STA 5106
Computational Methods in Statistics I

Department of Statistics
Florida State University

Class 14
October 10, 2019

9.4 Clustering

Reference: “Information Theory, Inference, and Learning Algorithms” by David J.C. MacKay.

Clustering

- **Clustering:** put a set of objects into groups that are similar to each other.

- We will discuss ways to take a set of N objects and group them into K clusters.

- We perform clustering because we believe the underlying cluster labels are meaningful, will lead to a more efficient description of our data, and will help us choose better actions.

K-Means Clustering

- The K-means algorithm is an algorithm for putting N data points in an I-dimensional space into K clusters. Each cluster is parameterized by a vector $m^{(k)}$ called its mean.

- The data points will be denoted by $x^{(n)}$ where the superscript n runs from 1 to the number of data points N.

- Each vector $x = (x_1, \ldots, x_i, \ldots, x_I)$ is a vector with I components. We will assume that the space that x lives in is a real space and that we have a metric that defines distances between points, for example,

$$d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

Initialization

- To start the K-means algorithm, the K means $\{m^{(k)}\}$ are initialized in some way, for example to random values.

- K-means is then an iterative two-step algorithm.

- In the *assignment step*, each data point $x^{(n)}$ is assigned to the nearest mean.

- In the *update step*, the means are adjusted to match the sample means of the data points that they are responsible for.

Algorithm

- **Initialization.** Set K means $\{m^{(k)}\}$ to random values.

- **Assignment step.** Each data point $x^{(n)}$ is assigned to the nearest mean. We denote our guess for the cluster $k^{(n)}$ that the point $x^{(n)}$ belongs to by

$$\hat{k}^{(n)} = \arg\min_k \{d(m^{(k)}, x^{(n)})\}$$

An alternative, equivalent representation of this assignment of points to clusters is given by "responsibilities", which are indicator variables as follows:

$$r_k^{(n)} = \begin{cases} 1 & \text{if } \hat{k}^{(n)} = k \\ 0 & \text{if } \hat{k}^{(n)} \neq k \end{cases}$$

Algorithm

- **Update step.** The model parameters, the means, are adjusted to match the sample means of the data points that they are responsible for

