

Soft computing



K.N. Toosi  
University of  
Technology

محاسبات نرم

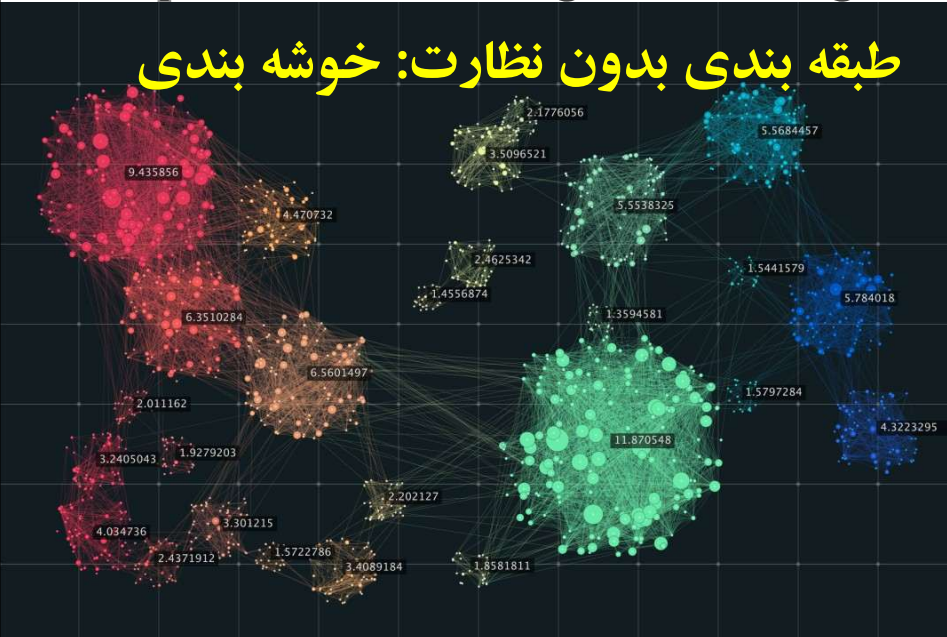


Hasan Ghasemzadeh  
<http://wp.kntu.ac.ir/ghasemzadeh>

Soft Computing

Unsupervised Learning: Clustering

طبقه بندی بدون نظارت: خوشه بندی



9.435856  
4.470732  
6.3510284  
2.011162  
3.2405043  
1.9279203  
4.034736  
2.4371912  
3.301215  
1.5722786  
3.4089184  
2.202127  
1.8581811  
2.1776056  
3.5096521  
2.4625342  
1.4556874  
5.5538325  
1.3594581  
11.870548  
1.5797284  
5.5684457  
1.5441579  
5.784018  
4.3223295

## یادگیری بدون نظارت

**Unsupervised learning:** The data have no target attribute.

We want to explore the data to find some intrinsic structures in them.

- ✓ در این قسمت داده‌ها برچسب ندارند.
- ✓ هزینه برچسب گذاری بسیار زیاد است.

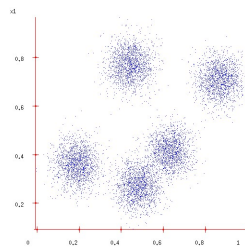
- ✓ خوشه‌بندی (clustering) K-means
- خوشه بندی طیفی
- خوشه‌بندی سلسله مراتبی
- DBSCAN
- کاهش ابعاد ✓
- تجزیه اصلی مؤلفه‌ها (PCA)
- t-SNE
- Autoencoders
- ✓ تشخیص ناهنجاری (Anomaly detection)
- ✓ مدل‌های مولد (Generative modelling)

Soft Computing

3

## خوشه بندی

- ✓ روشی که شباهت‌های بین گروهی از داده‌ها را در کل داده‌ها پیدا می‌کند خوشه بندی نام دارد.
- ✓ خوشه‌بندی یکی از روشهای بدون نظارت است.
- ✓ بدلیل تاریخچه آن معمولا خوشه بندی مترادف با یادگیری بدون نظارت استفاده می‌شود.
- ✓ روش خوشه بندی بیشترین استفاده را در داده کاوی دارد.



مثال

- افراد با سایز لباس مختلف
- ساختمانهایی که در هنگام زلزله خرابی مشابه دارند
- خوشه‌بندی بندی متن‌های مشابه

الگوریتم‌های معروف:

- خوشه بندی کی میانگین (K-means)
- خوشه بندی طیفی (Spectral clustering)
- خوشه‌بندی سلسله مراتبی (Hierarchical clustering)
- خوشه بندی بر مبنای دانسیته (DBSCAN: Density-based spatial clustering of applications with noise)

Soft Computing

4

## خوشه بندی

**Clustering:** one of the most utilized data mining techniques

- **A clustering algorithm**
  - Partitional clustering
  - Hierarchical clustering
- **A distance (similarity, or dissimilarity) function**
- **Clustering quality**
  - Inter-clusters distance  $\Rightarrow$  maximized
  - Intra-clusters distance  $\Rightarrow$  minimized

