K Means Clustering

Mr. Pritam Prakash Shete
Computer Division, BARC
Centre for Excellence in Basic Sciences

Topics

- Introduction
- Unsupervised learning
- Clustering applications
- K means clustering
- Optimization objective
- Random initialization
- Number of clusters
- Advantages and disadvantages

Supervised Learning

- Training set – { $(x^{(1)},y^{(1)})$, $(x^{(2)},y^{(2)})$, …, $(x^{(m)},y^{(m)})$}
- Labeled dataset

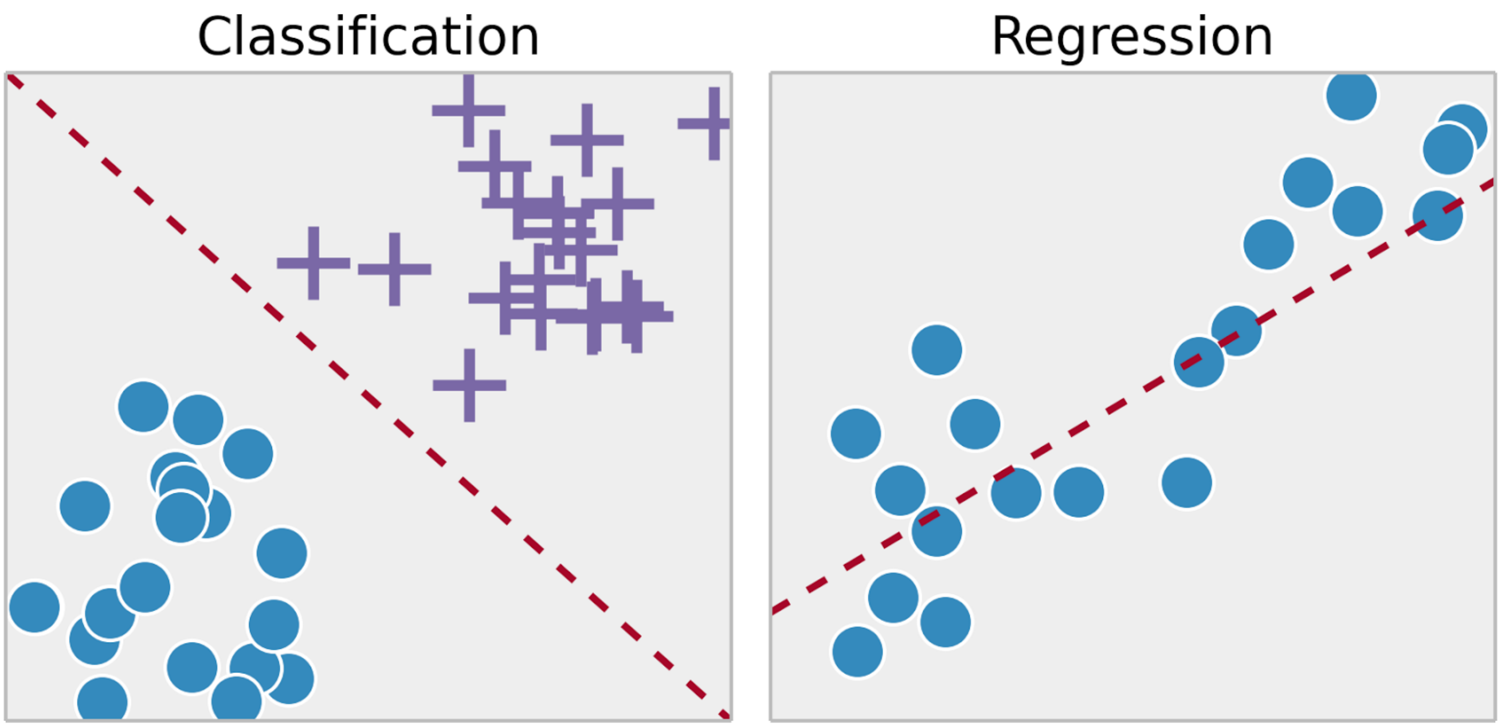


Image ack – Google

Unsupervised Learning – What?

