4 June, 2019

Lecture:

k-means Clustering and its Application

Dr. Kevin Kam Fung Yuen

PhD, Senior Lecturer, School of Business, Singapore University of Social Sciences

Learning Objectives

After this course, students should be able to

- Describe the theoretical issues and principles of K-means Clustering.
- explain the calculation procedure of k-mean algorithm with an example.
- develop a K-means algorithm with R code; compare with the built-in function for k-means.
- build k-means models for analysis of real world datasets.
- evaluate and discuss the results based on k-means model.

Reading

- Guan, C., and Yuen K.K.F., (2013) "Toward A Hybrid Approach of Primitive Cognitive Network Process and K-Means Clustering for Social Network Analysis", Proceedings of 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, pp.1267- 1271.

Acknowledgement

- The similar lecture was delivered at
- Xi'an Jiaotong-Liverpool University
  - Machine Learning (Using R)
    (Year 4)
- National Taiwan University of Science and Technology
  - Machine Learning Algorithms (Using R)
    (Postgraduate)

Why R?

2018

Language Rank	Types	Spectrum Ranking
1. Python		100.0
2. C++		99.7
3. Java		97.5
4. C		96.7
5. C#		89.4
6. PHP		84.9
7. R		82.9
8. JavaScript		82.6
9. Go		76.4
10. Assembly		74.1

2017

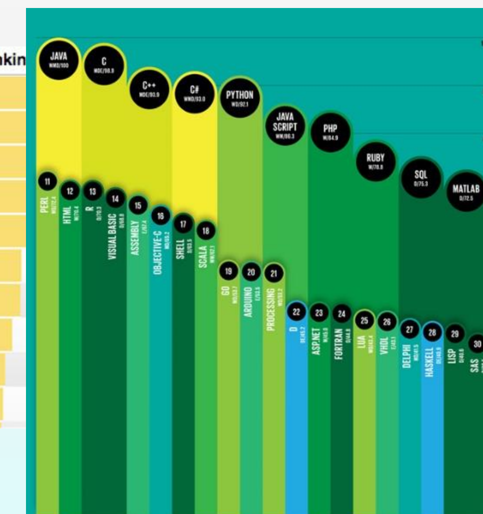
Language Rank	Types	Spectrum Ranking
1. Python		100.0
2. C		99.7
3. Java		99.4
4. C++		97.2
5. C#		88.6
6. R		88.1
7. JavaScript		85.5
8. PHP		81.4
9. Go		76.1
10. Swift		75.3

2016

Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

2015

Language Rank	Types	Spectrum Ranking	Spectrum Rankin
1. Java		100.0	100.0
2. C		99.9	99.3
3. C++		99.4	95.5
4. Python		96.5	93.5
5. C#		91.3	92.4
6. R		84.8	84.8
7. PHP		84.5	84.5
8. JavaScript		83.0	78.9
9. Ruby		76.2	74.3
10. Matlab		72.4	72.8



2014

https://spectrum.ieee.org/at-work/innovation/the-2018-top-programming-languages

Rank: 13->6->5->6->7

Supervised vs. Unsupervised Learning
Classification Vs. Clustering

Supervised Learning

Unsupervised Learning

```
head(iris)
```

X

Y

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

- Supervised Learning: both X and Y are known
- Unsupervised Learning: only X

Different Clustering Methods

- There are many different types of clustering methods
  - K-Means Clustering
  - Hierarchical Clustering
  - DBSCAN
  - …
- We will focus on K-means in this lecture.

Problem Definition

- Divide a data set, X=$\{x_1, \ldots, x_m\}$, into ***K*** clusters
- $C_1, \ldots, C_k$
- Such that satisfy two properties:
  - Each individual belongs to one of the *K* clusters: $C_1 \cup C_2 \cup \ldots \cup C_k = \{1, \ldots, m\}$.
  - No any individual belong to more than one clusters: $C_k \cap C_{k'} = \emptyset$ for all $\mathrm{k} \neq \mathrm{k}'$.

