Exemplo de Kmeans RESOLVIDO.

Questão 1

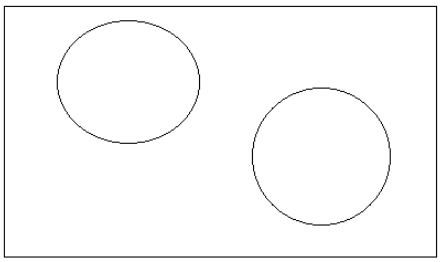
Considere o seguinte conjunto de 7 instâncias da base de dados abaixo.
A1 é o atributo 1 da base de dados
A2 é o atributo 2 da base de dados
E nesta base de dados temos 7 instâncias, representadas pelas linhas.

1) Aplique o algoritmo K-means para determinar uma partição adequada desses dados em 2 grupos. Realize **duas** execuções do algoritmo, partindo dos exemplos (centroides) **1 e 6** (estão marcados na tabela com a cor vermelha). Aplique a distância de Manhattan.

Exemplos	A1	A2
1	1,0	1,0
2	1,5	2,0
3	3,0	4,0
4	5,0	7,0
5	3,5	5,0
6	4,5	5,0
7	3,5	4,5

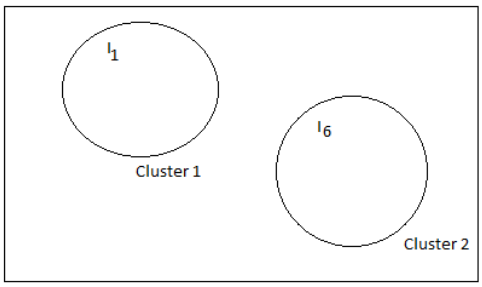
Passo 1 do algoritmo Kmeans:

Sabemos que temos que criar dois grupos usando o algoritmo **Kmeans**:



Exemplo de criação de 2 grupos usando o algoritmo Kmeans

O algoritmo inicia colocando a **instância 1 no cluster 1** e a **instância 6 no grupo 2**. Veja que estas instâncias são os centroides iniciais dos meus 2 grupos. Estes dois centroides, na primeira rodada do algoritmo, são selecionados de forma aleatória. No caso desta questão eu forneci como informação inicial (marcado em vermelho lá em cima).
Dito isso, temos então:



Exemplo de criação de 2 grupos usando o algoritmo Kmeans

Coloquei nome nos clusters (cluster 1 e 2) para facilitar aqui a referência a eles, ok? E coloquei também que a Instância 1 está em um grupo e que a Instância 6 está em outro grupo.

Agora precisamos saber em qual cluster as instâncias 2, 3, 4, 5 e 7, descritas na tabela acima , estarão. Ou seja, precisamos saber se:

- ✓ a instância 2 se parece mais com a instância 1 ou com a instância 6?
- ✓ a instância 3 se parece mais com a instância 1 ou com a instância 6?
- ✓ a instância 4 se parece mais com a instância 1 ou com a instância 6?
- ✓ a instância 5 se parece mais com a instância 1 ou com a instância 6?
- ✓ a instância 7 se parece mais com a instância 1 ou com a instância 6?

Para saber a semelhança entre as instâncias, usamos o conceito de distância. Quanto menor a distância, mais parecidas elas são entre si. Na literatura, temos várias distâncias: euclidiana, Manhattan, etc...
Nesta questão, pedi que utilizassem a distância de Manhattan (equação abaixo). Ela é mais simples. Não tem quadrado e nem raiz quadrado..

$$d(x,y) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_p - y_p|$$

Então vamos lá:
Vamos descobrir se a instância 2 se parece mais com a 1 ou com a 6
Sabemos que os valores das instâncias são:

