K-Means Clustering

… Or what to do when your data doesn't have labels

Some example use cases of Unsupervised Learning

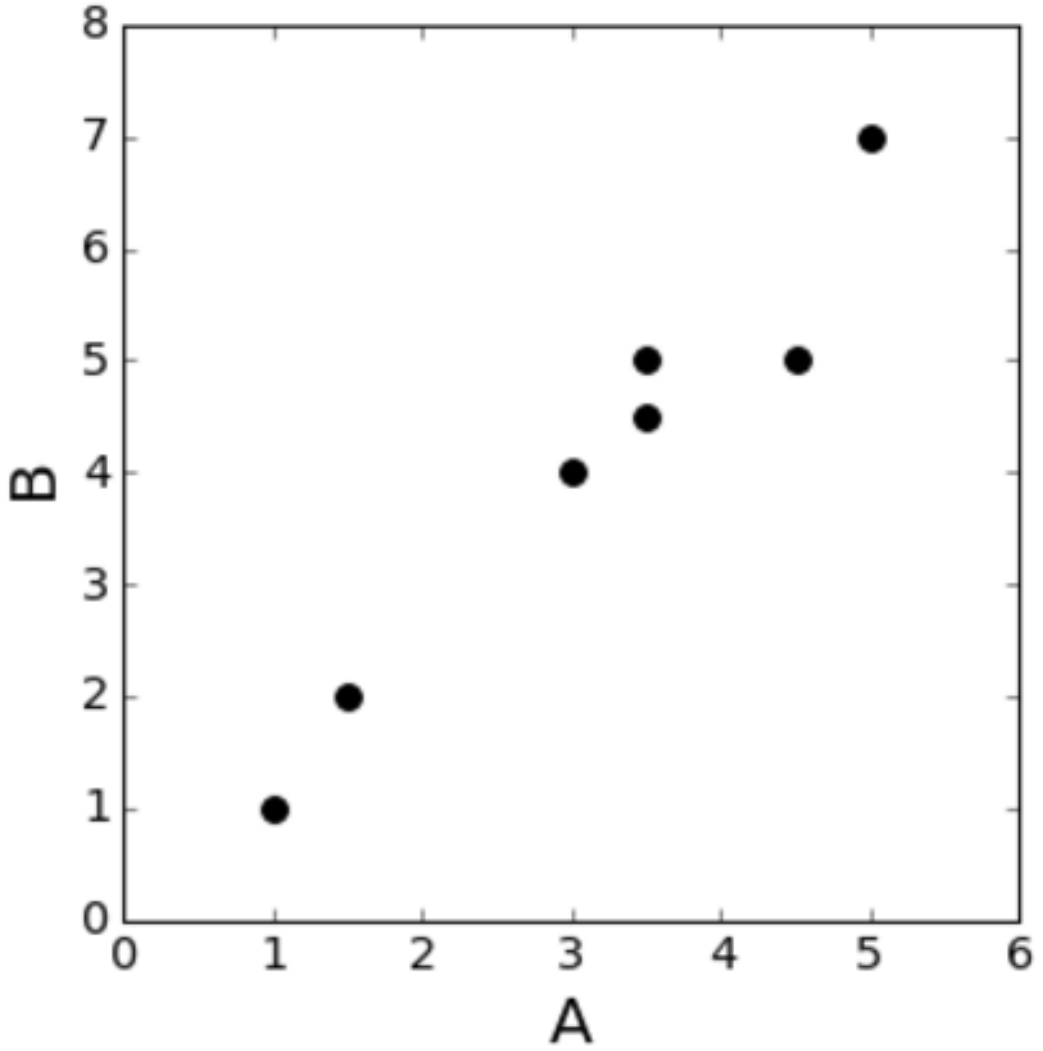
- Unsupervised learning is performed when you don't have labels for your data, but you would like to check if there are any patterns.

