k-means eficiente com uso de desigualdade triangular

Fabiano Fernandes dos Santos

Sumário

- Introdução
- Algoritmo tradicional
- *k*-means eficiente
- *Desigualdade triangular*
- *Exemplo*
- *Discussão dos resultados*

Fontes

- Elkan, C. **"Using the Triangle Inequality to Accelerate *k*-means"**, In *Proceedings of the Twentieth International Conference on Machine Learning* (ICML'03), pp. 147-153, 2003.
- Hamerly, G. **"Making *k*-means even faster"**, In *Proceedings of the Tenth SIAM International Conference on Data Mining* (SIAM'10), pp. 130-140, 2010.

Aspectos importantes

- Substituir a versão tradicional em qualquer cenário onde este possa ser aplicado
  - Iniciar com quaisquer protótipos
  - À partir dos mesmos protótipos iniciais, deve produzir exatamente o mesmos protótipos finais que a versão tradicional
  - Permitir o uso de qualquer medida de distância (dissimilaridade)

k-means tradicional

5

- Inicialização
  - Definir os *K* protótipos (centros) iniciais
- Para cada objeto ***X*** na base
  - Calcular a distância entre ***X*** e cada um dos centros ***C***
  - Atribuir ***X*** ao cluster do centro mais próximo
- Mover cada centro para a média dos objetos do cluster correspondente

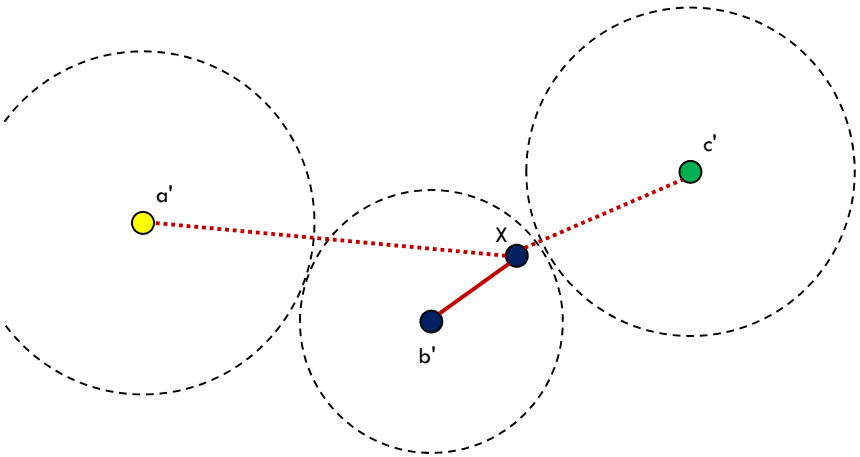
k-means tradicional

6

- Inicialização
  - Definir os *K* protótipos (centros) iniciais
- Para cada objeto ***X*** na base
  - Calcular a distância entre ***X*** e cada um dos centros ***C***
  - Atribuir ***X*** ao cluster do centro mais próximo
- Mover cada centro para a média dos objetos do cluster correspondente

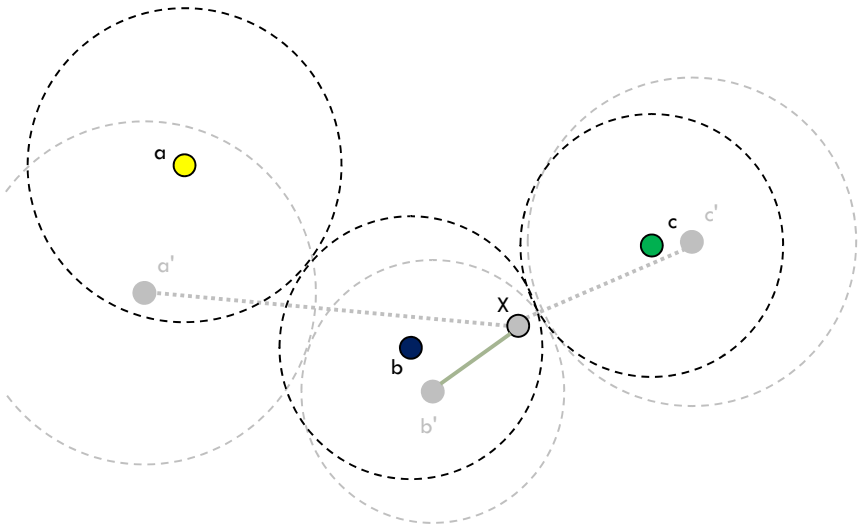
Calcular a distância entre X e cada um dos centros C

7



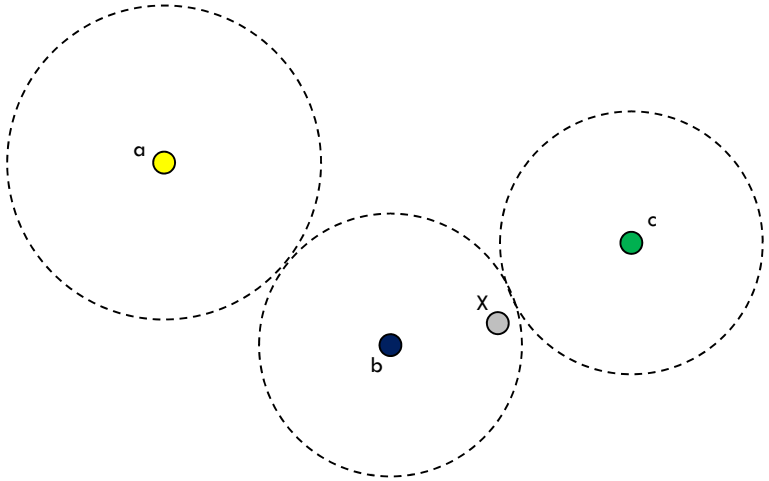
Calcular a distância entre X e cada um dos centros C

