Machine Learning for Data Science (CS4786)
Lecture 9

Clustering

March 1st, 2016

Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2016sp/

Clustering Criterion

1. Minimize within-cluster scatter

$$M_1 = \sum_{j=1}^{K} \sum_{\mathbf{x}_s, \mathbf{x}_t \in C_j} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2$$

2. Maximize between-cluster scatter

$$M_2 = \sum_{\mathbf{x}_s, \mathbf{x}_t : c(\mathbf{x}_s) \neq c(\mathbf{x}_t)} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2$$

3. Minimize weighted within-cluster variance: $\mathbf{r}_j = \frac{1}{n_j} \sum_{\mathbf{x} \in C_j} \mathbf{x}$

$$M_3 = \sum_{j=1}^{K} n_j \sum_{\mathbf{x}_t \in C_j} \|\mathbf{x}_t - \mathbf{r}_j\|_2^2$$

4. Maximize smallest between-cluster distance

$$M_4 = \min_{\mathbf{x}_s, \mathbf{x}_t : c(\mathbf{x}_s) \neq c(\mathbf{x}_t)} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2$$

5. Minimize largest within-cluster distance

$$M_4 = \max_{j \in [K]} \max_{\mathbf{x}_s, \mathbf{x}_t \in C_j} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2$$

6. Minimize within-cluster average scatter

$$M_6 = \sum_{j=1}^{K} \frac{1}{n_j} \sum_{\mathbf{x}_s, \mathbf{x}_t \in C_j} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2$$

Clustering Criterion

6 Minimize within-cluster average scatter

$$M_6 = \sum_{j=1}^{K} \frac{1}{n_j} \sum_{\mathbf{x}_s, \mathbf{x}_t \in C_j} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2$$

7 Minimize within-cluster variance: $\mathbf{r}_j = \frac{1}{n_j} \sum_{\mathbf{x} \in C_j} \mathbf{x}$

$$M_7 = \sum_{j=1}^{K} \sum_{\mathbf{x}_t \in C_j} \|\mathbf{x}_t - \mathbf{r}_j\|_2^2$$

Equivalence of Clustering Criteria

- $M_1 \equiv M_2$:

$$\sum_{s,t \in [n]} \|\mathbf{x}_t - \mathbf{x}_s\|_2^2 = \sum_{j=1}^{K} \sum_{\mathbf{x}_s, \mathbf{x}_t \in C_j} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2 + \sum_{\mathbf{x}_s, \mathbf{x}_t : c(\mathbf{x}_s) \neq c(\mathbf{x}_t)} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2$$

Hence, $M_1 = \text{Constant} - M_2$ (maximizing M_1 is same as minimizing M_2)

Equivalence of Clustering Criteria

- $M_1 \equiv M_3$:
  - Fact: $\forall j \in \{1, \ldots, K\}$, and for any $\mathbf{x} \in \mathbb{R}^d$,

$$\sum_{\mathbf{x}_t \in C_j} \|\mathbf{x}_t - \mathbf{x}\|_2^2 = \sum_{\mathbf{x}_t \in C_j} \|\mathbf{x}_t - \mathbf{r}_j\|_2^2 + n_j \|\mathbf{x} - \mathbf{r}_j\|_2^2$$

Proof:

$$\begin{aligned}
\sum_{\mathbf{x}_t \in C_j} \|\mathbf{x}_t - \mathbf{x}\|_2^2 &= \sum_{\mathbf{x}_t \in C_j} \|\mathbf{x}_t - \mathbf{r}_j + \mathbf{r}_j - \mathbf{x}\|_2^2 \\
&= \sum_{\mathbf{x}_t \in C_j} \|\mathbf{x}_t - \mathbf{r}_j\|_2^2 + \sum_{\mathbf{x}_t \in C_j} \|\mathbf{r}_j - \mathbf{x}\|_2^2 + 2 \sum_{\mathbf{x}_t \in C_j} (\mathbf{x}_t - \mathbf{r}_j)^\top (\mathbf{r}_j - \mathbf{x}) \\
&= \sum_{\mathbf{x}_t \in C_j} \|\mathbf{x}_t - \mathbf{r}_j\|_2^2 + n_j \|\mathbf{r}_j - \mathbf{x}\|_2^2 + 2 n_j \left(\frac{1}{n_j} \sum_{\mathbf{x}_t \in C_j} \mathbf{x}_t - \mathbf{r}_j \right)^\top (\mathbf{r}_j - \mathbf{x})
\end{aligned}$$

- $\mathbf{r}_j$ is the best cluster representative given cluster assignment.

