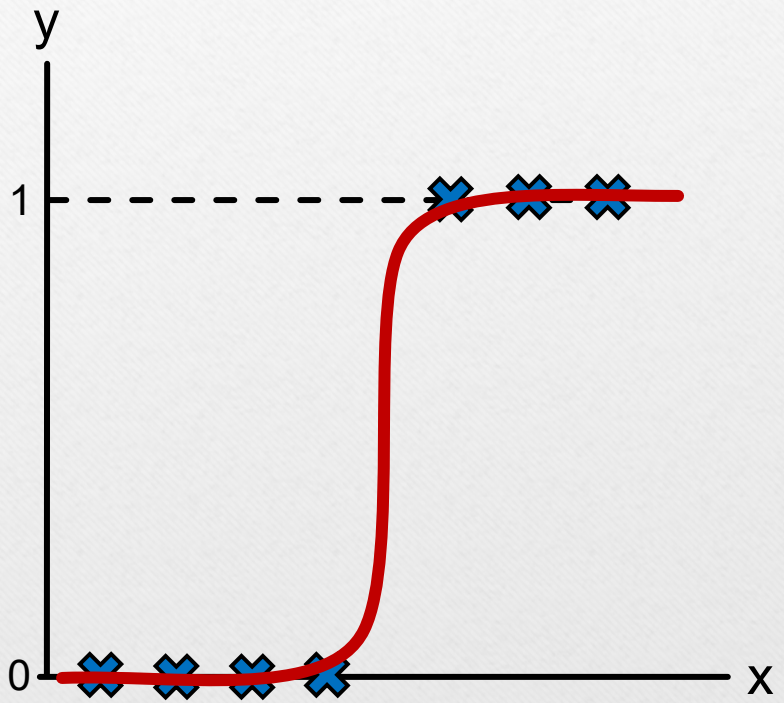
Clustering

ECON 4810

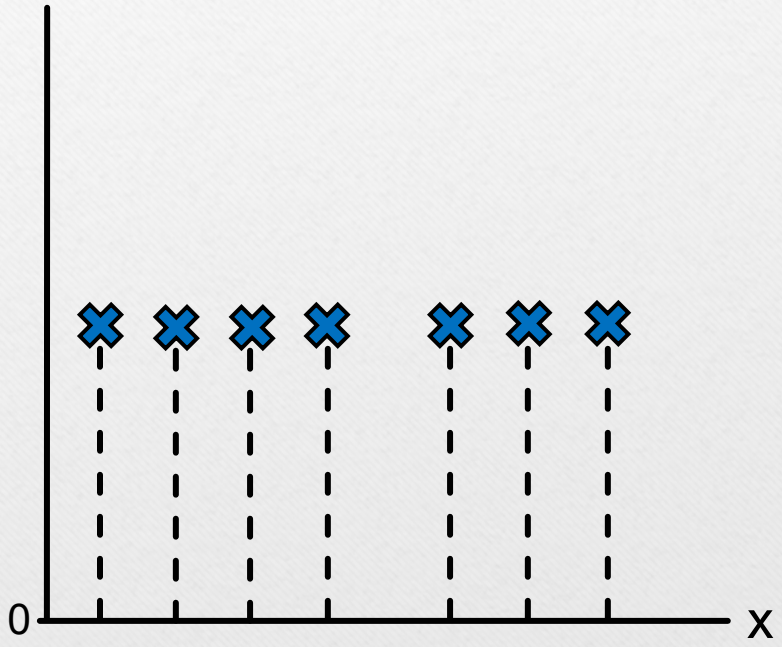
Classification

- Classification is a form of **supervised learning**. The model are given targets to learn from.



