Scribing: Clustering_L08

LAU, Trelan

SID:1155214344

Contents

1 Why We Need Clustering

Clustering is an unsupervised machine learning technique fundamental to data mining and exploratory data analysis. It addresses the need to find inherent structures in data without prior knowledge of the output labels.

1.1 Fundamental Purpose

The primary goal of clustering is to partition unlabeled data into groups of similar objects. This process of organizing data serves several key purposes:

- **Understanding and Gaining Insight:** Clustering is often a preliminary step in data analysis. As a stand-alone tool, it helps in understanding the underlying distribution and patterns within the data. By observing the characteristics of each cluster, analysts can formulate hypotheses about the data.
- **Data Organization and Summarization:** It provides a way to structure data. For example, grouping a large number of documents allows for better organization and faster searching (slide 11). Clustering can also be used for data summarization by reducing the size of large datasets. The cluster centroids can sometimes be used as representatives for all data points in that cluster, preserving significant information while reducing complexity.
- **Pre-processing Step:** Clustering can serve as an essential pre-processing step for other machine learning algorithms. For example, it can be used for feature engineering, where a cluster ID is added as a new feature to the dataset for a subsequent supervised learning task.

1.2 Improving Daily Life

The application of clustering has a tangible impact on our daily life by enabling smarter, more efficient systems:

- **Customer Segmentation:** In marketing, companies cluster customers based on their purchasing behavior, demographics, or browsing history. This allows for targeted marketing campaigns and personalized product recommendations, optimizing the business strategy based on the needs of different groups (slide 11).
- **Healthcare:** Patients can be clustered into different groups based on their clinical characteristics or genetic profiles. This facilitates the development of different treatment protocols for different groups, forming a cornerstone of personalized medicine (slide 11).
- **Information Management:** Search engines and news aggregators group related documents, web pages, or articles, allowing users to browse through topics of interest efficiently.

1.3 Applications in Biology

Clustering is indispensable in bioinformatics and computational biology for analyzing high-dimensional biological data (slide 12):

- **Gene Expression Analysis:** Clustering helps to identify groups of genes with similar expression patterns across different conditions.
  - **Co-expressed genes** are often involved in the same biological pathways or are co-regulated.
  - **Differentially expressed genes** can be related to specific diseases or cellular states.
- **Disease Sub-typing:** Clustering patient samples based on molecular data (e.g., gene expression) can reveal previously unknown disease sub-types. This is crucial in fields like oncology, where different sub-types of cancer may respond differently to treatments.
- **Cell Type Identification:** In single-cell RNA sequencing (scRNA-seq), clustering is used to group thousands of cells into distinct populations, enabling the discovery and characterization of new or rare cell types.

2 What is Clustering Analysis?

At its core, clustering analysis is about discovering groups in data.

2.1 Academic Definition

A widely accepted definition of clustering analysis is:

*"The process of finding groups of objects such that the objects in a group will be **similar** to one another and **different** from the objects in other groups."* (slide 5)

2.2 Keywords and Core Concepts

This definition is built on two fundamental concepts, which represent a dual objective:

- **Intra-cluster Similarity (Cohesion):** Objects within the same cluster should be as similar as possible. The "intra-cluster differences are small" (slide 5). This measures how tightly-knit the objects within a cluster are.

- **Inter-cluster Similarity (Separation):** Objects in different clusters should be as dissimilar as possible. The "inter-cluster differences are large" (slide 5). This measures how distinct and well-separated one cluster is from another.

The effectiveness of a clustering algorithm is judged by its ability to maximize intra-cluster similarity while minimizing inter-cluster similarity. The precise meaning of "similarity" and "dissimilarity" is crucial and depends on the chosen metric (e.g., Euclidean distance, Cosine similarity, Correlation) and the nature of the data itself.

2.3 Output of Clustering

The output of a clustering algorithm is a **cluster structure**, which typically includes:

- A partition of the data objects into a set of clusters.
- Each object is assigned a **cluster indicator** or label (e.g., an integer from 1 to K, where K is the number of clusters). (slide 19)
- In the case of *hierarchical clustering*, the output is a tree-like structure known as a **dendrogram**, which shows the nested grouping of objects and the sequence of merges or splits. (slide 13)

Ultimately, clustering adds labels to a previously unlabeled dataset, with the labels corresponding to the discovered groups.

2.4 Examples of Real-World Applications

Clustering analysis is applied across numerous domains (slide 14):

1. **Business Intelligence:** Grouping stocks with similar price fluctuations for portfolio management.
2. **Document Analysis:** Grouping related documents for browsing and topic modeling.
3. **Genomics:** Grouping genes and proteins that share similar functionality or evolutionary history.
4. **Image Processing:** Segmenting images by grouping pixels with similar properties.
5. **Social Network Analysis:** Identifying communities or cliques within social networks.