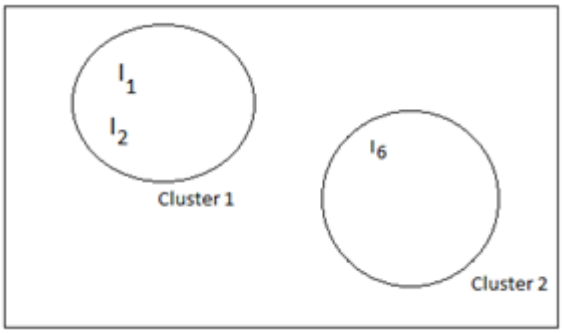
Instância	A1	A2
1	**1,0**	**1,0**
2	**1,5**	**2,0**

Distância (2,1) = |1,5-1| + |2-1| = 0,5+1 =**1,5**

Instância	A1	A2
6	**4,5**	**5,0**
2	**1,5**	**2,0**

Distância (2,6) = |1,5-4,5| + |2-5| = 3+3 =6

Ou seja, vemos que a distância entre as instâncias 2 e 1 é de **1,5** e a distância entre as instâncias 2 e 6 é 6.
Então podemos concluir que a instância 2 ficará no mesmo grupo que a distância 1, por ser mais semelhante (menor distância).
Veja:



Inserindo a instância 2 no cluster 1

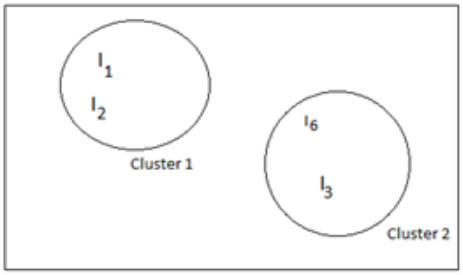
Agora vamos fazer o mesmo para a instância 3. Ou seja, com quem a instância 3 se parece mais? Com a 1 ou com a 6?

Instância	A1	A2
1	**1,0**	**1,0**
3	3,0	4,0
Distância (3,1) = |3-1| + |4-1| = 2+3 =**5**		

Instância	A1	A2
6	**4,5**	**5,0**
3	3,0	4,0
Distância (3,6) = |3-4,5| + |4-5| = 1,5+1 =2,5		

Ou seja, vemos que a distância entre as instâncias 3 e 1 é **5** e a distância entre as instâncias 3 e 6 é 2,5.

Então podemos concluir que a instância 3 ficará no mesmo grupo que a distância 6, por ser mais semelhante (menor distância).
Veja:



Inserindo a instância 3 no cluster 2

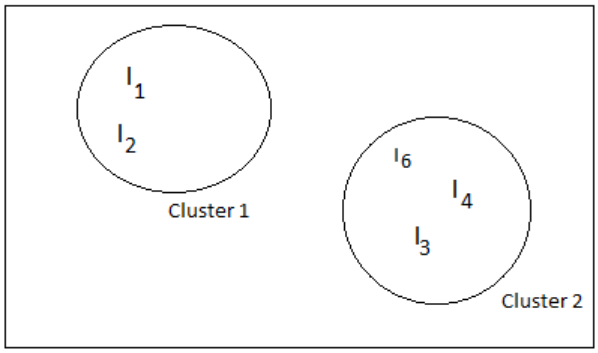
Faremos o mesmo para a instância 4. Ou seja, com quem a instância 4 se parece mais? Com a 1 ou com a 6?

Instância	A1	A2
1	**1,0**	**1,0**
4	5,0	7,0
Distância (3,1) = |5-1| + |7-1| = 4+6 =**10**		

Instância	A1	A2
6	**4,5**	**5,0**
4	5,0	7,0
Distância (3,6) = |5-4,5| + |7-5| = 0,5+2 =2,5		

Ou seja, vemos que a distância entre as instâncias 4 e 1 é **10** e a distância entre as instâncias 4 e 6 é 2,5.

