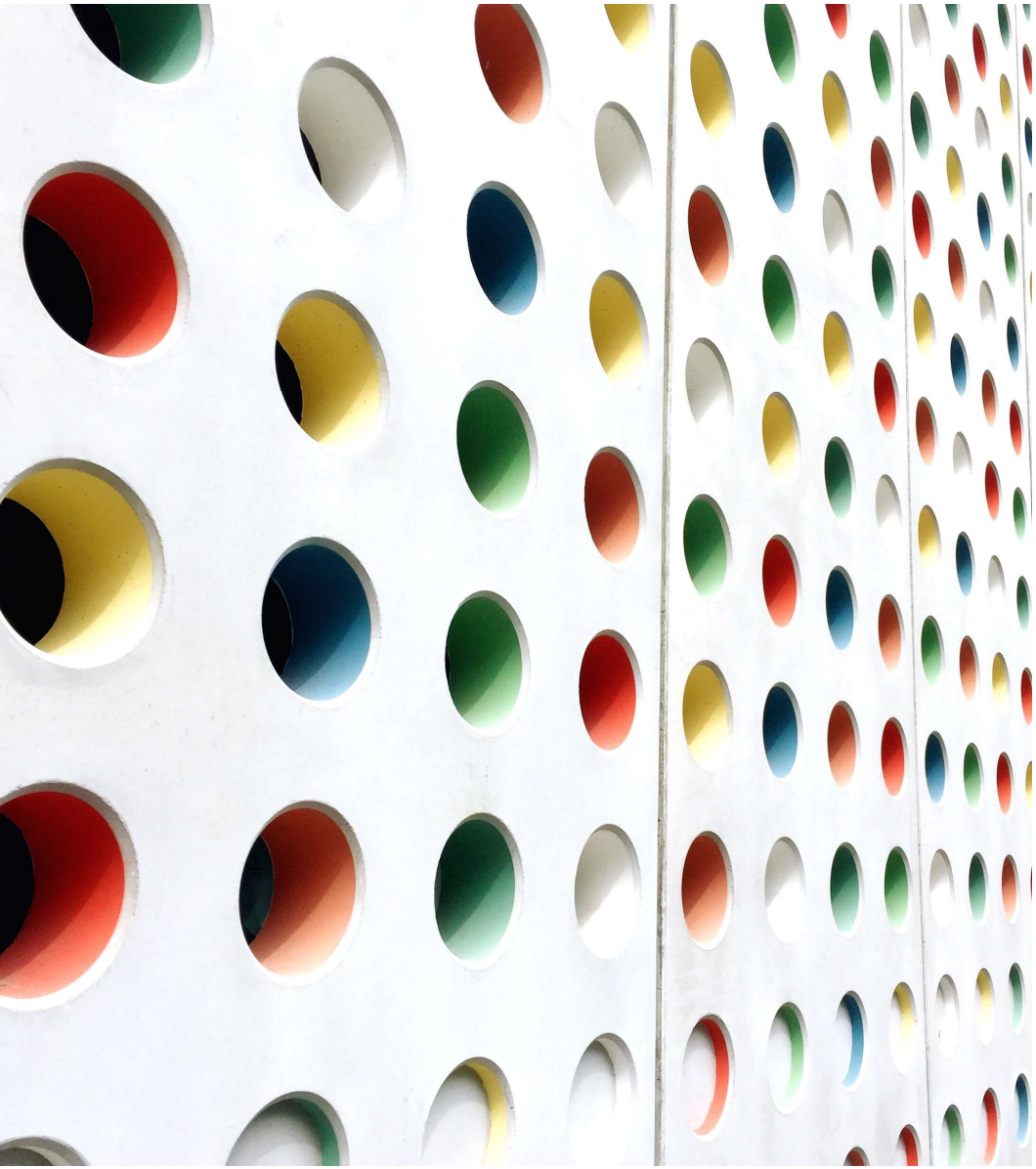
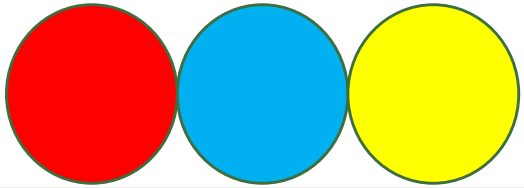


DAWSON COLLEGE

K-means clustering

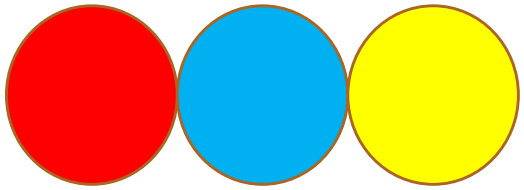
Garry Ka Lok CHU

K-means clustering



Clustering is one of the tasks performed by artificial intelligence. Given a set of training data using unsupervised learning, the system can assign all data points into clusters.

