Text Clustering and Topic Modeling

Applied Text Mining

Dr. Maryam Movahedifar

4-6 December 2024

University of Bremen, Germany
movahedm@uni-bremen.de

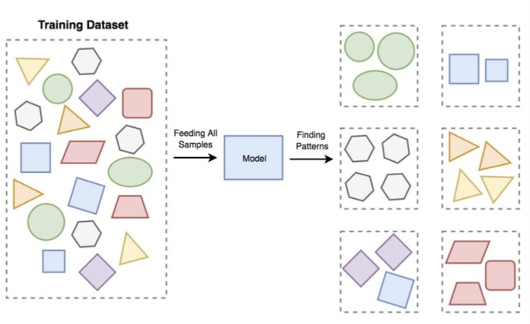
Outline

Introduction to Clustering

What is Clustering?

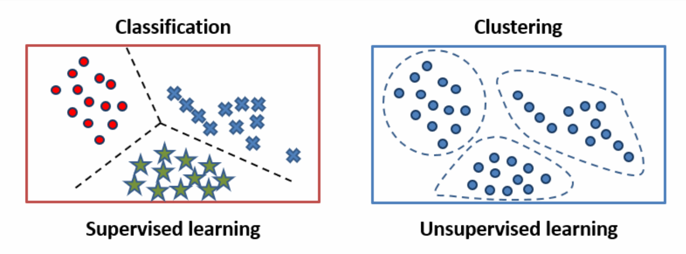
Clustering is the process of grouping similar objects into clusters without prior knowledge of the categories. It helps discover the natural structure in the data.

- **Criterion:** High intra-cluster similarity, low inter-cluster similarity.
- **Applications:** Grouping tweets, customer reviews, scientific articles, etc.



Clustering vs. Classification

- **Clustering:** Unsupervised learning where clusters are inferred from data without labeled input.
- **Classification:** Supervised learning that assigns predefined labels to data.



Clustering Methods

Clustering Algorithms Overview

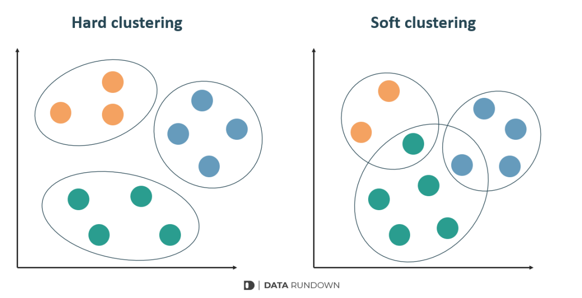
There are various clustering methods categorized by their approach:

- Hard vs. Soft Clustering:
  - **Hard Clustering:** Each data point belongs to exactly one cluster.
  - **Soft Clustering:** Each data point can belong to multiple clusters with varying probabilities.
- Partitional Clustering: Algorithms like K-means and K-medoids that divide data into non-overlapping clusters.
- Hierarchical Clustering: Builds a tree-like structure (dendrogram) to represent nested clusters.
- Topic Modeling:
  - Unsupervised learning to identify topics in a collection of documents.
  - Algorithms like LDA (Latent Dirichlet Allocation) are commonly used.
  - Each document is represented as a mixture of topics, and each topic is a distribution over words.

Hard vs. Soft Clustering

Hard vs. Soft Clustering

- Hard Clustering: Each document belongs to exactly one cluster.
- Soft Clustering: A document can belong to multiple clusters.



Partitional Clustering

Partitional Clustering: Overview

Partitional clustering divides n documents into K clusters by optimizing a partitioning criterion.

- Objective: Minimize intra-cluster distance and maximize inter-cluster distance.
- Challenges: Finding the globally optimal partition is intractable for many objective functions.
- Heuristic Methods:
  - **K-Means:** Assigns each document to the nearest centroid.