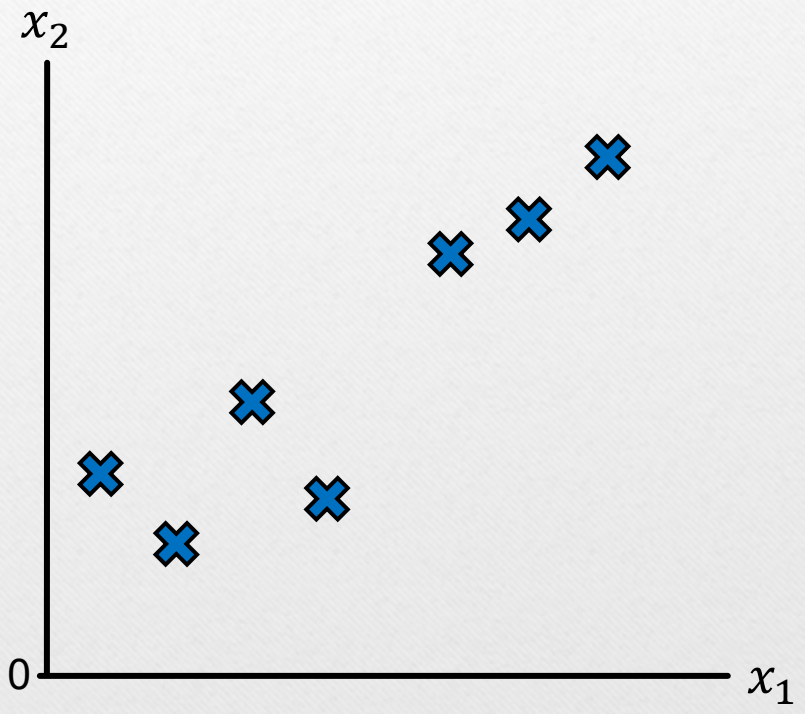
Clustering

- Clustering is a form of **unsupervised learning**. The model are *not* given targets to learn from.
- Samples are grouped by their features.



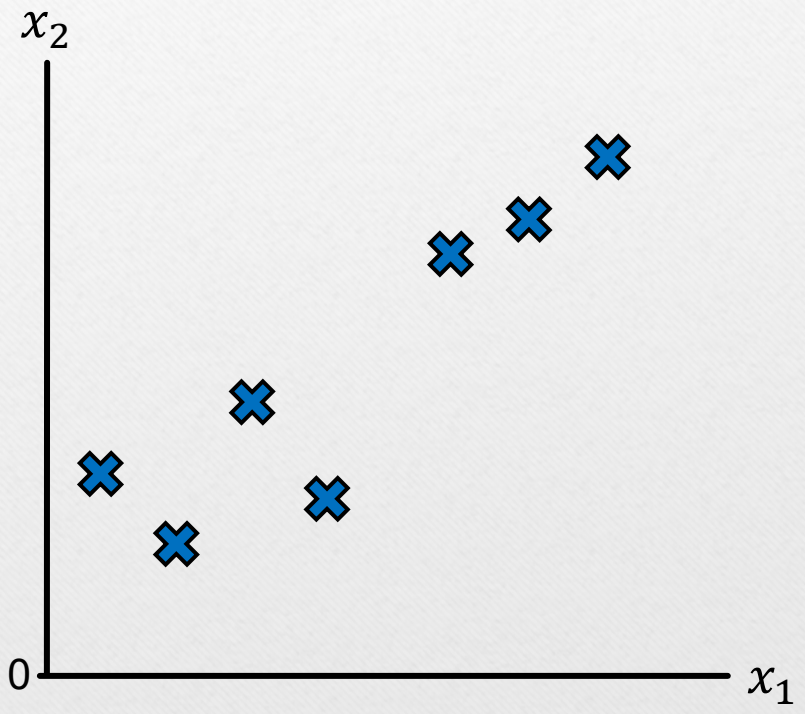
Clustering

- You will generally be clustering on multiple features.