- Training set – { $x^{(1)}$, $x^{(2)}$, …, $x^{(m)}$}
- Unlabeled dataset

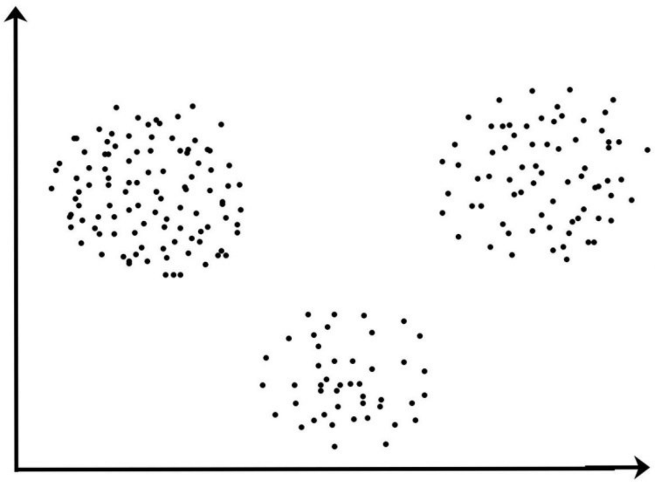


Image ack – Google

Unsupervised Learning – How?

- Find structure of data
- Group or cluster data
- Extract useful information about data

