14.3 – Cluster analysis in the book "The elements of statistical learning"

Wenjing Liao

School of Mathematics
Georgia Institute of Technology

Math 4803
Fall 2019

Clustering

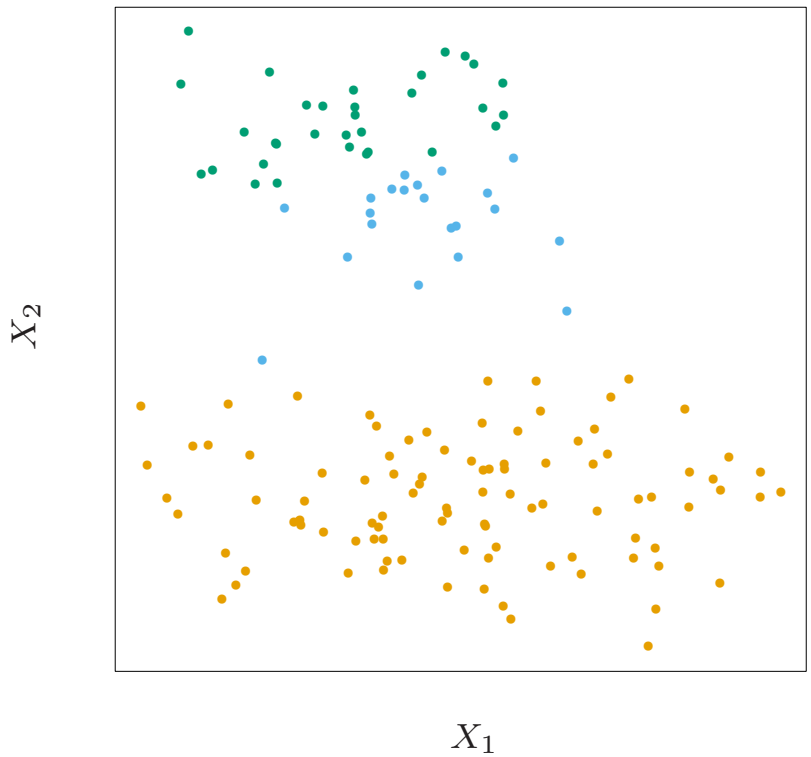

FIGURE 14.4. *Simulated data in the plane, clustered into three classes (represented by orange, blue and green) by the K-means clustering algorithm*

Outline

14.1 – Proximity matrices

Data: $\{x_i\}_{i=1}^N$

Proximity matrix: $\mathbf{D} \in \mathbb{R}^{N \times N}$, $d_{ii} = 0$

Symmetrization: If $\mathbf{D}$ is not symmetric, we can take $(\mathbf{D} + \mathbf{D}^T)/2$

Dissimilarities based on attributes:

$$D(x_i, x_{i'}) = \sum_{j=1}^{p} d_j(x_{ij}, x_{i'j})$$

where x_{ij} is the jth attribute of the ith data.

Dissimilarities

- Euclidean distance

$$d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2$$

- More general

$$d(x_i, x_{i'}) = l(|x_i - x_{i'}|)$$

  where $l(\cdot)$ is a monotone-increasing function.

- Based on correlation

$$\rho(x_i, x_{i'}) = \frac{\sum_j (x_{ij} - \bar{x}_i)(x_{i'j} - \bar{x}_{i'})}{\sqrt{\sum_j (x_{ij} - \bar{x}_i)^2 \sum_j (x_{i'j} - \bar{x}_{i'})^2}},$$

  where $\bar{x}_i = \sum_j x_{ij}/p$

Weighted average

$$D(x_i, x_{i'}) = \sum_{j=1}^{p} w_j \cdot d_j(x_{ij}, x_{i'j}); \quad \sum_{j=1}^{p} w_j = 1.$$