Então podemos concluir que a instância 4 também ficará no mesmo grupo que a distância 6, por ser mais semelhante (menor distância).
Veja:



Inserindo a I4 no cluster 2

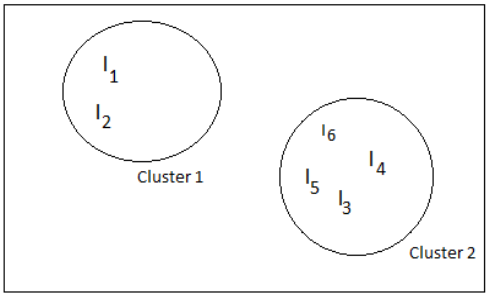
Faremos o mesmo para a instância 5. Ou seja, com quem a instância 5 se parece mais? Com a 1 ou com a 6?

Instância	A1	A2
1	**1,0**	**1,0**
5	3,5	5,0
Distância (5,1) = |3,5-1| + |5-1| = 2,5+4 =**6,5**		

Instância	A1	A2
6	**4,5**	**5,0**
5	3,5	5,0
Distância (5,6) = |3,5-4,5| + |5-5| = 1+0 =1		

Ou seja, vemos que a distância entre as instâncias 5 e 1 é **6,5** e a distância entre as instâncias 5 e 6 é 1.

Então podemos concluir que a instância 5 também ficará no mesmo grupo que a distância 6, por ser mais semelhante (menor distância).



Inserindo a I5 no cluster 2

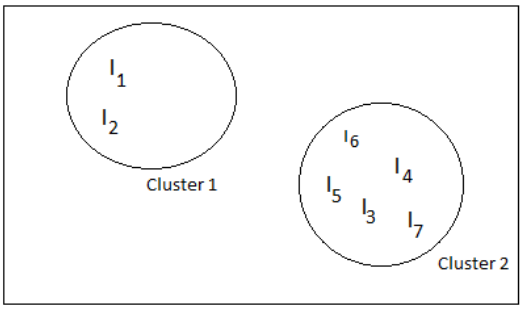
E finalmente, faremos o mesmo para a instância 7. Ou seja, com quem a instância 7 se parece mais? Com a 1 ou com a 6?

Instância	A1	A2
1	**1,0**	**1,0**
7	3,5	4,5
Distância (5,1) = |3,5-1| + |4,5-1| = 2,5+3,5 =**6**		

Instância	A1	A2
6	**4,5**	**5,0**
7	3,5	4,5
Distância (5,6) = |3,5-4,5| + |4,5-5| = 1+0,5 =1,5		

Ou seja, vemos que a distância entre as instâncias 7 e 1 é **6** e a distância entre as instâncias 7 e 6 é 1,5.

Então podemos concluir que a instância 7 também ficará no mesmo grupo que a distância 6, por ser mais semelhante (menor distância).

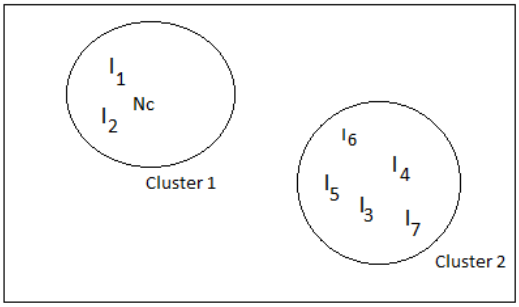


Inserindo a I7 no cluster 2

Pronto! A primeira rodada do algoritmo está pronta! Já temos os dois primeiros grupos formados.
Mas precisamos agora ir para a segunda rodada do algoritmo. Precisamos ajustar os centroides e recalcular tudo novamente!
Aqui vou ajudar vocês a recalcularem os centroides.
Lembrando que para recalcular o centroide, é só calcular a média aritmética de todos os pontos dos grupos. Vamos calcular?

