MONASH University

FIT1043 Introduction to Data Science

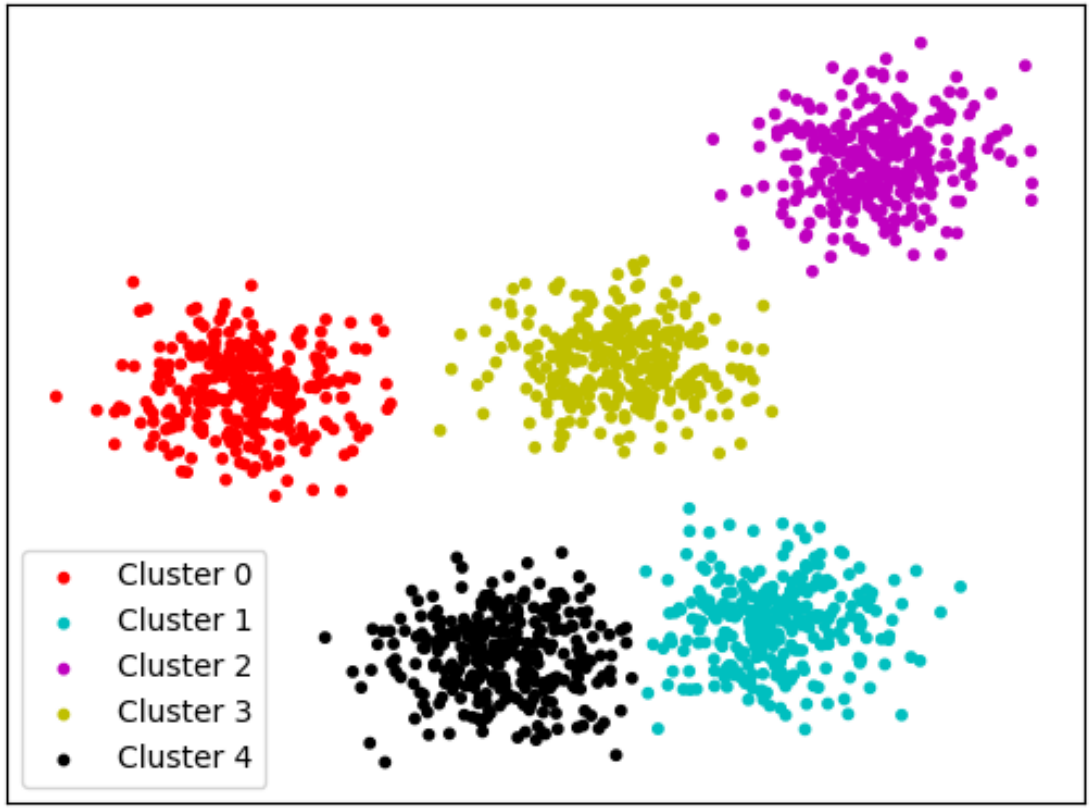
Week 7: Clustering

Ts. Dr. Sicily Ting

School of Information Technology
Monash University Malaysia

With materials from Wray Buntine, Mahsa Salehi

Clustering



MONASH University

What is Clustering?