  - Group emails or search results
  - Customer shopping patterns
  - Regions of images
  - Clusters of genes

Example: Data

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

Goal: Split the data into 2 classes
K=2

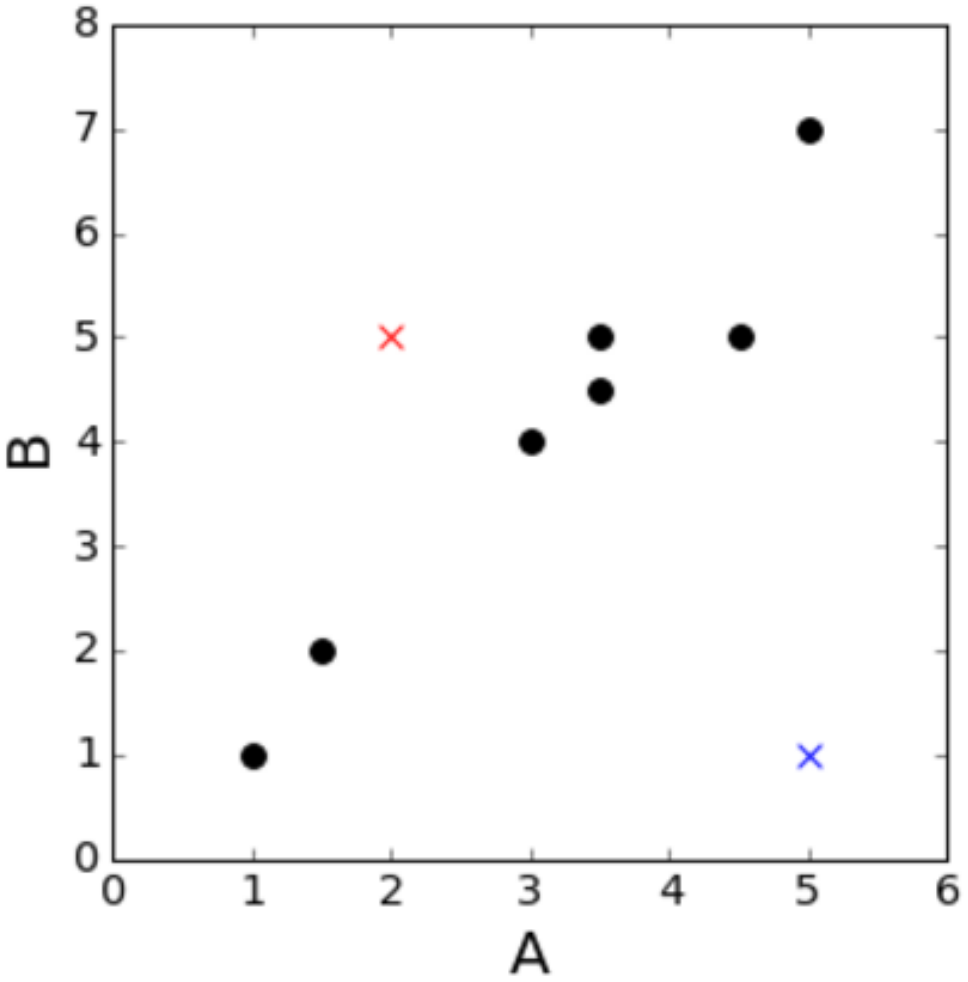


Step 1: Randomly initialize the cluster centroids

- The centroid refers to the “middle point” of the cluster in Euclidean space.
- You can pick any point that is in the range of your dataset.
- Our random number generator returns the following:

Step 1: Randomly initialize the cluster centroids

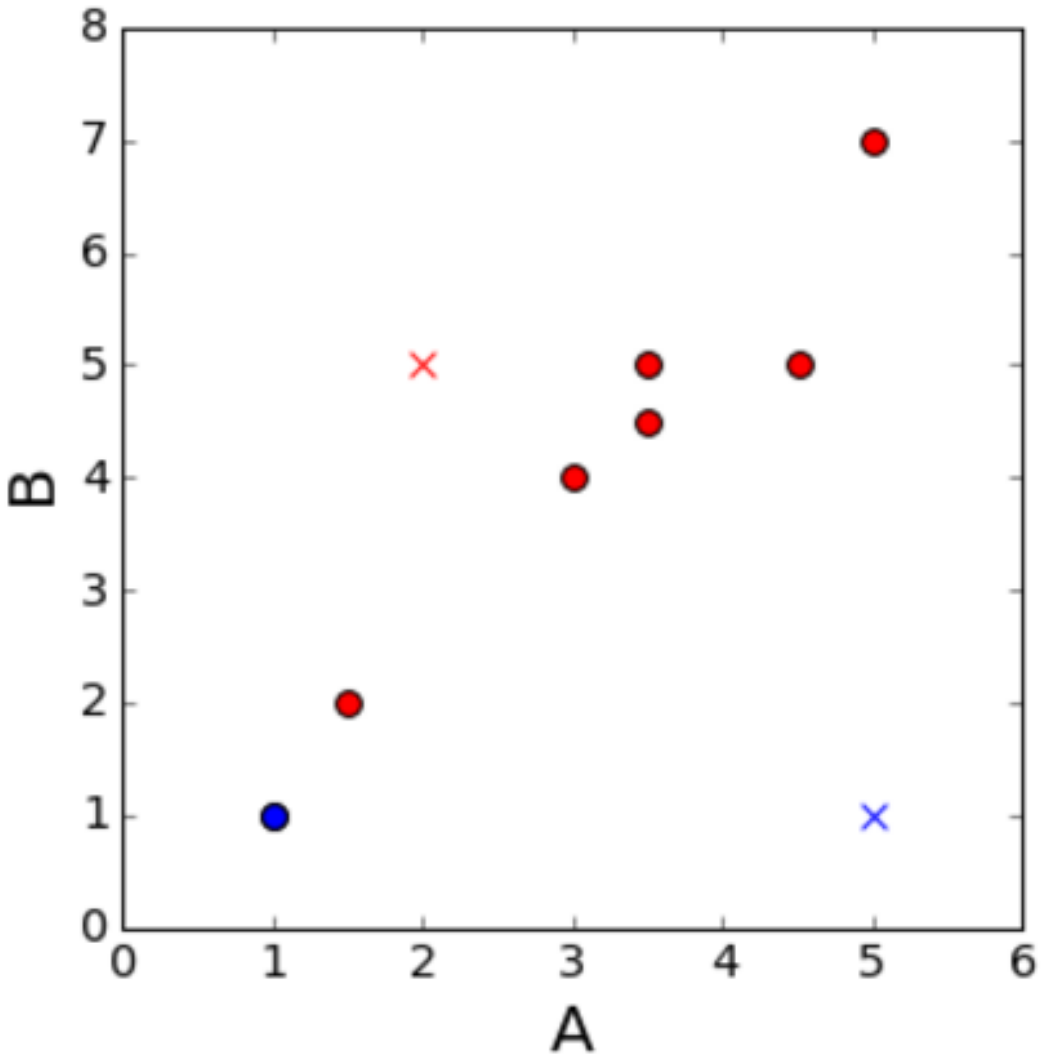
	Red Cluster (0)	Blue Cluster (1)
A	2	5
B	5	1



Step 2: Find the Euclidean Distance of Each Point from Each Cluster

Datapoint	Distance from centroid 0	Distance from centroid 1	Closest centroid
1	4.12	4	1
2	3.04	3.64	0
3	1.4	3.60	0
4	3.60	6.0	0
5	1.5	4.2	0
6	2.5	4.03	0
7	1.58	3.80	0

Give labels to your point: points get the label of the closest cluster



Recalculate the centroids: find the mean A & B for each cluster

- Centroid of cluster 0:
  A=mean(1.5,3.0,5.0,3.5,4.5,3.5)

  A= 3.5

  B= mean(2.0,4.0,7.0,5.0,5.0,4.5)

  B=4.58
- Centroid of cluster 1:

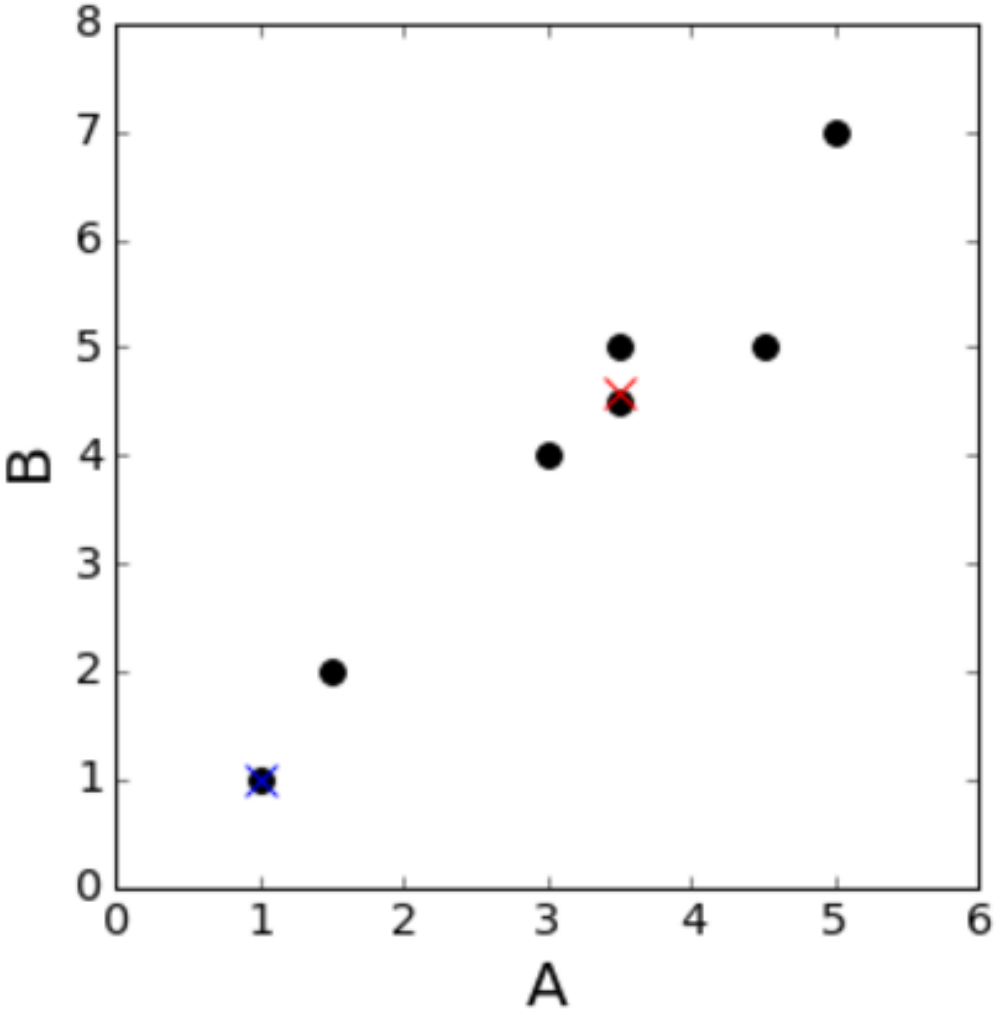
  A=mean(1)

  A=1

  B=mean(1)