$$m^{(k)} = \frac{\sum_n r_k^{(n)} x^{(n)}}{R^{(k)}}$$

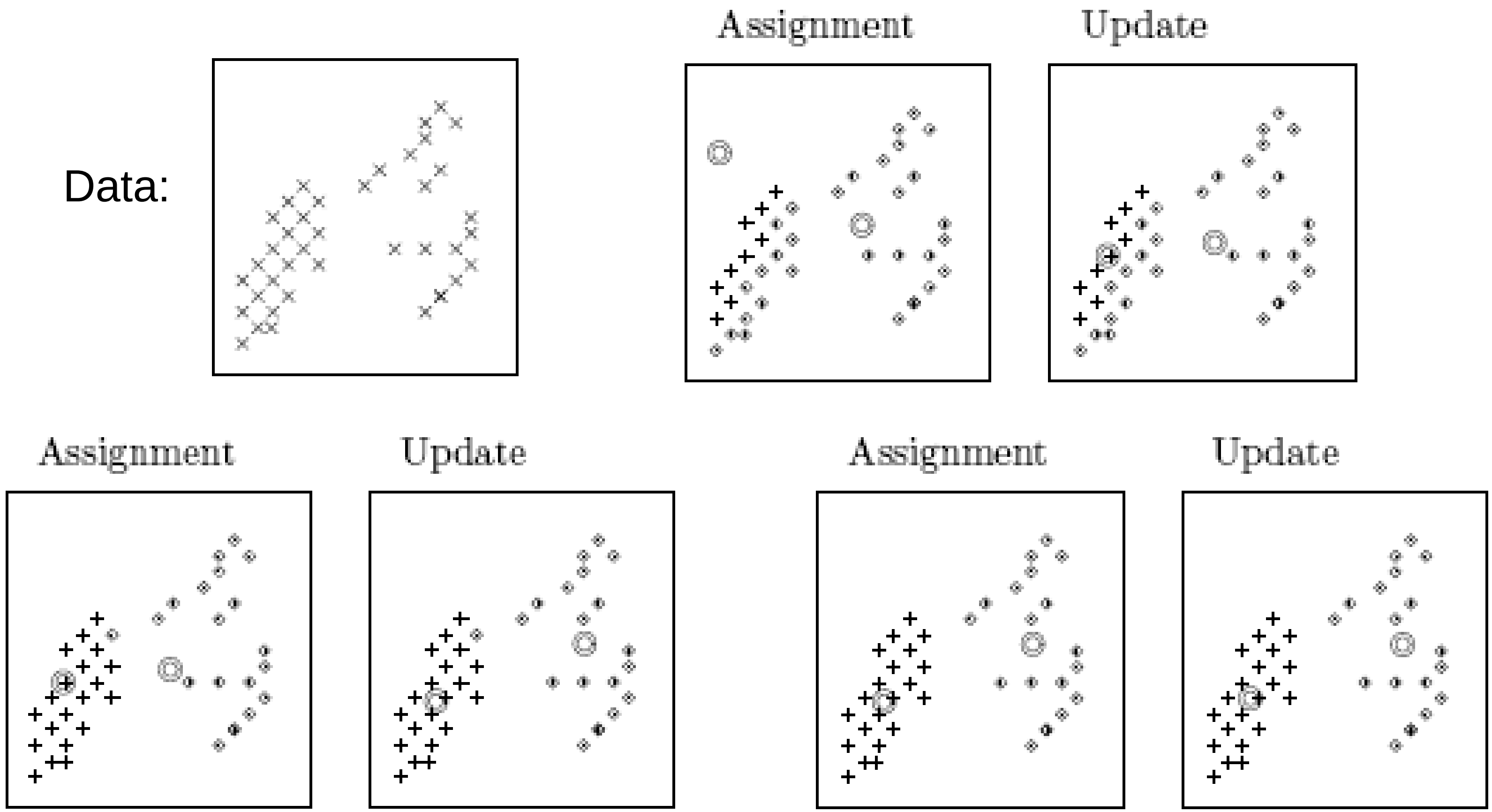
where $R^{(k)}$ is the total responsibility of mean k,

$$R^{(k)} = \sum_n r_k^{(n)}$$

- **Repeat the assignment step and update step** until the assignments do not change.

Example One

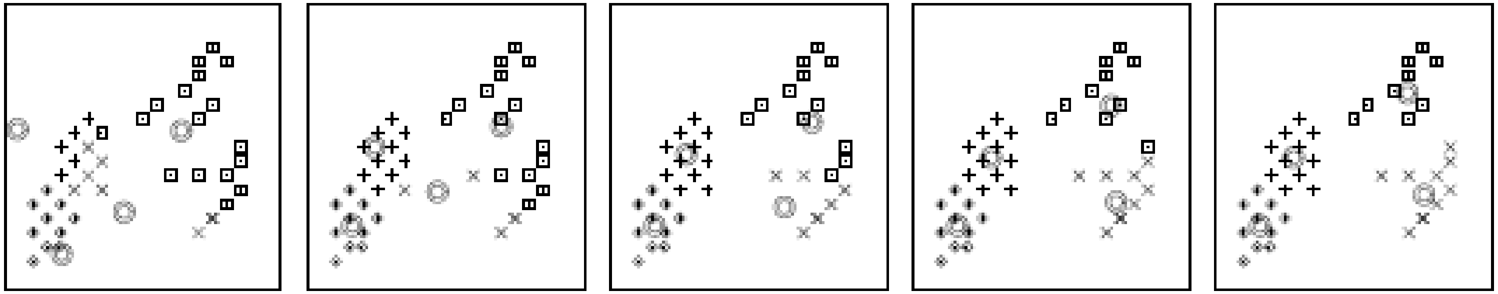
- K-means algorithm applied to a data set of 40 points. $K = 2$ means evolve to stable locations after three iterations.



