Section 9: Clustering

STA 141A – Fundamentals of Statistical Data Science

Instructor: Akira Horiguchi

Fall Quarter 2025 (Sep 24 – Dec 12)
MWF, 9:00 AM – 9:50 AM, TLC 1215
University of California, Davis

Overview

Based on Chapter 12 of ISL book James et al. (2021).

- For more R code examples, see R Markdown files in `https://www.statlearning.com/resources-second-edition`

Supervised data: predictors $X_1, \ldots, X_p$ and a response Y measured on n observations.

Unsupervised data: predictors $X_1, \ldots, X_p$ measured on n observations, but no response.

- Still useful to analyze the association between the predictors $X_1, \ldots, X_p$.
- Often performed as part of an exploratory data analysis.
- Harder to assess the results from an unsupervised learning method; there is no "truth" to compare to. (In contrast, in supervised learning the "truth" is the response Y.)

Common unsupervised learning task: find homogeneous subgroups (i.e., *clusters*) among observations.

- "Market segmentation" aims to identify subgroups of people who might be more receptive to certain kind of advertisements/products etc.
- Flow cytometry: group cells based on their biomarker values.

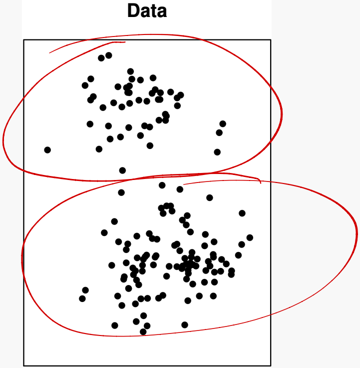


Figure 1: From James et al. (2021).

If we index the n observations by the integers $1, 2, 3, \ldots, n$, then

$$\text{cluster } n \text{ observations} \iff \text{cluster the integers } 1, 2, 3, \ldots, n$$

In other words, we want to partition the set $\{1, 2, 3, \ldots, n\}$.

Definition (Cluster)

Clusters are sets $C_1, \ldots, C_K$ with the following features:

- $C_1 \cup C_2 \cup \cdots \cup C_K = \{1, \ldots, n\}$ (each observation belongs to at least one cluster);
- $C_k \cap C_l = \emptyset$ for all $k \neq l$ (no observation belongs to more than one cluster).

There are almost K^n ways to partition n observations into K clusters.

- How to select "best" clustering of given data?
- A common algorithmic technique: *K-means clustering*

K-means clustering

The user chooses a positive integer K before performing *K-means clustering*.

- "Good" clustering: if the observations in each cluster are close to each other, i.e., if the *within-cluster variation* is relatively small.
- For observations $x_1, \ldots, x_n \in \mathbb{R}$, within-cluster variation of a cluster C defined by

Example
$x_1 = 1 \quad x_2 = 3 \quad x_3 = 7$
$W(C) = \frac{1}{3}\left[(1-3)^2 + (1-7)^2 + (3-7)^2\right]$

$$W(C) := \boxed{\frac{1}{\#C} \underbrace{\sum_{i,i' \in C}}_{\leftarrow \text{ all pairs of indices in } C} (x_i - x_{i'})^2} . \tag{1}$$

Here $\#C$ denotes the number of observations in cluster C.
- For observations $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, within-cluster variation of a cluster C defined by

$$W(C) := \boxed{\frac{1}{\#C} \sum_{i,i' \in C} \|\mathbf{x}_i - \mathbf{x}_{i'}\|_2^2} = \frac{1}{\#C} \sum_{i,i' \in C} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 . \tag{2}$$

Goal: We want to find clusters $C_1, \ldots, C_K$ that minimize

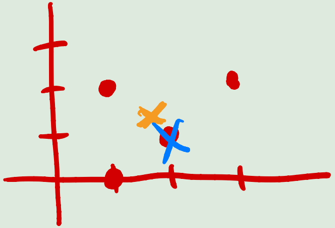
$$\sum_{k=1}^{K} W(C_k) \tag{3}$$

- It is very difficult to find the global minimizer, since there are almost K^n ways to partition n observations into K clusters.
- The following algorithm can be shown to provide a local minimizer.

Algorithm for K-*means* clustering

1. Randomly assign a number from 1 to K (K is pre-defined) to each observation.
2. Iterate steps (a) and (b) until the cluster assignments stop changing:
   (a) For each cluster, compute *cluster centroid* (mean of all observations in the cluster).
   (b) Assign each observation to the cluster whose centroid is the closest.

Example (Draw & compute the centroid of the points $(1,2),(2,1),(3,2),(1,0)$)



$$\frac{1}{4}\left[(1,2)+(2,1)+(3,2)+(1,0)\right]$$
$$=\frac{1}{4}\left(1+2+3+1,\ 2+1+2+0\right)$$
$$=\frac{1}{4}\left(7,\ 5\right)$$
$$=(1.75,\ 1.25)$$

Comments:

- Name: cluster centroids are computed as the *mean* of each cluster's observations.
- Step 2 will reduce (3) until at local minimum. The obtained value will depend on the initial (random) cluster assignment from Step 1.
- To reduce probability of choosing a "bad" local minimum, one should run the algorithm many times, and then choose clustering w/smallest value of (3).

Simulation of K-means clustering

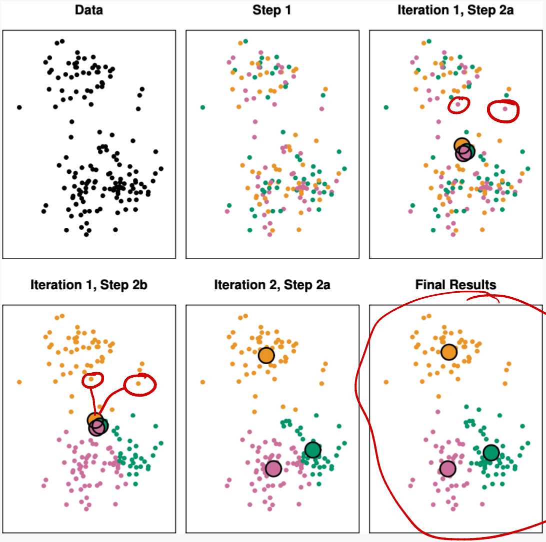


Figure 2: From James et al. (2021). 3-means clustering and 10 iterations.

Issues in clustering

Issues in clustering

- Should the features first be standardized in some way? E.g. maybe scale them to have standard deviation one?
- K-means clustering: how many clusters should we look for?

It is challenging to validate obtained clusters.

- Do obtained clusters represent true subgroups in the data, or are they a result of clustering noise?
- Outside scope of class; more details found in "sequel" book *The Elements of Statistical Learning*
- In practice, try several different choices, and look for the one with the most useful or interpretable solution.