Contents

1. Concept of Clustering Analysis
   - Main idea
   - Difference between Clustering and Classification

2. Distance Measures
   - Metric
   - Distance matrix

3,4. Hierarchical / K-means clustering
   - Main idea
   - Clustering Process

5,6. Quality measure / Others
   - Quality measure
   - Visualization approaches
Clustering Analysis (Today’s topic)

Check list

- What is the difference between Clustering and Classification?
- What are the statistical methods for defining “Different” and the analytical method for defining “Equivalent”.
- How to calculate ‘Similarity’
- What is the distance matrix?
- What is Hierarchical Clustering and K-means clustering method?
- What are the advantages and disadvantages of each clustering method?
- How do we evaluate clustering results?
- How do we apply the dimension reduction algorithm to clustering?
Concept of Clustering Analysis

Definition
What is ‘Clustering’
Motivation of Clustering Analysis

- Clustering analysis is the process of grouping a set of objects (Experimental subjects or their features).

- The commonest form of unsupervised learning.

- Definition of unsupervised learning is learning from raw data, as opposed to supervised data where a classification of examples is given.
What is ‘Clustering’
Motivation of Clustering Analysis

- How many clusters you can find in this data?
- For humans, this can be intuitively distinguished.
- Let’s formulate the process of finding!
What is ‘Clustering’
Motivation of Clustering Analysis

How many clusters?

➢ For clustering analysis without class information, there can be numerous answers.
‘Clustering’ vs ‘Classification’
Difference between ‘Clustering’ and ‘Classification’

- How many clusters based on the numeric variables? (Clustering analysis)
- When there is a characteristic value dependent on a numeric variable, you want to predict it based on the numeric variables (Classification analysis)
- What is ‘Class’.
‘Clustering’ vs ‘Classification’
Difference between ‘Clustering’ and ‘Classification’

➢ Clustering
  • Unsupervised learning
  • Classes are unknown a priori
  • Class discovery
    • Identification of new/unknown classes
    • Exploratory analysis

➢ Classification
  • Supervised learning
  • Classes are predefined
  • Class prediction
    • Find a classifier from a training data
    • Assignment of new cases into known classes
Clustering analysis
Definition of Clustering Analysis

➢ Grouping objects into subsets ("clusters")
  • Based on similarity between objects
  • High similarity within clusters
  • Low similarity between clusters
  • Find and visualize structures

➢ Used in the Quality control process
  • Identify outliers
  • Identify global pattern of missing values
Concept of "Similarity"
Similarity Measures
How do we define ‘Similarity’?
Distance Measures

- Recall that the goal is to group together “similar” data – but what does this mean?

- No single answer – it depends on what we want to find or emphasize in the data; this is one reason why clustering is an “art”

- The similarity measure is often more important than the clustering algorithm used – don’t overlook this choice!
Similarity and Dissimilarity
Fundamentals of all data science methods

➢ Similarity index
  • Numeric value that indicates how similar different objects are
  • In general, the higher the similarity of two objects, the higher the similarity.

➢ Dissimilarity index
  • Numeric value that indicates how different different objects are
  • In general, the higher the similarity between objects, the lower dissimilarity.
Diverse Distance Measures

Distance measures

- **Euclidean Distance**

\[ d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \]

- **Manhattan Distance**

\[ d_1(p, q) = \|p - q\|_1 = \sum_{i=1}^{n} |p_i - q_i|, \]
Diverse Distance Measures

Distance measures

- Pearson’s Correlation

\[
    r_{xy} = \frac{Cov(x, y)}{\sigma_x \sigma_y} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}
\]

- Spearman’s Correlation

\[
    r_s = \rho_{rg_X, rg_Y} = \frac{\text{cov}(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}}
\]
Example of Distance Matrix

Distance Matrix

<table>
<thead>
<tr>
<th>Individual</th>
<th>Var1</th>
<th>Var2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ind1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Ind2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Ind3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Ind4</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

Given Data

Using Euclidean Distance

<table>
<thead>
<tr>
<th></th>
<th>Ind1</th>
<th>Ind2</th>
<th>Ind3</th>
<th>Ind4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ind1</td>
<td>0</td>
<td>2.828</td>
<td>3.162</td>
<td>5.099</td>
</tr>
<tr>
<td>Ind2</td>
<td>2.828</td>
<td>0</td>
<td>1.414</td>
<td>3.162</td>
</tr>
<tr>
<td>Ind3</td>
<td>3.162</td>
<td>1.414</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Ind4</td>
<td>5.099</td>
<td>3.162</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
Clustering Algorithms