n criteria /Features /Factors /attributes

m individuals / records /objects /members

ID	A_1	A_2	…	A_j	…	A_n
1	x_{11}	x_{12}	….	x_{1j}	…	x_{1n}
2	x_{21}	x_{22}	….	x_{2j}	…	x_{2n}
…	…	…	….	…	…	…
i	x_{i1}	x_{i2}	….	x_{ij}	…	x_{in}
…	…	…	….	…	…	…
m	x_{m1}	x_{m2}	…	x_{mj}	…	x_{mn}

m x n data matrix

Which object should belong to which cluster?

$$C_1, \ldots, C_K$$

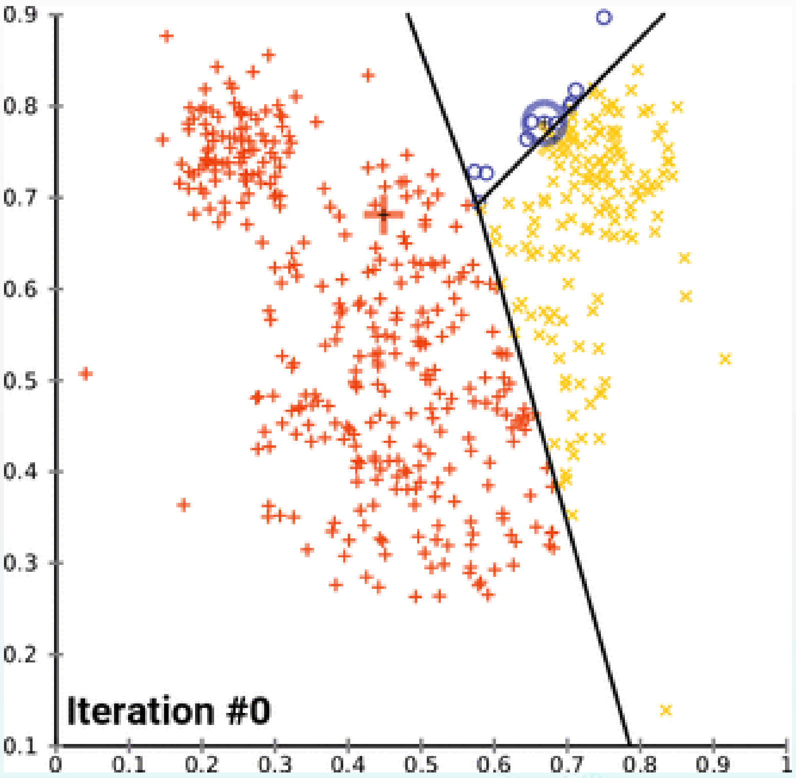
- K means is to attemp to find the least within group variances by finding the ideal means.

$$\min \sum_{x,x`\in C_i} \left\| x - x` \right\|$$

is equivalent to

$$\min \sum_{i=}^{k} \sum_{x\in C_i} \left\| x - \mu_i \right\|$$

Gif file



https://en.wikipedia.org/wiki/K-means_clustering

Weighted K-Means Clustering

- Step 1: Data matrix normalization
  - Normalization for data of different scale
- Step 2: Attributes' weights determination
- Step 3: data partition by k-means approach
  - Choose centers & compute the Euclidean distances
  - Assign each object to the cluster
  - Update the centers of clusters
  - Loop until clusters have no changes.

(Guan and Yuen, 2013)

The classical K-means treat each feature equally

Step 1: Data matrix normalization

n criteria

m individuals / records /objects /members

ID	C_1	C_2	...	C_j	...	C_n
T_1	x_{11}	x_{12}	....	x_{1j}	...	x_{1n}
T_2	x_{21}	x_{22}	....	x_{2j}	...	x_{2n}
...	...	...	....	...	...	...
T_i	x_{i1}	x_{i2}	....	x_{ij}	...	x_{in}
...	...	...	....	...	...	...
T_m	x_{m1}	x_{m2}	...	x_{mj}	...	x_{mn}

m x n data matrix

Scale normalization

A function *f* rescales the data source to the interval of [0,1].

$$f : x_{ij} \rightarrow x'_{ij}, x'_{ij} \in [0,1]$$

Ordinal/ interval / ratio scale

$$x'_{ij} = \frac{x_{ij} - \min x_{kj}}{\max x_{kj} - \min x_{kj}}, k \in \{1, 2, \ldots, m\}$$

Binary scale

The value is either 0 or 1

Social Network Analysis example
Friends Referral

6 x 3 data matrix

ID	C_1	C_2	C_3
1	15	High school	Yes
2	22	Undergraduate	No
3	17	High school	No
4	40	Doctor	Yes
5	23	Undergraduate	Yes
6	25	Postgraduate	Yes

C_1 : Age,
C_2 : Education
C_3: Music preference

How can we organize the data for clustering?

