PCS5024 - Statistical Machine Learning

Escola Politecnica, USP

Anna H. Reali Costa, Fabio G. Cozman

Studant : Angel Felipe M. de Paula
USP : 11030561

Exercise 1.1 : Implement K-means manually

Distance function: Euclidean distance
K: 3
Instances: 10
Dimension: 2

Initial centroids:	X	Y
Red - C1	6.2	3.2
Green - C2	6.6	3.7
Blue - C3	6.5	3.0

= Minimum distance

X=

Data Points

	X	Y
A	5.9	3.2
B	4.6	2.9
C	6.2	2.8
D	4.7	3.2
E	5.5	4.2
F	5.0	3.0
G	4.9	3.1
H	6.7	3.1
I	5.1	3.8
J	6.0	3.0

Distances

Red	Green	Blue
0.300	0.860	0.632
1.628	2.154	1.903
0.400	0.985	0.361
1.500	1.965	1.811
1.221	1.208	1.562
1.217	1.746	1.500
1.304	1.803	1.603
0.510	0.608	0.224
1.253	1.503	1.612
0.283	0.922	0.500

Clusters

Red	Green	Blue
A	E	C
B		H
D		
F		
G		
I		
J		

New centroids:	X	Y
Red - C1	5.171	3.171
Green - C2	5.500	4.200
Blue - C3	6.450	2.950

X=

Data Points

	X	Y
A	5.9	3.2
B	4.6	2.9
C	6.2	2.8
D	4.7	3.2
E	5.5	4.2
F	5.0	3.0
G	4.9	3.1
H	6.7	3.1
I	5.1	3.8
J	6.0	3.0

Distances

Red	Green	Blue
0.729	1.077	0.604
0.633	1.581	1.851
1.094	1.565	0.292
0.472	1.281	1.768
1.080	0.000	1.570
0.242	1.300	1.451
0.281	1.253	1.557
1.530	1.628	0.292
0.633	0.566	1.595
0.846	1.300	0.453

Clusters

Red	Green	Blue
B	E	A
D	I	C
F		H
G		J

	X	Y
Red - C1	4.800	3.050
Green - C2	5.300	4.000
Blue - C3	6.200	3.025

X=

Data Points

	X	Y
A	5.9	3.2
B	4.6	2.9
C	6.2	2.8
D	4.7	3.2
E	5.5	4.2
F	5.0	3.0
G	4.9	3.1
H	6.7	3.1
I	5.1	3.8
J	6.0	3.0

Distances

Red	Green	Blue
1.110	1.000	0.347
0.250	1.304	1.605
1.422	1.500	0.225
0.180	1.000	1.510
1.346	0.283	1.368
0.206	1.044	1.200
0.112	0.985	1.302
1.901	1.664	0.506
0.808	0.283	1.346
1.201	1.221	0.202

Clusters

Red	Green	Blue
B	E	A
D	I	C
F		H
G		J

New centroids:

	X	Y
Red - C1	4.800	3.050
Green - C2	5.300	4.000
Blue - C3	6.200	3.025

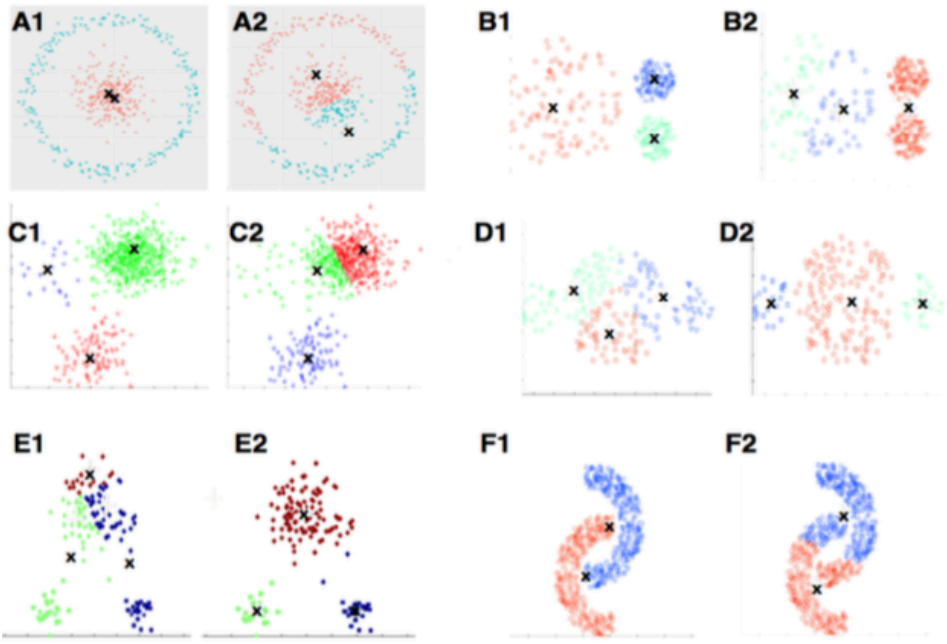
Stopping Criterion : No re-assignments of data points to different clusters and No-change of centroids.

Ans:

		X	Y
1)	Red - C1	5.171	3.171
2)	Green - C2	5.300	4.000
3)	Blue - C3	6.200	3.025

4) 3 iterationds

Exercise 1.2 : Application of K-means



Ans:

1) A2
2) B2
3) C2
4) D1
5) E2
6) F2

Exercise 1.3 : Hierarchical clustering

Distance function: Euclidean distance
Instances: 8
Dimension: 2

Data Points

X=

	X	Y
A	4.7	3.2
B	4.9	3.1
C	5	3
D	4.6	2.9
E	5.9	3.2
F	6.7	3.1
G	6	3
H	6.2	2.8

Distance between all points

	A	B	C	D
E	1.200	1.005	0.922	1.334
F	2.002	1.800	1.703	2.110
G	1.315	1.105	1.000	1.404
H	1.552	1.334	1.217	1.603

1) **Distance Between Futher members (Complete link):** 2.1095
2) **Distance Between two closest members (Single link):** 0.9220
3) **Average distance between all pairs :** 1.4129
4) **Robust to noise:** Avareage

Exercise 1.4 : Translation of the word Standardi

Collins Dicrionary

Standardization (n): Padronização

Cambridge Dictionary

Standardization (n): The process of making things of the same type have the same basic features:

Ex: The standardization of the internet may facilitate mergers and acquisitions by making corporate systems instantly compatible.

A padronização da internet pode facilitar fusões e aquisições tornando os sistemas corporativos instantaneamente compatíveis

Exercise 1.5 : Z-score

In the the calculus of z-score :

$$Z = \frac{X - \mathrm{E}[X]}{\sigma(X)}$$

it is necessary to calculate the mean absolute deviation of attribute f, denoted by σ(x), that is computed as follows:

$$\sigma(X) = \sqrt{\mathrm{Var}(X)}$$

$$\mathrm{var}(X) = \mathrm{E}((X - \mu)^2).$$

Ans: So the formula in the professor slide is correct.