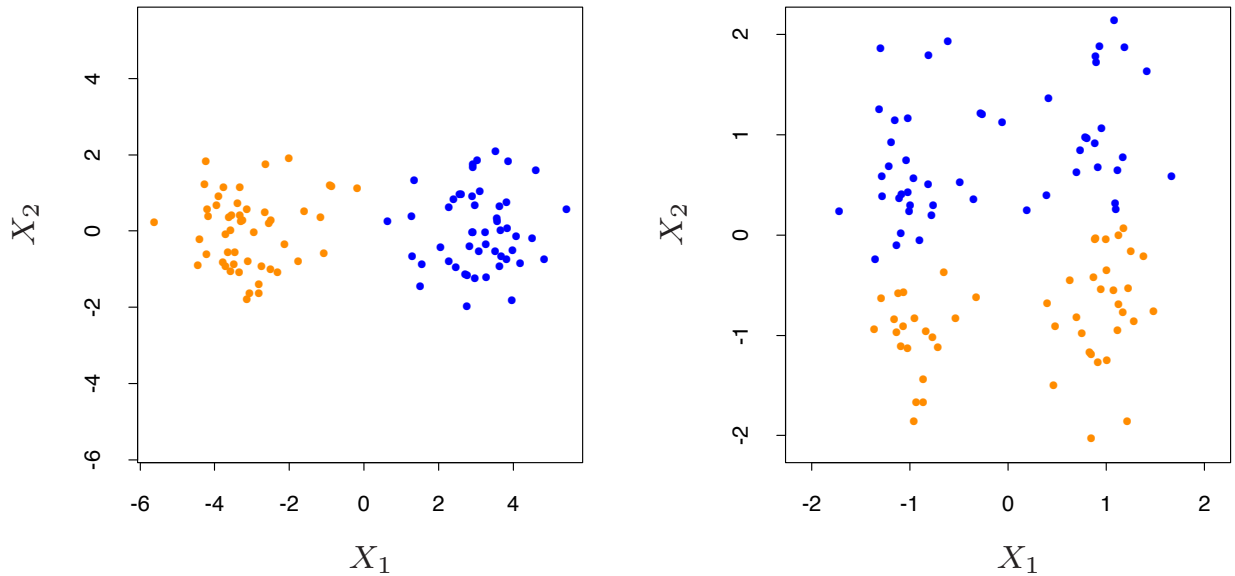

FIGURE 14.5. *Simulated data: on the left, K-means clustering (with $K=2$) has been applied to the raw data. The two colors indicate the cluster memberships. On the right, the features were first standardized before clustering. This is equivalent to using feature weights $1/[2 \cdot \mathrm{var}(X_j)]$. The standardization has obscured the two well-separated groups. Note that each plot uses the same units in the horizontal and vertical axes.*

Outline

Clustering algorithms

- N points: $x_i, i \in \{1, \ldots, N\}$
- K clusters $k \in \{1, \ldots, K\}$
- Encoder $k = C(i)$, that assigns the ith point to the kth cluster

Energy functions:

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d(x_i, x_{i'}).$$

$$B(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')\neq k} d_{ii'}$$

$$T = \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} d_{ii'} = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \left(\sum_{C(i')=k} d_{ii'} + \sum_{C(i')\neq k} d_{ii'} \right)$$

$$T = W(C) + B(C),$$

W is the in-cluster point scatters, and W is the between-cluster point scatters. T is the total point scatters, which is a constant given the data, independent of cluster assignment.

Clustering algorithms

Clustering: maximize $W(C)$ or minimize $B(C)$

Problem: solving this optimization problem directly is combinatorial. The number of distinct assignments is

$$S(N,K) = \frac{1}{K!}\sum_{k=1}^{K}(-1)^{K-k}\binom{K}{k}k^N.$$

$S(10,4) = 34,105$ and $S(19,4) \approx 10^{10}$

K-means

Euclidean distance:

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = ||x_i - x_{i'}||^2$$

Within-point scatter:

$$\begin{aligned} W(C) &= \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} ||x_i - x_{i'}||^2 \\ &= \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - \bar{x}_k||^2, \end{aligned}$$

where $\bar{x}_k = (\bar{x}_{1k}, \ldots, \bar{x}_{pk})$ is the mean vector associated with the kth cluster and $N_k = \sum_{i=1}^{N} I(C(i) = k)$.

Iterative descent algorithm

Minimization:

$$C^* = \min_C \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - \bar{x}_k||^2$$

where for any subset S

$$\bar{x}_S = \operatorname{argmin}_m \sum_{i \in S} ||x_i - m||^2.$$

Enlarged minimization:

$$\min_{C, \{m_k\}_1^K} \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - m_k||^2.$$

K-means algorithm

Algorithm 14.1 *K-means Clustering.*

1. For a given cluster assignment C, the total cluster variance (14.33) is minimized with respect to $\{m_1, \ldots, m_K\}$ yielding the means of the currently assigned clusters (14.32).