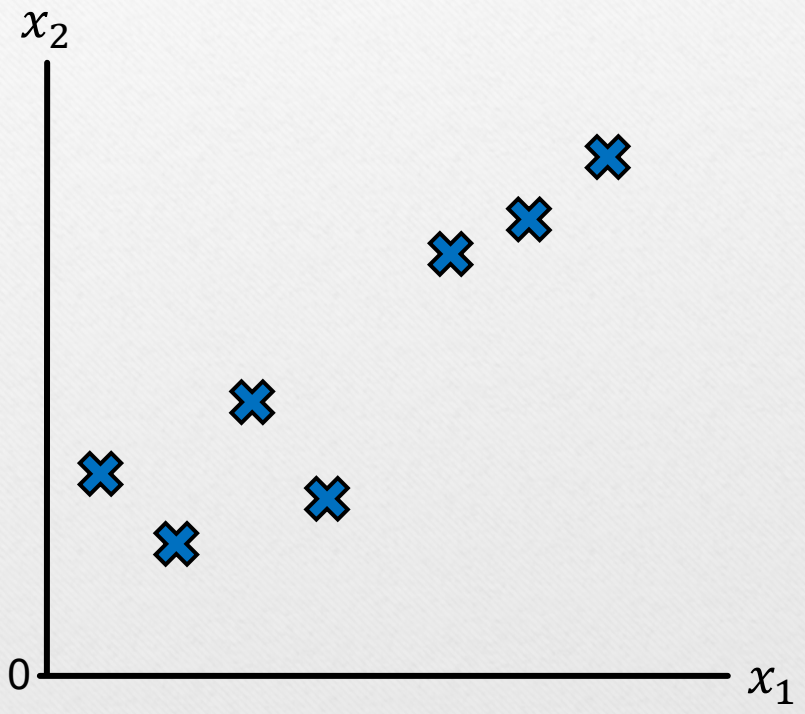
K-Means

- K-Means minimizes the sum-of-squares distance of each point from the mean in a group.
- The mean of a group is called a **centroid**.



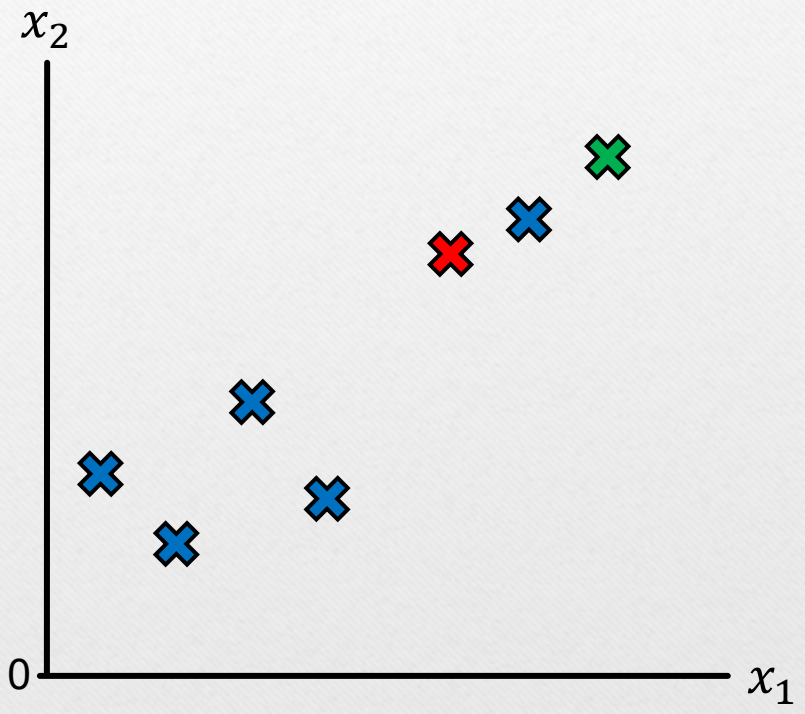
K-Means

- You first specify the number of groups you want.