3 How to Do Clustering

The process of clustering follows a logical workflow, underpinned by mathematical measures of similarity and a chosen algorithm.

3.1 Logical Flow of a Clustering Project

A clustering task is not a single-step process. It involves several stages, from data preparation to the final analysis of clusters (slides 25, 49).

1. **Data Collection and Preparation:**
    - **Collect Data:** Gather the raw data (e.g., gene expression matrix).
    - **Data Cleaning:** Handle missing data, remove duplicates, and denoise data if necessary. This step ensures data quality.
    - **Data Normalization:** Scale features to a common range. This is critical for distance-based algorithms like Euclidean distance to prevent features with large scales from dominating the clustering process.
2. **Core Clustering Components:**
    - **Select a Similarity/Distance Metric:** Choose a mathematical formula to quantify how similar or different two data points are. This is arguably the most critical step as the "correct" clustering depends heavily on this choice.
    - **Select a Clustering Algorithm:** Choose an algorithm (e.g., Hierarchical, K-Means, DBSCAN) that fits the data's characteristics and the analysis's goal.
    - **Perform Clustering:** Run the algorithm on the prepared data to generate cluster assignments.
3. **Exploration and Validation:**
    - **Visualization:** Use plots (e.g., scatter plots for low-dimensional data, heatmaps, dendrograms) to visualize the results.
    - **Interpretation:** Analyze the resulting clusters. What are the defining characteristics of each group? This stage turns the numerical output into actionable insights.

3.2 Mathematical Formulas for Similarity and Distance

The choice of metric defines what "similar" means for your dataset.

Minkowski Distance A generalized distance metric. For two vectors p and q of dimension m:

$$\text{dist}(p, q) = \left(\sum_{k=1}^{m} |p_k - q_k|^r \right)^{\frac{1}{r}}$$

Why: Its flexibility allows it to adapt to different geometric interpretations of distance. Key special cases include:

- **r=1 (Manhattan Distance):** Measures the distance as the sum of absolute differences along each dimension ("city block" distance). Useful when movement is restricted to a grid.
- **r=2 (Euclidean Distance):** The shortest straight-line distance between two points. It is the most common and intuitive distance metric.
- **r=∞ (Supremum Distance):** The maximum absolute difference between any single dimension.

Pearson Correlation Measures the linear relationship between two vectors X and Y. It ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation).

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Why: It is ideal for finding objects that have similar trends or patterns, regardless of their absolute magnitudes. In gene expression analysis (slides 17, 43), it helps identify genes that are up-regulated or down-regulated together, suggesting they are part of the same biological process.

Mahalanobis Distance A sophisticated metric that accounts for the correlations within the data. For vectors p and q with a covariance matrix Σ:

$$D_M(p, q) = \sqrt{(p - q)^T \Sigma^{-1} (p - q)}$$

Why: Unlike Euclidean distance, it is scale-invariant and considers the covariance of the data distribution. It effectively measures the distance in terms of standard deviations, providing a more robust measure when features are correlated and have different variances (slide 27). This is useful for identifying outliers in a multivariate distribution.

3.3 Hierarchical Clustering: Gene Expression Example

Hierarchical clustering builds a hierarchy of nested clusters, often visualized as a tree called a **dendrogram** (slide 13). The **agglomerative** (bottom-up) approach is most common.

The Agglomerative Algorithm:

1. Start by treating each data point (e.g., each gene) as its own cluster.

2. Compute a similarity matrix between all clusters.
3. **Repeat**:
    - (a) Merge the two closest (most similar) clusters.
    - (b) Update the similarity matrix to reflect the merge.
4. Continue until only one cluster remains.

How to Update the Similarity Matrix (Linkage Criteria): When merging clusters, the similarity to other clusters must be re-calculated. Common criteria include (slide 15):

- **Single Linkage (MIN):** The similarity is the maximum similarity between any two points in the two clusters. It can handle non-elliptical shapes but is sensitive to noise.
- **Complete Linkage (MAX):** The similarity is the minimum similarity between any two points. It produces more compact clusters.
- **Average Linkage:** The similarity is the average similarity between all pairs of points.

Walkthrough: Clustering Gene Expression Data
Let's follow the example from the slides (18-22) using the provided gene data and Pearson Correlation as the similarity metric. We will use **single linkage** for updating.

