Wednesday, September 22, 2021

Cs506 lec 4

Clustering

- definition: a clustering is a grouping/assignment of objects
- Application: outlier detection/anomaly detection filling gaps in the data
- We want similar data points In the same cluster dissimilar data in different clusters
- Types of clustering: 1. Partitional 2. Hierarchical 3. Density-based 4.soft clustering
- Partitional clustering:

Cost function

Randomly pick k centers assign each point in the dataset to its closest center

Calculate the means repeat the 2 steps several times until convergence

Does it always converge:

Yes it does

Does it always end in the optimal solution:

No depends on the initial k point picked

So find a better way to initialize the first k points

D(x)^a chose next center with probability promotional to its distance

a=2

Limitations: tends to prefer clusters of the same times

Does not handle diff density points well

Doesn't not handle points with weird shapes

1

Wednesday, September 22, 2021

How do we chose the right Ks:

Iterate though different values of k s

Use empirical/domain-specific knowledge

K-Medians (use L norm/Manhattan distance)

K-medoids

Weighted k-means (each point has a diff weight when computing the mean)