From lecture notes by Andrew Ng

Grouping a set of data points into different subgroups based on their similarity

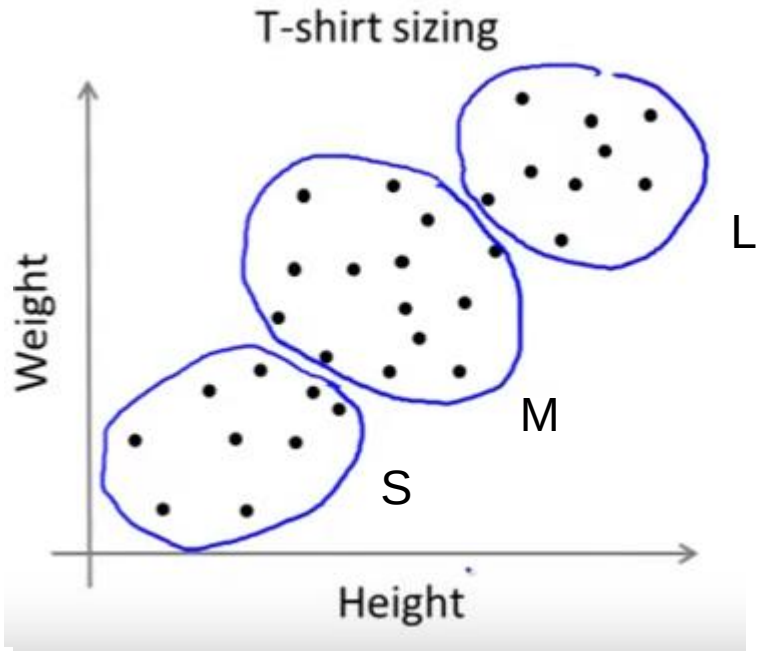
called clusters

k-means

T-shirt manufacturer

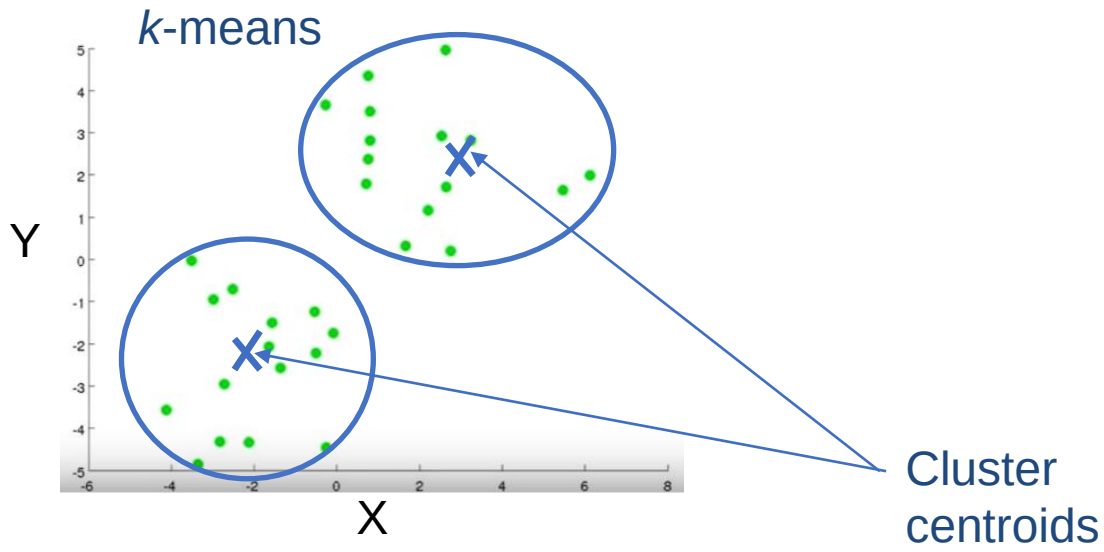
Group into 3 sizes:

- Small,
- Medium, and
- Large



k-means Clustering

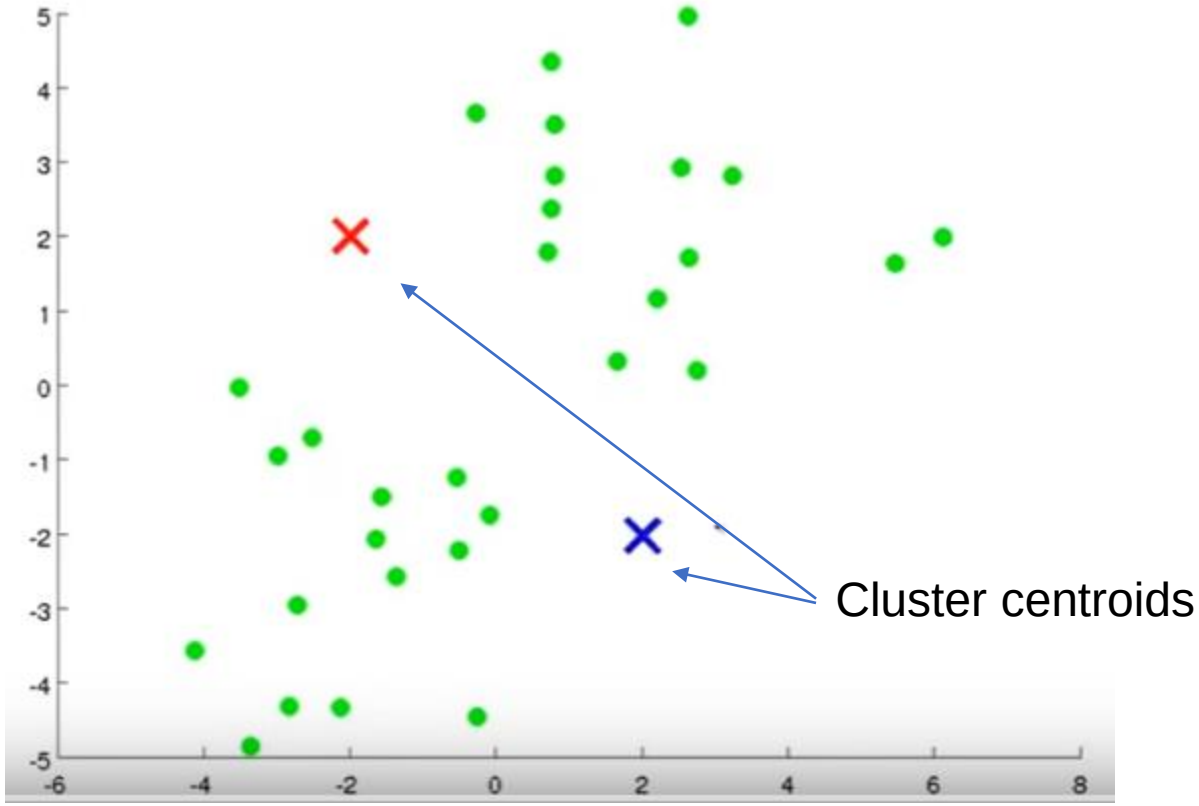
Example: Partition into two clusters based on similarity



k-means

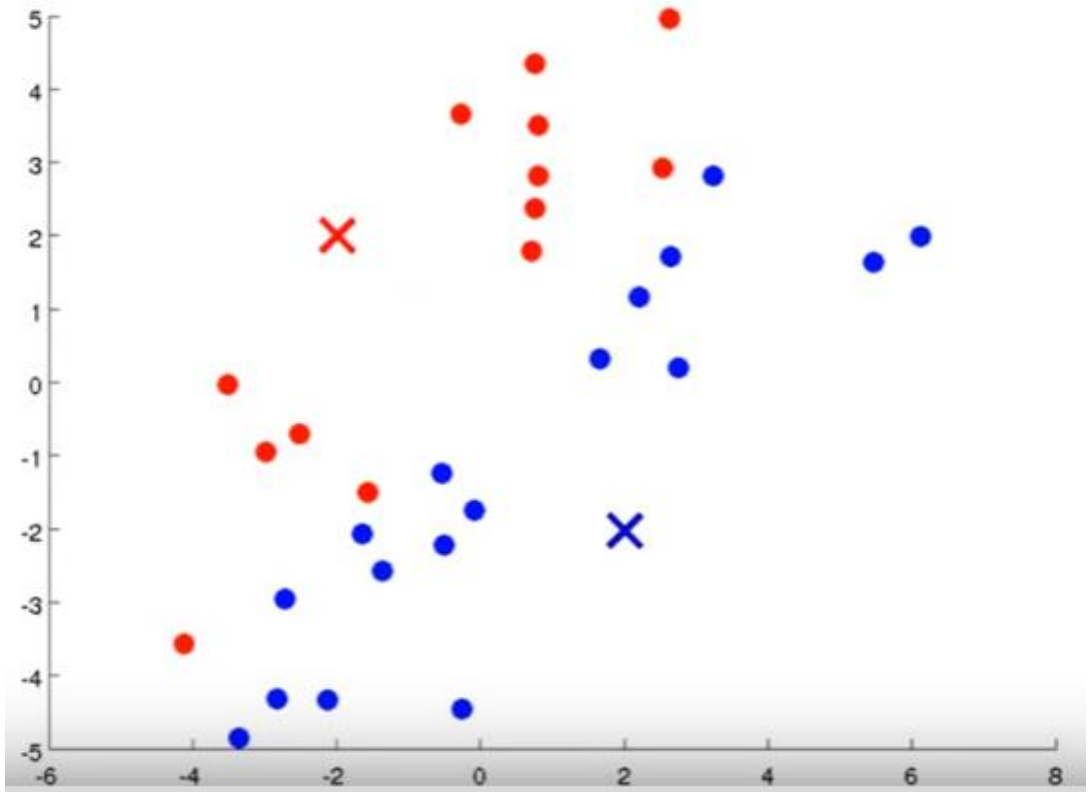
- *k* = the number of clusters
- Cluster centroid= The **mean (average)** of the location of all data points in a cluster

