Unsupervised Learning

Lecture 14: Text Clustering

Harito

September 15, 2025

Recap: Supervised vs. Unsupervised Learning

Supervised Learning (e.g., Text Classification)

- Requires labeled data (input-output pairs).
- Learns a mapping from input to output.
- Goal: Predict labels for new, unseen data.

Unsupervised Learning (e.g., Clustering)

- Works with unlabeled data.
- Learns patterns or structures within the data.
- Goal: Discover hidden groupings or relationships.

What is Text Clustering?

Definition

Text Clustering is the task of grouping a set of text documents in such a way that documents in the same group (cluster) are more similar to each other than to those in other groups.

Analogy: Imagine you have a pile of unsorted books. Clustering is like sorting them into piles based on their content (e.g., fiction, non-fiction, sci-fi, history) without knowing the categories beforehand.

Applications of Text Clustering

- **Topic Discovery**: Automatically finding themes or topics in a large collection of documents (e.g., news articles, research papers).
- **Document Organization**: Structuring and navigating large archives of text data.
- **Anomaly Detection**: Identifying unusual or outlier documents that don't fit into any established group.
- **Customer Segmentation**: Grouping customer feedback or reviews to identify common issues or sentiments.
- **Information Retrieval**: Improving search results by grouping similar documents.

K-Means Clustering Algorithm

K-Means is one of the simplest and most popular clustering algorithms.
Goal: Partition ‘n‘ data points into ‘k‘ clusters, where each data point belongs to the cluster with the nearest mean (centroid).
Algorithm Steps:

1. **Initialization**: Choose ‘k‘ initial centroids randomly.
2. **Assignment Step**: Assign each data point to the cluster whose centroid is closest.
3. **Update Step**: Recalculate the centroids as the mean of all data points assigned to that cluster.
4. **Iteration**: Repeat steps 2 and 3 until the centroids no longer change significantly or a maximum number of iterations is reached.

Document Representation for Clustering

Just like with classification, text documents must be converted into numerical vectors before clustering.

- **TF-IDF Vectors** (Lab 3): Effective for capturing term importance and distinguishing documents.
- **Averaged Word Embeddings** (Lab 4): Can capture semantic similarity between documents, even if they don't share exact words.

The choice of representation can significantly impact the quality of the clusters.

Evaluating Clustering

Evaluating clustering is challenging because there are no ground truth labels.

Internal Evaluation Metrics:

- **Silhouette Score**: Measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation).
  - Score ranges from -1 to +1.
  - +1: Well-separated clusters.
  - 0: Indifferent, overlapping clusters.
  - -1: Data points assigned to the wrong cluster.
- **Inertia (Sum of Squared Distances)**: Measures how internally coherent clusters are. Lower is better, but can be misleading.

External Evaluation Metrics: Require ground truth labels (e.g., Adjusted Rand Index, Normalized Mutual Information).

Next Steps

Time for Lab 14!

Objective:

- Represent documents using TF-IDF or embeddings.
- Implement K-Means clustering for text data.
- Analyze the resulting clusters.