K-means Clustering

Tuesday, May 2, 2017 9:53 AM

$\{x_i\}_{i=1}^n \quad x_i \in \mathbb{R}^p$

Task: Learn $z_i \in \{1, \dots, K\}$ (cluster assignments)

eg $C_1 = \{x_1, x_3, x_7\}$ $\quad z_1 = 1, \; z_2 = 2, \; z_3 = 1$

$C_2 = \{x_2, x_5, x_6\}$ $\quad z_4 = 3$

$C_3 = \{x_4, x_8, x_9\}$

also learn $m_k \in \mathbb{R}^p$, cluster centers, $k = 1, \dots, K$

Objective of K-means

$$\min_{z, m} J(z, m) = \sum_{i=1}^{n} \| x_i - m_{z_i} \|_2^2$$

↳ distortion

Write this: $J(z, m) = \sum_{k=1}^{K} \underbrace{\sum_{i : z_i = k} \| x_i - m_k \|_2^2}$

fix z_i's min m_k = average w/in clusters

Lloyd's Algorithm

(1) Initialize m_k arbitrarily

(2) Alternate

(a) Update $z_i \leftarrow \underset{k}{\operatorname{argmin}} \| x_i - m_k \|_2^2$

(b) $m_k \leftarrow \sum_{i=1}^{n} 1\{z_i = k\} x_i \Big/ \sum_{i=1}^{n} 1\{z_i = k\}$

Hierarchical Clustering

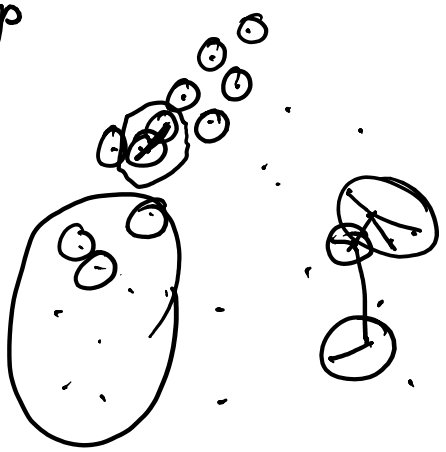
Tuesday, May 2, 2017 9:54 AM

Agglomerative Clustering: Bottom-up

(1) Start w/ all data points in own cluster

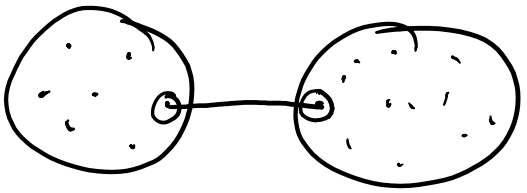
(2) Find clusters most similar: $C_1 \& C_2$ (*)

(3) Merge $C_1 \& C_2$ goto 2



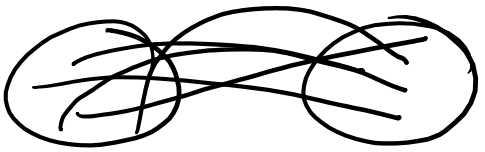
Cluster similarity

Single linkage:



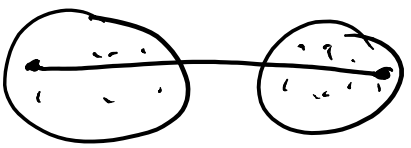
$$d_{sl}(C_1, C_2) = \min_{x \in C_1, y \in C_2} d(x, y)$$

Average linkage:



$$d_{al}(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{x \in C_1, y \in C_2} d(x, y)$$

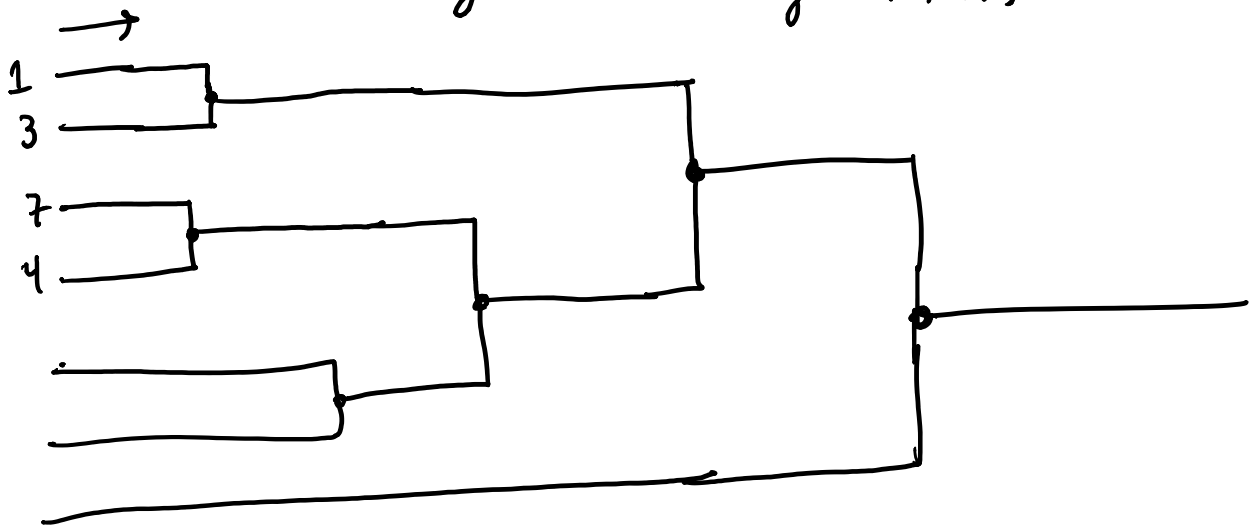
Complete linkage: $d_{cl}(C_1, C_2) = \max_{x \in C_1, y \in C_2} d(x, y)$



▷ general observations: sl tends to yield "unbalanced

clusters" and cl yields "balanced clusters"

Dendrogram Visualization tool where we start w/ n liners and merge when merge clusters



Divisive Clustering Top-down

Option 1: recursively apply k-means w/ k=2

Con 1: initialization of k-means really matters

Con 2: algorithm can violate monotonicity of distortion

Option 2: greedy approach

(1) Start w/ 1 cluster

(2) Repeat

- Choose a cluster G
- Remove point most dissimilar from G

Starts cluster H

repeat {
- remove $x^* = \underset{x \in G}{\operatorname{argmax}} \frac{1}{|G-1|} \sum_{g \in G \setminus \{x\}} d(x,g) - \frac{1}{|H|} \sum_{h \in H} d(x,h)$
- add $x^* \to H$
}

Document Vectorization

Tuesday, May 2, 2017 9:59 AM

Bag-of-words models

- Construct a dictionary of words: enumerate all words in our corpus (all documents)
- vectorize a document, D,

X_{Di} = measure of frequency of word i in D.

eg $X_{Di} = 1\{i \in D\}$ (binary FF)

$$\|X_D - X_{D'}\|_2^2 = \sum_i \left(1\{i \in D\} - 1\{i \in D'\}\right)^2$$

$$= \sum_i 1\{i \in D, i \notin D' \text{ or } i \in D', i \notin D\}$$

other frequencies: count of word, $\frac{\text{count}}{\text{total \# of words in } D}$

Inverse document freq

$$idf(i) = \log\left(\frac{\#\text{ of doc's in corpus}}{\{D : i \in D\}}\right)$$

Term-frequency Inverse-document-frequency

$$tfidf(i, D) = tf(i, D) \cdot idf(i)$$