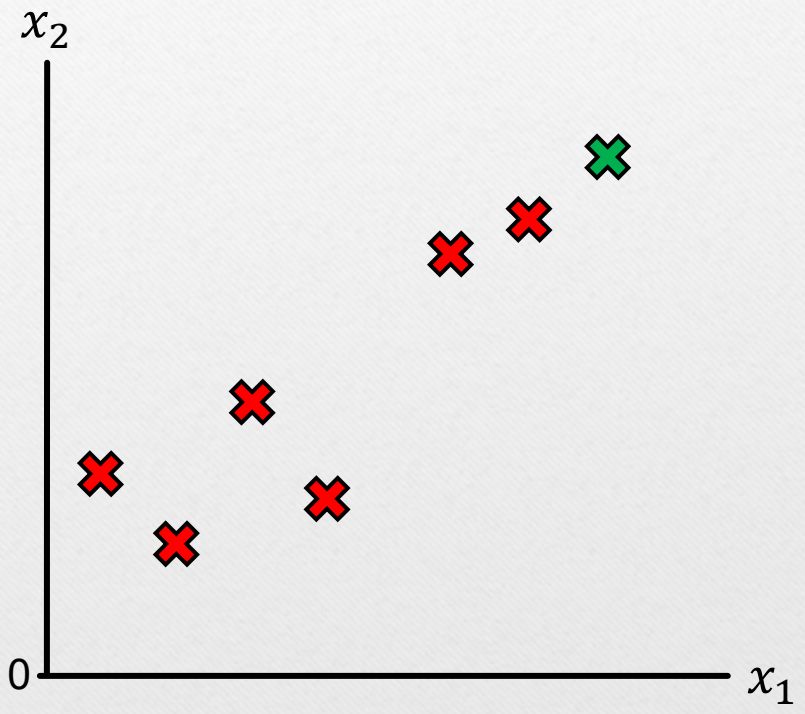
K-Means

- You first specify the number of groups you want.
- Random samples are chosen as initials centroids.



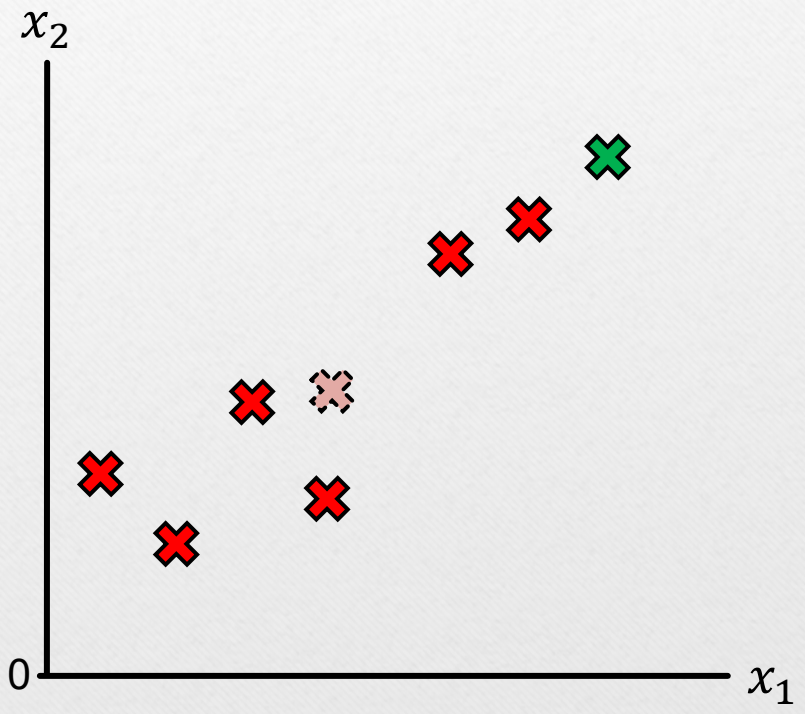
K-Means

- You first specify the number of groups you want.
- Random samples are chosen as initials centroids.
- Samples are assigned to the nearest centroid.



