Clustering Analysis with Documents (part I/2)

CS4422/7263 Information Retrieval
Lecture 10

Jiho Noh

Department of Computer Science
Kennesaw State University

CS4422/7263 Spring 2025

KENNESAW STATE
UNIVERSITY
COLLEGE OF COMPUTING AND
SOFTWARE ENGINEERING

What is Clustering?

- Clustering is the process of grouping a set of documents into clusters of similar items.
  - Items within a cluster should be similar.
  - Items from different clusters should be dissimilar.
- Clustering is the most representative form of **Unsupervised Learning**.
  - Unsupervised = *"There are no labeled or annotated data."*

Clustering for Data Analysis and Applications

- Grouping similar texts or documents together and discovering **patterns**
- Identifying recurring support issues and discovering new content to drive SEO practices
- Detecting topic trends in social media
- Discovering duplicate content
- Allows for creativity in finding new applications
- Can be used as a quick method for **exploratory data analysis**

Goals of Clustering

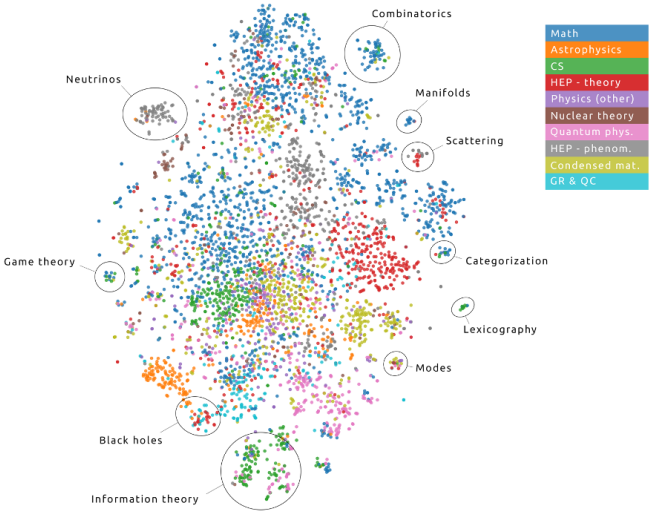
- **General goal**:
  - Put related items in the same cluster
  - Put unrelated items in different clusters
- **Secondary goals**:
  - Avoid very small and very large clusters
  - Define clusters that are easy to explain to the user
- Number of Clusters
  - The number of clusters should be appropriate for the data set we are clustering.
  - Initially, we will assume the number of clusters k is given.
  - Later, Semi-automatic methods for determining k.

Summary

- Questions?
- Discussion?

Document Clustering

- arXiv abstracts
- on 2d using t-SNE



How to Measure the Quality of Clustering

- Similarity is expressed in terms of a distance function.
- **Distance functions** differ for different types of variables:
  - Interval-scaled
  - Boolean
  - Categorical
  - Ordinal ratio
- Quality of clustering:
  - A separate "quality" function should measure the "goodness" of a cluster.
  - Defining "goodness" of a cluster is subjective.

Considerations for Cluster Analysis

- **Partitioning method**
  - Single level (e.g., k-means), hierarchical, density-based, etc.
- **Separation of clusters (hard vs. soft clustering)**
  - Can an item belong to only one cluster or multiple clusters?
- **Similarity measure**
  - Distance-based (Euclidean, Manhattan distance, cosine similarity) or Connectivity-based (density or contiguity)
- **Number of clusters**
- **Initialization methods**
- ...

Clustering Algorithms

- **Partitioning**
  - K-means, K-medoids, PAM, CLARA, CLARANS
- **Hierarchy**
  - BIRCH, CURE, ROCK, Chameleon
- **Density**
  - DBSCAN, OPTICS, Mean-shift
- **(Distribution) Model**
  - COBWEB, GMM, SOM, ART, DBCLASD
- **Graph theory**
  - Louvain, Affinity propagation, Spectral clustering, InfoMap, Density peaks
- **Grid-based**
  - STING, WaveCluster, CLIQUE
- **Fractal theory**

Partitioning Algorithms

Partitioning a dataset D into a set of k clusters, such that the sum of squared distances is minimized

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} (p - c_i)^2$$

where c_i is the centroid of cluster C_i.