k-means Initial Setup



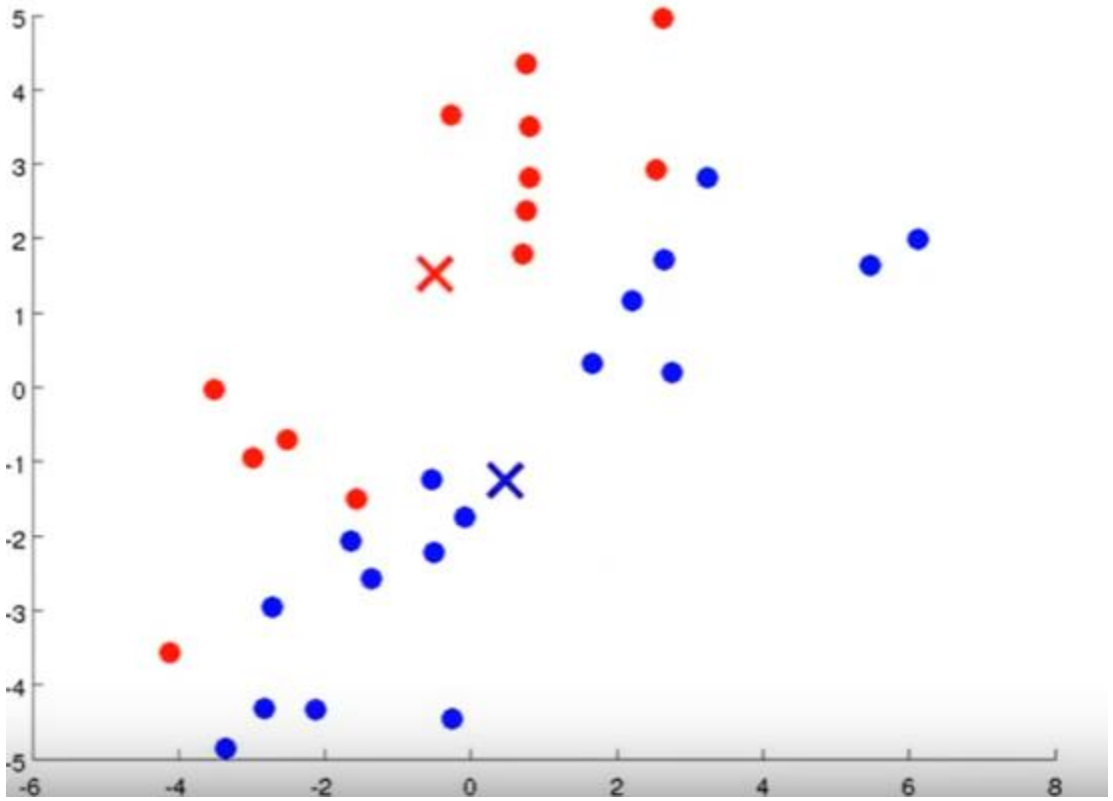
k-means Two Main (Iterative) Steps

1. **Cluster assignment**
2. Move centroid



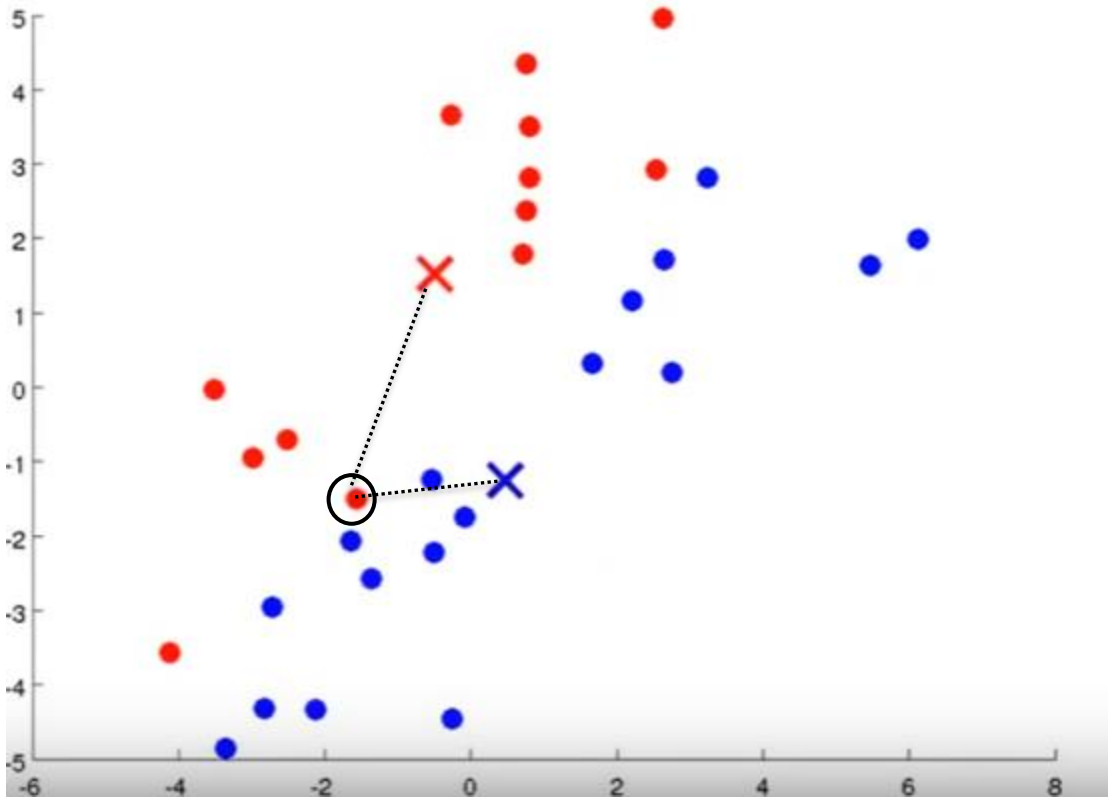
k-means Two Main (Iterative) Steps