Diverse approaches

- **Bottom-up agglomerative methods**
  - Hierarchical Clustering (HCL) (Eisen et al. 1998)

- **Top-down partitioning methods**
  - K-means (Herwig et al. 1999)
  - Self-Organizing Map (SOM) (Tamayo et al. 1999)

- **Projection methods**
  - Principal component analysis (PCA) (Raychaudhuri et al., 2000)
  - Multi-dimensional scaling (MDS)
Hierarchical Clustering
Dendrogram approach
Hierarchical Clustering (HC)

- HC produces a binary tree or dendrogram
- The final cluster is the root and each data item is a leaf
- The height of the bars indicate how close the items are
Hierarchical Clustering (HC)

Bottom-up approach

- Bottom-up agglomerative method
- Joining similar genes into groups

Steps
  - Let q objects as q clusters (#clusters = q)
  - Join two clusters which have smallest distance (#clusters = q – 1)
  - Iterate this process with resulting clusters until all clusters are connected in a hierarchical tree (Dendrogram)
**Hierarchical Clustering (HC)**

**Step-by-Step Clustering**

- **g1** is most like **g8**

- **g4** is most like \{**g1**, **g8**\}

![Hierarchical Clustering Diagram](image-url)
Hierarchical Clustering (HC)

Step-by-Step Clustering

- g5 is most like g7
- \{g5, g7\} is most like \{g1, g4, g8\}
Hierarchical Clustering (HC)

Step-by-Step Clustering
Hierarchical Clustering (HC)

Linkage Method

- We already know about distance measures between data items, but what about between a data item and a cluster or between two clusters?

- We just treat a data point as a cluster with a single item, so our only problem is to define a linkage method between clusters.

- As usual, there are lots of choices…
Hierarchical Clustering (HC) 
Linkage Method

- The **single linkage** method is based on minimum distance, or the nearest neighbor rule.

- The **complete linkage** method is based on the maximum distance or the furthest neighbor approach.

- The **average linkage** method the distance between two clusters is defined as the average of the distances between all pairs of objects
Hierarchical Clustering (HC)

Linkage Method

- Distance or similarity between clusters \((G, H)\)

- Single linkage
  
  \[
  d_{S}(G, H) = \min_{i \in G, j \in H} d(z_i, z_j)
  \]

- Complete linkage
  
  \[
  d_{C}(G, H) = \max_{i \in G, j \in H} d(z_i, z_j)
  \]

- Average linkage
  
  \[
  d_{A}(G, H) = \frac{1}{n_G n_H} \sum_{i \in G, j \in H} d(z_i, z_j)
  \]
Hierarchical Clustering (HC)

Linkage Method

- Distance or similarity between clusters \((G, H)\)

- Single linkage
  \[
  d_S(G, H) = \min_{i \in G, j \in H} d(z_i, z_j)
  \]

- Complete linkage
  \[
  d_C(G, H) = \max_{i \in G, j \in H} d(z_i, z_j)
  \]

- Average linkage
  \[
  d_A(G, H) = \frac{1}{n_G n_H} \sum_{i \in G, j \in H} d(z_i, z_j)
  \]
Hierarchical Clustering (HC)  
Linkage Method

- Distance or similarity between clusters \((G, H)\)

- Single linkage
  \[ d_S(G, H) = \min_{i \in G, j \in H} d(z_i, z_j) \]

- Complete linkage
  \[ d_C(G, H) = \max_{i \in G, j \in H} d(z_i, z_j) \]

- Average linkage
  \[ d_A(G, H) = \frac{1}{n_G n_H} \sum_{i \in G, j \in H} d(z_i, z_j) \]
Hierarchical Clustering (HC)

Linkage Method

- Distance or similarity between clusters \((G, H)\)

- **Single linkage**
  \[
  d_S(G, H) = \min_{i \in G, j \in H} d(z_i, z_j)
  \]

- **Complete linkage**
  \[
  d_C(G, H) = \max_{i \in G, j \in H} d(z_i, z_j)
  \]

- **Average linkage**
  \[
  d_A(G, H) = \frac{1}{n_G n_H} \sum_{i \in G, j \in H} d(z_i, z_j)
  \]
Hierarchical Clustering (HC)