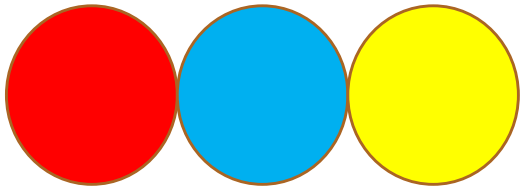
K-means clustering



First, determine the value of k. In this case, k = 2 is used.

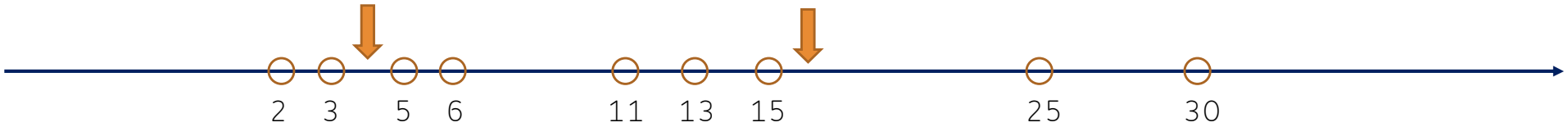
Data		2	3	5	6	11	13	15	25	30
K = 2										
1st iteration			Mean 1 =				Mean 2 =			
Distance to Mean 1										
Distance to Mean 2										
Cluster Number										
			C1 total =				C2 total =			
			C1 count =				C2 count =			
			Mean 1 =				Mean 2 =			

K-means clustering

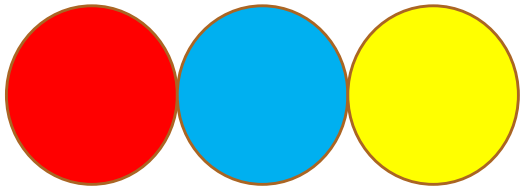


Choose a reasonable mean for each cluster to start the first iteration.

Data	2	3	5	6	11	13	15	25	30
K = 2									
1st iteration		Mean 1 =	4			Mean 2 =	16		
Distance to Mean 1									
Distance to Mean 2									
Cluster Number									
		C1 total =				C2 total =			
		C1 count =				C2 count =			
		Mean 1 =				Mean 2 =			

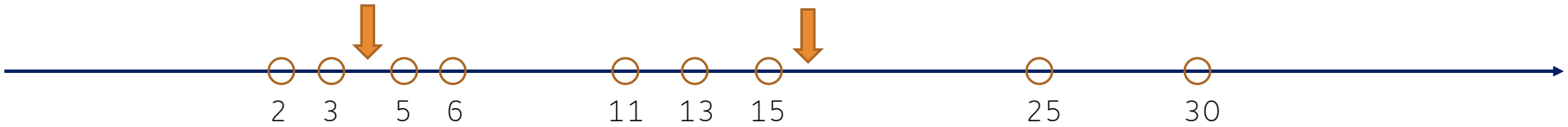


K-means clustering

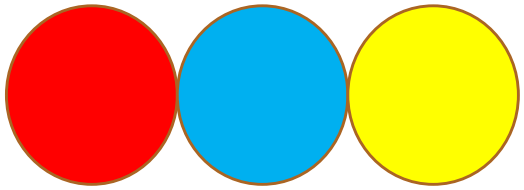


Then, find the distance between each point and each mean.

Data	2	3	5	6	11	13	15	25	30
K = 2									
1st iteration		Mean 1 =	4			Mean 2 =	16		
Distance to Mean 1	2	1	1	2	7	9	11	21	26
Distance to Mean 2	14	13	11	10	5	3	1	9	14
Cluster Number									
		C1 total =				C2 total =			
		C1 count =				C2 count =			
		Mean 1 =				Mean 2 =			