K-Means Algorithm: Mathematical Explanation

- Objective Function: Minimize the sum of squared distances between each point $\mathbf{x}_i$ and its assigned cluster centroid μ_j:

$$J = \sum_{j=1}^{K} \sum_{\mathbf{x}_i \in C_j} \|\mathbf{x}_i - \mu_j\|^2$$

- Algorithm Steps:
  - **Initialization:** Randomly select K centroids $\mu_1, \mu_2, \ldots, \mu_K$.
  - **Assignment Step:** Assign each data point $\mathbf{x}_i$ to the nearest centroid:

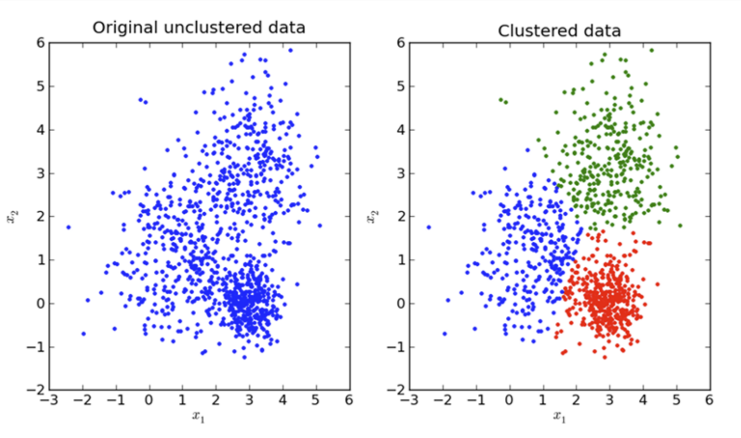
$$C_j = \{\mathbf{x}_i : \|\mathbf{x}_i - \mu_j\|^2 \leq \|\mathbf{x}_i - \mu_k\|^2 \, \forall k\}$$

  - **Update Step:** Recalculate the centroids for each cluster:

$$\mu_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i$$

  - Repeat the assignment and update steps until convergence.

Example of K-Means

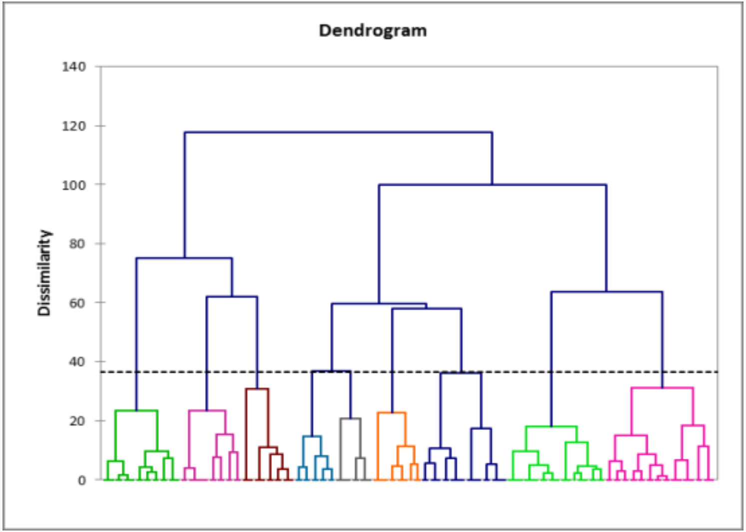


Hierarchical Clustering Overview

Hierarchical clustering builds a tree-based hierarchical taxonomy (dendrogram) to group data into clusters based on their similarity.

- Dendrogram: A tree structure representing the nested clustering of documents.
- Clustering is obtained by cutting the dendrogram at a chosen level, with each connected component forming a cluster.
- Hierarchical clustering does not require specifying the number of clusters beforehand.