Linkage Method

- **Single Linkage**
  - Minimum Distance
  - Cluster 1
  - Cluster 2

- **Complete Linkage**
  - Maximum Distance
  - Cluster 1
  - Cluster 2

- **Average Linkage**
  - Average Distance
  - Cluster 1
  - Cluster 2
How to determine Clusters
From Dendrogram

- Cut and assign as many clusters as you want ($k$ clusters).
K-means Clustering

Widely used approach
Partitioning Algorithms
Representative methods for clustering analysis

- **Top-down divisive methods**
  - K-means, K-medians, SOM
  - Combinatorial optimization problem

- **Partition objects into K groups**
  - Pre-specified tuning parameter: $K = \#\text{clusters}$
  - Minimize a dispersion measure within clusters
  - Iteratively reallocate objects to clusters until some criterion is met

- **Weakness**
  - Global optimum will not be found
  - No information on similarity within cluster
K-means Clustering

Motivation

- How many clusters do you think there are in this data? How might it have been generated?
K-means Clustering

Main idea

- **Choose** a number of clusters \( k \)

- **Initialize** cluster centers \( \mu_1, \ldots, \mu_k \)
  - Could pick \( k \) data points and set cluster centers to these points
  - Or could *randomly assign* points to clusters and take means of clusters

- For each data point, compute the cluster center it is closest to (using some distance measure) and assign the data point to this cluster

- Re-compute cluster centers (mean of data points in cluster)

- Stop when there are no new re-assignments
K-means Clustering

Formula

Let \( m_k(t) = \text{centroid of k-cluster at step t} \)
Let \( C_i(t) = \text{cluster where i^{th} object belongs at step t} \)

- Specify \#clusters (K)
- Start with random initial centroids of clusters
  \[ \{m_1(0), m_2(0), \ldots, m_K(0)\} \]
- Assign objects to nearest clusters
  \[ C_i(0) = \arg\min_{1\leq k \leq K} d(z_i, m_k(0)) \quad i = 1, 2, \ldots, q \]
K-means Clustering

Formula (Cont.)

- Update centroids based on clusters at (t-1)th step
  \[ \{m_1(t), m_2(t), \ldots, m_K(t)\} \]

- Reassign objects to nearest clusters
  \[ C_i(t) = \arg\min_{1 \leq k \leq K} d(z_i, m_k(t)) \quad i = 1, 2, \ldots, q \]

- Repeat until Tth step or until total intra-cluster variance does not change
  \[ V(t) = \sum_{k=1}^{K} \sum_{i:C_i(t)=k} (z_i - m_k(t))^2 \]
K-means Clustering
Visualization of Process for K-means Clustering

- Learn about K-means Clustering with $(k=3)$.
K-means Clustering
Visualization of Process for K-means Clustering

- Minimize the sum of distances in each cluster
- Representative example of EM (Expectation-Maximize) algorithm
K-means Clustering

Issues

➢ Random initialization means that you may get different clusters each time

➢ Data points are assigned to only one cluster (hard assignment)

➢ Implicit assumptions about the “shapes” of clusters

➢ You have to pick the number of clusters
What is a Good Clustering

Measures for quality of the clustering
Cluster Assessment
What is Cluster Assessment

➢ How many clusters?
  • Tibshirani et al. (2002) Gap statistics
  • Kaufman and Rousseeuw (1990) Average silhouette width
  • Dudoit and Fridyand (2001) Clest

➢ How reliable are clustering results?
  • Kaufman and Rousseeuw (1990)
  • Rand (1971)
  • Fowlkes and Mallows (1983)
Cluster Assessment
What is Cluster Assessment

- **Internal criterion:** A good clustering will produce high quality clusters in which:

  - the intra-class (that is, intra-cluster) similarity is high
  - the inter-class similarity is low
  - The measured quality of a clustering depends on both the document representation and the similarity measure used
Cluster Assessment
Silhouette Width

- **Silhouette width** = a measure of cluster validity

- Objects with large silhouette width well clustered

- Objects with small silhouette width tend to lie between clusters

\[
s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad s(i) = 0, \text{ if } |C_i| = 1
\]

\[
a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)
\]

\[
b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)
\]

\[-1 \leq s(i) \leq 1\]

d(i,j) is the distance between data points i and j in the Cluster C_i
Cluster Assessment
Interpretation of Silhouette Width

$$-1 \leq s(i) \leq 1$$

- For Silhouette width to be close to 1, $a(i) \downarrow$ and $b(i) \uparrow$.
  - Thus an Silhouette width close to 1 means that the data is appropriately clustered.