  B=1

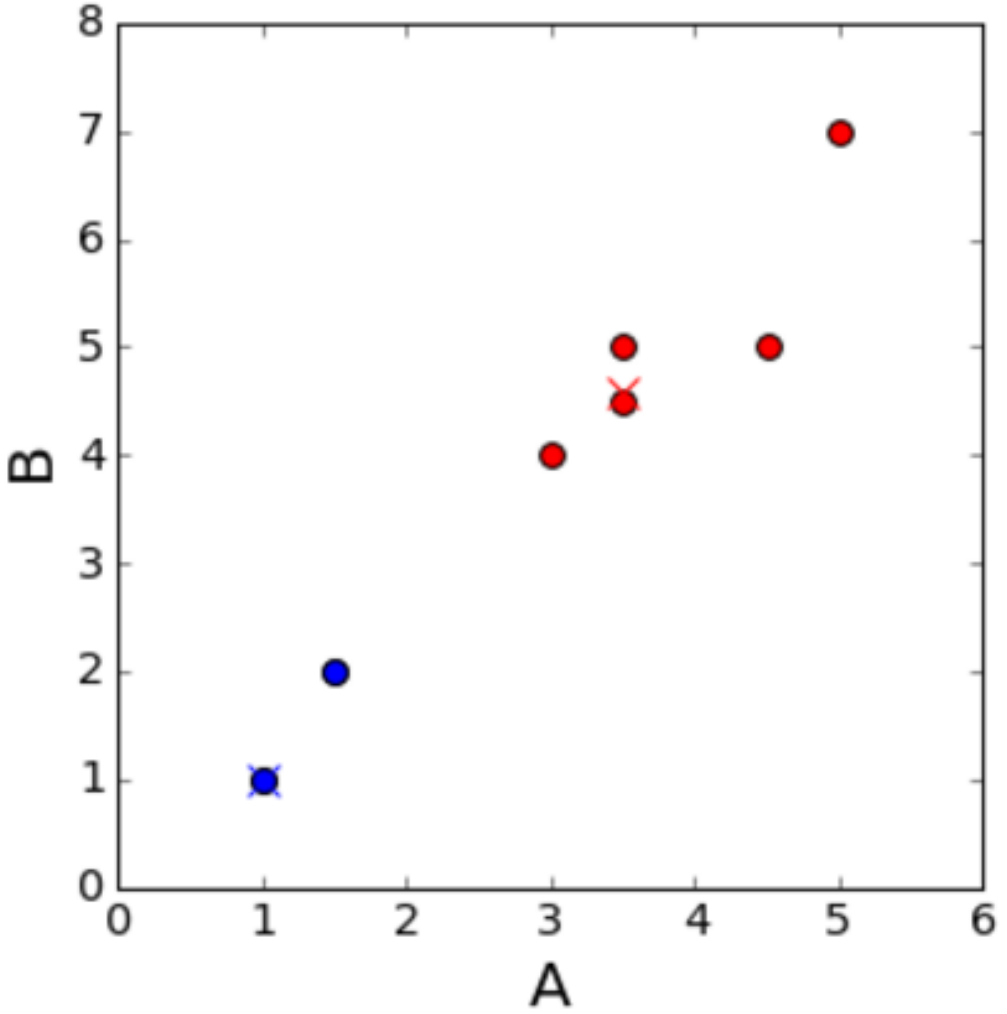
Our new centroids!



Repeat the two steps until your labels don't change anymore

- Reassign your labels:

Datapoint	Distance from centroid 0	Distance from centroid 1	Closest centroid
1	4.36	0	1
2	3.26	1.11	1
3	0.76	3.60	0
4	2.84	7.21	0
5	0.419	4.71	0
6	1.08	5.31	0
7	0.08	4.3	0


8
7
6
5
4
3
2
1
0
B
0
1
2
3
4
5
6
A

Repeat the two steps until your labels don't change anymore

- Recompute your centroids
- Centroid of cluster 0:
  A=mean(3.0,5.0,3.5,4.5,3.5)

  A= 3.89

  B= mean(4.0,7.0,5.0,5.0,4.5)