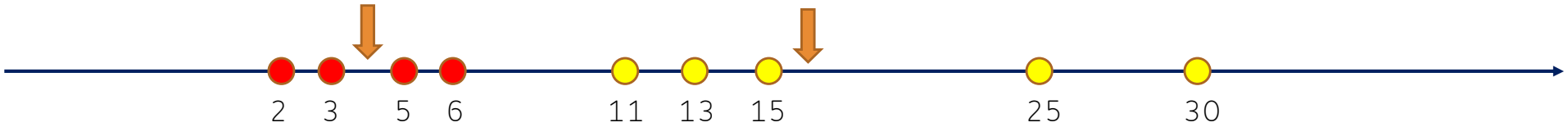


K-means clustering

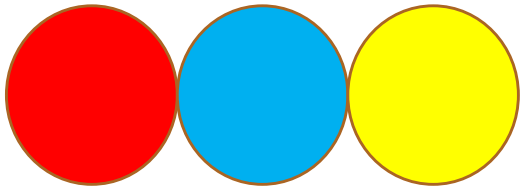


Use the shortest distance to assign a point to a cluster.

Data	2	3	5	6	10	13	15	25	30
K = 2									
1st iteration		Mean 1 =	4			Mean 2 =	16		
Distance to Mean 1	2	1	1	2	7	9	11	21	26
Distance to Mean 2	14	13	11	10	5	3	1	9	14
Cluster Number	C1	C1	C1	C1	C2	C2	C2	C2	C2
		C1 total =				C2 total =			
		C1 count =				C2 count =			
		Mean 1 =				Mean 2 =			

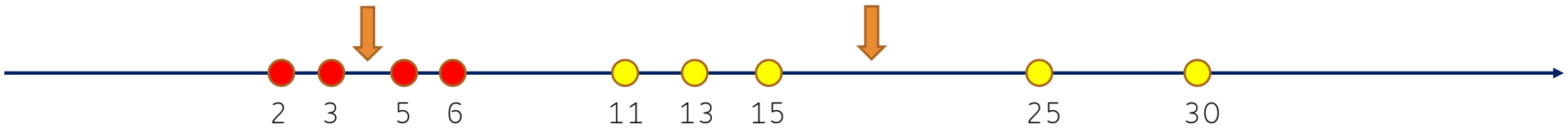


K-means clustering

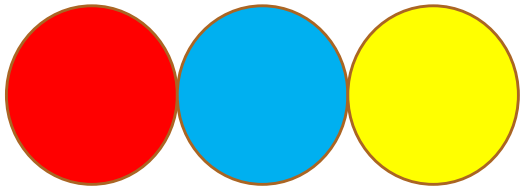


Update the mean of each cluster.

Data	2	3	5	6	10	13	15	25	30
K = 2									
1st iteration		Mean 1 =	4			Mean 2 =	16		
Distance to Mean 1	2	1	1	2	7	9	11	21	26
Distance to Mean 2	14	13	11	10	5	3	1	9	14
Cluster Number	C1	C1	C1	C1	C2	C2	C2	C2	C2
		C1 total =	16			C2 total =	94		
		C1 count =	4			C2 count =	5		
		Mean 1 =	4			Mean 2 =	18.8		

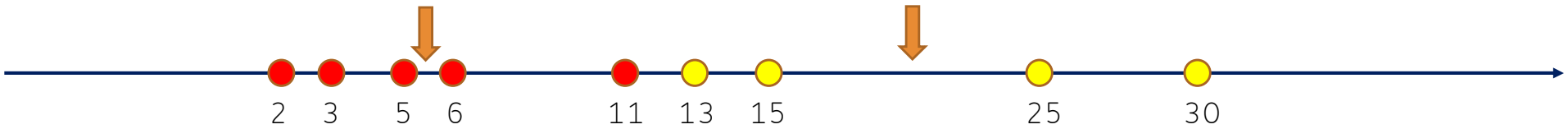


K-means clustering

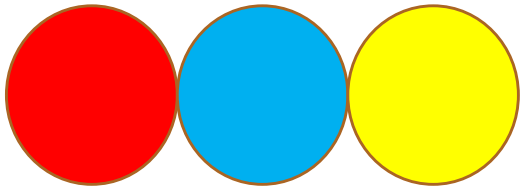


Repeat the iteration until convergence.