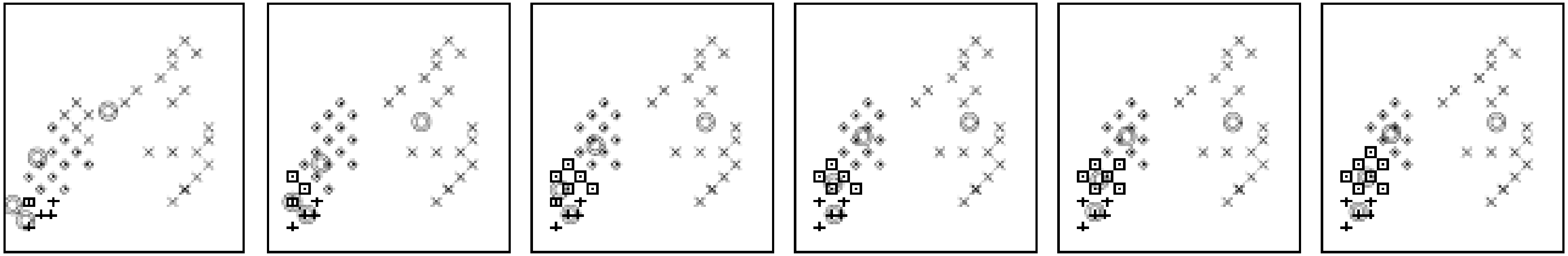
Example Two

- Same data set. Two separate runs, both with $K = 4$ means, reach different solutions.

Run 1



Run 2



Minimization

- **Distortion Measure (Sum of Squares):**

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_k^{(n)} d(x^{(n)}, m^{(k)})^2$$
$$= \sum_{n=1}^{N}\sum_{k=1}^{K} r_k^{(n)} \| x^{(n)} - m^{(k)} \|^2$$

- J is a quadratic function of $m^{(k)}$, and it can be minimized by setting its derivatives with respect to $m^{(k)}$ as zero, that is,