کیفیت خوشه‌بندی به الگوریتم، تابع فاصله و نوع مساله بستگی دارد

Soft Computing

5

## K-means clustering

- K-means is a **partitional clustering** algorithm
- Set of data points  $D$ 

$$D = \{x_1, x_2, \dots, x_n\} \quad x_i = (x_{i1}, x_{i2}, \dots, x_{ir})$$

$n$  the number of data  
 $r$  the number of attributes (dimensions) in the data.
- The  $k$ -means partitions the data into  $k$  clusters.
  - **Center** of each cluster, called **centroid**.
  - $k$  is specified by the user ( $k < n$ )

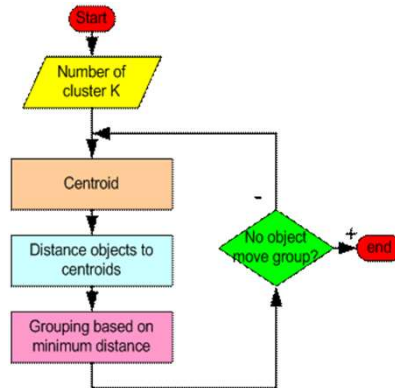
Soft Computing

6

## K-means clustering

- the *k-means* algorithm:

- 1) Randomly choose  $k$  data points (**seeds**) to be the initial **centroids**, cluster centers
- 2) Assign each data point to the closest **centroid**
- 3) Re-compute the **centroids** using the current cluster memberships.
- 4) If a convergence criterion is not met, go to 2).



Soft Computing

7

## Stopping/convergence criterion

1. Minimum or no re-assignments of data points to different clusters,
2. minimum or no change of centroids
3. minimum decrease in the **sum of squared error** (SSE),

$$SSE = \sum_{j=1}^k \sum_{i=1}^{\text{no of } x \in C_j} \text{dist}(x, m_j)^2$$

- $m_j$  is the centroid of cluster  $C_j$
- $\text{dist}(x, m_j)$  is the distance between data point  $x$  and centroid  $m_j$ .

Soft Computing

8

## مثال

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.datasets import make_blobs
from sklearn.preprocessing import StandardScaler

# Generate synthetic 2D data
np.random.seed(42)
X, _ = make_blobs(n_samples=150, centers=3, n_features=2,
                  cluster_std=2.5)

# Standardize the data
scaler = StandardScaler()
X = scaler.fit_transform(X)

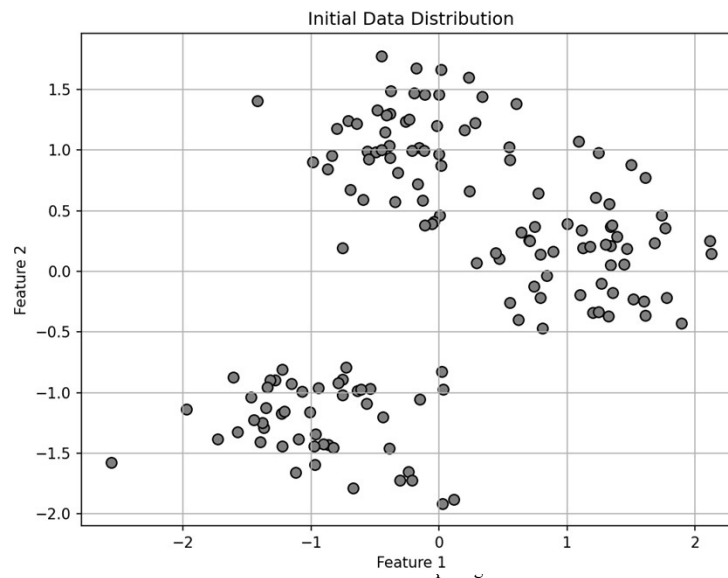
# Plot the initial data distribution
plt.figure(figsize=(8, 6))
plt.scatter(X[:, 0], X[:, 1], c='gray', edgecolor='k', s=50)
plt.title('Initial Data Distribution')
plt.xlabel('Feature 1')
plt.ylabel('Feature 2')
plt.grid(True)
plt.show()
```

12 clustering k means.py

Soft Computing

9

## مثال خوشه بندی لغات



10

## مثال

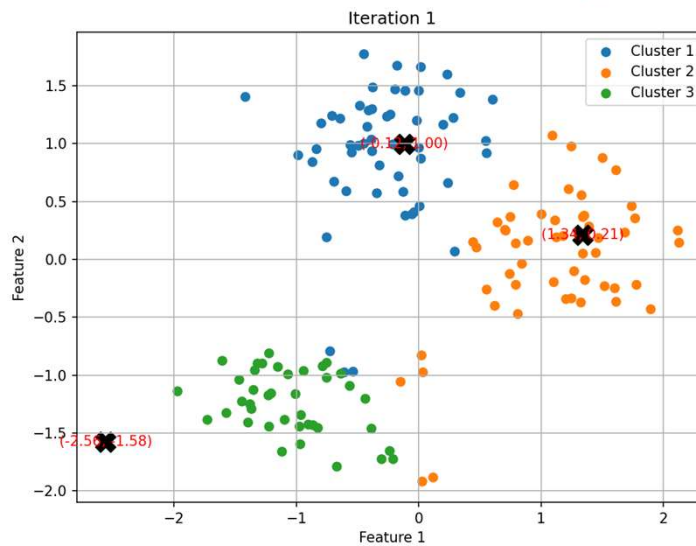
```
def kmeans_clustering(X, n_clusters, max_iters=10):
    # Randomly initialize centroids
    np.random.seed(42)
    centroids = X[np.random.choice(X.shape[0], n_clusters, replace=False)]