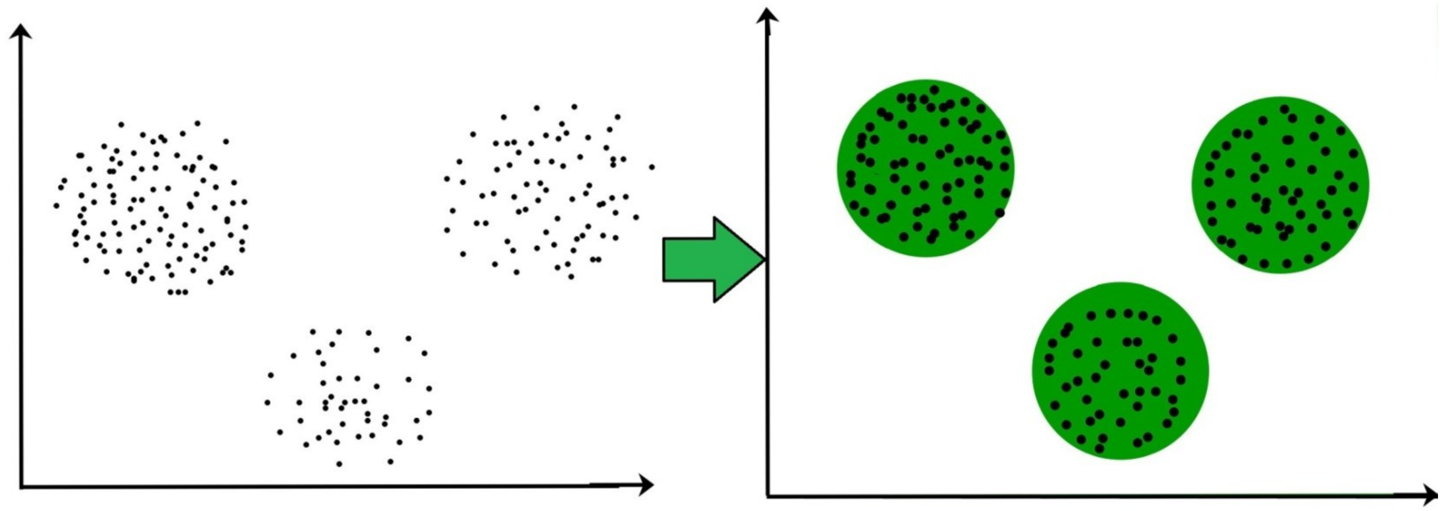


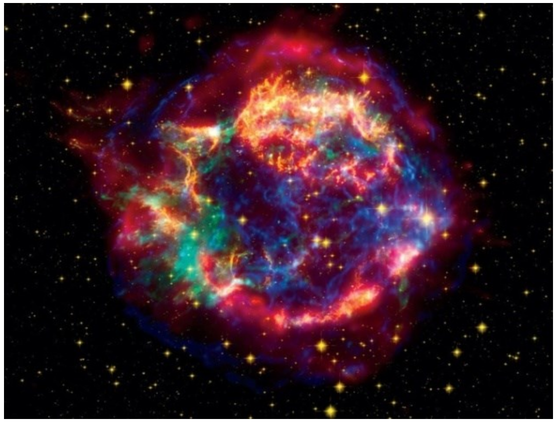
Image ack – Google

Applications

Market segmentation

Social network analysis

Computing cluster organization



Astronomical data analysis

Image ack – Google

Market Segmentation

- Customer database
- Group into market segments
- Serve market segments differently

Image ack – Google

Social Network Analysis

- Users send mails frequently
- Users receive mails frequently
- Coherence group of users

Image ack – Google

Computing Cluster Organization

- Compactness – Similarity within a cluster
- Separation – Difference between clusters
- Different nodes
  - Compute node
  - Data node
  - GPU vs CPU nodes
- Different applications

Image ack – Google

Astronomical Data Analysis

- Identify
  - Star clusters
  - Cosmic structures
- Anomaly detection
  - Brightness
  - Spectral characteristics

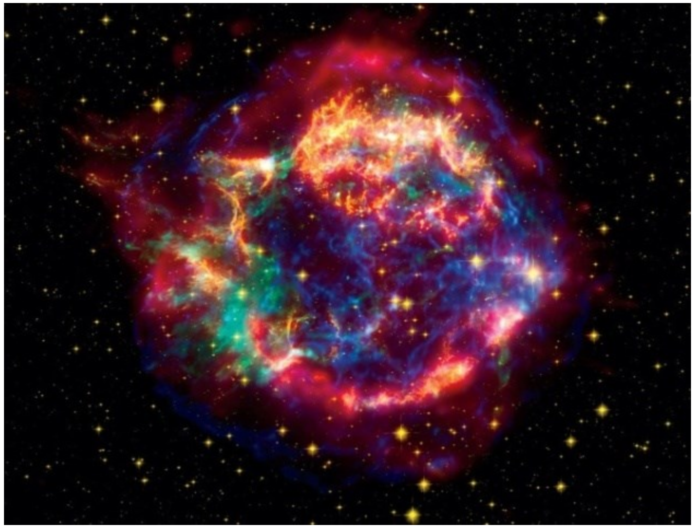


Image ack – Google

K Means Clustering

- Unsupervised learning – Unlabeled dataset
- Centroid based algorithm
- Iterative algorithm
- K – Number of pre-defined clusters
- Divide dataset into K different clusters
- Decrease distance between samples from same cluster
- Increase distance between samples from different clusters

Optimization Objective

- WCSS – Within Cluster Sum of Squares
- Variations within a cluster

$$\text{WCSS} = \sum_{c=1}^{K} \sum_{p=1}^{Pn} (\text{Centroid}_c - \text{Point}_p)^2$$

- Objective – Minimize WCSS

K Means Clustering Algorithm

1. Select number of clusters – K
2. Select random K points as centroids
3. Cluster assignment
   - Assign each data point to closest centroid
4. Centroid movement
   - Compute centroids for new clusters
5. Repeat Steps 3 to 5 until
   - Maximum number of iterations
   - Minimum variation in cluster centers
   - No change in cluster centers

