STA141C: Big Data & High Performance Statistical Computing

Lecture 11: Clustering

Cho-Jui Hsieh
UC Davis

May 24, 2018

Outline

- Kmeans Clustering
- Graph Clustering

Supervised versus Unsupervised Learning

Supervised Learning:

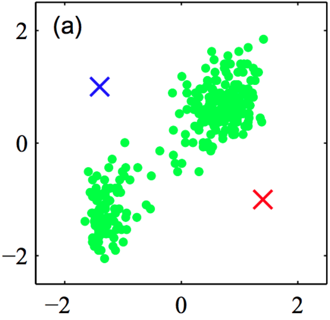
- Learning from labeled observations
- Classification, regression, ...

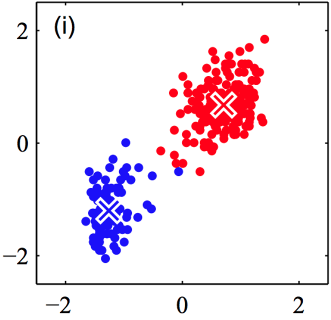
Unsupervised Learning:

- Learning from unlabeled observations
- Discover hidden patterns
- Clustering (today)

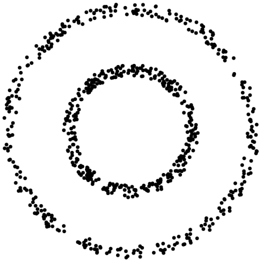
Clustering

- Given $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ and K (number of clusters)
- Output $A(\boldsymbol{x}_i) \in \{1, 2, \ldots, K\}$ (cluster membership)



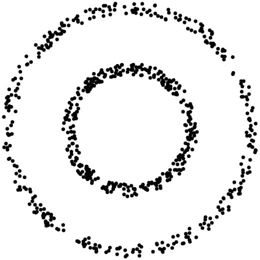


Two circles

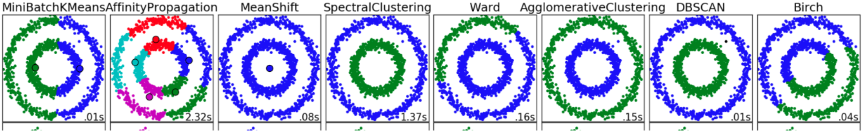


Can we split the data into two clusters?

Two circles

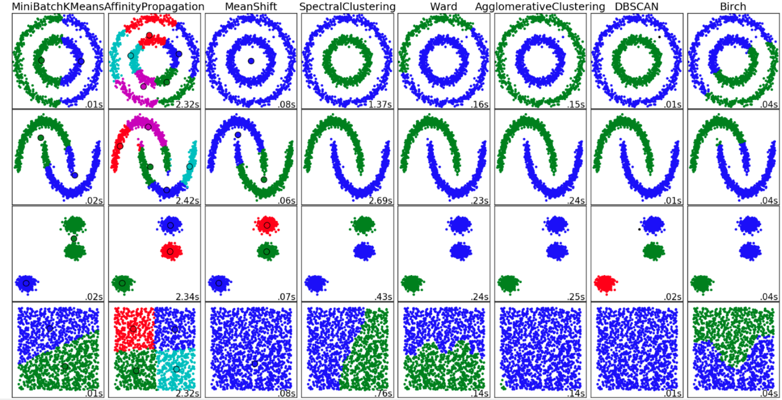


Can we split the data into two clusters?



Clustering is Subjective

- Non-trivial to say one clustering is better than the other
- Each algorithm has two parts:
  - Define the objective function
  - Design an algorithm to minimize this objective function



K-means Objective Function

- Partition dataset into $C_1, C_2, \ldots, C_K$ to minimize the following objective:

$$J = \sum_{k=1}^{K} \sum_{\boldsymbol{x} \in C_k} \|\boldsymbol{x} - \boldsymbol{m}_k\|_2^2,$$

where $\boldsymbol{m}_k$ is the mean of C_k.

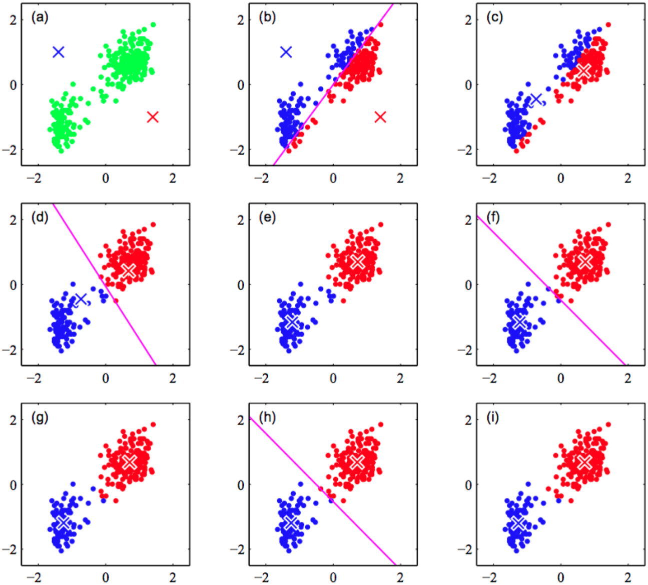
K-means Objective Function

- Partition dataset into $C_1, C_2, \ldots, C_K$ to minimize the following objective:

$$J = \sum_{k=1}^{K} \sum_{\boldsymbol{x} \in C_k} \|\boldsymbol{x} - \boldsymbol{m}_k\|_2^2,$$

  where $\boldsymbol{m}_k$ is the mean of C_k.
- Multiple ways to minimize this objective
  - Hierarchical Agglomerative Clustering
  - Kmeans Algorithm (Today)
  - ...

K-means Algorithm



K-means Algorithm

- Re-write objective:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\boldsymbol{x}_n - \boldsymbol{m}_k\|_2^2,$$

where $r_{nk} \in \{0, 1\}$ is an indicator variable

$$r_{nk} = 1 \ \text{ if and only if } \ \boldsymbol{x}_n \in C_k$$

- Alternative optimization between $\{r_{nk}\}$ and $\{\boldsymbol{m}_k\}$
  - Fix $\{\boldsymbol{m}_k\}$ and update $\{r_{nk}\}$
  - Fix $\{r_{nk}\}$ and update $\{\boldsymbol{m}_k\}$

K-means Algorithm

- Step 0: Initialize $\{\boldsymbol{m}_k\}$ to some values

K-means Algorithm

- Step 0: Initialize $\{\boldsymbol{m}_k\}$ to some values
- Step 1: Fix $\{\boldsymbol{m}_k\}$ and minimize over $\{r_{nk}\}$:

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\boldsymbol{x}_n - \boldsymbol{m}_j\|_2^2 \\ 0 & \text{otherwise} \end{cases}$$

K-means Algorithm

- Step 0: Initialize $\{\boldsymbol{m}_k\}$ to some values
- Step 1: Fix $\{\boldsymbol{m}_k\}$ and minimize over $\{r_{nk}\}$:

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\boldsymbol{x}_n - \boldsymbol{m}_j\|_2^2 \\ 0 & \text{otherwise} \end{cases}$$

- Step 2: Fix $\{r_{nk}\}$ and minimize over $\{\boldsymbol{m}_k\}$:

$$\boldsymbol{m}_k = \frac{\sum_n r_{nk}\boldsymbol{x}_n}{\sum_n r_{nk}}$$

K-means Algorithm

- Step 0: Initialize $\{\boldsymbol{m}_k\}$ to some values
- Step 1: Fix $\{\boldsymbol{m}_k\}$ and minimize over $\{r_{nk}\}$:

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\boldsymbol{x}_n - \boldsymbol{m}_j\|_2^2 \\ 0 & \text{otherwise} \end{cases}$$

- Step 2: Fix $\{r_{nk}\}$ and minimize over $\{\boldsymbol{m}_k\}$:

$$\boldsymbol{m}_k = \frac{\sum_n r_{nk}\boldsymbol{x}_n}{\sum_n r_{nk}}$$

- Step 3: Return to step 1 unless stopping criterion is met

K-means Algorithm

Equivalent to the following procedure:

- Step 0: Initialize centers $\{\boldsymbol{m}_k\}$ to some values
- Step 1: Assign each $\boldsymbol{x}_n$ to the nearest center:

$$A(\boldsymbol{x}_n) = \arg\min_j \|\boldsymbol{x}_n - \boldsymbol{m}_j\|_2^2$$

  Update clusters:

$$C_k = \{\boldsymbol{x}_n : A(\boldsymbol{x}_n) = k\} \;\; \forall k = 1, \ldots, K$$

- Step 2: Calculate mean of each cluster C_k:

$$\boldsymbol{m}_k = \frac{1}{|C_k|} \sum_{\boldsymbol{x}_n \in C_k} \boldsymbol{x}_n$$

- Step 3: Return to step 1 unless stopping criterion is met

More on K-means Algorithm

- Always decrease the objective function for each update
- Objective function will keep unchanged when step 1 doesn't change cluster assignment $\Rightarrow$ Converged

More on K-means Algorithm

- Always decrease the objective function for each update
- Objective function will keep unchanged when step 1 doesn't change cluster assignment $\Rightarrow$ Converged
- May not converge to global minimum

  Sensitive to initial values

More on K-means Algorithm

- Always decrease the objective function for each update
- Objective function will keep unchanged when step 1 doesn't change cluster assignment $\Rightarrow$ Converged
- May not converge to global minimum

  Sensitive to initial values
- Kmeans++: A better way to initialize the clusters

Coming up

- Clustering

Questions?