Data	2	3	5	6	10	13	15	25	30
K = 2									
2nd iteration		Mean 1 =	4			Mean 2 =	18.8		
Distance to Mean 1	2	1	1	2	7	9	11	21	26
Distance to Mean 2	16.8	15.8	13.8	12.8	7.8	5.8	3.8	6.2	11.2
Cluster Number	C1	C1	C1	C1	C1	C2	C2	C2	C2
		C1 total =	27			C2 total =	83		
		C1 count =	5			C2 count =	4		
		Mean 1 =	5.4			Mean 2 =	20.75		

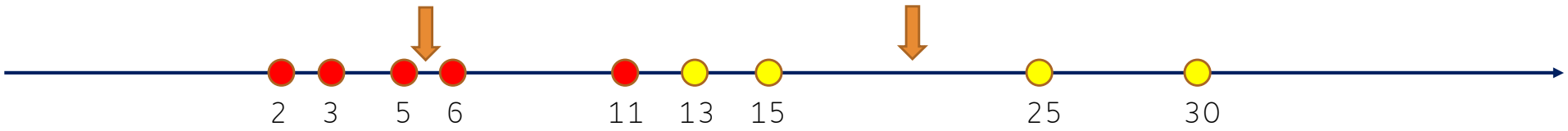


K-means clustering

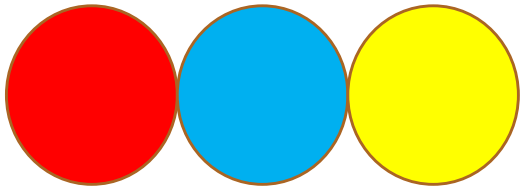


The means stay the same at the 3rd iteration.

Data	2	3	5	6	10	13	15	25	30
K = 2									
3rd iteration		Mean 1 =	5.4			Mean 2 =	20.75		
Distance to Mean 1	2	1	1	2	7	9	11	21	26
Distance to Mean 2	18.75	17.75	15.75	14.75	9.75	7.75	5.75	4.25	9.25
Cluster Number	C1	C1	C1	C1	C1	C2	C2	C2	C2
		C1 total =	27			C2 total =	83		
		C1 count =	5			C2 count =	4		
		Mean 1 =	5.4			Mean 2 =	20.75		

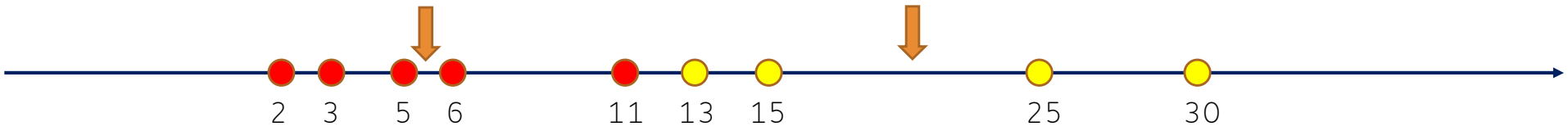


K-means clustering

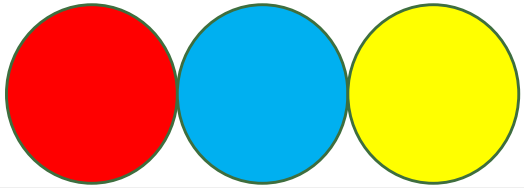


Conclusion: First 5 points are assigned to cluster #1 and others to #2.

Data	2	3	5	6	10	13	15	25	30
K = 2									
3rd iteration		Mean 1 =	5.4			Mean 2 =	20.75		
Distance to Mean 1	2	1	1	2	7	9	11	21	26
Distance to Mean 2	18.75	17.75	15.75	14.75	9.75	7.75	5.75	4.25	9.25
Cluster Number	C1	C1	C1	C1	C1	C2	C2	C2	C2
		C1 total =	27			C2 total =	83		
		C1 count =	5			C2 count =	4		
		Mean 1 =	5.4			Mean 2 =	20.75		



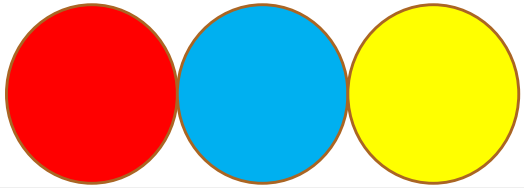
K-means clustering



Advanced Questions:

- Is there any method to determine the optimal value of k?
- Other than absolute value (Manhattan distance), may we use other distance formula?

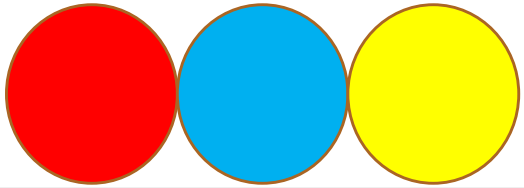
K-means clustering



Applications:

- Image compression
- Document analysis
- Market segmentation

K-means clustering



Further Topics:

- K-Medoids Clustering
- Hierarchical Clustering
- Hard/Soft Clustering
- Clustering with Outliers