Example of Hierarchical Clustering



Types of Hierarchical Clustering

Hierarchical clustering is divided into two main types:

- Top-down Divisive Clustering:
  - Start with all data in one cluster.
  - Repeatedly split the remaining clusters into two smaller clusters.
  - Suited for datasets with clear large-to-small groupings.
- Bottom-up Agglomerative Clustering (HAC):
  - Start with each document in a separate cluster.
  - Iteratively merge the closest pair of clusters until only one cluster remains.
  - Forms a hierarchy visualized as a binary tree (dendrogram).

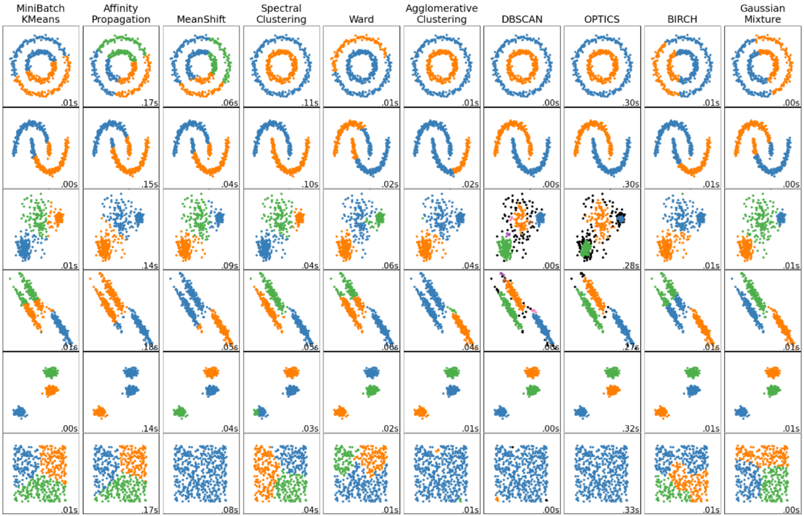
Linkage Methods in Agglomerative Clustering

- Single-Link: Uses the smallest distance between any two points in two clusters (nearest neighbor).
- Complete-Link: Uses the largest distance between any two points in two clusters (farthest neighbor).
- Centroid: Merges clusters based on the distance between their centroids (average position).
- Average-Link: Uses the average distance between all pairs of points from two clusters.
- Ward's Linkage: Minimizes the variance within clusters by merging the pair that results in the smallest increase in total variance.

What is `scikit-learn`?

- `scikit-learn` is a widely-used open-source Python library for machine learning.
- It provides simple and efficient tools for data mining, data analysis, and machine learning.
- Built on top of popular libraries like `NumPy`, `SciPy`, and `matplotlib`.
- Supports various machine learning tasks, including:
  - Supervised Learning: Classification and Regression.
  - Unsupervised Learning: Clustering, Dimensionality Reduction.
  - Model selection and evaluation.
- Well-documented with extensive examples for practical use.