    for i in range(max_iters):
        # Assign clusters based on closest centroid
        distances = np.linalg.norm(X[:, np.newaxis] - centroids, axis=2)
        labels = np.argmin(distances, axis=1)
        # Plot current state
        plt.figure(figsize=(8, 6))
        for j in range(n_clusters):
            plt.scatter(X[labels == j][:, 0], X[labels == j][:, 1], label=f'Cluster {j+1}')
            plt.scatter(centroids[j, 0], centroids[j, 1], s=200, c='black', marker='x')
            # Add centroid coordinates as text
            plt.text(centroids[j, 0], centroids[j, 1], f'({centroids[j, 0]:.2f},
{centroids[j, 1]:.2f})',
                    fontsize=10, color='red', ha='center', va='center')
        plt.title(f'Iteration {i+1}')
        plt.xlabel('Feature 1')
        plt.ylabel('Feature 2')
        plt.legend()
        plt.grid(True)
        plt.show()
        # Update centroids
        new_centroids = np.array([X[labels == j].mean(axis=0) for j in range(n_clusters)])
        # Check for convergence
        if np.all(centroids == new_centroids):
            break
        centroids = new_centroids
    # Perform K-means clustering and visualize iterations
    kmeans_clustering(X, n_clusters=3, max_iters=10)
```

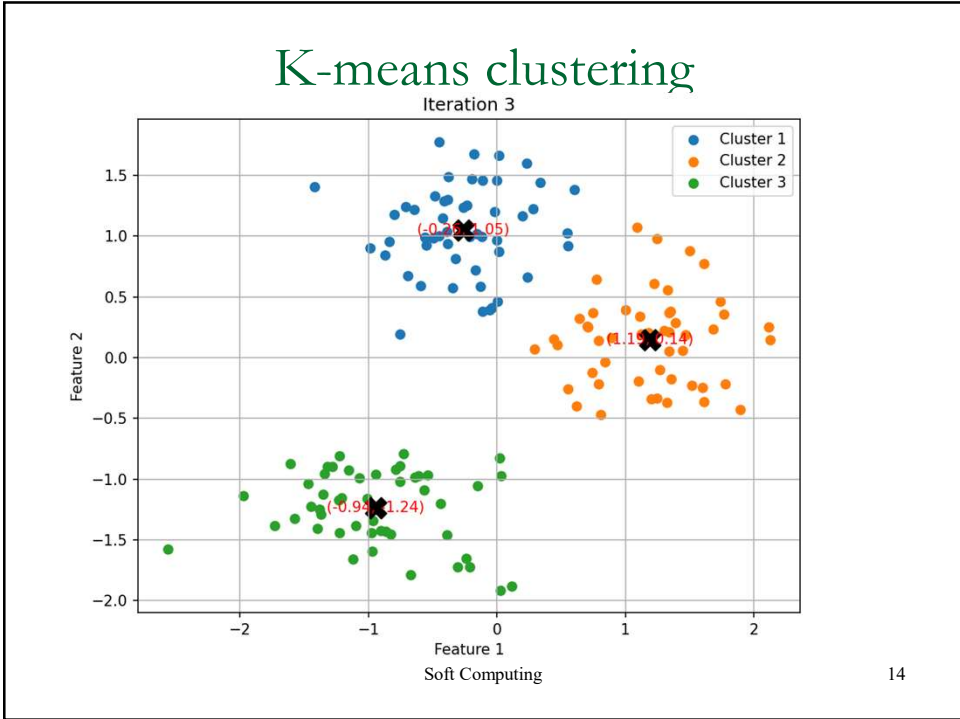
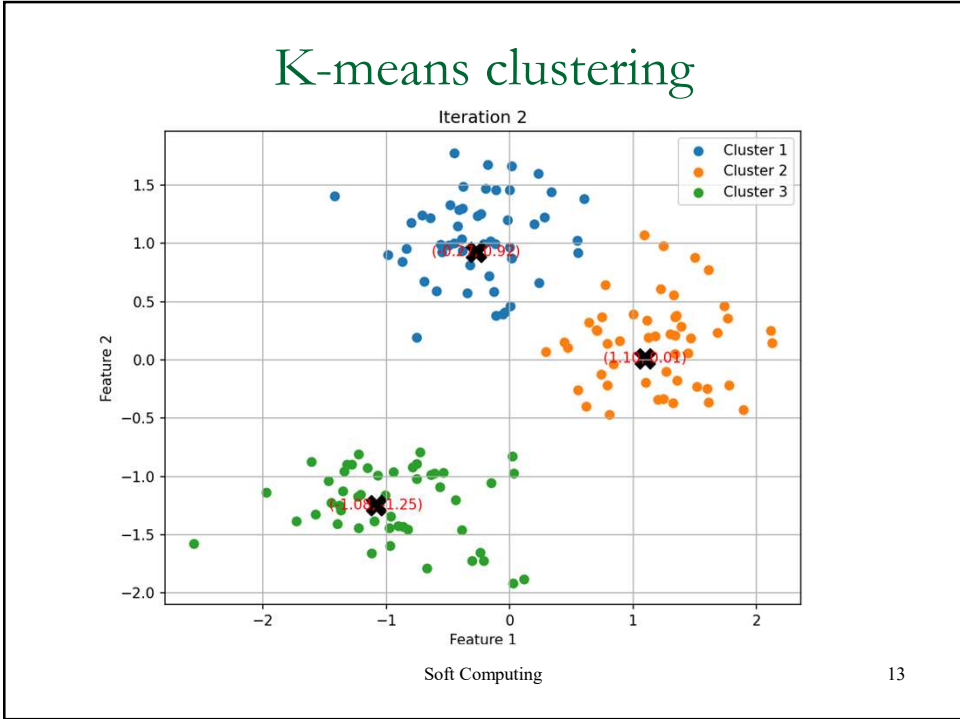
Soft Computing

11

## K-means clustering



12



## مثال خوشه بندی لغات