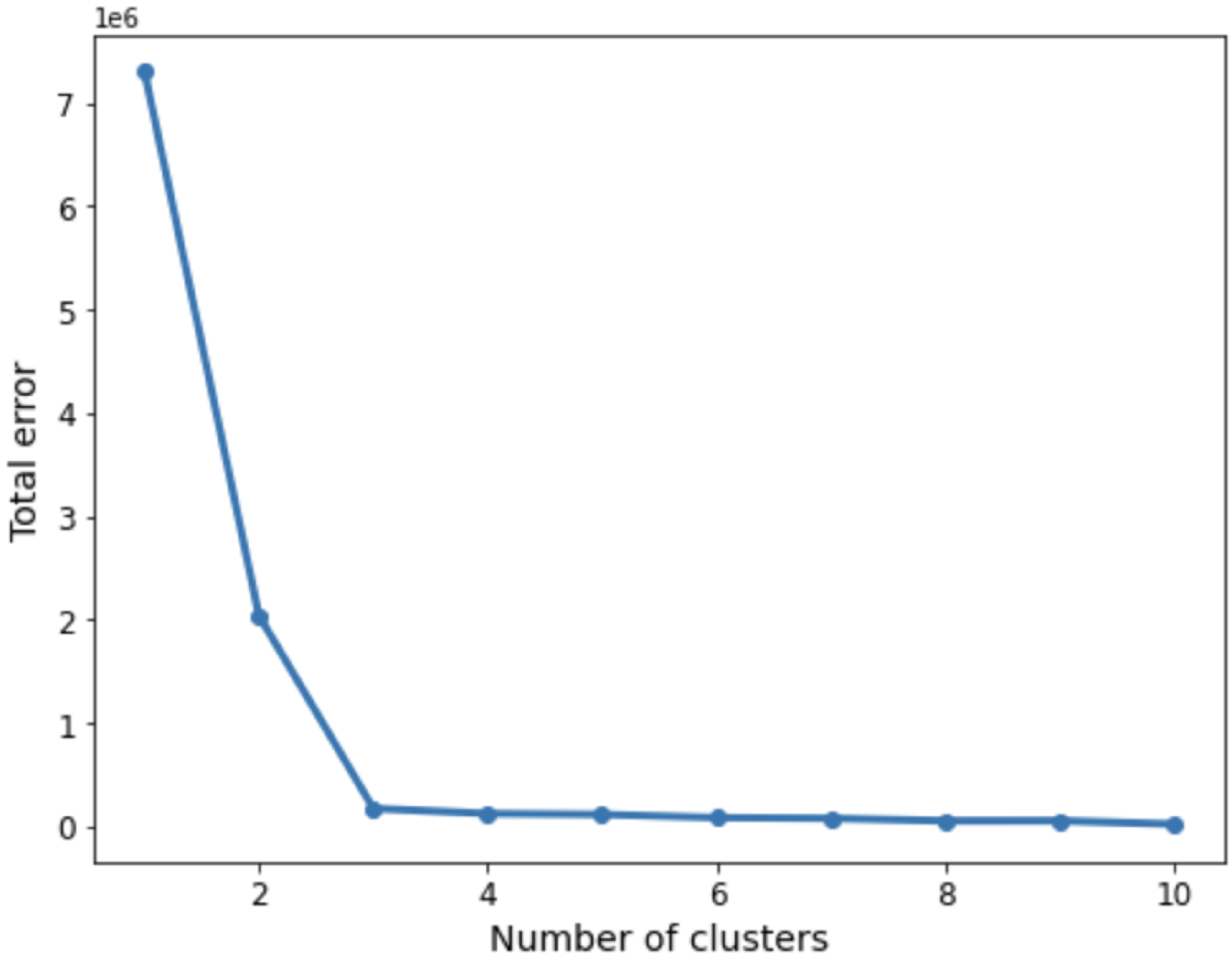
Random Initialization

1. Select maximum number of iterations
2. For each iteration
   - Select random K points as centroids
   - Compute K clusters
   - Compute WCSS value
   - Keep cluster centroids with minimum WCSS
3. Use cluster centroids with minimum WCSS

Elbow Method – Number Of Clusters

1. Select range of values for K
2. For each value of K
   – Compute WCSS value
3. Plots curve between
   – Calculated WCSS values
   – Number of clusters K
4. Best value of K – Sharp reduction in WCSS

Elbow Method – Number Of Clusters



Advantages

- Easy to implement
- Computationally faster
- Works well with spherical clusters

Disadvantages

- Difficult to predict number of clusters K
- Random initialization – Strong impact
- Sensitive to outliers
- Asymmetric clusters
- Good for spherical clusters only

Questions?

Thank you