8



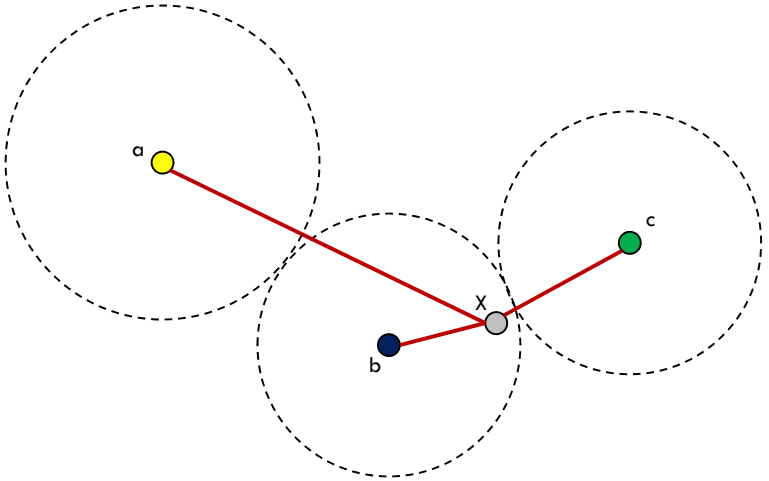
Calcular a distância entre X e cada um dos centros C

9



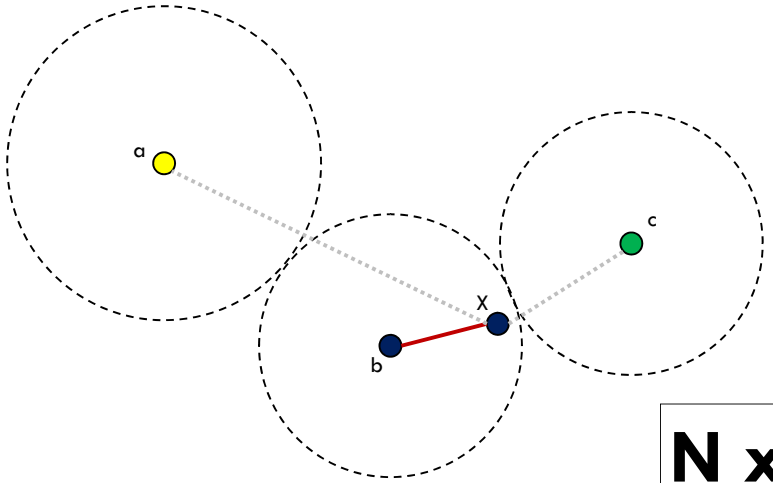
Calcular a distância entre X e cada um dos centros C

10



Calcular a distância entre X e cada um dos centros C

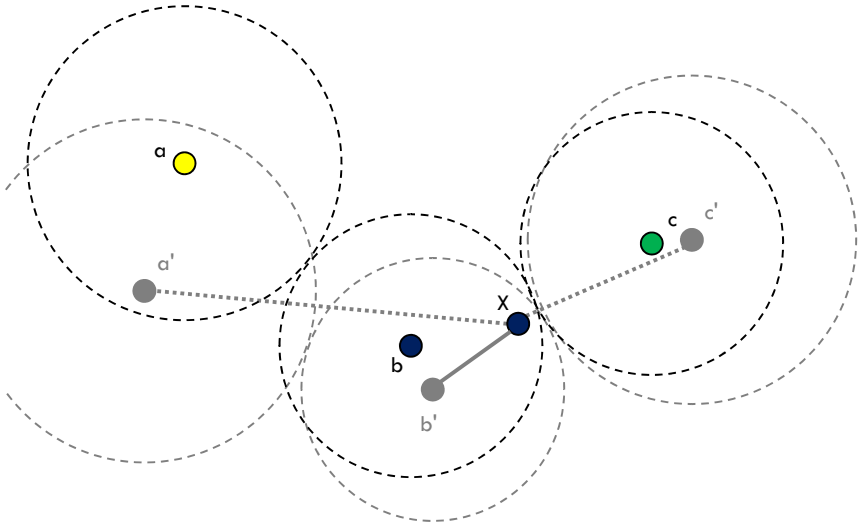
11



N x k

Calcular a distância entre X e cada um dos centros C

12



Calcular a distância entre X e cada um dos centros C

- **O objeto deve mudar de grupo?**
  - Caso negativo, não é preciso calcular a distância entre o objeto e os protótipos dos outros grupos!
- **Se mudar, para qual grupo ele deve ir?**
  - É possível evitar o cálculo da distância entre o objeto e os ***K*** protótipos

k-means eficiente

- Inicialização
  - Definir os *K* protótipos (centros) iniciais
  - Calcular a distância entre todos os protótipos
  - Calcular a distância entre ***X*** e cada um dos centros **C**
  - Atribuir ***X*** ao cluster do centro mais próximo
  - Calcular estimativas para cada objeto
- Para cada objeto ***X*** na base
  - O objeto ***X*** deve mudar de grupo?
    - Caso a resposta seja positiva, atribuir ***X*** ao cluster do centro mais próximo
    - Caso Contrário, não faça nada
- Mover cada centro para a média dos objetos do cluster correspondente
  - Calcular a distância entre todos os protótipos

k-means eficiente

- Inicialização
  - Definir os *K* protótipos (centros) iniciais
  - Calcular a distância entre todos os protótipos
  - Calcular a distância entre ***X*** e cada um dos centros **C**
  - Atribuir ***X*** ao cluster do centro mais próximo
  - Calcular estimativas para cada objeto