$$\sum_{n=1}^{N} 2 r_k^{(n)} (x^{(n)} - m^{(k)}) = 0,$$

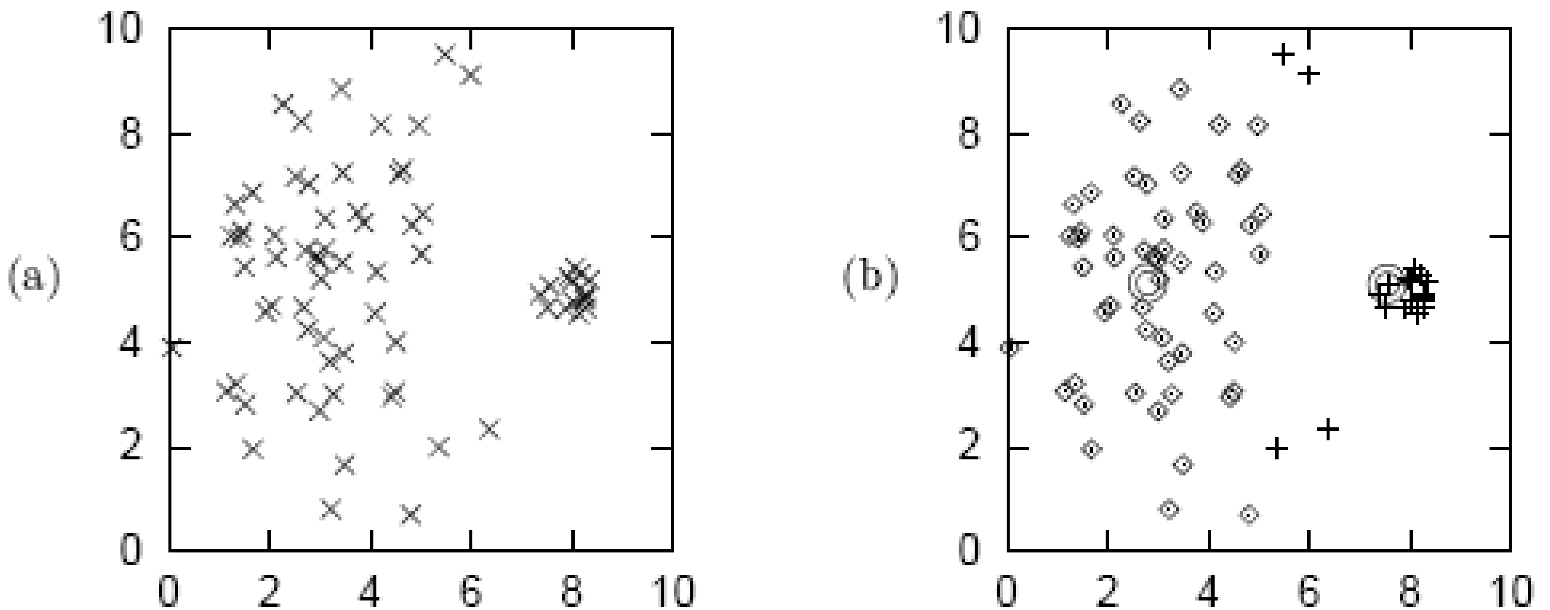
or

$$m^{(k)} = \sum_{n=1}^{N} r_k^{(n)} x^{(n)} \Big/ \sum_{n=1}^{N} r_k^{(n)}$$

K-Means May Fail

- **Case 1:** The K-means algorithm takes account only of the distance between the means and the data points; it has no representation of the weight or breadth of each cluster.

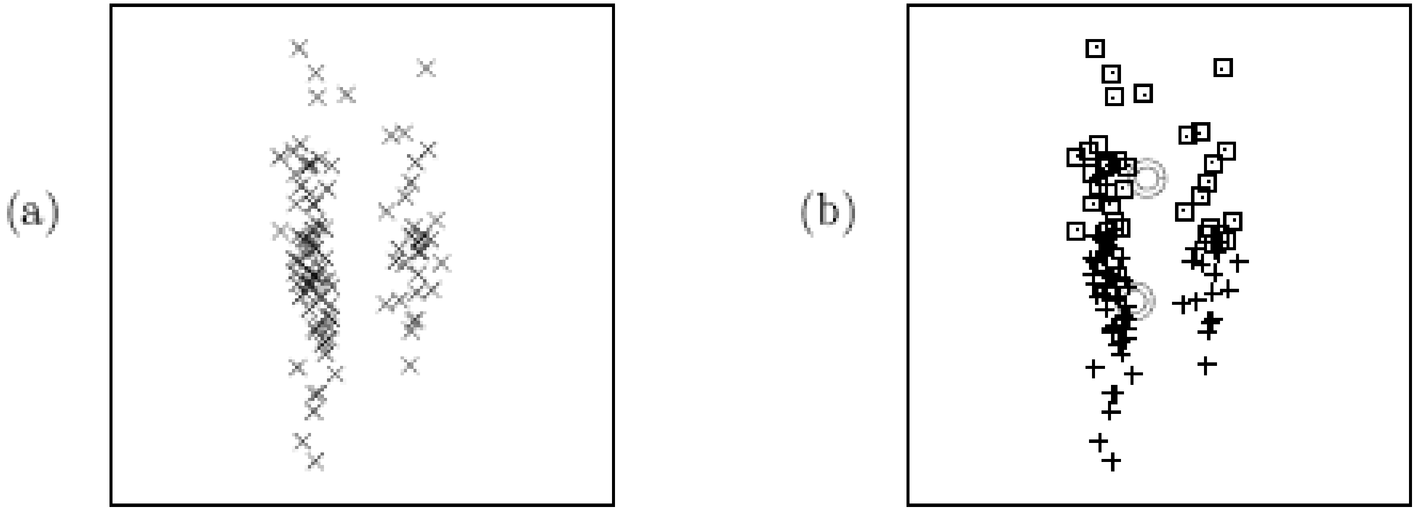
- Example:



K-Means May Fail

- **Case 2:** The K-means algorithm has no way of representing the size or shape of a cluster.

- Example:



Announcement

- Tuesday, 10/15

  Review for the Midterm
- Thursday, 10/17

  No Class
- Tuesday, 10/22

  Midterm Presentations (Group 1)
- Thursday 10/24

  Midterm Presentations (Group 2)

  Midterm Report Due, HW 7 out

Midterm Presentation Schedule

Tuesday (10/22)	Thursday (10/24)
Tingan Chen	Tianyuan Cheng
Harshita Dogra	Ke Han
Shuai Hao	Hanwen Hu
Taka Iguchi	Seyedkamyar Kazemi
Rufeng Liu	Pengfei Lyu
Xiaoxiao Ma	Yijia Ma
Sayantika Nag	Jario Pena Hidalgo
Sudipto Saha	Changhee Suh
Michael Wilson	Ka Chun Wong
Tao Xu	Zhou Xinyu

Each student will have up to 7 minutes for presentation.