```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

# Load the uploaded file
file_path = "family_words.csv"
data = pd.read_csv(file_path)

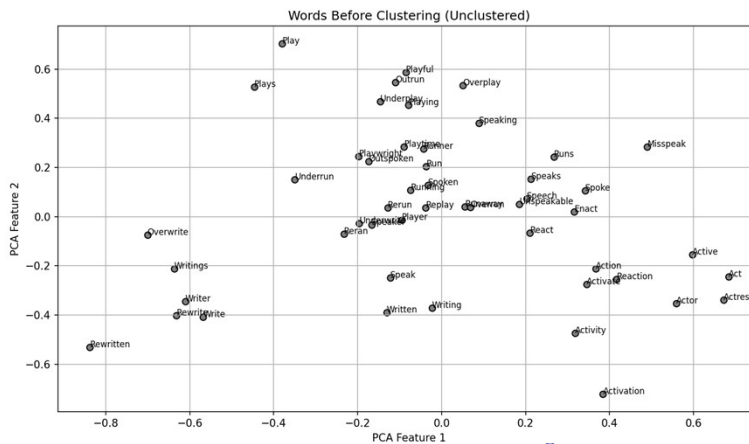
# Extract the words from the file (correcting the column
name)
words = data['Words_Family']
```

12 Hierarchical Clustering.py

Soft Computing

15

## ترسیم به روش کاهش بعد



برای دسته بندی لغات ابتدا آنها به بردار تبدیل شده اند ابعاد این بردارها ۱۹۴ بوده و برای رسم بردارهای نقاط داده از یکی از روشهای معروف یادگیری بدون نظارت - کاهش ابعاد استفاده شده است. نام الگوریتم استفاده شده تجزیه اصلی مؤلفه‌ها (PCA: Principal Component Analysis) است.

Soft Computing

16



## مثال خوشه بندی لغات

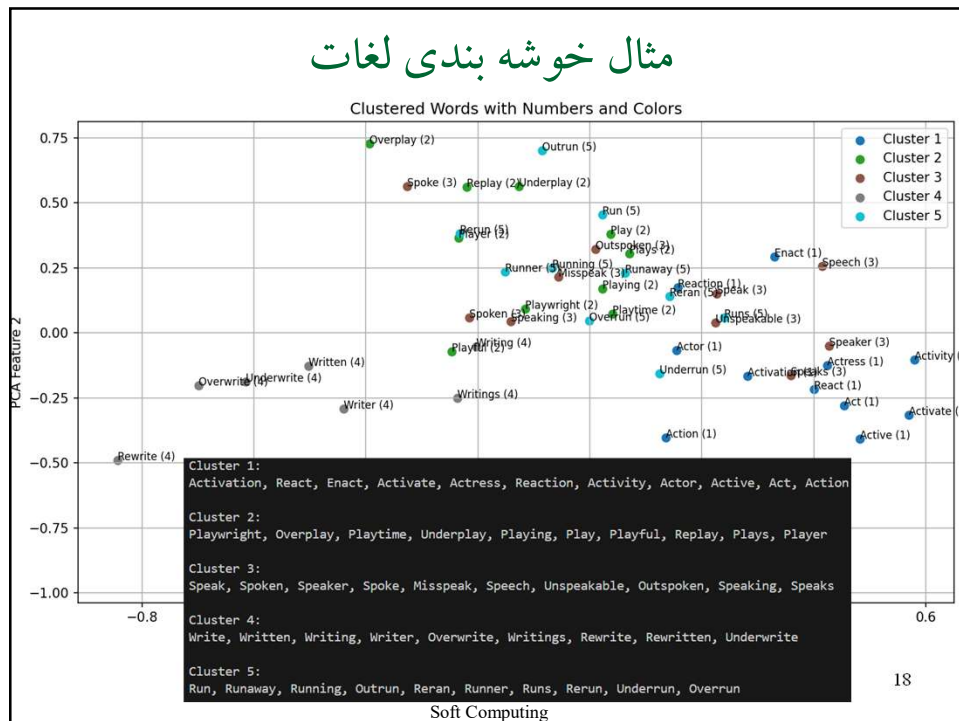
```
# Vectorize the words using TF-IDF
vectorizer = TfidfVectorizer(analyzer='char', ngram_range=(2, 3))
X = vectorizer.fit_transform(words)

