Week 4

Sunday, October 6, 2019 10:30 AM

K-Means clustering

Intro to Clustering

- Clustering gen segmentation

- What is clustering?

What is a cluster?

A group of objects that are **similar to other objects** in the cluster, and **dissimilar to data points** in other clusters.

- Clustering vs classification

- Clustering applications

- RETAIL/MARKETING:
  - Identifying buying patterns of customers
  - Recommending new books or movies to new customers
- BANKING:
  - Detecting patterns in credit card use
  - Identifying clusters of customers (e.g. loyal)
- INSURANCE:
  - Fraud detection in claims analysis
  - Insurance risk of customers
- PUBLICATION:
  - Auto-categorizing news based on their content
  - Recommending similar news articles
- WEB:
  - Characterizing patient behavior
- BIOLOGY:
  - Clustering genetic markers to identify family ties

- Why clustering?
  - EDA
  - Summary generation
  - Outlier detection
  - Finding duplicates
  - Pre processing step

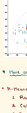
- Clustering algorithms

- Partitional-based Clustering
  - Often is random, fast
  - E.g. K-Means, K-Median, Fuzzy c-Means
- Hierarchical Clustering
  - Slow, deterministic
  - E.g. Agglomerative, Ward
- Density-based Clustering
  - Finds non-linear shaped clusters
  - E.g. DBSCAN

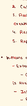
Intro to K-Means

K-Means algorithm

- Partitioning Clustering
- It divides divides the data into non-overlapping subsets (clusters) without any cluster-interior data points
- Exemplars within a cluster are very similar
- Exemplars across different clusters are very different



- Describe the similarity or distinctness



- K-Means clustering
  1) Initialize k centroids randomly
  2) Distance calculation

  3) Assign each point to the closest centroid

  4) Compute new centroids for each cluster
  5) Repeat until there are no more changes

Pick an K-Means

- K-Means clustering algorithm
  1. Randomly pick K centroids, one for each cluster
  2. Calculate the distance of each point from each centroid
  3. Assign each data point (object) to its closest centroid, creating a cluster
  4. Recalculate the centroids of K clusters
  5. Repeat the steps 2-4, until the centroids no longer move

- K-Means accuracy
  - External approach
    - Compare the clusters with the ground truth, if it is available.
  - Internal approach
    - Average the distance b/w data points within a cluster.

- Choosing K

- K-Means relies
  - Random and very sensitive initial centroids regarding
  - Produces spherical clusters
  - Fixed number of clusters (K)

Hierarchical clustering

Intro to Hierarchical Clustering

- Hierarchical clustering

Hierarchical clustering algorithm builds a hierarchy of clusters where each node is a cluster consisting of the clusters of its daughter nodes.
Cluster
Cluster
Cluster