N x k

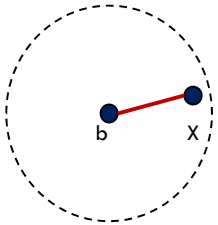
- Para cada objeto ***X*** na base
  - O objeto ***X*** deve mudar de grupo?
    - Caso a resposta seja positiva, atribuir ***X*** ao cluster do centro mais próximo
    - Caso Contrário, não faça nada
- Mover cada centro para a média dos objetos do cluster correspondente
  - Calcular a distância entre todos os protótipos

k-means eficiente

- Para cada objeto X na base
  - O objeto X deve mudar de grupo?
- Caso a resposta seja positiva, atribuir ***X*** ao cluster do centro mais próximo

Desigualdade triangular
Lema 1

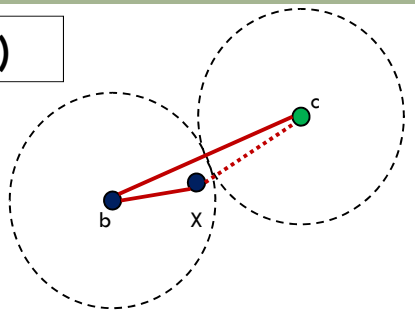
17



Desigualdade triangular
Lema 1

18

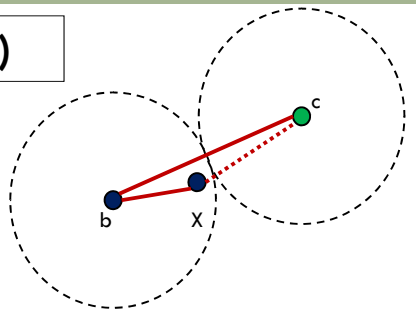
d(b,c) ≤ d(X,c) + d(X,b)



Desigualdade triangular
Lema 1

19

d(b,c) ≤ d(X,c) + d(X,b)



Seja:

d(X,b) ≤ ½ d(b,c)

d(b,c) ≥ 2d(X,b)

Então:

$2d(X,b) \leq d(b,c) \leq d(X,c) + d(X,b)$

$2d(X,b) \leq d(X,c) + d(X,b)$

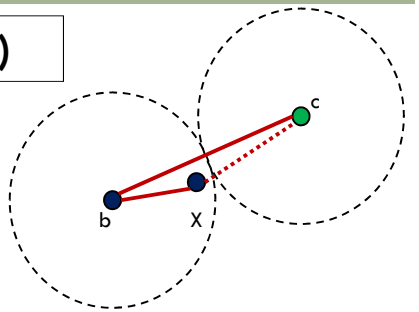
$2d(X,b) - d(X,b) \leq d(X,c) + d(X,b) - d(X,b)$

$d(X,b) \leq d(X,c)$

Desigualdade triangular
Lema 1

20

d(b,c) ≤ d(X,c) + d(X,b)



Se:

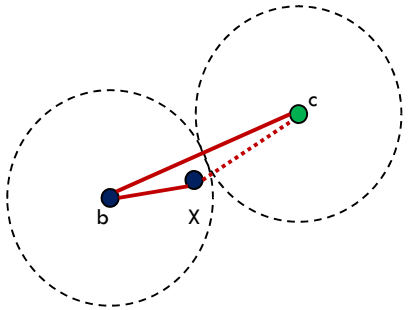
d(X,b) ≤ ½ d(b,c)

Então:

d(X,c) ≥ d(X,b)

Desigualdade triangular
Lema 2

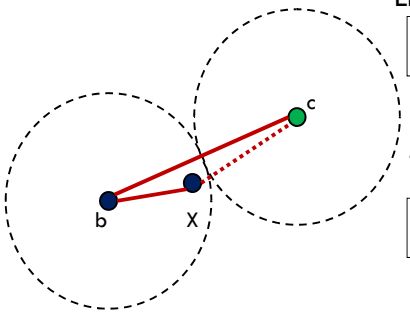
21



Desigualdade triangular
Lema 2

22

$$d(X,b) \leq d(X,c) + d(b,c)$$



Então:

$$d(X,c) \geq d(X,b) - d(b,c)$$

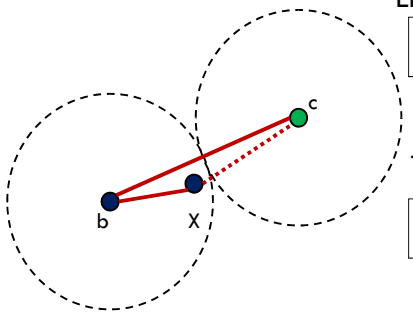
Também:

$$d(X,c) \geq 0$$

Desigualdade triangular
Lema 2

23

$$d(X,b) \leq d(X,c) + d(b,c)$$



Então:

$$d(X,c) \geq d(X,b) - d(b,c)$$

Também:

$$d(X,c) \geq 0$$

Portanto:

$$d(X,c) \geq \max\{0, d(X,b) - d(b,c)\}$$

Desigualdade triangular

24

Limite Superior (U)

Se:

$$d(X,b) \leq \tfrac{1}{2}\, d(b,c)$$

Então:

$$d(X,c) \geq d(X,b)$$

Limite Inferior (L)

$$d(X,c) \geq \max\{0, d(X,b) - d(b,c)\}$$

Desigualdade triangular

Limite Superior (U)

Se:

$d(X,b) \leq \frac{1}{2}\, d(b,c)$

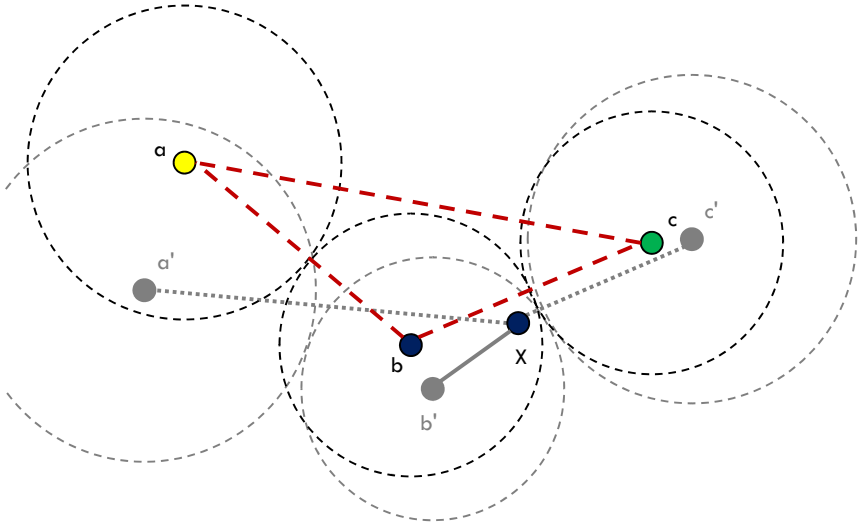
Então:

$d(X,c) \geq d(X,b)$

Limite Inferior (L)

$d(X,c) \geq \max\{0, d(X,b) - d(b,c)\}$

k-means eficiente



Desigualdade triangular

Limite Superior (U)

Seja:

$U \geq d(X,b)$

Então, se:

$U \leq \frac{1}{2}\, d(b,c)$

Continua válido que:

$d(X,c) \geq d(X,b)$

Desigualdade triangular

Limite Inferior (L)

$d(X,c) \geq \max\{0, d(X,b) - d(b,c)\}$

Seja: $d(X,b) \geq L$

Então:

$d(X,c) \geq \max\{0, L - d(b,c)\}$

k-means eficiente

29

Inicialização

- Inicialização
  - Definir os K protótipos (centros) iniciais
  - Calcular a distância entre todos os protótipos
  - Calcular a distância entre X e cada um dos centros C
  - Atribuir X ao cluster do centro mais próximo
  - Calcular estimativas para cada objeto

k-means eficiente

30

Inicialização

Para cada objeto:

- U(X): Estimativa da distância entre o objeto e o protótipo do grupo de ***X***, tal que:
  - $U(X) = d(X, c(X))$
- L(X): Estimativa da distância entre o objeto e o protótipo do grupo vizinho mais próximo, tal que:
  - $L(X) = \min_{j \neq c(X)} d(X, c(j))$

k-means eficiente

31

- Para cada protótipo **k**
  - $S(k) = \min_{j' \neq j} d(k(j'), k(j))$
- Para cada objeto ***X*** na base
  - $M = \max (\frac{1}{2} S(c(X)) , L(X))$
    - Se $U(X) \leq M$, Então não faça nada
    - Se $U(X) > M$
      - Atualizar estimativa de U(X) com valor exato
      - Se $U(X) > M$, Então atualizar ***X*** como no algoritmo tradicional
        - Atualizar U(X) e L(X) com valores exatos

k-means eficiente

32

- Mover cada centro para a média dos objetos do cluster correspondente
  - Armazenar distância que cada protótipo deslocou em relação ao passo anterior (p(***k***))
  - Calcular a distância entre todos os protótipos

k-means eficiente

33

- Atualizar estimativas U(X) e L(X) para todos os objetos tal que:
  - Determinar os dois protótipos que mais deslocaram: r' e r"
  - Para cada objeto X
    - U(X) = U(X) + p(c(X))
    - Se r' = c(X)
      - L(X) = L(X) – p(r")
    - Caso contrário
      - L(X) = L(X) – p(r')

Exemplo

34

Objeto	X	Y
1	1	2
2	2	1
3	1	1
4	2	2
5	8	9
6	9	8
7	9	9
8	8	8
9	1	15
10	2	15
11	1	14
12	2	14

K	X	Y
1	6	6
2	4	6
3	5	10

Exemplo

35

- Inicialização
  - Distância entre os protótipos

	1	2	3
1	0	2	4,12
2		0	4,12
3			0

Protótipos mais próximos

[1,2]
S(1) =2

[2,1]
S(2) =2

[3,1]
S(3) = 4,12

Exemplo

36

Objeto	X	Y	D(C1)	D(C2)	D(C3)	Cluster
1	1	2	6,403124	2,040223	8,0825	2
2	2	1	6,403124	3,026964	7,112805	2
3	1	1	7,071068	3,185465	7,122304	2
4	2	2	5,656854	2,029224	7,997795	2
5	8	9	3,605551	5,54377	4,669312	1
6	9	8	3,605551	4,661908	5,51722	1
7	9	9	4,242641	5,299841	4,760787	1
8	8	8	2,828427	5,104789	5,355261	1
9	1	15	10,29563	11,809	5,596086	3
10	2	15	9,848858	11,65391	5,123168	3
11	1	14	9,433981	10,57318	4,470876	3
12	2	14	8,944272	10,42443	3,967042	3

Exemplo

37

Objeto	X	Y	U(X)	L(X)
1	1	2	2,040223	6,403124
2	2	1	3,026964	6,403124
3	1	1	3,185465	7,071068
4	2	2	2,029224	5,656854
5	8	9	3,605551	4,669312
6	9	8	3,605551	4,661908
7	9	9	4,242641	4,760787
8	8	8	2,828427	5,104789
9	1	15	5,596086	10,29563
10	2	15	5,123168	9,848858
11	1	14	4,470876	9,433981
12	2	14	3,967042	8,944272

Exemplo

38

- Para cada objeto ***X*** na base
- M = max (½ S(c(X)) , L(X))
  - Se U(X) ≤ M, Então não faça nada
  - Se U(X) > M, Então atualizar

Objeto	X	Y	U(X)	L(X)	½ S(c(X))	M	Atualizar
1	1	2	2,040223	6,403124	2,061552813	2,061552813	Não
2	2	1	3,026964	6,403124	2,061552813	2,061552813	Sim

Exemplo

39

Objeto	X	Y	U(X)	L(X)	½ S(c(X))	M	Atualizar
1	1	2	2,040223	6,403124	2,061552813	2,061552813	Não
2	2	1	3,026964	6,403124	2,061552813	2,061552813	Sim