2. Given a current set of means $\{m_1, \ldots, m_K\}$, (14.33) is minimized by assigning each observation to the closest (current) cluster mean. That is,

$$C(i) = \underset{1 \le k \le K}{\operatorname{argmin}} ||x_i - m_k||^2. \tag{14.34}$$

3. Steps 1 and 2 are iterated until the assignments do not change.

Algorithm iteration

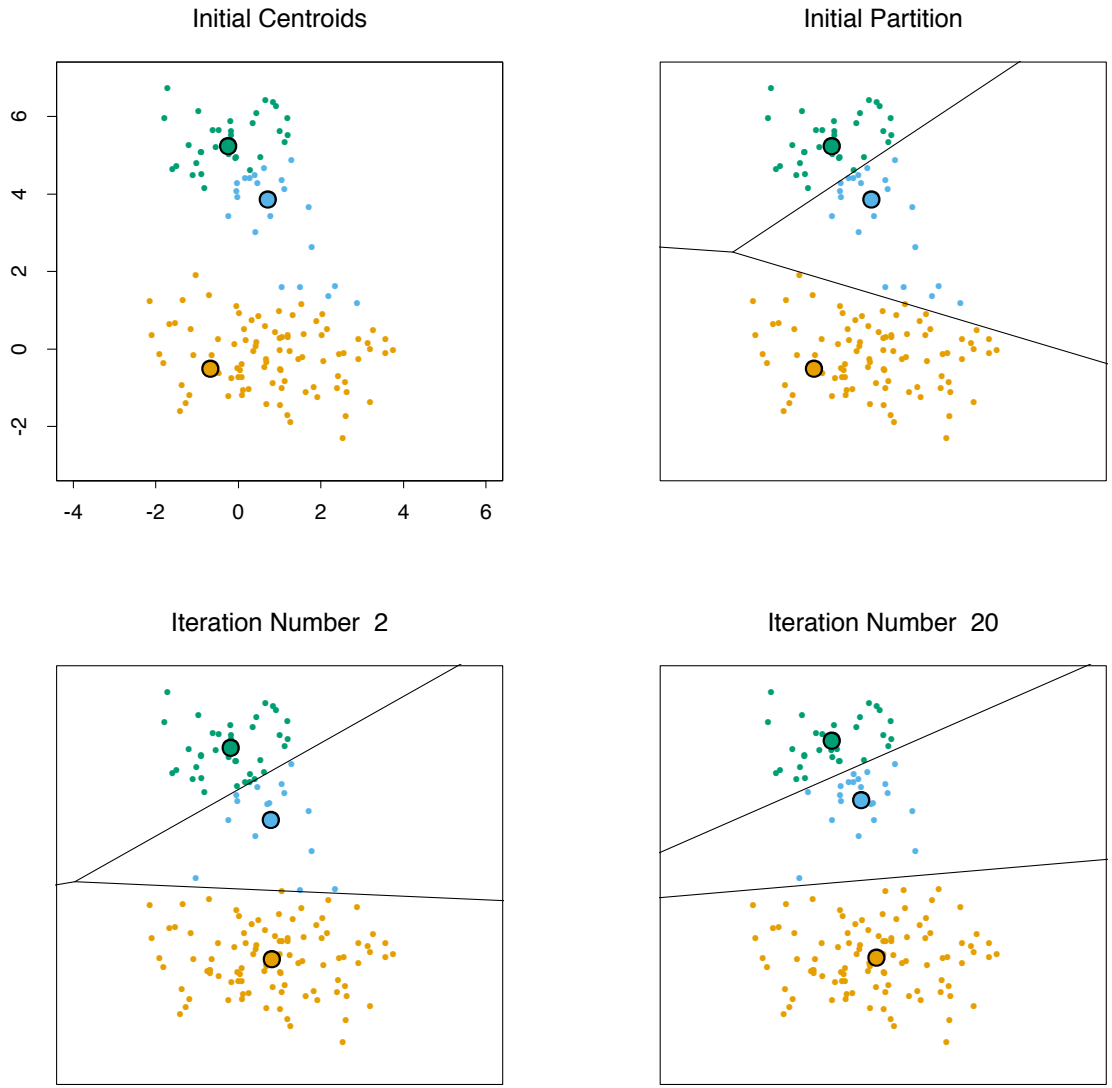

FIGURE 14.6. *Successive iterations of the K-means clustering algorithm for the simulated data of Figure 14.4.*