- **k-means clustering**
  - Each cluster is represented by the center of the cluster.
  - Vector Quantization; we use the vector space model.
  - Relatedness between vectors is measured by Euclidean distance.
  - Euclidean distance vs. cosine similarity?

Lloyd's Algorithm

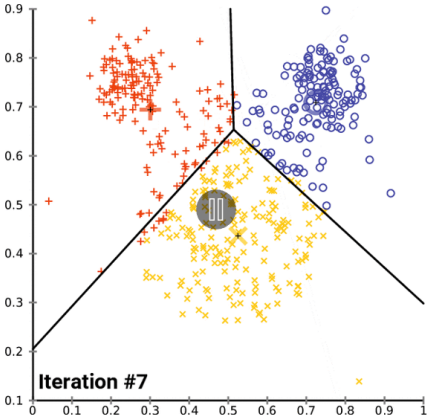
Specify the number k of clusters to assign.
Randomly initialize k centroids.

while *The centroid positions change* **do**
  expectation: Assign each point to its closest centroid.
  maximization: Compute the new centroid of each cluster.
end

Algorithm 1: k-means algorithm

- The Expectation-Maximization (EM) algorithm
  - **E-step**: Computes the expected value given the observed data.
  - **M-step**: Maximizing the expectation computed in E-step.

k-means Clustering Example



k-means clustering iterations

Does it converge?

1. Residual sum of squares (RSS) decreases during each reassignment step because each vector is moved to a closer centroid

$$RSS = \sum_{k=1}^{k} \sum_{x \in C_k} |x - \mu_k|^2$$

2. There is only a finite number of clusters.
3. Thus, we must reach a fixed point.
4. A finite set & monotonically decreasing evaluation function implies convergence.

Initialization of k-means

- **Random seed selection** is just one of many ways K-means can be initialized.
  - Random seed selection is not very robust; Cluster assignment converges, but it can be sub-optimal.
- We need better ways of computing initial centroids:

Methods of Initializing K-means

- **K-mean++**: selects initial cluster centroids using sampling based on an empirical probability distribution of the points' contribution to the overall inertia.
- **RP**: Randomly selected point.
- **RGC**: The data points are partitioned randomly.
- **SIMFP**: Farthest points (simple selection); the first centroid is selected as a random case. The second centroid is selected as the case maximally distant from the first. Continues
- **Hierarchical Clustering Initialization**
- **Multiple Random Initialization**

K-means++ Initialization

1. Choose one center uniformly at random among the data points.
2. For each data point x not chosen yet, compute $D(x)$, the distance between x and the nearest center that has already been chosen.
3. Choose one new data point at random as a new center, using a weighted probability distribution where a point x is chosen with probability proportional to $D(x)^2$.
4. Repeat Steps 2 and 3 until k centers have been chosen.

Hierarchical Clustering Initialization

- First, perform hierarchical clustering.
- The K clusters with the largest dissimilarity between them are selected as the initial centroids.
- Effective with complex structure.

Multiple Random Initialization

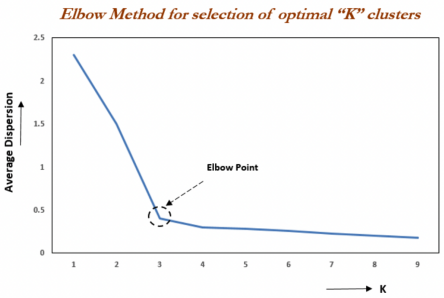
- Run K-means multiple times with different random initialization
- Select the clustering with the lowest RSS (Residual Sum of Squares)

How to find K? The Elbow Method

- Most well-known method
- Calculate the **Within-Cluster-Sum of Squared Errors (WSS)** for different values of K

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

where S_k is the set of observations in the k-th cluster.