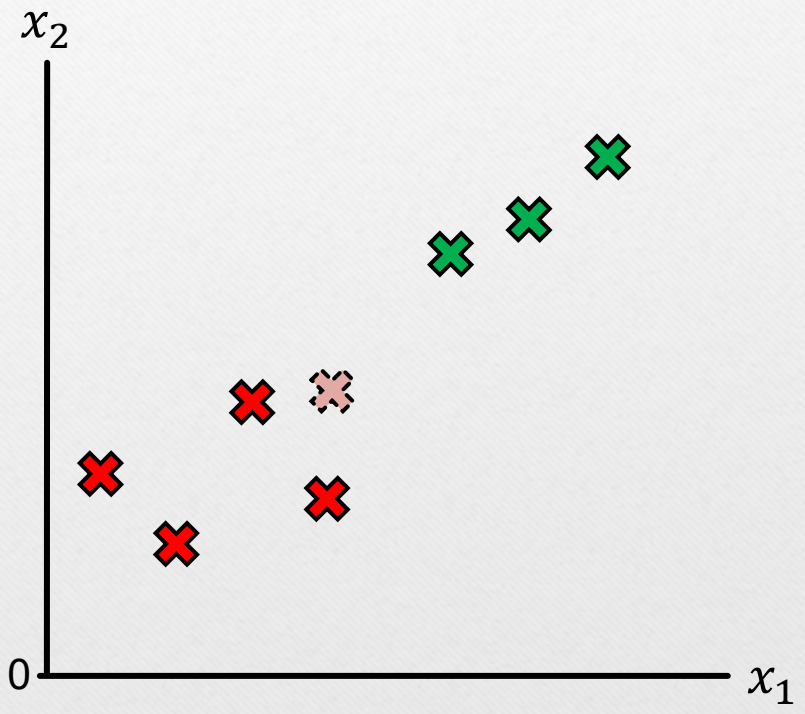
K-Means

- New centroids are then computed.



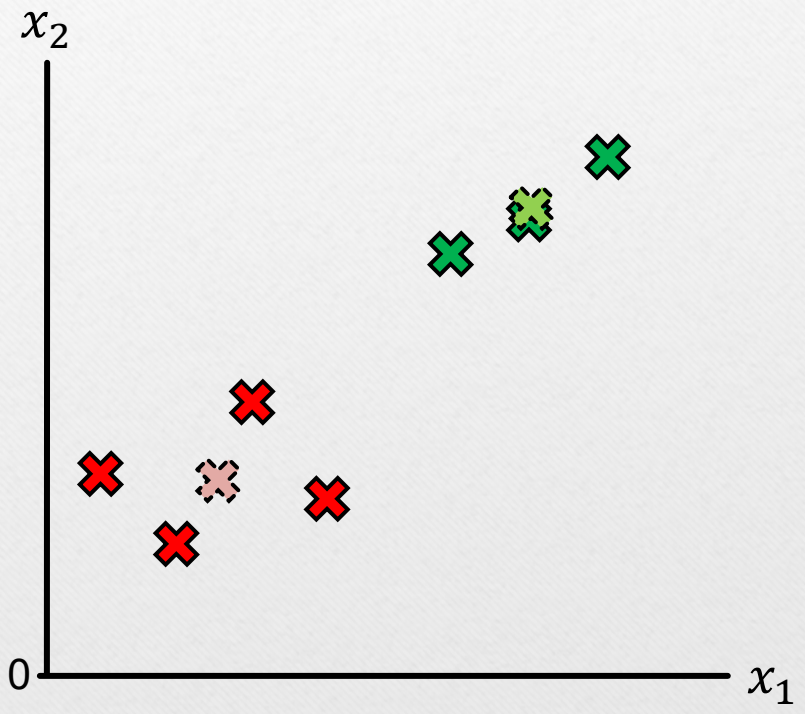
K-Means

- New centroids are then computed.
- Assign samples to nearest centroid.