Equivalence of Clustering Criteria

- Hence,

$$
\begin{aligned}
M_1 &= \sum_{j=1}^{K} \left(\sum_{\mathbf{x}_s, \mathbf{x}_t \in C_j} \|\mathbf{x}_t - \mathbf{x}_s\|_2^2 \right) \\
&= \sum_{j=1}^{K} \left(\sum_{\mathbf{x}_s \in C_j} \left(\sum_{\mathbf{x}_t \in C_j} \|\mathbf{x}_t - \mathbf{r}_j\|_2^2 + n_j \|\mathbf{x}_s - \mathbf{r}_j\|_2^2 \right) \right) \\
&= 2 \sum_{j=1}^{K} \left(n_j \sum_{\mathbf{x}_t \in C_j} \|\mathbf{x}_t - \mathbf{r}_j\|_2^2 \right) \\
&= 2\, M_3
\end{aligned}
$$

CLUSTERING

- Multiple clustering criteria all equally valid
- Different criteria lead to different algorithms/solutions
- Which notion of distances or costs we use matter

K-means Clustering

- For all $j \in [K]$, initialize cluster centroids $\hat{\mathbf{r}}_j^1$ randomly and set $m = 1$
- Repeat until convergence (or until patience runs out)
  1. For each $t \in \{1, \ldots, n\}$, set cluster identity of the point

$$\hat{c}^m(\mathbf{x}_t) = \underset{j \in [K]}{\operatorname{argmin}} \|\mathbf{x}_t - \hat{\mathbf{r}}_j^m\|$$

  2. For each $j \in [K]$, set new representative as

$$\hat{\mathbf{r}}_j^{m+1} = \frac{1}{|\hat{C}_j^m|} \sum_{\mathbf{x}_t \in \hat{C}_j^m} \mathbf{x}_t$$

  3. $m \leftarrow m + 1$

K-means Convergence

- K-means algorithm converges to local minima of objective

$$O\left(c;\mathbf{r}_1,\ldots,\mathbf{r}_K\right)=\sum_{j=1}^{K}\sum_{c(\mathbf{x}_t)=j}\|\mathbf{x}_t-\mathbf{r}_j\|_2^2$$

- Proof:
  Clustering assignment improves objective:

$$O\left(\hat{c}^{m-1};\mathbf{r}_1^m,\ldots,\mathbf{r}_K^m\right)\geq O\left(\hat{c}^{m};\mathbf{r}_1^m,\ldots,\mathbf{r}_K^m\right)$$

  (By definition of $\hat{c}^m(\mathbf{x}_t)$)
  Computing centroids improves objective:

$$O\left(\hat{c}^{m};\mathbf{r}_1^m,\ldots,\mathbf{r}_K^m\right)\geq O\left(\hat{c}^{m};\mathbf{r}_1^{m+1},\ldots,\mathbf{r}_K^{m+1}\right)$$

  (By the fact about centroid)

Single Link Clustering

- Initialize n clusters with each point $\mathbf{x}_t$ to its own cluster
- Until there are only K clusters, do
  1. Find closest two clusters and merge them into one cluster
  2. Update between cluster distances (called proximity matrix)

Single Link Clustering Demo

Single Link Objective

Objective for single-link:

$$M_4 = \min_{\mathbf{x}_s, \mathbf{x}_t : c(\mathbf{x}_s) \neq c(\mathbf{x}_t)} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2$$

Single link clustering is optimal for above objective!