- Atualizar estimativa de U(X) com valor exato
  - U(2) = d(2, c(2)) = 3,026964
- Se U(X) > M, Então atualizar X como no algoritmo tradicional

Objeto	X	Y	D(C1)	D(C2)	D(C3)	Cluster
2	2	1	6,403124	3,026964	7,112805	2

  - U(2) = d(2, c(2)) = 3,026964
    L(2) = min $_{j \neq c(X)}$ d(2, c(j)) = 6,403124

Exemplo

40

Objeto	X	Y	U(X)	L(X)	½ S(c(X))	M	Atualizar
1	1	2	2,040223	6,403124	2,061552813	2,061552813	Não
2	2	1	3,026964	6,403124	2,061552813	2,061552813	Sim
3	1	1	3,185465	7,071068	2,061552813	2,061552813	Sim
4	2	2	2,029224	5,656854	2,061552813	2,061552813	Não
5	8	9	3,605551	4,669312	1	1	Sim
6	9	8	3,605551	4,661908	1	1	Sim
7	9	9	4,242641	4,760787	1	1	Sim
8	8	8	2,828427	5,104789	1	1	Sim
9	1	15	5,596086	10,29563	2,061552813	2,061552813	Sim
10	2	15	5,123168	9,848858	2,061552813	2,061552813	Sim
11	1	14	4,470876	9,433981	2,061552813	2,061552813	Sim
12	2	14	3,967042	8,944272	2,061552813	2,061552813	Sim

Exemplo

41

- Mover cada centro para a média dos objetos do cluster correspondente
  - Armazenar distância que cada protótipo deslocou em relação ao passo anterior (p(***k***))

K	X	Y	p(K)
1	3,570543	5,152577	2,573012
2	6,383543	2,570469	4,176477
3	9,630685	11,11513	4,763062

  - Calcular a distância entre todos os protótipos

Exemplo

42

- Distância entre os protótipos

	1	2	3
1	0	3,8	8,5
2		0	9,1
3			0

Protótipos mais próximos

[1,2]
S(1) =3,8

[2,1]
S(2) =3,8

[3,1]
S(3) = 8,5

Exemplo

43

- Atualizar estimativas U(X) e L(X) para todos os objetos tal que:
  - r' = 3 (4,763062)
  - r" = 2 (4,176477)
  - Para cada objeto X
    - U(X) = U(X) + p(c(X))
    - Se r' = c(X)
      - L(X) = L(X) – p(r")
    - Caso contrário
      - L(X) = L(X) – p(r')

Discussão dos resultados

44

Complexidade do algoritmo

	init. time	time/iteration	memory
k-d tree	$nd + n\log(n)$	-	nd
elkan	$ndk + dk^2$	dk^2	$nk + k^2$
hamerly	ndk	dk^2	n

Porcentagem do tempo "evitando" o laço interno

dataset	rand2	rand8	rand32	rand128
elkan	0.56	0.01	0.00	0.00
hamerly	0.97	0.88	0.91	0.83

dataset	birch	covtype	kddcup	mnist50
elkan	0.52	0.34	0.18	0.22
hamerly	0.94	0.89	0.82	0.82

Discussão dos resultados

45

- Tempo de execução
  - Para dados com d<=50 é melhor
  - Algoritmo de Elkans é recomendado nos outros casos
- Uso de memória
  - Muito próximo ao do algoritmo tradicional, por adicionar pouca informação extra

Discussão dos resultados

46

- Algoritmo de Elkan reduz quantidade de cálculos de distância
  - Melhor quando se usa uma medida de distância mais complexa
- Se K >> N, não é recomendado

47

Obrigado

Questões?

k-means eficiente

48

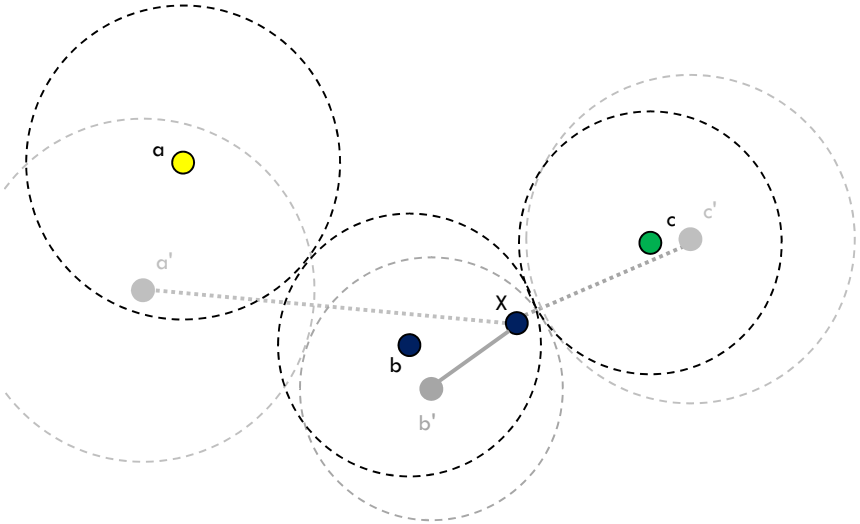
Inicialização

Para cada objeto:

- U(X): Estimativa da distância entre o objeto e o protótipo do grupo de ***X***, tal que:
  - $U(X) = d(X, c(X))$
- L(X): Estimativa da distância entre o objeto e o protótipo do grupo vizinho mais próximo, tal que:
  - $L(X) = \min_{j \neq c(X)} d(X, c(j))$

k-means eficiente

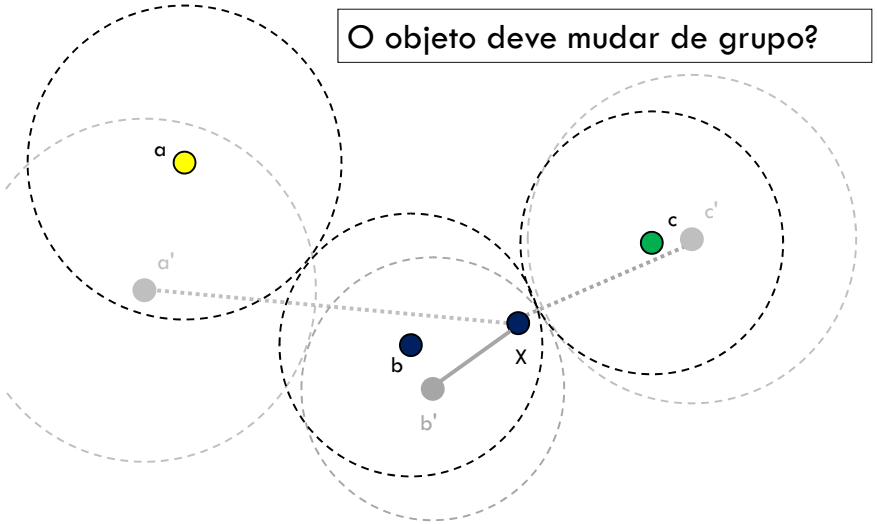
49



k-means eficiente

50

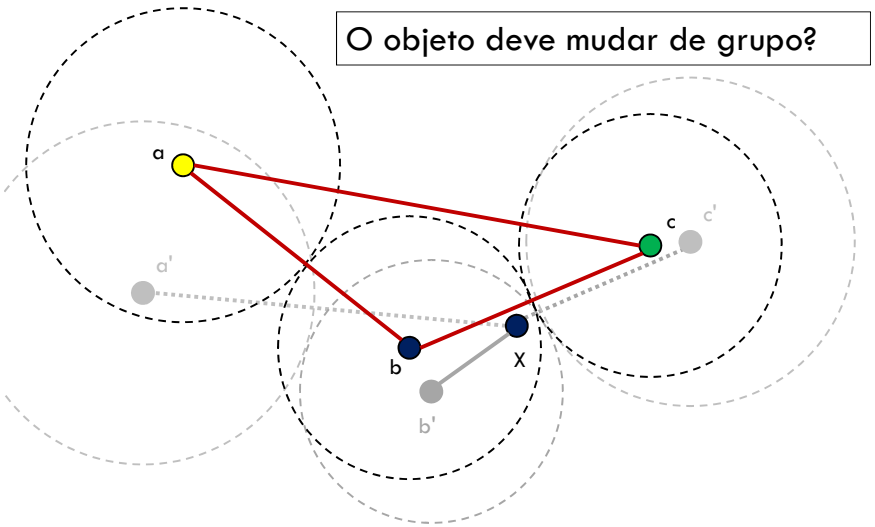
O objeto deve mudar de grupo?



k-means eficiente

51

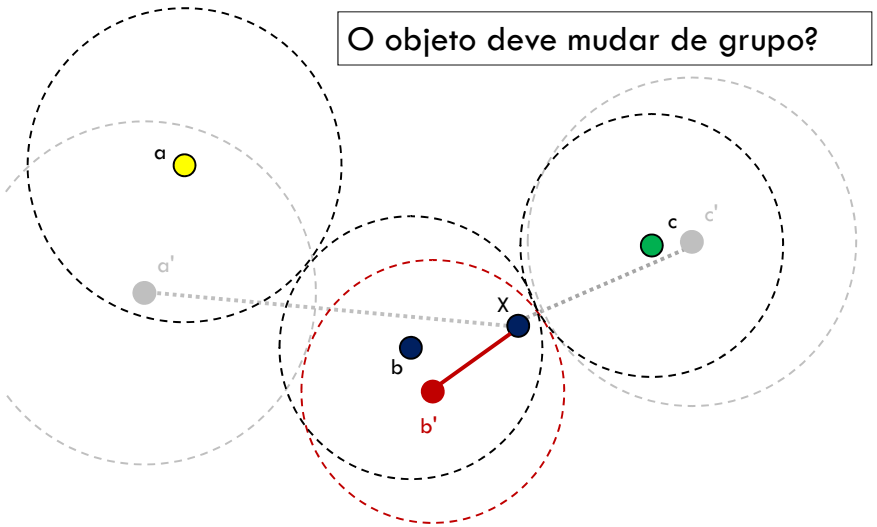
O objeto deve mudar de grupo?



k-means eficiente

52

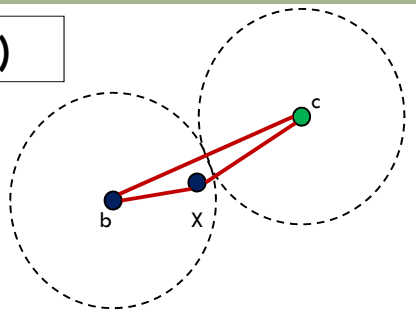
O objeto deve mudar de grupo?



Desigualdade triangular

53

$$d(b,c) \leq d(X,c) + d(X,b)$$



Se:

$$d(X,b) \leq \frac{1}{2}\, d(b,c)$$

Então:

$$d(X,c) \geq d(X,b)$$

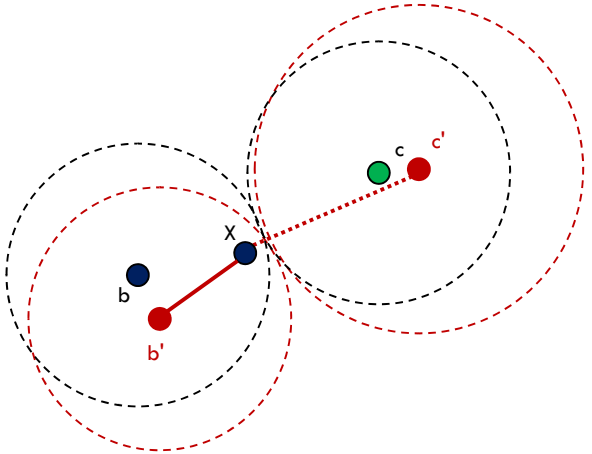
Seja ***U*** uma estimativa do valor de d(X,b) tal que:

$$U(X) \geq d(X,b)$$

Desigualdade triangular

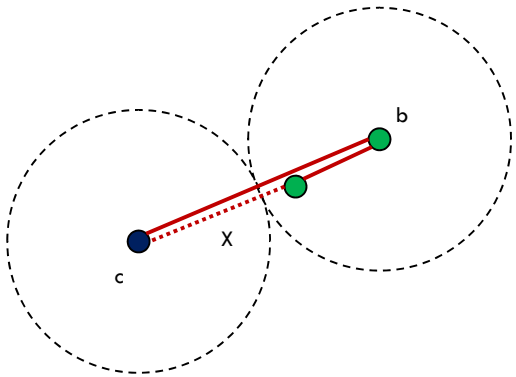
54

O objeto deve mudar de grupo?



Desigualdade triangular

55



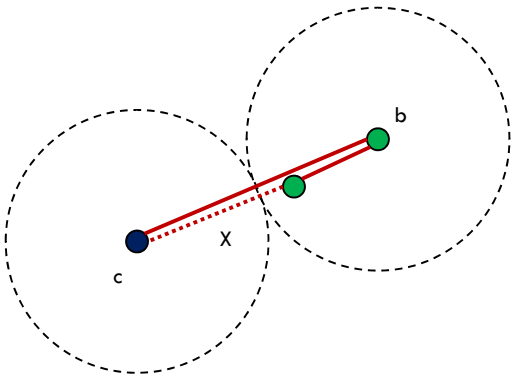
Valores conhecidos:

d(b,c)

d(X,b)

Desigualdade triangular

56



Valores conhecidos:

d(b,c)

d(X,b)

Se:

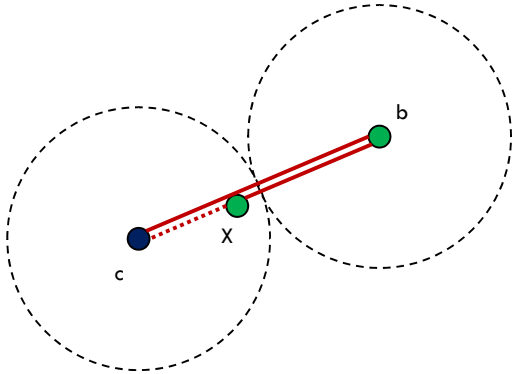
$$d(b,c) \geq 2d(X,b)$$

Então:

$$d(X,c) \geq d(X,b)$$

Desigualdade triangular

57



Valores conhecidos:

d(b,c)

d(X,b)

Se:

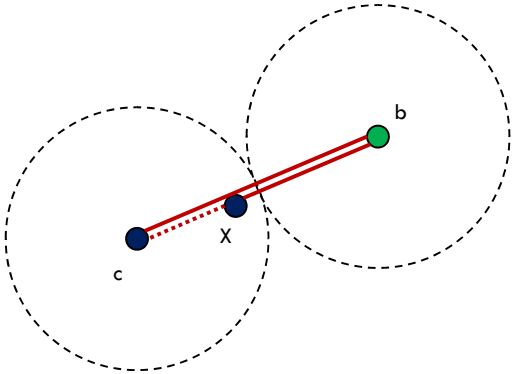
d(b,c) < 2d(X,b)

Então:

d(X,c) < d(X,b)

Desigualdade triangular

58



Valores conhecidos:

d(b,c)

d(X,b)

Se:

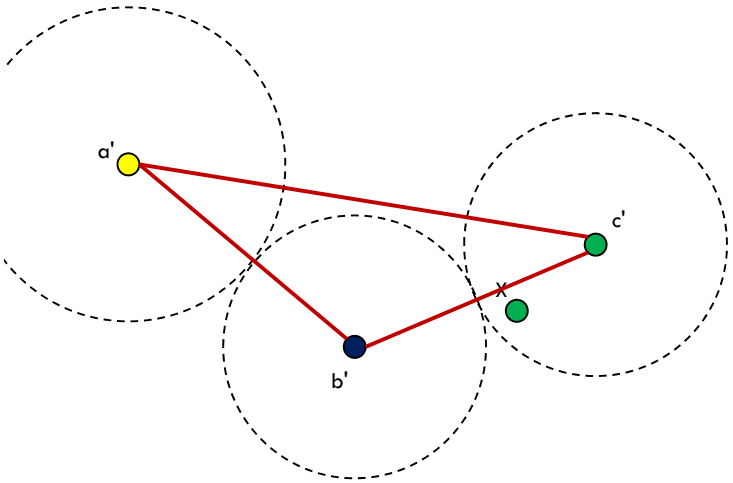
d(b,c) < 2d(X,b)

Então:

d(X,c) < d(X,b)

Inicialização

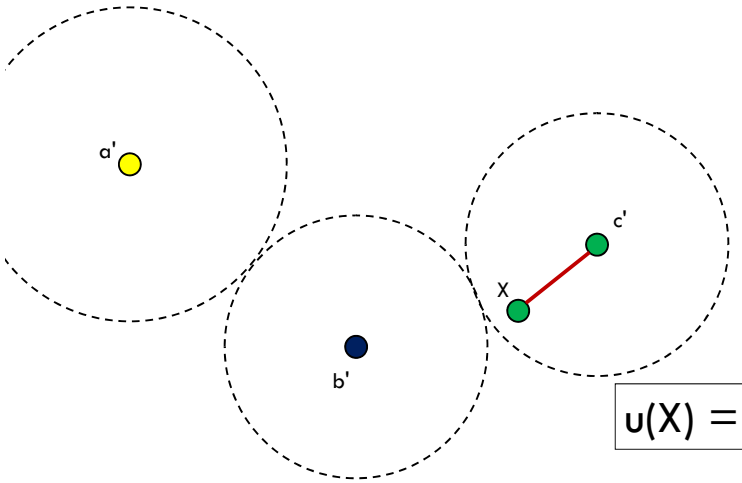
59



d(a',b')
d(a',c')
d(b',c')