	Cluster 1	
I1 I2	1 1,5	1 2
Média	(1+1,5)/2= **1,25**	(1+2)/2 = **1,5**

Portanto o novo centroide do cluster 1 agora será o ponto (1,25 ; 1,5).
Veja que o centroide, pessoal, tem o mesmo número de atributos do conjunto original, ou seja, tem 2 atributos. O primeiro tem valor 1,25 e o segundo tem valor 1.5
Podemos colocar este novo centroide no cluster 1 de uma vez:

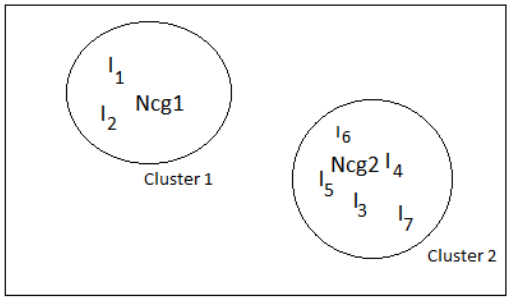


Mostrando que o cluster 1 agora tem novo centroide (1,25; 1,5)

Vamos agora recalcular o novo centroide do cluster 2:

	Cluster 2	
I6	4,5	5
I3	3	4
I4	5	7
I5	3,5	5
I7	3,5	4,5
Média	(4,5+3+5+3,5+3,5)/5= **3,9**	(5+4+7+5+4,5)/5 = **5,1**

Portanto o novo centroide do cluster 2 agora será o ponto (3,9 ; 5,1).
Podemos colocar este novo centroide no cluster 2 de uma vez:



Mostrando que o cluster 2 agora tem novo centroide (3,9; 5,1)
Ncg1 = Novo centroide do grupo 1
Ncg2 = Novo centroide do grupo 2

Agora precisamos fazer os cálculos todos novamente! Precisamos saber se as instâncias 1, 2, ...7 estão com o centroide do grupo 1 ou o centroide do grupo 2! Pense que como os novos centroides foram gerados, pode ser que as instâncias agora mudem de local! E normalmente isso ocorrerá! Por quê? Porque estes agrupamentos que foram feitos foram obtidos com os centroides selecionados de forma aleatória! E pode ser que esta seleção não tenha sido boa o suficiente! Ou seja, quem disse que a instância 1 e 6 realmente estavam em grupos separados, entende?
Agora precisamos recalcular tudo novamente e ver como ficarão.

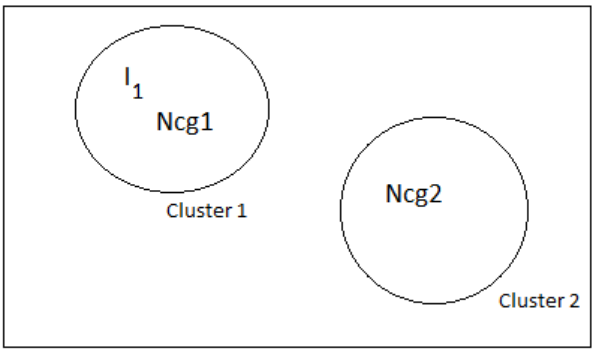
Vou fazer apenas o primeiro passo, ok?
Vamos ver com quem a instância 1, está? Será se a instância 1 se parece mais com o Ncg1 ou com o Ncg2?

Instância	**A1**	**A2**
Ncg1	1,25	1,5
1	1	1
Distância (2,1) = |1-1,25| + |1-1,5| = 0,25+0,5 =**0,75**		

Instância	**A1**	**A2**
Ncg2	3,9	5,1
1	1	1
Distância (2,6) = |1-3,9| + |1-5,1| = 2,9+4,1 =7		

Ou seja, vemos que a distância entre as instâncias 1 e o Ncg1 é de **0,75** e a distância entre as instâncias 1 e Ncg2 é 7

Então podemos concluir que a instância 1 continuará no mesmo grupo que estava antes... ou seja, a instância 1 está mais perto do Ncg1



Inserindo a instância 1 no cluster 1

Agora é necessário calcular a distância das instâncias 2, 3...7 aos 2 centroides e ver se elas ficarão no mesmo grupo.
Fiz de cabeça aqui e parece que todos as instâncias vão permanecer no mesmo lugar... ou seja, nenhuma instância mudará de grupo. Isso foi coincidência, tá? Neste caso, o chute dos centroides iniciais foi perfeito! E claro que isso não acontece sempre.