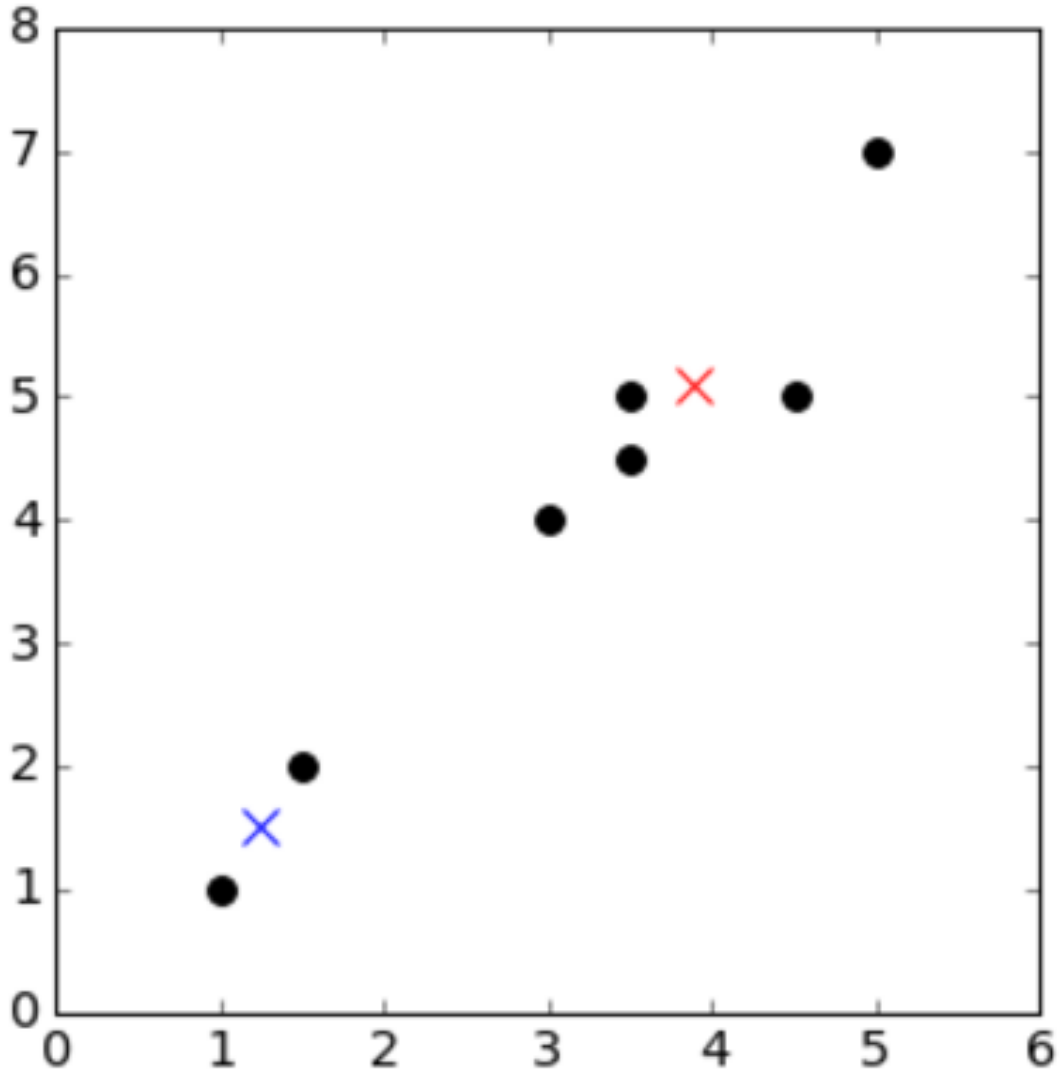
  B=5.09
- Centroid of cluster 1:

  A=mean(1,1.5)

  A=1.25

  B=mean(1,2)