K-means on human gene data

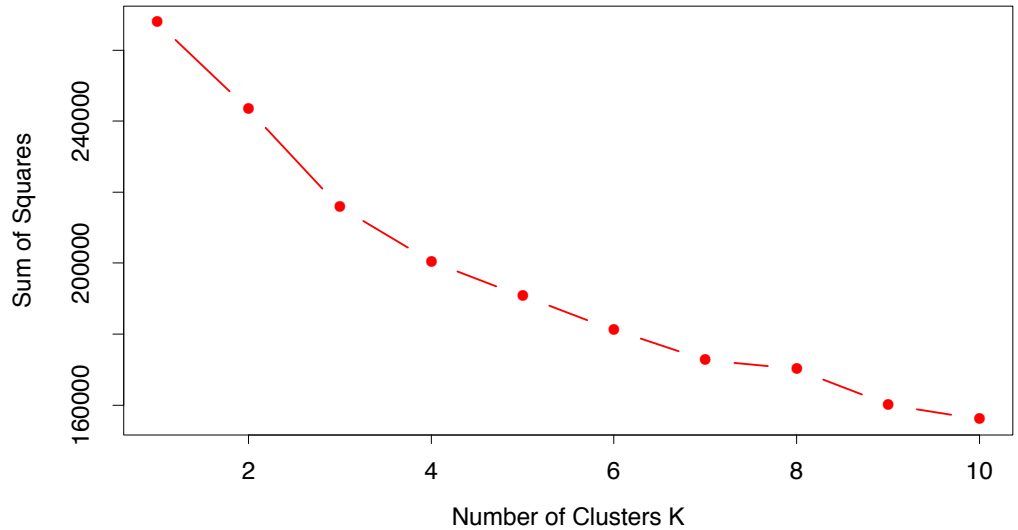

FIGURE 14.8. *Total within-cluster sum of squares for K-means clustering applied to the human tumor microarray data.*

TABLE 14.2. *Human tumor data: number of cancer cases of each type, in each of the three clusters from K-means clustering.*

Cluster	Breast	CNS	Colon	K562	Leukemia	MCF7
1	3	5	0	0	0	0
2	2	0	0	2	6	2
3	2	0	7	0	0	0
Cluster	Melanoma	NSCLC	Ovarian	Prostate	Renal	Unknown
1	1	7	6	2	9	1
2	7	2	0	0	0	0
3	0	0	0	0	0	0

K-medoids

About K-means

- The squared Euclidean distance is used: $d(x_i, x_{i'}) = \|x_i - x_{i'}\|^2$, which places the highest influence on the largest distances.
- Not robust against outliers with large distances.

K-medoids: an alternating algorithm that tries to solve

$$\min_{C,\ \{i_k\}_1^K} \sum_{k=1}^{K} \sum_{C(i)=k} d_{ii_k}.$$

K-medoids

Algorithm 14.2 *K-medoids Clustering.*

1. For a given cluster assignment C find the observation in the cluster minimizing total distance to other points in that cluster:

$$i_k^* = \underset{\{i:C(i)=k\}}{\operatorname{argmin}} \sum_{C(i')=k} D(x_i, x_{i'}). \tag{14.35}$$

   Then $m_k = x_{i_k^*}$, $k = 1, 2, \ldots, K$ are the current estimates of the cluster centers.