1. Cluster assignment
2. **Move centroid**



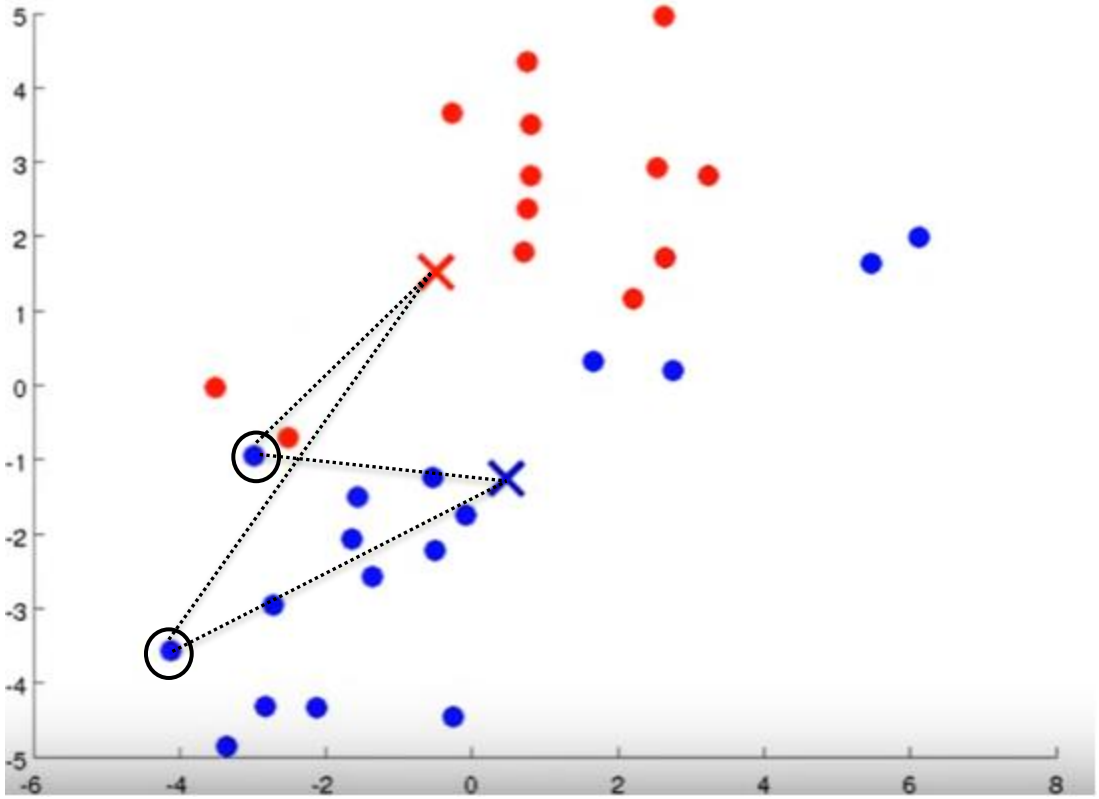
k-means Two Main (Iterative) Steps

1. **Cluster assignment**
2. Move centroid



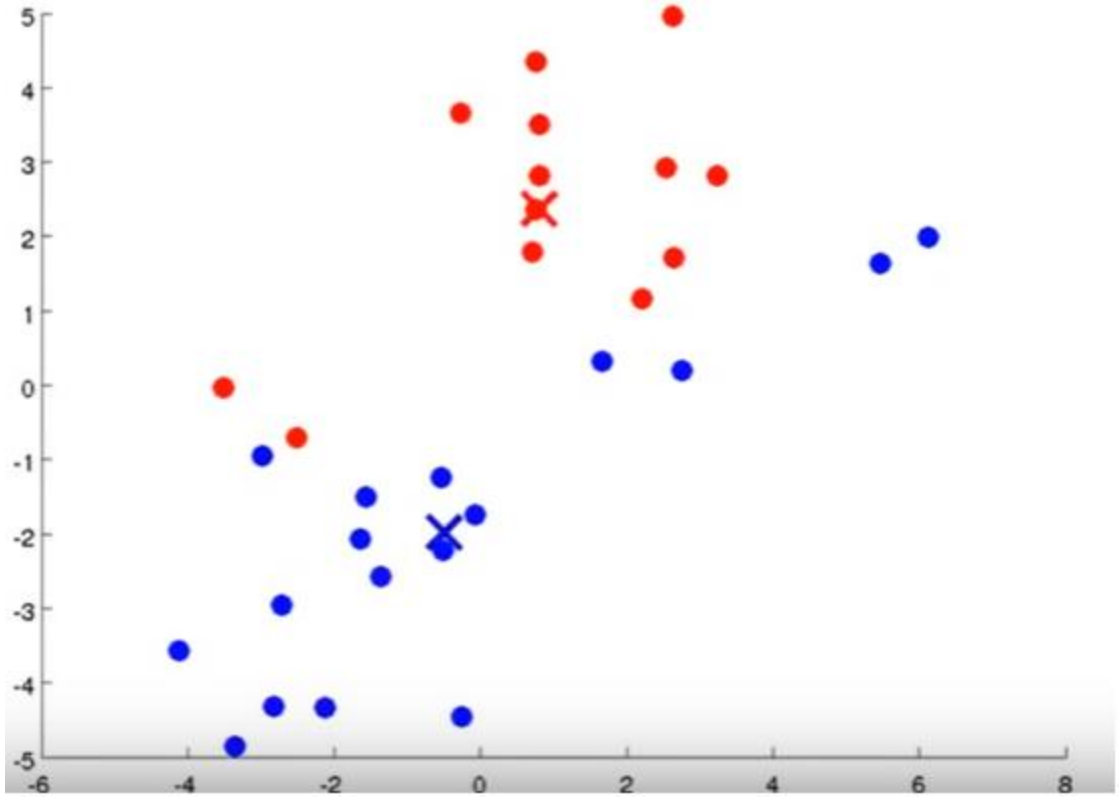