Step 0: Initial Data and Similarity Matrix
We begin with 5 genes, each its own cluster. The initial pairwise similarity matrix is computed (slide 18):

Table 1: Initial Similarity (Correlation) Matrix

Gene	**At4g35770**	**At1g30720**	**At4g27450**	**At2g34930**	**At2g05540**
At4g35770	1				
At1g30720	0.9733	1			
At4g27450	-1.0000	-0.9733	1		
At2g34930	0.9493	0.9909	-0.9493	1	
At2g05540	0.5774	0.5620	-0.5774	0.4528	1

Iteration 1: First Merge (slide 19)

1. **Merge:** The highest similarity is **0.9909** between **At2g34930** and **At1g30720**. We merge them into a new cluster: `(C1: At1g30720, At2g34930)`.
2. **Update:** Using single linkage, the similarity between C1 and any other gene (e.g., At4g35770) is the *maximum* of the individual similarities:
    - sim(C1, At4g35770) = max(0.9733, 0.9493) = **0.9733**

- sim(C1, At4g27450) = max(-0.9733, -0.9493) = **-0.9493**
- sim(C1, At2g05540) = max(0.5620, 0.4528) = **0.5620**

Iteration 2: Second Merge (slide 20) The updated matrix is:

Table 2: Matrix After First Merge

Cluster	**At4g35770**	**C1**	**At4g27450**
At4g35770	1		
C1	0.9733	1	
At4g27450	-1.0000	-0.9493	1
At2g05540	0.5774	0.5620	-0.5774

1. **Merge:** The new highest similarity is **0.9733** between **At4g35770** and cluster **C1**. We merge them into a new, larger cluster: `(C2: At4g35770, At1g30720, At2g34930)`.
2. **Update:** We repeat the update step with the new cluster C2.

...This process continues, merging clusters based on the highest similarity at each step, until all genes are grouped into a single cluster. The sequence of merges is what forms the final dendrogram (slide 22).

3.4 Programming Logic in Python

Scikit-learn is a popular Python library for machine learning (slide 23). While SciPy is often used for dendrograms, Scikit-learn provides an easy-to-use interface for hierarchical clustering.

Key Steps:

1. **Import Libraries:**

```
import numpy as np
from sklearn.cluster import AgglomerativeClustering
import matplotlib.pyplot as plt
from scipy.cluster.hierarchy import dendrogram, linkage
```

2. **Prepare Data:** The data should be in a numerical format, typically a NumPy array. Let's call it `X`.
3. **Perform Clustering (Scikit-learn):**

```
# Initialize the model
# affinity='correlation' cannot be used with linkage='ward'
# Use 'euclidean' for this example or 'precomputed' for custom matrices
cluster_model = AgglomerativeClustering(
    n_clusters=3,             # The desired number of clusters
    affinity='euclidean',     # The distance metric
    linkage='ward'            # The linkage criterion
)

# Fit the model and get cluster labels
labels = cluster_model.fit_predict(X)
```

4. **Visualize as a Dendrogram (SciPy):** Scikit-learn does not have a built-in function to plot dendrograms, so we use SciPy for visualization, which shows the full hierarchy (slide 24).

```
# Generate the linkage matrix
linked_matrix = linkage(X, method='ward', metric='euclidean')

# Plot the dendrogram
plt.figure(figsize=(10, 7))
dendrogram(linked_matrix)
plt.title('Hierarchical Clustering Dendrogram')
plt.xlabel('Data Points')
plt.ylabel('Distance')
plt.show()
```

4 Broader Applications of Clustering

Clustering algorithms are general-purpose tools that have catalyzed advancements across numerous fields beyond biology and business.

Cybersecurity: In network anomaly detection, normal network traffic patterns form dense clusters. Any traffic that falls outside these clusters (outliers) can be flagged as a potential intrusion or attack, allowing for real-time threat identification.

Urban Planning: Municipalities can cluster neighborhoods based on census data, utility usage, and traffic flow. This helps in identifying areas with similar needs, optimizing public transportation routes, and planning for new infrastructure like parks, schools, and emergency services.

Climate Science: Climatologists cluster vast amounts of satellite imagery and meteorological data to identify regions with similar climate patterns. This is essential for modeling climate change, predicting extreme weather events like hurricanes, and understanding global weather systems like El Niño.

Astronomy: Researchers apply clustering to astronomical survey data to automatically group celestial objects. Galaxies can be clustered based on their morphology, stars can be grouped into stellar populations or associations, and this helps to test theories about cosmic structure formation and stellar evolution.

Robotics and Computer Vision: Clustering is fundamental to image segmentation, where pixels are grouped by color, texture, or intensity. This allows a robot's vision system to distinguish different objects in its environment, enabling it to navigate and interact with the physical world.