Comparison of Clustering Algorithms in `scikit-learn`



A comparison of the clustering algorithms in scikit-learn

Topic Modeling

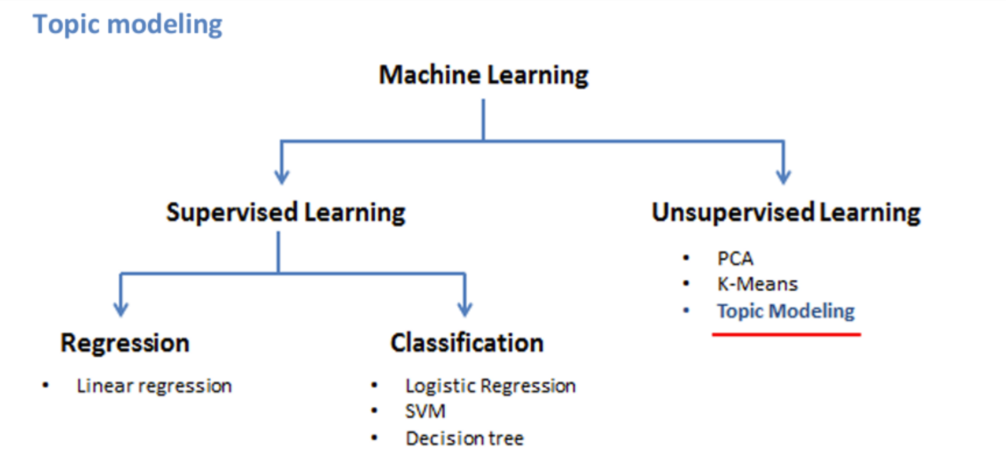


Figure 2: Overview of Machine Learning: Supervised vs. Unsupervised Learning

Topic Modeling and LDA

Topic modeling discovers latent topics in a collection of documents. Latent Dirichlet Allocation (LDA) is one of the most common methods.

- Documents have a probability distribution over topics.
- Topics have a probability distribution over words.

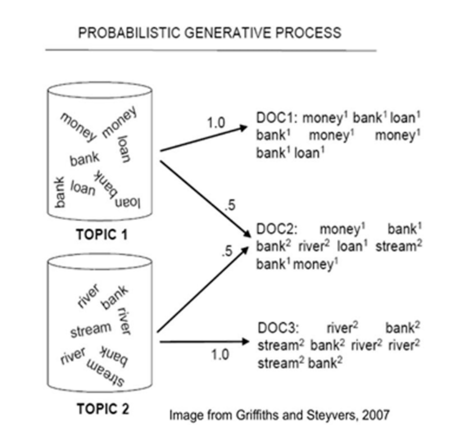


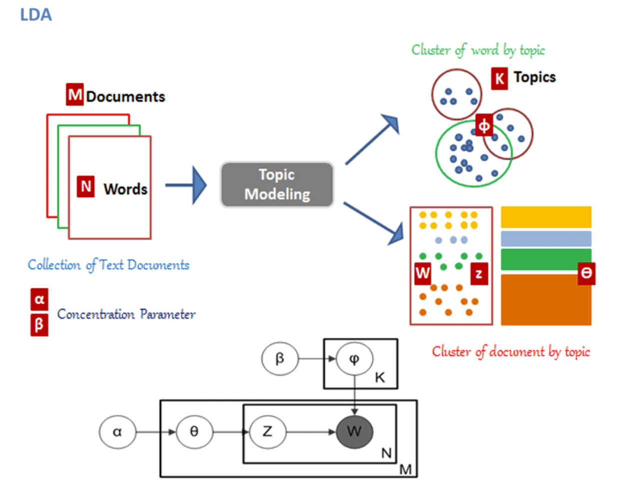
Image from Griffiths and Steyvers, 2007

This diagram illustrates the probabilistic generative process of LDA:

- Topic 1: Words like "money," "bank," and "loan."
- Topic 2: Words like "river," "stream," and "bank."

LDA: Process Overview

This slide provides an overview of how LDA clusters words and documents by topics.



Posterior Inference and New Data Integration

Posterior Inference:

- Identify topics describing a collection of documents.
- Estimate the probability of each topic for a document using the posterior distribution.

New Data Integration:

- For new documents, determine their fit within the existing topic structure.
- Use pre-trained model parameters to compute the document's topic distribution.

Example: *"I enjoy eating broccoli while watching football."*

- Topic 1 (Food): Keywords like "broccoli," "banana" Topic 2 (Sports): Keywords like "football," "tennis"
- LDA assigns probabilities: - 70% Topic 1 (Food) - 30% Topic 2 (Sports)

Iterative Word Reassignment in LDA

Key Steps:

- Each word w in a document is initially assigned a random topic.
- For each word w in each document d, update the topic assignment based on:
  - $p(\text{topic } t|\text{document } d)$: Proportion of words in d currently assigned to topic t.
  - $p(\text{word } w|\text{topic } t)$: Proportion of assignments to t across all documents for word w.