Inicialização

60

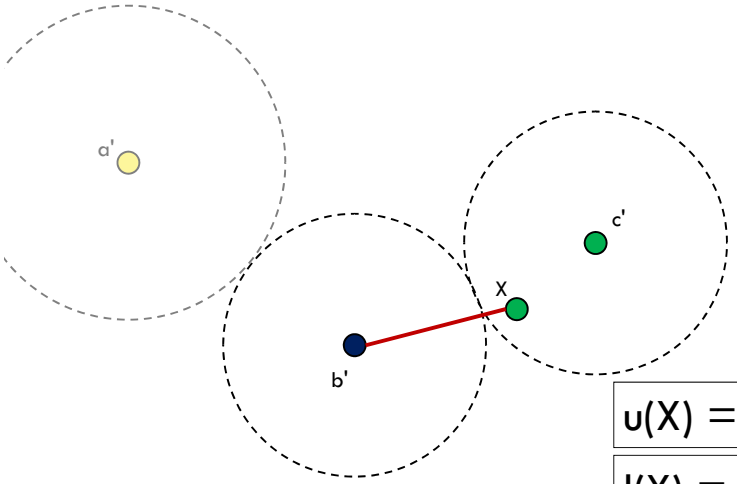


d(a',b')
d(a',c')
d(b',c')

u(X) = d(X,c')

Inicialização

61



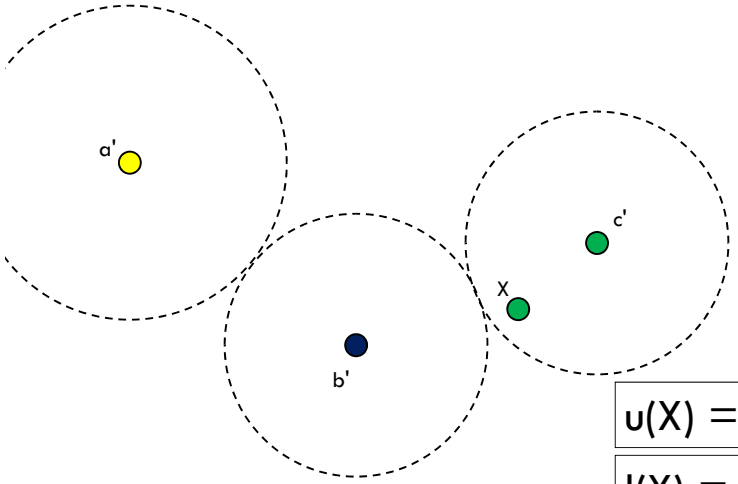
d(a',b')
d(a',c')
d(b',c')

u(X) = d(X,c')

l(X) = d(X,b')

Inicialização

62



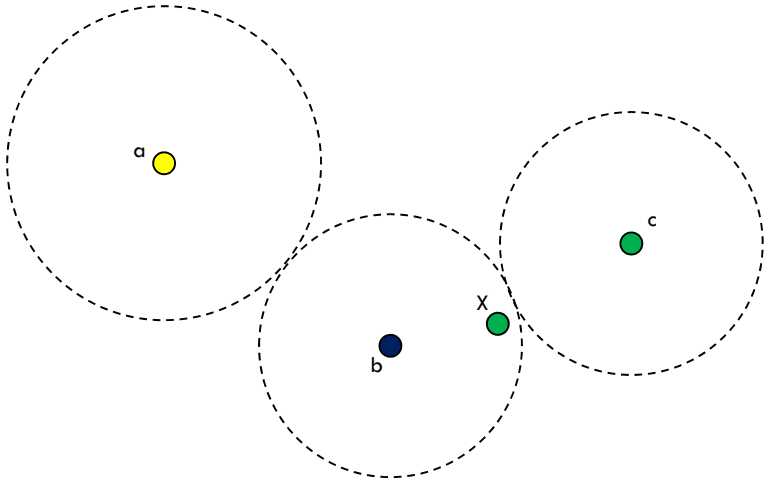
d(a',b')
d(a',c')
d(b',c')

u(X) = d(X,c')

l(X) = d(X,b')

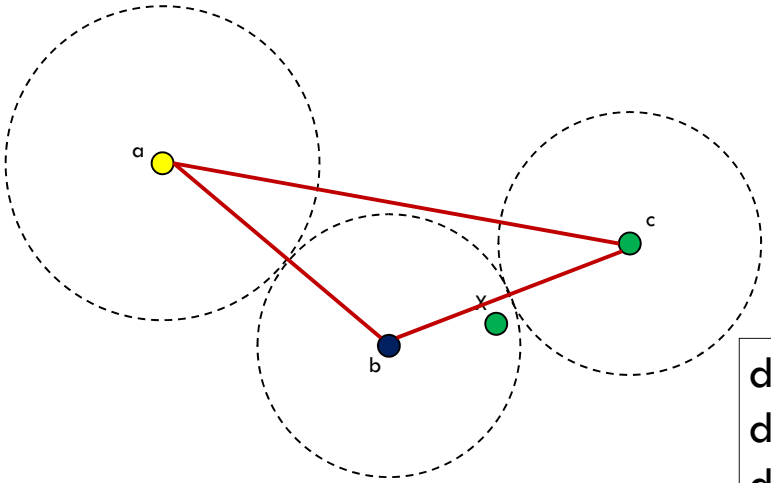
Calcular a distância entre X e cada um dos centros C

63



Calcular a distância entre X e cada um dos centros C

64



d(a,b)
d(a,c)
d(b,c)