Lecture 13 Notes

Monday, May 9, 2016 11:53 AM

Finding K clusters

$\{x_i\}_{i=1}^n$ $x_i \in \mathbb{R}^p$ Learn $\underline{z_i \in \{1, \dots, K\}}$ (cluster assignments)

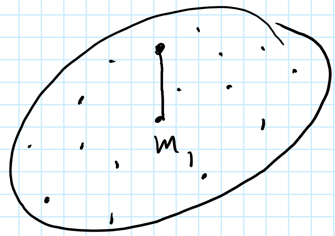
eg $C_1 = \{x_1, x_3, x_7\}$ $z_1 = 1$, $z_2 = 2$, $z_3 = 1$, $z_4 = 3, \dots$

$C_2 = \{x_2, x_5, x_6\}$ $K = 3$

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

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

Objective of K-means



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

▷ called the distortion

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

Lloyd's algorithm

(1) Initialize m_k arbitrarily

(2) Alternate

(a) Update $z_i \leftarrow \arg\min_k \|x_i - m_k\|_2^2 \quad \forall i$

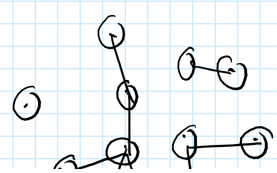
(b) Update $m_k \leftarrow \arg\min_{m \in \mathbb{R}^p} \sum_{i: z_i = k} \|x_i - m\|_2^2 \quad \forall k$

$$▷ \quad m_k = \frac{\sum_{i=1}^n \mathbb{1}\{z_i = k\} x_i}{\sum_{i=1}^n \mathbb{1}\{z_i = k\}}$$

(b') $m_k \leftarrow \arg\min_{x_j} \sum_{i: z_i = k} \|x_i - x_j\|_2^2$ K-medoids

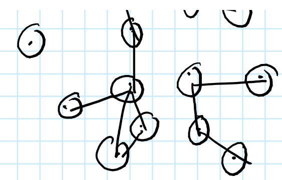
Hierarchical Clustering

Agglomerative clustering: bottom-up



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

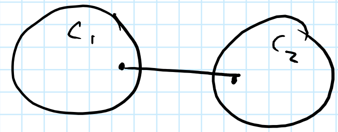
Agglomerative Clustering : bottom-up



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

(2) Find clusters C_1 & C_2 most similar (*)

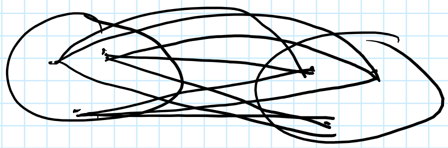
(3) Merge C_1, C_2 goto (2)

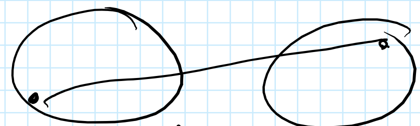


Cluster similarity

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

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

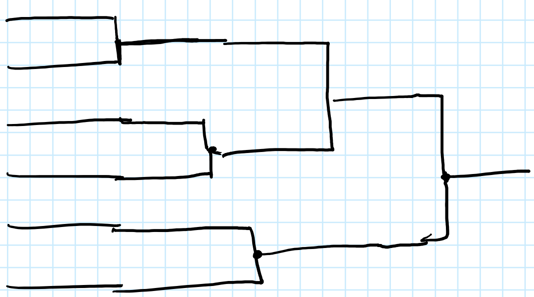




Complete linkage: $d_{cl}(C_1, C_2) = \max_{x \in C_1, y \in C_2} d(x, y)$

▷ general observation: sl tends to yield "unbalanced" clusterings (some v. large clusters & some v. small) and cl tends to yield "balanced" clusterings.

Dendrogram Visualization tool where we start w/ n lines then merge lines corresponding to cluster mergers



Divisive Clustering Top-down

Option 1: recursively apply k-means w/ k=2 to selected clusters.

Con 1: initialization really matters

Con 2: algorithm can violate monotonicity of objective

Option 2: greedy approach

1. start w/ one big cluster
2. repeat until all clusters are singletons
   - Choose a cluster G
   - Remove point most dissimilar from average of other pts
     Starts a new cluster H
     - repeat until objective non-positive
       remove $x^* = \underset{x \in G}{\arg\max} \; \frac{1}{|G-1|} \sum_{g \in G \setminus \{x\}} d(x,g) - \frac{1}{|H|} \sum_{h \in H} d(x,h)$
       add x^* to H