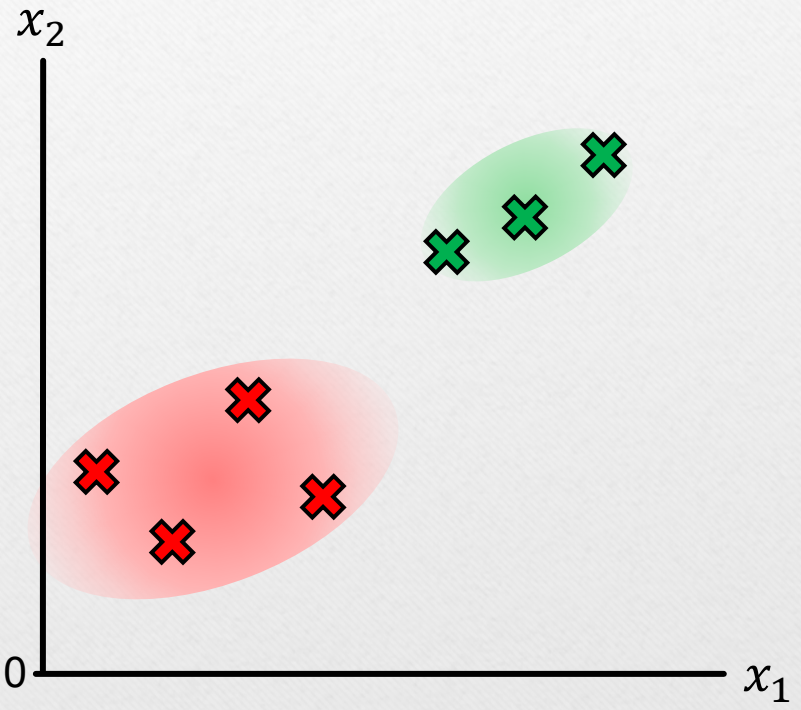
K-Means

- New centroids are then computed.
- Assign samples to nearest centroid.
- Repeat these two steps until the centroids are stable.



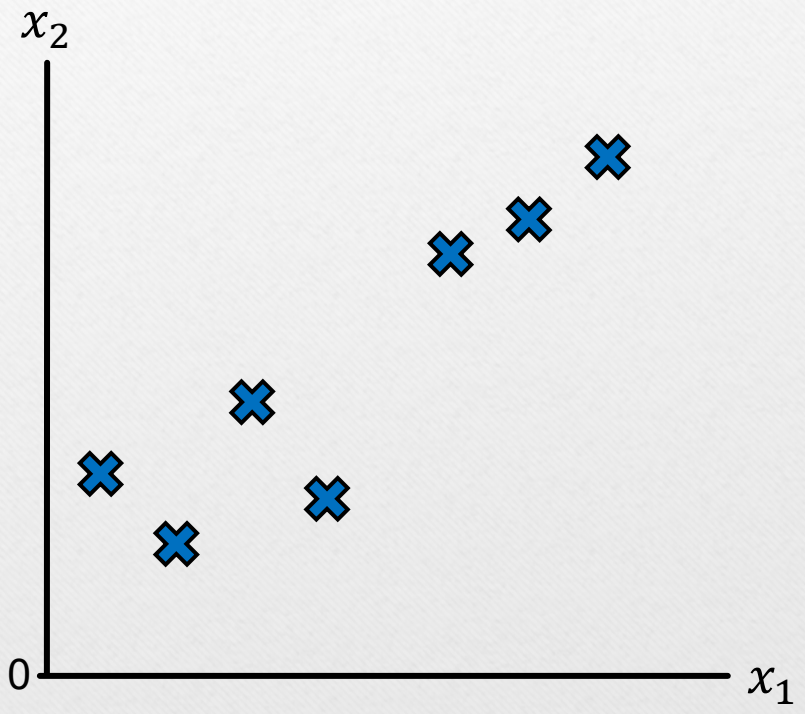
Gaussian Mixture

- Gaussian Mixture generalizes K-Means by assuming that each group is generated from a multinomial distribution.