ID	C_1	C_2	C_3
1	15	High school	Yes
2	22	Undergraduate	No
3	17	High school	No
4	40	Doctor	Yes
5	23	Undergraduate	Yes
6	25	Postgraduate	Yes

Interval scale

Ordinal scale

Binary scale

ID	C_1	C_2	C_3
1	15	1	1
2	22	2	0
3	17	1	0
4	40	4	1
5	23	2	1
6	25	3	1

ID	C_1	C_2	C_3
1	15	1	1
2	22	2	0
3	17	1	0
4	40	4	1
5	23	2	1
6	25	3	1

$\frac{x-15}{40-15}$ ↓ ↓ $\frac{x-1}{4-1}$ ↓ No change

ID	C_1	C_2	C_3
1	0	0.00	1
2	0.28	0.33	0
3	0.08	0.00	0
4	1	1.00	1
5	0.32	0.33	1
6	0.4	0.67	1

Step 2: Attributes' weights determination

- Classical K-means method assumes each criterion has the equal weight, but this may not reflect the decision maker's preference. Cognitive pairwise comparison can adjust the weight of the criteria /variables.
- Suppose that the weights are as below.

	Weight
C1	0.32
C2	0.28
C3	0.40

Step 3:data partition by k-means

1)Choose objects successively, and compute the weighted Euclidean distances from objects to all the centers by

$$d_{ij} = \sqrt{w_1 \left| x_{i1} - x_{j1} \right|^2 + w_2 \left| x_{i2} - x_{j2} \right|^2 + \cdots + w_n \left| x_{in} - x_{jn} \right|^2}$$

Note: If Wj = 1, it is the classical k-means approach. That is

$$d_{ij} = \sqrt{\left| x_{i1} - x_{j1} \right|^2 + \left| x_{i2} - x_{j2} \right|^2 + \cdots + \left| x_{in} - x_{jn} \right|^2}$$

SNA Example

C	C_1	C_2	C_3
W	*0.32*	*0.28*	*0.40*
1	0	0.00	1
2	0.28	0.33	0
3	0.08	0.00	0
4	1	1.00	1
5	0.32	0.33	1
6	0.4	0.67	1

Question.
We need to classify the above members into two clusters, A and B.
Supposed ID2 is chosen for Cluster A, and ID6 is chosen for Cluster B,
What is the Euclidean distance for each member to each cluster?

	C_1	C_2	C_3	Distance To 2	Distance To 6
W	*0.32*	*0.28*	*0.40*		
1	0	0.00	1	0.68	0.42
2	0.28	0.33	0		
3	0.08	0.00	0		
4	1	1.00	1		
5	0.32	0.33	1		
6	0.4	0.67	1		

$$d_{12} = \sqrt{0.32|0-0.28|^2 + 0.28|0-0.33|^2 + 0.40|1-0|^2} = 0.68$$

$$d_{16} = \sqrt{0.32|0-0.40|^2 + 0.28|0-0.67|^2 + 0.40|1-1|^2} = 0.42$$

Similarly, we compute the Euclidean distances for the other individuals

W	C_1	C_2	C_3	*Distance To 2*	*Distance To 6*
W	*0.32*	*0.28*	*0.40*		
1	0	0.00	1	0.68	0.42
2	0.28	0.33	0	0.00	0.66
3	0.08	0.00	0	0.21	0.75
4	1	1.00	1	0.83	0.38
5	0.32	0.33	1	0.64	0.18
6	0.4	0.67	1	0.66	0.00

2) Assign each object to the cluster of minimum distance from this object to the cluster center

Cluster A Cluster B

	C_1	C_2	C_3	*Distance To 2*	*Distance To 6*	**Cluster A (center is 2)**	**Cluster B (center is 6)**
W	*0.32*	*0.28*	*0.40*				
1	0	0.00	1	0.68	0.42	{}	{1}
2	0.28	0.33	0	0.00	0.66	{2}	{1}
3	0.08	0.00	0	0.21	0.75	{2, 3}	{1}
4	1	1.00	1	0.83	0.38	{2, 3}	{1, 4}
5	0.32	0.33	1	0.64	0.18	{2, 3}	{1, 4, 5}
6	0.4	0.67	1	0.66	0.00	{2, 3}	{1, 4, 5, 6}

3)Calculate the mean of each cluster, then update the centers of clusters by the mean values.