- Based on the Silhouette width, we can determine optimal number of clusters, $k$.

- It's also possible to compare different clustering methods.
Cluster Assessment
Semi-supervised method

- Quality measured by its ability to discover some or all of the hidden patterns or latent classes in gold standard data

- Assesses a clustering with respect to **ground truth** ... requires **labeled data (Class information)**.

- Assume documents with $C$ gold standard classes, while our clustering algorithms produce $K$ clusters, $\omega_1, \omega_2, \ldots, \omega_K$ with $n_i$ members.
Cluster Assessment
Purity index

- Simple measure: purity, the ratio between the dominant class in the cluster $\pi_i$ and the size of cluster $\omega_i$

$$Purity(\omega_i) = \frac{1}{n_i} \max_j (n_{ij}) \quad j \in C$$

- Biased because having $n$ clusters maximizes purity

- Others are entropy of classes in clusters (or mutual information between classes and clusters)
Cluster Assessment
Example of Purity index

- Cluster I: Purity = \( \frac{1}{6} \times \max(5, 1, 0) \) = \( \frac{5}{6} \)
- Cluster II: Purity = \( \frac{1}{6} \times \max(1, 4, 1) \) = \( \frac{4}{6} \)
- Cluster III: Purity = \( \frac{1}{5} \times \max(2, 0, 3) \) = \( \frac{3}{5} \)
Cluster Assessment

Rand Index

<table>
<thead>
<tr>
<th>Number of points</th>
<th>Same Cluster in clustering</th>
<th>Different Clusters in clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same class in ground truth</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>Different classes in ground truth</td>
<td>20</td>
<td>72</td>
</tr>
</tbody>
</table>

- Accuracy = Rand Index =
  \[
  \frac{20 + 72}{(20 + 24 + 20 + 72)} = 0.6764 = 67.64\%
  \]
Cluster Assessment
Determining the “correct” number of clusters

- We’d like to have a measure of cluster quality $Q$ and then try different values of $k$ until we get an optimal value for $Q$.

- But, since clustering is an unsupervised learning method, we can’t really expect to find a “correct” measure $Q$.

- So, once again there are different choices of $Q$ and our decision will depend on what dissimilarity measure we’re using and what types of clusters we want.
Projection Algorithms
Principal Component Analysis
Projection Algorithms
Projection algorithm can be used for clustering analysis

- Principal Component Analysis (PCA)
  - Raychaudhuri et al. (2001)

- Gene-Shaving (GS)
  - Tibshirani et al. (2002)

- Multi Dimensional Scaling (MDS)

- Independent Component Analysis (ICA)
  - Liebermeister (2002)
Principal Component Analysis (PCA)
Representative algorithm for dimensionality reduction

- Reduce dimension of variables while retaining as much of variation in data as possible

- Principal components (PCs)
  - New variables which are linear combinations of original variables
  - $(i+1)^{st}$ PC = direction with maximal variance among those orthogonal to first $i$ PCs
Principal Component Analysis (PCA)
Representative algorithm for dimensionality reduction

“Cloud” of data points in multi-dimensional space

Data points resolved along 3 principal component axes.
Comparison among Algorithms
Pros and Cons of each clustering methods

- **Hierarchical clustering**
  - Fast computation (for agglomerative clustering)
  - Rigid (unable to correct previous bad decisions)

- **Partitioning methods (e.g. K-means)**
  - Provide clusters that satisfy approximately an optimality criterion
  - Difficult to specify tuning parameters
  - Long computing time
  - No unique results

- **Projection methods (e.g. PCA, MDS)**
  - Dimension reduction
  - Difficult to interpret artificially generated vector
What is Projection methods?

Next topic

- **Hierarchical clustering**
  - Fast computation (for agglomerative clustering)
  - Rigid (unable to correct previous bad decisions)

- **Partitioning methods (e.g. K-means)**
  - Provide clusters that satisfy approximately an optimality criterion
  - Difficult to specify tuning parameters
  - Long computing time
  - No unique results

- **Projection methods (e.g. PCA, MDS)**
  - Dimension reduction
  - Difficult to interpret artificially generated vector