MONASH University

ETC3250

# Business Analytics

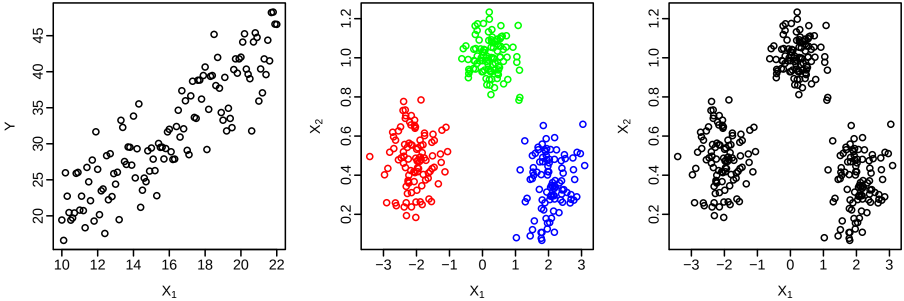
**Week 12**
**Clustering**

17 October 2016

# Outline

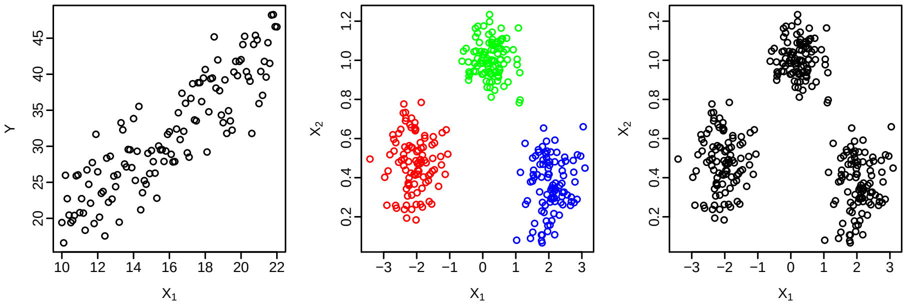
| Week | Topic | Chapter | Lecturer |
|---|---|---|---|
| 1 | Introduction to business analytics & R | 1 | Souhaib |
| 2 | Statistical learning | 2 | Souhaib |
| 3 | Regression for prediction | 3 | Souhaib |
| 4 | Resampling | 5 | Souhaib |
| 5 | Dimension reduction | 6,10 | Souhaib |
| 6 | Visualization | | Di |
| 7 | Visualization | | Di |
| 8 | Classification | 4,8 | Di |
| 9 | Classification | 4,9 | Di |
| | - | | |
| 10 | Classification | 8 | Souhaib |
| 11 | Advanced regression | 6 | Souhaib |
| 12 | Clustering | 10 | Souhaib |

# Learning problems




→ Clustering is an **unsupervised learning method**

# Learning problems




→ Clustering is an **unsupervised learning method**

# Unsupervised learning

- **Unsupervised learning** is often performed as part of an **exploratory data analysis**.
- Unsupervised learning is often much **more challenging than supervised learning**. The exercise tends to be more **subjective**, and there is no simple goal for the analysis, such as prediction of a response
- **Hard to assess** the results obtained from unsupervised learning methods
- Techniques for unsupervised learning are of **growing importance** in a number of fields
- Examples of unsupervised learning methods?

# Unsupervised learning methods

Both **PCA** and **clustering** seek to simplify the data via a small number of summaries, but their mechanisms are different:

- **PCA** (unsupervised dimension reduction method) looks to find a low-dimensional representation of the observations that explain a good fraction of the variance
- **Clustering** looks to find homogeneous subgroups among the observations.

# Clustering

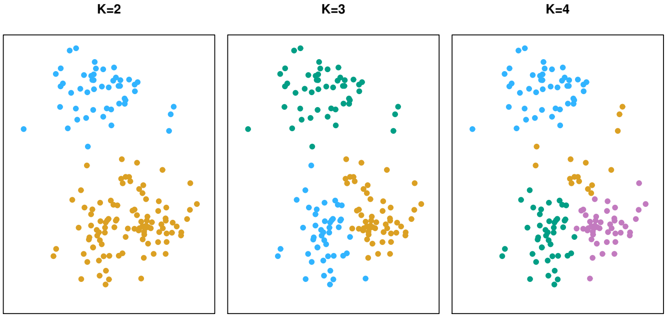
Since clustering is popular in many fields, there exist a great number of clustering methods.

- **K-means clustering**
  - We seek to partition the observations into $K$ clusters.
- **Hierarchical clustering**
  - We do not know in advance how many clusters we want. We consider all possible number of clusters, from 1 to $n$.

# K-means clustering

Find $K$ clusters $C_1, \ldots, C_K$ where

- $C_1 \cup C_2 \cup C_K = \{1, \ldots, n\}$
- $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$



# K-means clustering

- The **within-cluster variation** for cluster $C_k$ is a measure $W(C_k)$ of the amount by which the **observations within a cluster differ from each other**
- The idea behind K-means clustering is that **a good clustering is one for which the within-cluster variation, summed over all** $K$ **clusters, is as small as possible**

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

# K-means within-cluster variation

There are many possible ways to define the **within-cluster variation**, but by far the most common choice involves **squared Euclidean distance**:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2,$$

# K-means optimization problem

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

- There are almost $K^n$ ways to partition $n$ observations into $K$ clusters. This is a huge number unless $K$ and $n$ are tiny.
- Fortunately, a very simple algorithm can be shown to provide a local optimum—a pretty good solution—to the K-means optimization problem

# K-means optimization problem

**Algorithm 10.1** *$K$-Means Clustering*

1. Randomly assign a number, from 1 to $K$, to each of the observations. These serve as initial cluster assignments for the observations.

2. Iterate until the cluster assignments stop changing:

   (a) For each of the $K$ clusters, compute the cluster *centroid*. The $k$th cluster centroid is the vector of the $p$ feature means for the observations in the $k$th cluster.

   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

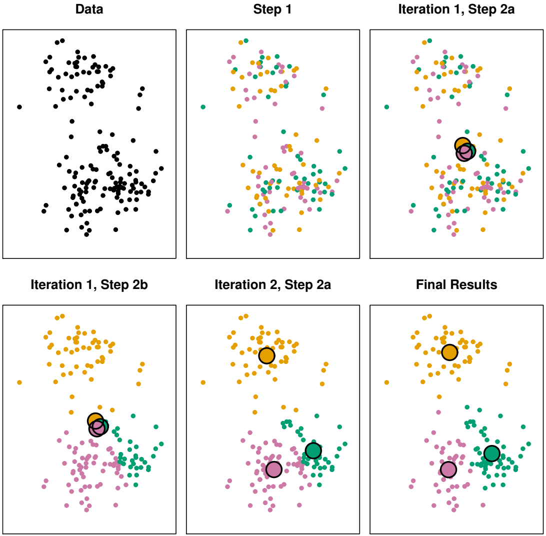
# K-means optimization problem

The K-means algorithm is **guaranteed to decrease the value of the objective at each step**. The following identity helps to understand:

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$$

$$\bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i\in C_k} x_{ij}$$

# K-means clustering



# K-means clustering

Different random initial configurations