# Apply K-means clustering
n_clusters = 5 # You can adjust the number of clusters
kmeans = KMeans(n_clusters=n_clusters, random_state=42)
kmeans.fit(X)
# Assign clusters to words
data['Cluster'] = kmeans.labels_
# Visualize the clustering results
for cluster in range(n_clusters):
    cluster_words = data[data['Cluster'] ==
cluster][ 'Words_Family' ].values
    print(f"Cluster {cluster + 1}:")
    print(", ".join(cluster_words))
    print()
# Save the clustered data for further use
clustered_file_path = "family_words_clustered.csv"
data.to_csv(clustered_file_path, index=False)
```

17

Soft Computing

## مثال خوشه بندی لغات



18

Soft Computing

## K-means clustering – Disk version

وقتی داده‌ها خیلی زیاد است نمی‌توان روی حافظه رایانه تمام داده‌ها را بارگذاری نمود در این حالت بایستی از ویرایش دیسک برای خوشه‌بندی استفاده نمود

در این روش بسته‌های کوچکی (mini-batches) از داده در هر مرحله انتخاب می‌شود و سپس خوشه‌بندی (با سرعت زیاد و اشغال حافظه کم) بر روی آنها انجام می‌شود. تعداد تکرارها محدود است و ممکن است کمی خطا ایجاد شود (جواب به نقطه بهینه محلی ختم شود و به نقطه بهینه کلی نرسیم)

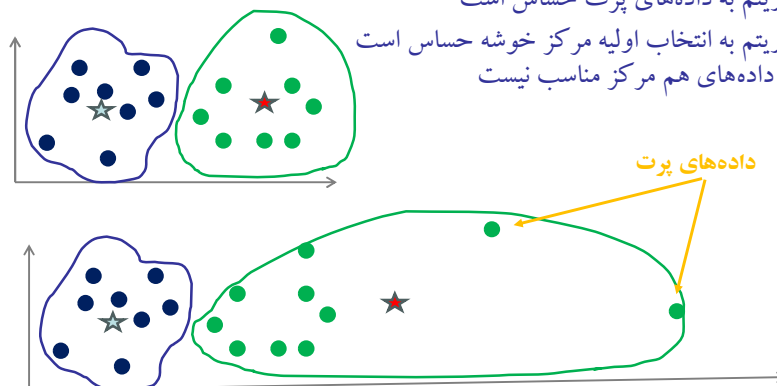
We need to limit the number of iterations (< 50 normally).  
There are other scale-up algorithms, e.g., BIRCH  
(Balanced Iterative Reducing and Clustering using Hierarchies)

Soft Computing

19

## K-means نقاط ضعف

- مقدار  $k$  توسط کاربر بایستی تعیین شود
- مرکز خوشه‌ها (سنتر وید) تحت تاثیر داده‌های با فراوانی زیاد قرار می‌گیرد
- الگوریتم به داده‌های پرت حساس است
- الگوریتم به انتخاب اولیه مرکز خوشه حساس است
- برای داده‌های هم مرکز مناسب نیست



Soft Computing

20

## نقاط ضعف K-means

داده‌های هم مرکز

یک روش استفاده از خوشه بندی طیفی است که فاصله بین نقاط را اندازه می گیرد.

Soft Computing 21

## Hierarchical Clustering

خوشه بندی سلسله مراتبی

یک توالی تو در تو از خوشه ها یا دیاگرام درختی (دندروگرام) تولید می کند

خوشه بندی سلسله مراتبی، سلسله ای از پارتیشن ها ارائه می کند

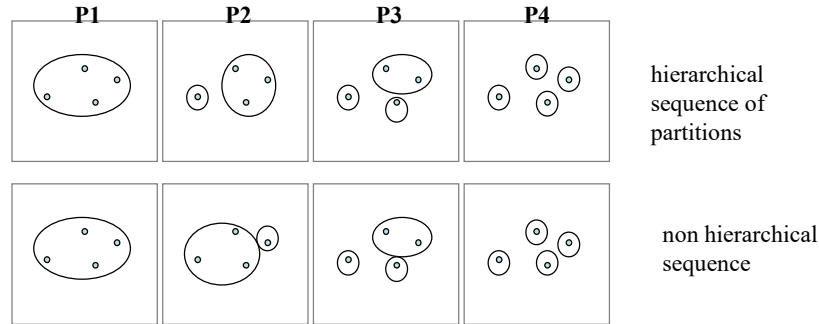
- a partition  $P_1$  into 1 clusters (the entire collection)
- a partition  $P_2$  into 2 clusters
- ...
- a partition  $P_n$  into n clusters (each object forms its own cluster)

It is then up to the user to decide which of the partitions reflects actual sub-populations in the data.

Soft Computing 22

## Hierarchical Clustering

در صورتی دنباله‌ای از پارتیشن‌بندی سلسله‌مراتبی است که هر خوشه در یک پارتیشن معین اجتماعی از خوشه‌ها در پارتیشن بزرگتر بعدی باشد



Soft Computing

23

## Hierarchical Clustering

### روشهای خوشه‌بندی سلسله‌مراتبی

#### Agglomerative methods:

روش تجمیعی (از پایین به بالا)

- Start with partition  $P_n$ , where each object forms its own cluster.
- Merge the two closest clusters, obtaining  $P_{n-1}$ .
- Repeat merge until only one cluster is left.

#### Divisive methods:

روش تقسیمی (از بالا به پایین)

- Start with  $P_1$ .
- Split the collection into two clusters that are as homogenous (and as different from each other) as possible.
- Apply splitting procedure recursively to the clusters.

در روش تجمیعی نیاز به قاعده‌ای برای تجمیع خوشه‌ها و در روش تقسیمی نیاز به قاعده‌ای برای تقسیم خوشه‌ها است که معمولاً با استفاده از مقدار فاصله انجام می‌شوند.

Soft Computing

24

## خوشه‌بندی سلسله مراتبی تجمیعی

### تعیین فاصله بین خوشه‌ها

$d(P,Q)$  فاصله بین خوشه‌های P و Q بر اساس فاصله مشاهدات در دو خوشه  $d(x,y)$  تعیین می‌شود

For  $x$  in P,  $y$  in Q

1.  $d_1(P,Q) = \min d(x,y)$  ( single linkage )
  2.  $d_2(P,Q) = \text{ave } d(x,y)$  ( average linkage )
  3.  $d_3(P,Q) = \max d(x,y)$  ( complete linkage )
  4.  $d_4(P,Q) = \|\bar{x}_P - \bar{x}_Q\|$  ( centroid method )
  5.  $d_5(P,Q) = 2 \frac{|P||Q|}{|P|+|Q|} \|\bar{x}_P - \bar{x}_Q\|^2$  ( Ward's method )