k-means Two Main (Iterative) Steps

1. **Cluster assignment**
2. Move centroid



k-means Two Main (Iterative) Steps

1. Cluster assignment
2. **Move centroid**

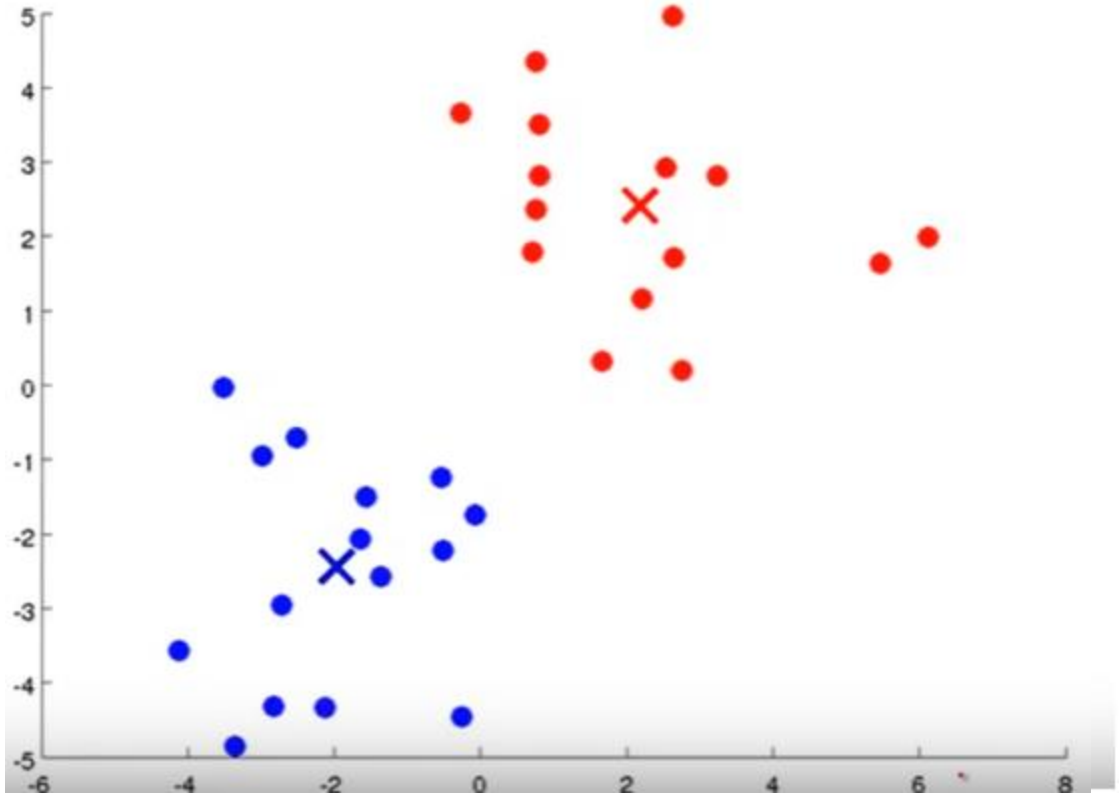


Iterate until there are no changes

k-means Two Main (Iterative) Steps

Iterate until there are no changes

1. Cluster assignment
2. Move centroid



k-means Algorithm

Input:

A set of data points

The number of clusters (K)

Method:

Select K initial random points

Repeat

Cluster assignment

Move the cluster centroids to the mean value of data points in the cluster

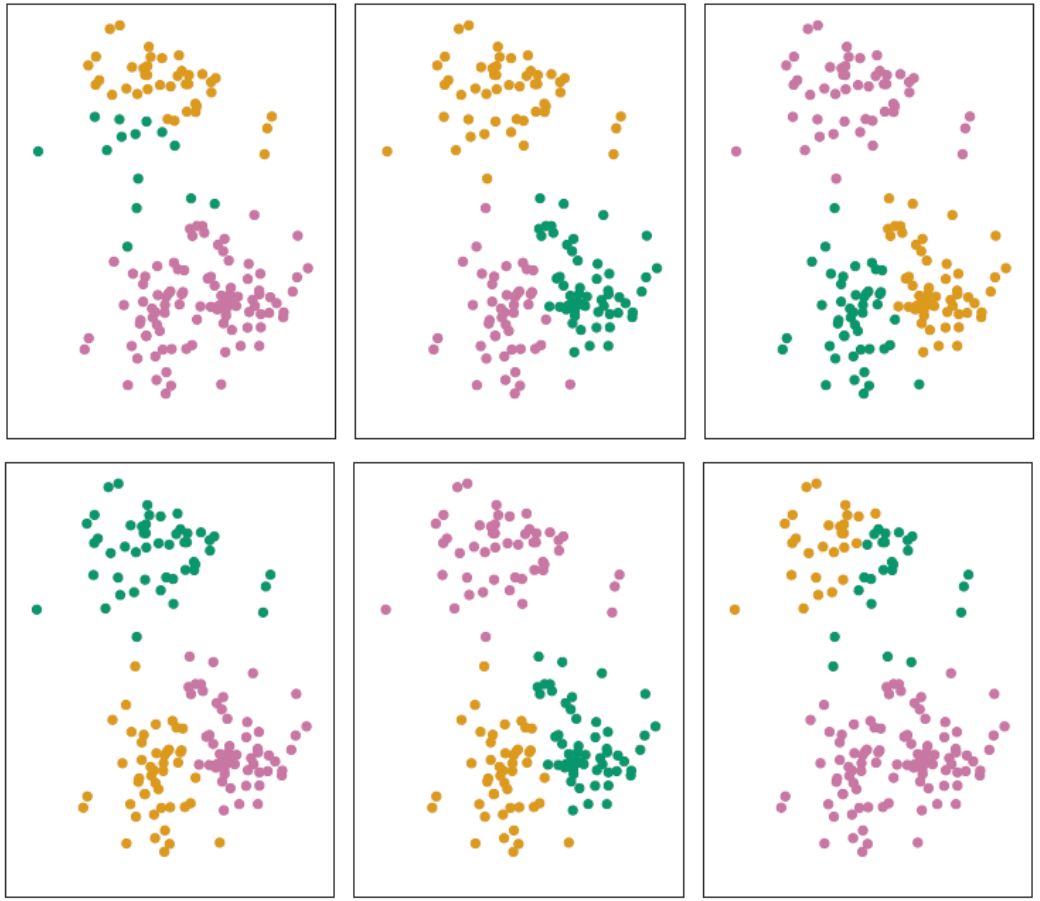
Until no change

Impact of Random Initial Points

How to choose K?
A priori knowledge about application domain
◦ There are two kinds of people in the world: k = 2
◦ There are five different types of bacteria: k = 5
◦ There are three different sizes of T-shirt: k = 3

Search for a good k
◦ Try different values of k and evaluate quality of results
◦ Run hierarchical clustering on subset of data

Random Initial Points
◦ *k* random data points are selected from the dataset
◦ highly volatile and provides for a scenario where the selected centroids are not well positioned throughout the entire data space.



Two Key Messages that We Learnt

1. Steps of *k*-means clustering

2. Importance of initial step in *k*-means

Learning Outcomes

Week 7

By the end of this week you should be able to:

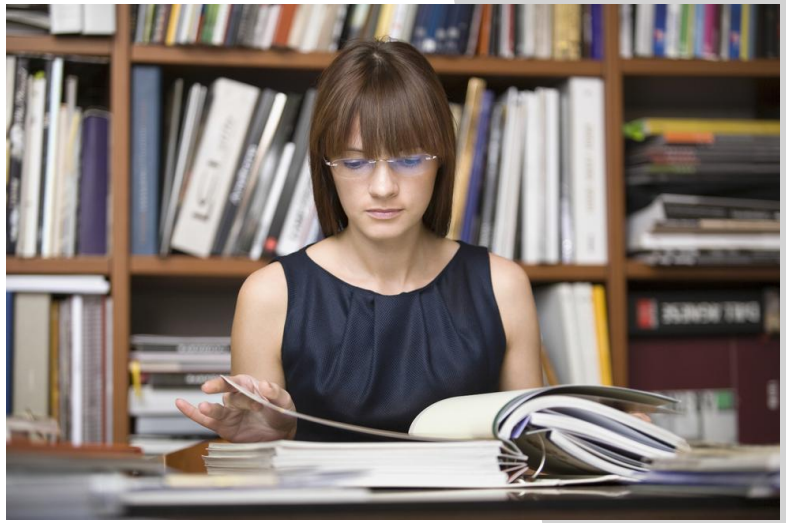
- Differentiate between classification and regression models
- Analyse confusion matrix and how to calculate prediction accuracy
- Differentiate between different classification metrics
- Explain how decision trees and regression trees work
- Explain how random forest works
- Explain how *k*-means clustering works

Home Activities

Suggested Activities for the week

Videos

Video (55 mins on evaluating a classification model but you can watch it at 1.5x to 1.75x speed): https://www.youtube.com/watch?v=85dtiMz9tSo&list=PL5-da3qGB5ICeMbQuqbbCOQWcS6OYBr5A&index=9



Articles

Read The Star article on 5th April 2020 and understand the importance of being able to interpret sensitivity, specificity and accuracy.

MONASH University