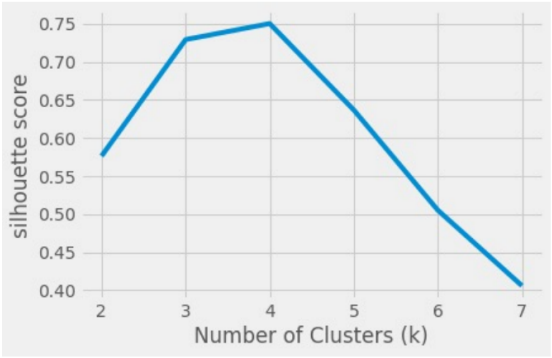
How to find K? The Silhouette Method

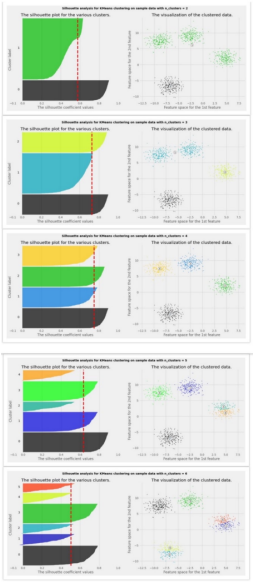
- The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).
- The range of the Silhouette value is between +0 and -1 (a high value is desirable).

Algorithm:

- $a(i)$: The average distance of that point with all other points in the same cluster.
- $b(i)$: The average distance of that point with all the points in the closest cluster to its cluster.
- $s(i)$: The silhouette value.

$$s(i) = \frac{b(i) - a(i)}{\max(b(i), a(i))}$$


silhouette score
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
2
3
4
5
6
7
Number of Clusters (k)



Hierarchical Clustering

- Use distance matrix as clustering criteria.
- This method does not require the number of clusters k as an input, but needs a termination condition.

Agglomerative (bottom-up)

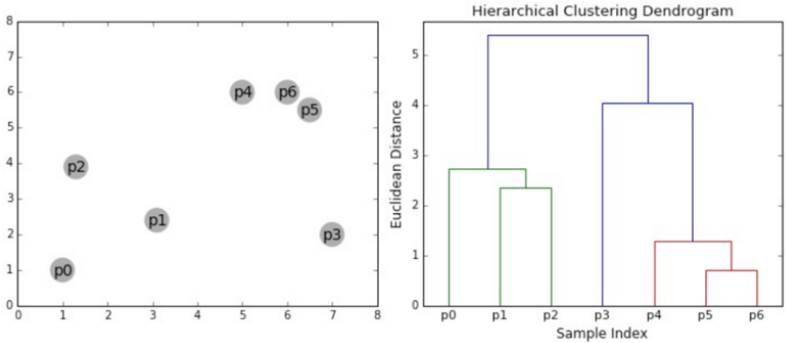
- Assign each data point as one cluster.
- Iteratively combine sub-clusters.
- Eventually, all data points is a part of 1 cluster.

Divisive (top-down)

- Assign all data points to the same cluster.
- Iteratively divide into smaller groups.
- Eventually each data point forms its own cluster.

AGNES (Agglomerative Nesting)

1. Assign each data point to its own cluster
2. Compute similarity between clusters
3. Merge two most similar clusters to form one cluster



Cluster Similarity

- How do we compute similar clusters?
  - Distance between two points in the clusters?
  - Distance from means of two clusters?
  - Distance between two closest points in the clusters?
- Different similarity metric could produce different types of cluster
- Common similarity metric used
  - Single linkage
  - Complete linkage
  - Average group linkage

Similarity between Clusters

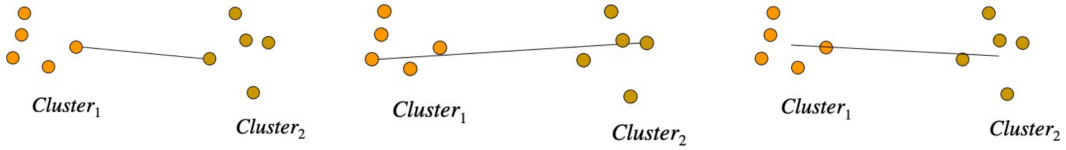


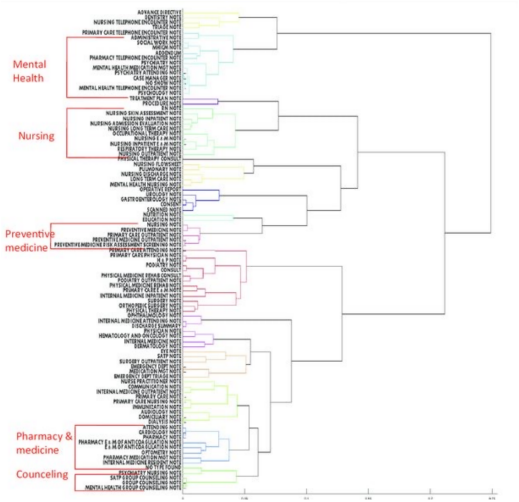
Figure: **Single-** (left), **Complete-** (middle), and **Average-Linkage** (right)