- $d_5$  is called Ward's distance.

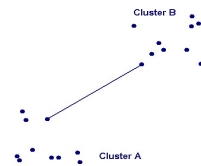
Soft Computing

25

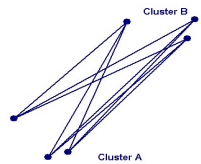
## خوشه‌بندی سلسله مراتبی تجمیعی

### تعیین فاصله بین خوشه‌ها

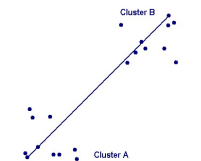
$$d_1(P,Q) = \min d(x,y) \quad (\text{single linkage})$$



$$d_2(P,Q) = \text{ave } d(x,y) \quad (\text{average linkage})$$



$$d_3(P,Q) = \max d(x,y) \quad (\text{complete linkage})$$



Soft Computing

26

## خوشه‌بندی سلسله مراتبی تجمیعی

تعیین فاصله بین خوشه‌ها – فاصله Ward

Let  $\mathbf{P}_k = P_1, \dots, P_k$  be a partition of the observations into  $k$  groups.

- Measure goodness of a partition by the sum of squared distances of observations from their cluster means:

$$RSS(\mathbf{P}_k) = \sum_{i=1}^k \sum_{j \in P_i} \|x_j - \bar{x}_{P_i}\|^2$$

- Consider all possible  $(k-1)$ -partitions obtainable from  $\mathbf{P}_k$  by a merge
- Merging two clusters with smallest Ward's distance optimizes goodness of new partition.

Soft Computing

27

## خوشه‌بندی سلسله مراتبی تقسیمی

There are divisive versions of single linkage, average linkage, and Ward's method.

### Divisive version of single linkage:

- Compute minimal spanning tree (graph connecting all the objects with smallest total edge length).
- Break longest edge to obtain 2 subtrees, and a corresponding partition of the objects.
- Apply process recursively to the subtrees.

Agglomerative and divisive versions of single linkage give identical results (more later).

Soft Computing

28

## خوشه‌بندی سلسله مراتبی تقسیمی

### Divisive version of Ward's method.

Given cluster R.

Need to find split of R into 2 groups P,Q to minimize

$$RSS(P, Q) = \sum_{i \in P} \|\mathbf{x}_i - \bar{\mathbf{x}}_P\|^2 + \sum_{j \in Q} \|\mathbf{x}_j - \bar{\mathbf{x}}_Q\|^2$$

or, equivalently, to maximize Ward's distance between P and Q.

**Note:** No computationally feasible method to find optimal P, Q for large |R|. Have to use approximation.

Soft Computing

29

## خوشه‌بندی سلسله مراتبی تقسیمی

### Iterative algorithm to search for the optimal Ward's split

Project observations in R on largest principal component.

Split at median to obtain initial clusters P, Q.

Repeat {   Assign each observation to cluster with closest mean  
                   Re-compute cluster means  
 } Until convergence

**Note:**

- Each step reduces  $RSS(P, Q)$
- No guarantee to find optimal partition.

### Divisive version of average linkage

Algorithm Diana, Struyf, Hubert, and Rousseuw

Soft Computing

30

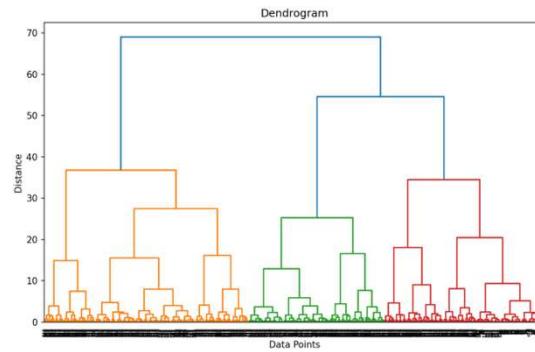
## خوشه‌بندی سلسله مراتبی

خوشه بندی سلسله مراتبی بوسیله دیاگرام درختی (دندروگرام) نمایش داده می‌شود

Result of hierarchical clustering can be represented as binary tree:

- Root of tree represents entire collection
- Terminal nodes represent observations
- Each interior node represents a cluster
- Each subtree represents a partition

y-coordinate of vertex =  
distance between daughter  
clusters.



Soft Computing

31

## بررسی خوشه‌بندی

### Silhouette Score

It measures how well each data point fits within its assigned cluster compared to other clusters.

For a single data point  $i$ , the silhouette score  $s(i)$  is calculated as:

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

$a(i)$ : The mean intra-cluster distance

$b(i)$ : The mean nearest-cluster distance

The score ranges from **-1 to 1**, where:

$s(i) = 1$  indicates that the data point is well-clustered and far from neighboring clusters.

$s(i) = 0$  indicates that the data point lies on the boundary between clusters.

$s(i) = -1$  indicates that the data point may be assigned to the wrong cluster.

Soft Computing

32



## خوشه‌بندی سلسله مراتبی

مثال

