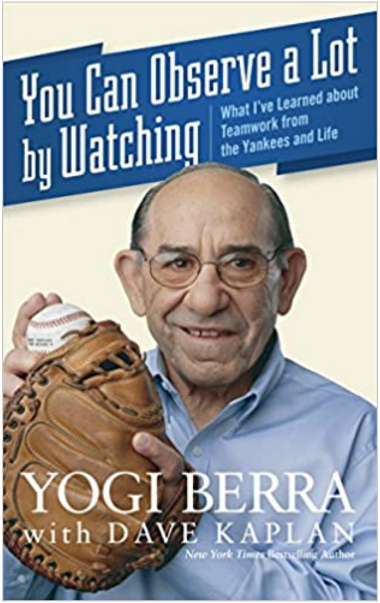
kMeans

23 March 2023

Sumit Kumar Yadav

Department of Management Studies
Indian Institute of Technology, Roorkee


You Can Observe a Lot by Watching
What I've Learned about Teamwork from the Yankees and Life
YOGI BERRA
with DAVE KAPLAN

Clustering

Definition Attempt 1 - "subset of points that are closer to each other than to all other data points

Definition Attempt 1 - "subset of points that are closer to each other than to all other data points

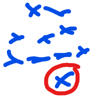
Definition Attempt 2 - Represent a cluster by its center/mean. Points in a cluster are closer to center/mean of their own cluster than to the mean of other clusters. (Circular definition beacuse??)

Clustering

- View points as union of k disjoint clusters - $C_1, C_2, ..., C_k$
- Each point lies in exactly one

k-means Clustering problem

- Let the points be $x_1, x_2, ..., x_n$
- Mean of the j^{th} cluster =



$$c_j = \frac{1}{m_j} \sum_{i \in C_j} x_i$$

m_j is the number of points in the j^{th} cluster

- Define cost of a cluster as - sum of squared distance from the points to the mean -

$$\sum_{i \in C_j} ||x_i - c_j||^2$$

- k-means problem : Partition points into k clusters so as to minimize sum of cluster costs - $\sum_{j=1}^{k} \sum_{i \in C_j} ||x_i - c_j||^2$

k-Means algorithm

- Maintain clusters $C_1, C_2, ..., C_k$
- Compute the cluster centers for these clusters
- Iteration - For each point, assign it to the c_j that it is closest to. Update $C_1, C_2, ..., C_k$ and proceed to the next iteration

How to evaluate clustering?

Finding the value of K

- Elbow Method

Finding the value of K

- Elbow Method
- DB Index
  Define cluster dispersion for the j^{th} cluster as -
  $$d_j = \sqrt{\frac{1}{m_j}\sum_{i \in C_j} ||x_i - c_j||^2}$$
- Define cluster similarity between 2 clusters j and l as -
  $$S_{jl} = \frac{d_j + d_l}{||c_j - c_l||}$$
- $$V_{DB} = \frac{1}{K}\sum_{i=1}^{K} \max_{l \neq i} S_{il}$$

Gaussian Mixture Models

Thank you for your attention