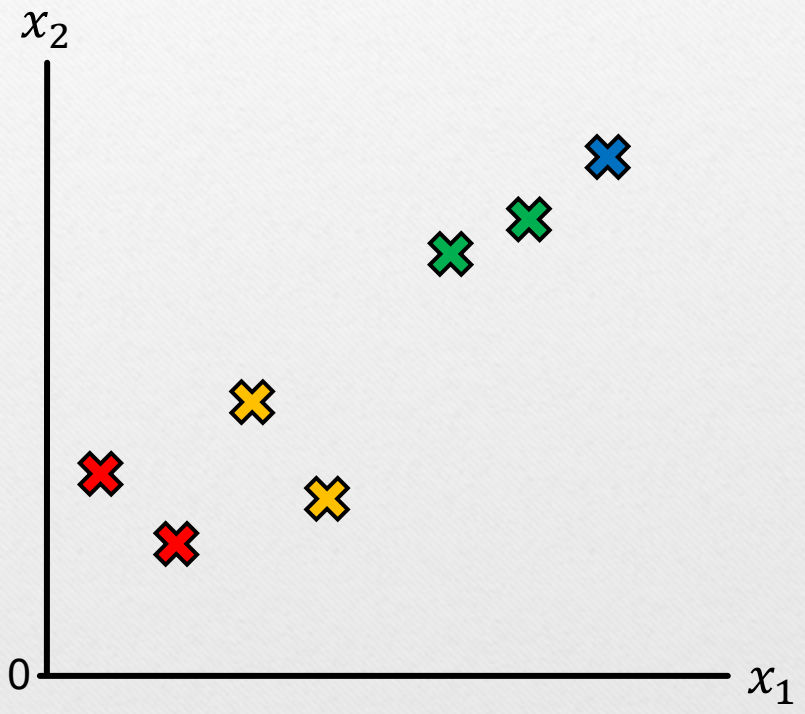
Agglomerative Clustering

- Each sample starts as its own cluster in Agglomerative clustering.



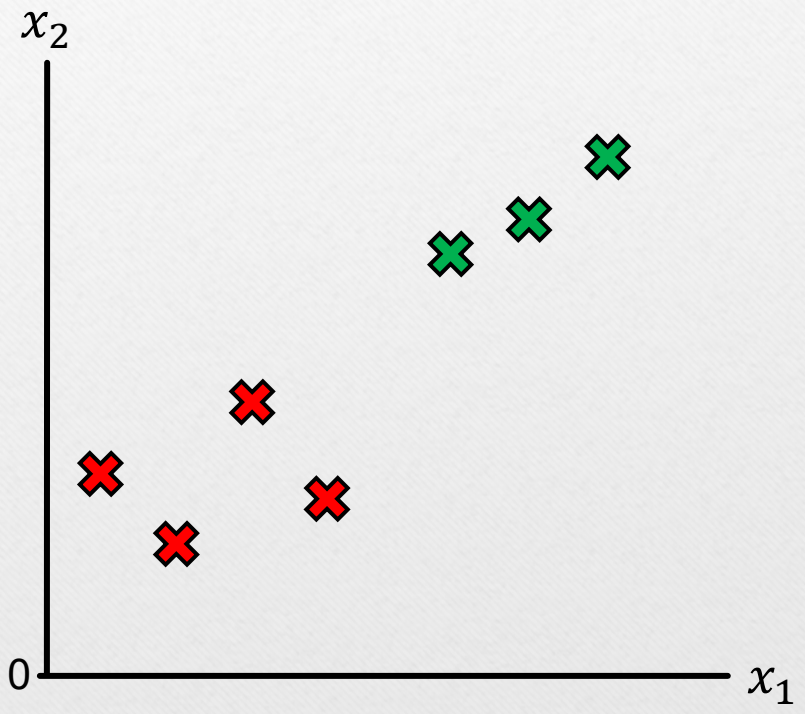
Agglomerative Clustering

- Each sample starts as its own cluster in Agglomerative clustering.
- Closest clusters are repeatedly merged until there are as many clusters as you specify.



Agglomerative Clustering

- Each sample starts as its own cluster in Agglomerative clustering.
- Closest clusters are repeatedly merged until there are as many clusters as you specify.



Clustering Algorithms in Scikit-learn