```

import matplotlib.pyplot as plt
from sklearn.datasets import make_blobs
from scipy.cluster.hierarchy import dendrogram, linkage, fcluster
from sklearn.metrics import silhouette_score
from sklearn.cluster import AgglomerativeClustering

# Generate artificial data using make_blobs
X, y = make_blobs(n_samples=500, centers=3, cluster_std=1.0, random_state=42)

# Perform hierarchical clustering
# 1. Compute linkage matrix
linkage_matrix = linkage(X, method='ward') # 'ward' linkage minimizes variance

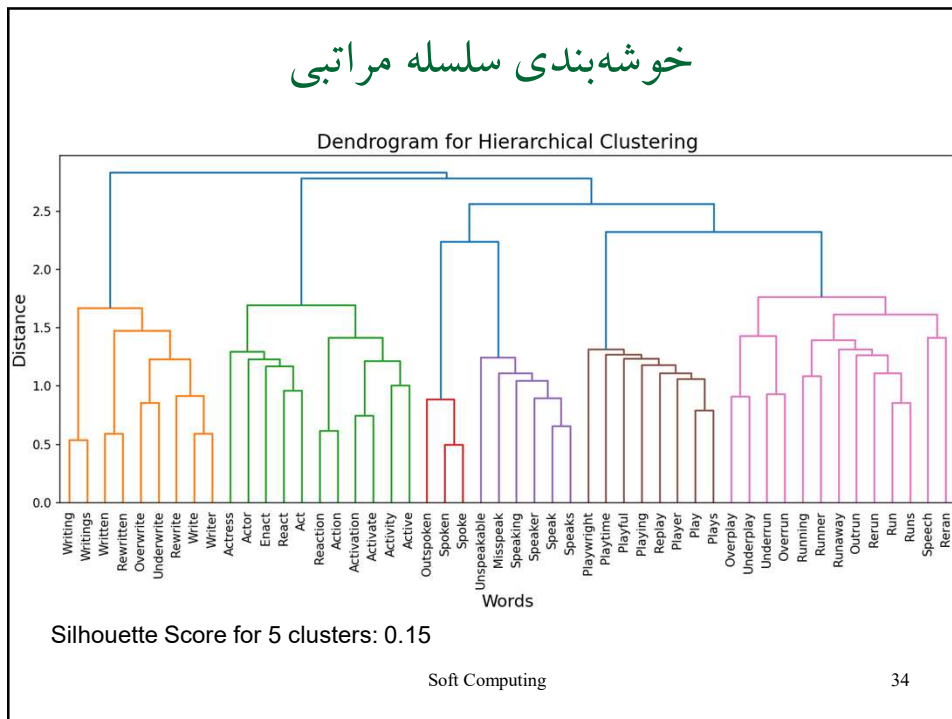
# 2. Plot the dendrogram
plt.figure(figsize=(12, 8))
dendrogram(linkage_matrix, truncate_mode='level', p=5) # Display only the top 5
levels
plt.title("Dendrogram")
plt.xlabel("Data Points")
plt.ylabel("Distance")
plt.show()