	C_1	C_2	C_3	*Distance To 2*	*Distance To 6*	Cluster A (center is 2)	Cluster B (center is 6)
W	*0.32*	*0.28*	*0.40*				
1	0	0.00	1	0.68	0.42	{}	{1}
2	0.28	0.33	0	0.00	0.66	{2}	{1}
3	0.08	0.00	0	0.21	0.75	{2, 3}	{1}
4	1	1.00	1	0.83	0.38	{2, 3}	{1, 4}
5	0.32	0.33	1	0.64	0.18	{2, 3}	{1, 4, 5}
6	0.4	0.67	1	0.66	0.00	{2, 3}	{1, 4, 5, 6}

New center for cluster A

Member of Cluster A	C1	C2	C3
2	0.28	0.33	0
3	0.08	0.00	0
Mean(New Center)	0.180	0.165	0

3)Calculate the mean of each cluster, then update the centers of clusters by the mean values.

	C_1	C_2	C_3	*Distance To 2*	*Distance To 6*	Cluster A (center is 2)	Cluster B (center is 6)
W	*0.32*	*0.28*	*0.40*				
1	0	0.00	1	0.68	0.42	{}	{1}
2	0.28	0.33	0	0.00	0.66	{2}	{1}
3	0.08	0.00	0	0.21	0.75	{2, 3}	{1}
4	1	1.00	1	0.83	0.38	{2, 3}	{1, 4}
5	0.32	0.33	1	0.64	0.18	{2, 3}	{1, 4, 5}
6	0.4	0.67	1	0.66	0.00	{2, 3}	{1, 4, 5, 6}

New center for cluster B

Member of Cluster B	C1	C2	C4
1	0.00	0.00	1
4	1.00	1.00	1
5	0.32	0.33	1
6	0.40	0.67	1
Mean (New Center)	0.430	0.500	1.000

4) the next loop will not generate new centers until the current centers are the same as ones of previous loop.

	C_1	C_2	C_3	*Distance To* [0.180; 0.165; 0]	*Distance To* [0.430; 0.500; 1]	Cluster A (Center [0.180; 0.165; 0])	Cluster B (Center [0.430; 0.500; 1])
W	*0.32*	*0.28*	*0.40*				
1	0	0.00	1	0.65	0.36	{}	{1}
2	0.28	0.33	0	0.11	0.65	{2}	{1}
3	0.08	0.00	0	0.10	0.72	{2, 3}	{1}
4	1	1.00	1	0.90	0.42	{2, 3}	{1, 4}
5	0.32	0.33	1	0.65	0.11	{2, 3}	{1, 4, 5}
6	0.4	0.67	1	0.70	0.09	{2, 3}	{1, 4, 5, 6}

$$d_{1A} = \sqrt{0.32|0-0.18|^2 + 0.28|0-0.165|^2 + 0.40|1-0|^2} = 0.65$$

$$d_{1B} = \sqrt{0.32|0-0.43|^2 + 0.28|0-0.50|^2 + 0.40|1-1|^2} = 0.36$$

Exercise 1

	C1	C2	C3	C4
weight	0.109	0.352	0.305	0.234
P1	0.090	0.201	0.149	0.194
P2	0.135	0.170	0.233	0.108
P3	0.174	0.094	0.125	0.132
P4	0.174	0.167	0.201	0.174
P5	0.135	0.194	0.170	0.153
P6	0.292	0.174	0.122	0.240

Find two clusters using the above matrix by weighted K-means Clustering. Select P3 and P5 as the two clusters. Show your steps.

Labs

- Develop R codes for k-means algorithms
- Limitation of k-means
- Comparisons with the other algorithms.