2. Given a current set of cluster centers $\{m_1, \ldots, m_K\}$, minimize the total error by assigning each observation to the closest (current) cluster center:

$$C(i) = \underset{1 \le k \le K}{\operatorname{argmin}} D(x_i, m_k). \tag{14.36}$$

3. Iterate steps 1 and 2 until the assignments do not change.

Example: country dissimilarities

TABLE 14.3. *Data from a political science survey: values are average pairwise dissimilarities of countries from a questionnaire given to political science students.*

	BEL	BRA	CHI	CUB	EGY	FRA	IND	ISR	USA	USS	YUG
BRA	5.58										
CHI	7.00	6.50									
CUB	7.08	7.00	3.83								
EGY	4.83	5.08	8.17	5.83							
FRA	2.17	5.75	6.67	6.92	4.92						
IND	6.42	5.00	5.58	6.00	4.67	6.42					
ISR	3.42	5.50	6.42	6.42	5.00	3.92	6.17				
USA	2.50	4.92	6.25	7.33	4.50	2.25	6.33	2.75			
USS	6.08	6.67	4.25	2.67	6.00	6.17	6.17	6.92	6.17		
YUG	5.25	6.83	4.50	3.75	5.75	5.42	6.08	5.83	6.67	3.67	
ZAI	4.75	3.00	6.08	6.67	5.00	5.58	4.83	6.17	5.67	6.50	6.92

Clustering by K-medoids

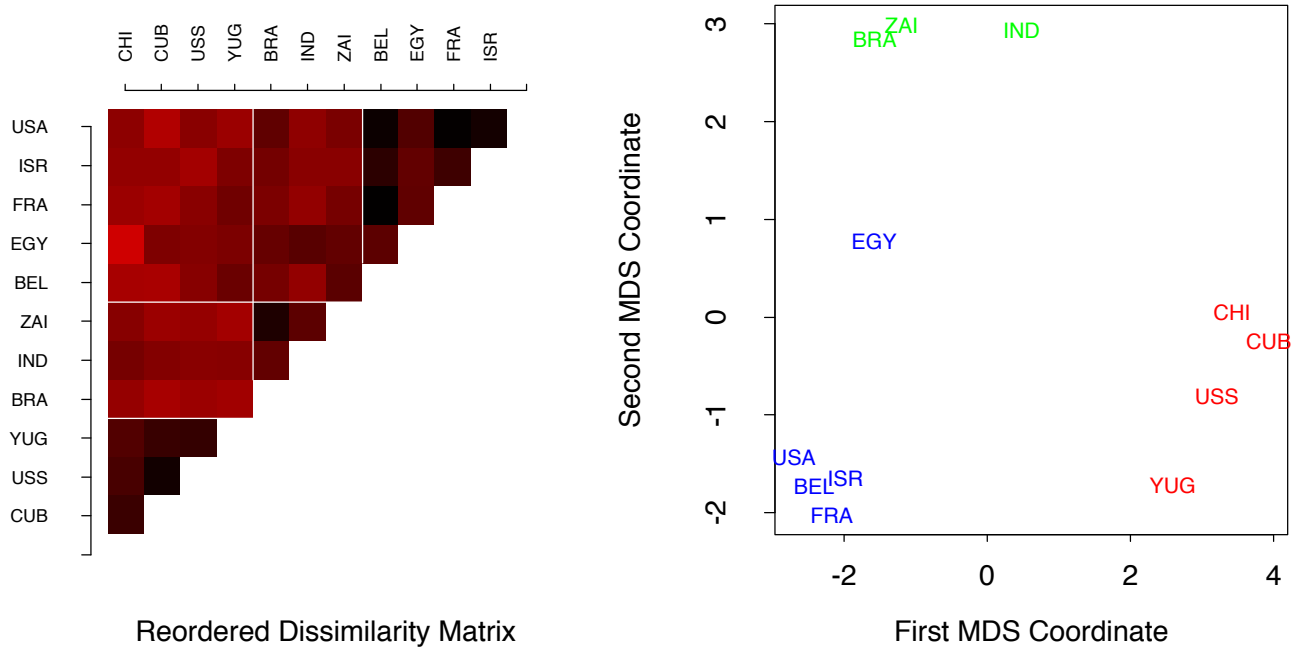

FIGURE 14.10. *Survey of country dissimilarities. (Left panel:) dissimilarities reordered and blocked according to 3-medoid clustering. Heat map is coded from most similar (dark red) to least similar (bright red). (Right panel:) two-dimensional multidimensional scaling plot, with 3-medoid clusters indicated by different colors.*

Reference

Chapter 14: Trevor Hastie, Robert Tibshirani, The Elements of Statistical Learning, Second Edition.