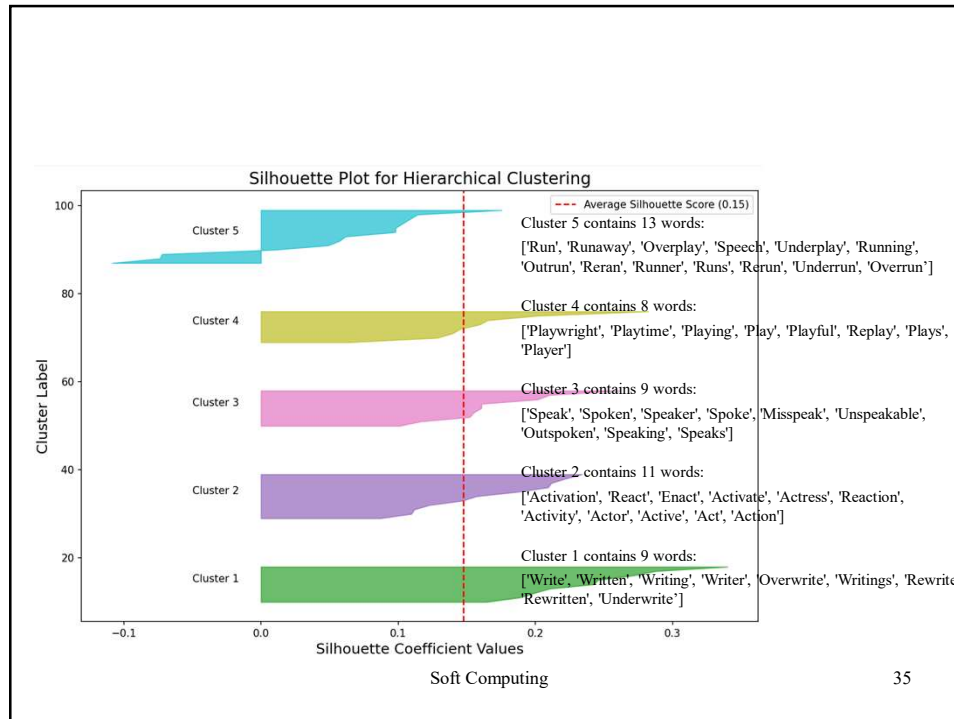
# Apply Agglomerative Clustering
n_clusters = 3 # Number of clusters
agg_clustering = AgglomerativeClustering(n_clusters=n_clusters, linkage='ward')
labels = agg_clustering.fit_predict(X)

```

12 Hierarchical Clustering.py Soft Computing 33

## خوشه‌بندی سلسله مراتبی





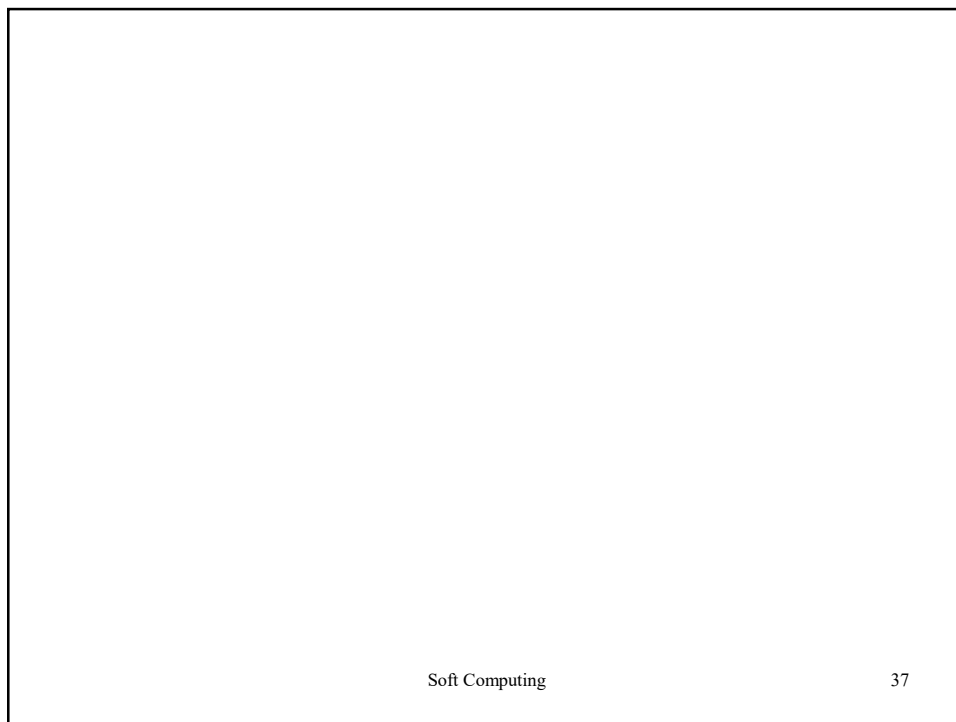
## تمرین برنامه نویسی

تمرین ششم: یک برنامه به زبان پایتون بنویسید که یک فایل داده را خوانده و خوشه‌بندی نماید.

۱- به روش K means

۲- با استفاده از خوشه‌بندی سلسله مراتبی

۳- بررسی صحت نتایج برای دو حالت فوق



Soft Computing

37