidos



re-calcular
vetores
médios



FINALIZAÇÃO



Quando não houver mais reatribuição de dados a novos centros

Ou

Quando o deslocamento de um centro for menor que um limite

- (no caso de muitos e muitos dados)

POR SE TRATAR DE APRENDIZADO NÃO-SUPERVISIONADO

A interpretação do que significa cada grupo é feita por um “agente inteligente”

Que pode, justamente pela similaridade, descrever cada grupo pelos atributos que apresentam valores próximos entre si

Mas a atribuição de cada objeto a um grupo foi feita sem supervisão

Apenas observando a estrutura dos dados

automovel	peso(kg)	potencia(cv)	bagagem(litros)	Grupo
A	1250	110	650	g2
B	800	80	300	g1
C	900	90	450	g4
D	750	100	400	g4
E	1100	90	350	g1
F	1050	90	600	g3
G	750	70	500	g3

MAIS UM PROBLEMINHA, HOUSTON...

Como resolver qual o melhor valor de *k*,

Ou seja

Qual a melhor partição que "separa bem" os dados?

Critérios:

- Mínima distância intracluster
- Máxima distância intercluster

Próximos slides...