Convergence and Steady State in LDA

- LDA iterates the reassignment process until a steady state is reached.
- **Steady State:** Topic assignments stabilize, resulting in consistent topic distributions.
- **Estimations:**
  - **Topic Mixtures:** Proportion of each document's words assigned to each topic.
  - **Topic-Word Distributions:** Frequency of each word within each topic across the corpus.

LDA: Identifying Structure in Text

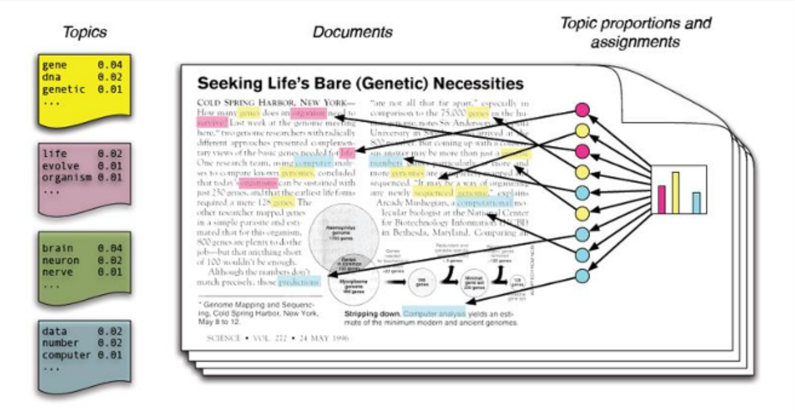


Figure 3: Overview of Identifying Structure in Text

Variations of Latent Dirichlet Allocation (LDA)

LDA has evolved into several variations to address different modeling needs:

- Hierarchical LDA (hLDA): Automatically discovers hierarchical relationships among topics, forming a tree-like structure.
- Supervised LDA (sLDA): Integrates class labels during training to learn topics aligned with specific categories or outcomes.
- Hybrid LDA: Combines LDA with additional information extraction, merging topic modeling with other analyses.
- LDA & BERT: Explores the integration of LDA with deep learning models like BERT.

Cluster Validation

Cluster Evaluation

- Internal Validation: Measures coherence within clusters using metrics like Davies-Bouldin Index.
- External Validation: Compares clusters to known labels using metrics like Rand Index.

Clustering Performance Evaluation in scikit-learn

scikit-learn provides various metrics for clustering evaluation:

- **Rand Index:** Measures the similarity between the clustering result and a ground truth classification.
- **Mutual Information Scores:** Captures the amount of shared information between clusters and the ground truth.
- **Homogeneity, Completeness, and V-measure:** Evaluate how well clusters contain only members of a single class (homogeneity) and how well all members of a given class are assigned to the same cluster (completeness).
- **Fowlkes-Mallows Score:** Measures the similarity between true clusters and predicted clusters by evaluating the pairwise precision and recall.
- **Silhouette Coefficient:** Measures how similar an object is to its own cluster compared to other clusters.

Clustering Evaluation Metrics (Continued)

Additional metrics for clustering evaluation in scikit-learn:

- **Calinski-Harabasz Index:** Measures the ratio of the sum of between-clusters dispersion to within-cluster dispersion.
- **Davies-Bouldin Index:** Evaluates the average similarity ratio between each cluster and the cluster that is most similar to it.
- **Contingency Matrix:** A matrix that shows the overlap between the true labels and the predicted clusters.
- **Pair Confusion Matrix:** Measures pairwise similarity, detailing true positives, false positives, true negatives, and false negatives in clustering assignments.

Summary of Text Clustering and Evaluation

Key takeaways for text clustering:

- Clustering is an unsupervised learning method where clusters are inferred from data without human input.
- The outcome of clustering can be influenced by:
  - Number of clusters.
  - Similarity measure used (e.g., cosine similarity, Euclidean distance).
  - Representation of documents (e.g., TF-IDF, embeddings).
- Evaluation is crucial to ensure meaningful clustering results.

Practical 4