  B=1.5

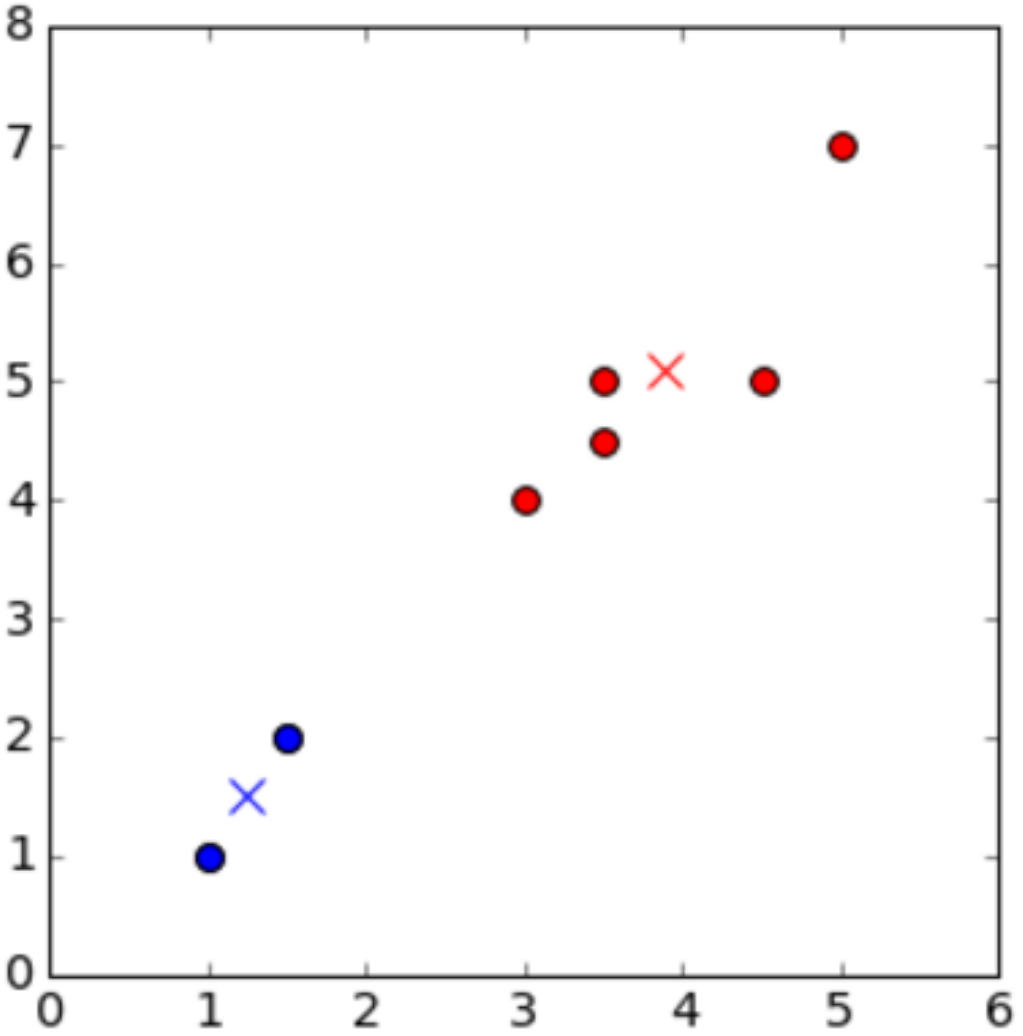

8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6

Repeat the two steps until your labels don't change anymore

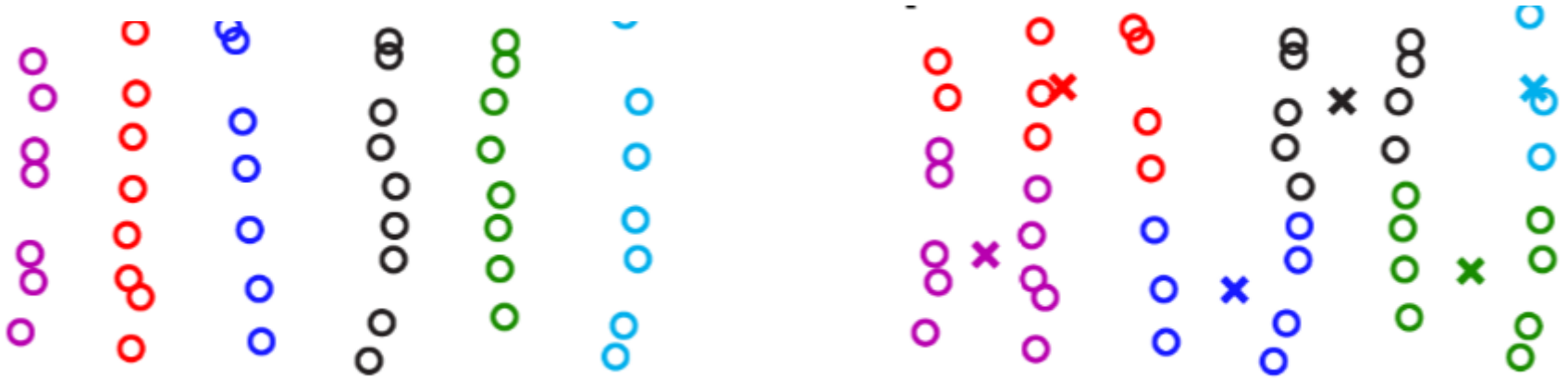
- Reassign your labels:

Datapoint	Distance from centroid 0	Distance from centroid 1	Closest centroid
1	5.00	0.55	1
2	3.91	0.55	1
3	1.40	3.05	0
4	2.21	6.65	0
5	0.400	4.16	0
6	0.61	4.77	0
7	0.70	3.75	0

Our labels have not changed since the last iteration, so we're done!



When will K-Means fail?



When will K-means fail?