DIANA (DIvisive ANAlysis)

- Introduced by Kaufman and Rousseeuw in 1990
- The algorithm starts with all data points in one cluster
- Clusters are recursively divided into smaller sub-clusters
- Division continues until each data point is in its own cluster
- Based on a chosen dissimilarity measure

Dendrogram

- A tree diagram that is used to represent hierarchical relationships between data points or objects.
- It shows the similarities and differences between groups.
- A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.



Density-Based Clustering Methods

- Clustering is based on the density (local cluster criterion) of data points in the feature space.
- *How do you measure the density?*
  - By looking at the number of data points within a certain radius.
- Features
  - Discover clusters of arbitrary shape
  - Handle noise and outliers
  - Do not require the number of clusters to be specified in advance
  - Need density parameters as termination condition
- Methods:
  - DBSCAN, OPTICS, DENCLUE, CLIQUE

Density-based Clustering: Concepts

Two important parameters: ϵ (eps) and MinPts

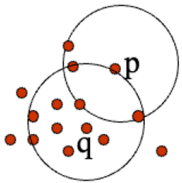
- ϵ (eps): The maximum distance between two data points to be considered in the same neighborhood
- MinPts: The minimum number of data points required to form a dense region (core point)

$$N_{\text{Eps}}(p) : \{q \in D \mid \text{dist}(p,q) \leq \text{Eps}\}$$

Core Point: A data point with at least MinPts data points in its ϵ-neighborhood

$$\left|N_{\text{Eps}}(q)\right| \geq \text{MinPts}$$

Border Point: A data point that is not a core point but is within the ϵ-neighborhood of a core point



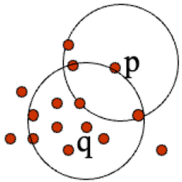
MinPts = 5

Eps = 1 cm

Density-based Clustering: Concepts (Cont.)

Directly Density-Reachability: A data point p is density-reachable from another data point q if

- $p \in N_\epsilon(q)$
- core point condition: $|N_\epsilon(q)| \geq$ MinPts

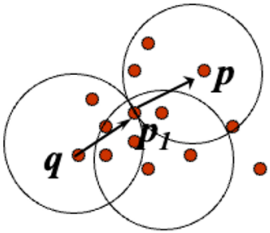


MinPts = 5

Eps = 1 cm

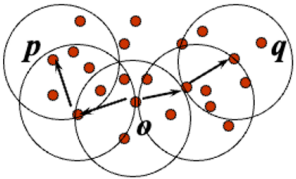
Density-based Clustering: Concepts (Cont.)

Density-Reachable: A data point p is density-reachable from another data point q if there is a chain of core points directly density-reachable among them.



Density-based Clustering: Concepts (Cont.)

Density-Connected: A data point p is density-reachable from another data point q if there is a point o such that both, p and q are density-reachable from o w.r.t. Eps and $MinPts$



DBSCAN Algorithm

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a popular density-based clustering algorithm.
- It groups data points together that are closely packed and marks outliers as noise.

The algorithm works by:

1. Finding core points by identifying data points with at least MinPts data points in their ϵ-neighborhood
2. Expanding the clusters by finding density-reachable points from the core points
3. Assigning border points to the clusters
4. Continue the process until all of the points have been processed
5. Marking noise points as outliers

Measuring Clustering Quality — Instrinsic

- An external reference is not needed
- Unsupervised
- Methods
  - **Silhouette Score**
  - **Davies-Bouldin Index**
  - **Dunn Index**

Measuring Clustering Quality — Extrinsic

- Compare a clustering against the ground truth
- Supervised
- Methods
  - **Adjusted Rand Index**
  - **Fowlkes-Mallows Index**
  - **Jaccard Index**

1. Clustering Methods
   - Basic Concepts
   - Partitioning Methods
   - Hierarchical Methods
   - Density-based Methods
2. Evaluation of Clustering
3. Topic Modeling
4. **LLM-based Approaches**