K-Means

Marcos Castro

Agrupamento (clustering)

- Objetivo de um agrupamento: separar os objetos em grupos.
- Para separá-los, leva-se em conta as características dos objetos.
- Ideia: colocar em um mesmo grupo objetos que sejam similares.
- Qual o critério?
  - Baseia-se em uma função de dissimilaridade.
  - Essa função retorna a distância entre dois objetos.
- Exemplo de medida de dissimilaridade:
  - Distância euclidiana: $d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{ip}-x_{jp}|^2)}$

Aplicações

- Bioinformática – agrupar sequências.
- Marketing – grupos de clientes.
- Web – agrupamento de documentos semanticamente similares.
- Etc.

K-Means

- Trata-se de uma técnica de agrupamento não hierárquico.
- É uma heurística.
- Busca minimizar a distância dos elementos a um conjunto de k centros iterativamente. K é o número de clusters (grupos).
- Tem-se um conjunto de clusters onde cada cluster tem o seu centro.
- Dado um objeto, é calculada a distância (euclidiana por exemplo) desse objeto ao centro de cada cluster para, então, determinar a qual cluster pertence esse objeto.

K-Means

- O centro de cada grupo vai mudando.
- Para calcular o centro de cada grupo, basta calcular a média (mean) dos valores dos objetos que estão naquele grupo.
- Esse algoritmo é muito rápido.
- O parâmetro K é definido pelo usuário.

K-Means

- Algoritmo
  - 1) Escolhe-se k distintos valores para os centros dos grupos (pode ser aleatório).
  - 2) Associar cada ponto ao centro mais próximo.
    - Pode-se usar a distância euclidiana.
  - 3) Recalcular o centro de cada grupo.
    - Utiliza-se a média.
  - 4) Repetir os passos 2-3 até nenhum elemento mudar de grupo.

K-Means

- Vamos agrupar os seguintes dados em 2 (K = 2) grupos:

Subject	A	B
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

K-Means

- Primeiro fazemos a inicialização (imagem à direita):

Subject	A	B
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

	Individual	Mean Vector (centroid)
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

K-Means

- Passos da execução do algoritmo (imagem à direita):

Subject	A	B
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

	Cluster 1		Cluster 2	
Step	Individual	Mean Vector (centroid)	Individual	Mean Vector (centroid)
1	1	(1.0, 1.0)	4	(5.0, 7.0)
2	1, 2	(1.2, 1.5)	4	(5.0, 7.0)
3	1, 2, 3	(1.8, 2.3)	4	(5.0, 7.0)
4	1, 2, 3	(1.8, 2.3)	4, 5	(4.2, 6.0)
5	1, 2, 3	(1.8, 2.3)	4, 5, 6	(4.3, 5.7)
6	1, 2, 3	(1.8, 2.3)	4, 5, 6, 7	(4.1, 5.4)

K-Means

- O ponto 3 está mais próximo do centroide do cluster 2 do que do cluster 1, portanto, o 3 vai para o cluster 2.

Subject	A	B
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

	Cluster 1		Cluster 2	
Step	Individual	Mean Vector (centroid)	Individual	Mean Vector (centroid)
1	1	(1.0, 1.0)	4	(5.0, 7.0)
2	1, 2	(1.2, 1.5)	4	(5.0, 7.0)
3	1, 2, 3	(1.8, 2.3)	4	(5.0, 7.0)
4	1, 2, 3	(1.8, 2.3)	4, 5	(4.2, 6.0)
5	1, 2, 3	(1.8, 2.3)	4, 5, 6	(4.3, 5.7)
6	1, 2, 3	(1.8, 2.3)	4, 5, 6, 7	(4.1, 5.4)

	Individual	Mean Vector (centroid)
Cluster 1	1, 2	(1.3, 1.5)
Cluster 2	3, 4, 5, 6, 7	(3.9, 5.1)

Implementação em C++

- O código encontra-se no GitHub:
  - https://github.com/marcoscastro/kmeans/

Contato

mcastrosouza@live.com

www.geeksbr.com

http://github.com/marcoscastro

www.youtube.com/c/marcoscastrosouza

https://about.me/mcastrosouza