بنام خدا

جلسه یازدهم داده کاوی – مرور جلسه قبل در صفحه بعد

الگوریتم خوشه‌بندی DBScan، ازجمله الگوریتم‌های خوشه‌بندی معروف است. نام این الگوریتم که برگرفته از حروف اول عبارت Density Based Spatial Clustering of Applications with Noise می‌باشد، بر اساس چگالی نقاط موجود در اطراف هر نقطه پایه‌گذاری شده و خصوصاً برای یافتن خوشه‌ها در مجموعه داده‌هایی با خوشه‌هایی به اشکال دلخواه ارائه‌شده است. این الگوریتم نیازمند دو پارامتر Eps و MinPts است. قابل‌ذکر است که Eps یک عدد حقیقی مثبت که بیان‌کننده اندازه شعاع برای محاسبه چگالی نقاط موجود در اطراف هر نقطه به‌اندازه آن پارمتر می‌باشد. پارامتر MinPts نیز یک عدد طبیعی است که اصولاً بزرگ‌تر از ۱ انتخاب می‌شود. الگوریتم DBScan با بهره‌گیری از تعاریف و مفاهیم زیر که بر مبنای پارامترهای Eps و MinPts هستند، طراحی می‌شود.

- *همسایگی Eps یک نقطه*[1]: همسایگی Eps یک نقطه مانند p شامل مجموعه نقاطی مانند q از مجموعه داده‌ای است طوری که فاصله q از p بیشتر از Eps نباشد.
- *نقاط هسته*[2]: نقطه q، یک نقطه هسته نامیده می‌شود اگر تعداد نقاط موجود در محدوده به شعاع Eps آن با احتساب خود q کمتر از MinPts نباشد.
- مجموعه نقاط محدوده یک نقطه: نقطه p در مجموعه محدوده نقاط q یا به‌اختصار محدوده q است اگر فاصله آن تا q بیشتر از Eps نباشد.
- *قابل‌دسترس چگال به طرز مستقیم* یا اختصاراً *دسترس چگال مستقیم*[3]: نقطه p در دسترس چگال مستقیم q است اگر نقطه q هسته بوده و p نیز در محدوده q باشد.
- *دسترس چگال*[4]: نقطه p در دسترس چگال q است اگر زنجیره $p_1,p_2,\ldots p_n$ موجود باشد که $p_1=q$، $p_n=p$ و p_{i+1} در دسترس چگال مستقیم p_i باشد.
- *پیوستگی چگال*[5]: نقطه p با q پیوستگی چگال دارد اگر نقطه‌ای مانند o موجود باشد، طوری که p و q در دسترس چگال o باشند.
- خوشه: زیرمجموعه C از مجموعه داده‌ای D یک خوشه محسوب می‌شود اگر:
  ۱. به از هر دونقطه مانند p و q از مجموعه داده‌ای D، اگر p عضو C بوده و q در دسترس چگال p باشد در این صورت q نیز عضو C باشد.
  ۲. به از هر دونقطه مانند p و q از زیرمجموعه C، p باید با q پیوستگی چگال داشته باشد.

- نقطه *پرت*[6]نقطه‌ای است که نقطه هسته نبوده و متعلق به هیچ خوشه‌ای نباشند. ازآنجایی‌که در الگوریتم DBScan هر نقطه هسته، الزاماً تشکیل‌دهنده یک خوشه است می‌توان تعریف نقطه پرت را به این صورت بیان کرد که نقطه پرت هسته نیست و در دسترس چگال از یک نقطه هسته دیگر قرار ندارد. اگر نقطه‌ای، نقطه هسته نبوده اما در دسترس چگال از یک نقطه هسته باشد، نقطه مرزی[7] نامیده می‌شود.

حال با مفاهیم فوق می‌توان الگوریتم DBScan را بیان کرد. این الگوریتم هر نقطه را با توجه به محدوده آن بررسی می‌کند. هرکدام از نقطه‌ها در یکی از سه گروه هسته، مرزی یا پرت قرار می‌گیرند. در ادامه الگوریتم، هسته‌هایی که در دسترسی چگال هم هستند، ادغام‌شده و خوشه واحدی می‌شوند. با این اوصاف هر خوشه از حداقل یک هسته و تعدادی نقاط مرزی تشکیل می‌شود.

```
from sklearn.cluster import DBSCAN


...
X = iris_data[['sepal.length', 'sepal.width',
'petal.length', 'petal.width']]

dbscan = DBSCAN(eps=0.5, min_samples=5)
labels = dbscan.fit_predict(X)

noise = X[labels == -1]
```

خوشه‌بندی سلسله‌مراتبی یکی از روش‌های گروه‌بندی داده‌ها است که بر اساس ساختار دسته‌بندی تدریجی کار می‌کند. این روش داده‌ها را به صورت ساختار درختی مرتب می‌کند تا ارتباط بین گروه‌های مختلف را مشخص کند.

شروع با داده‌های مجزا :۱. بتدا هر داده به عنوان یک خوشه‌ی مستقل در نظر گرفته می‌شود .۲ .ادغام خوشه‌های مشابه :در هر مرحله، دو خوشه‌ای که بیشترین شباهت را دارند، با یکدیگر ترکیب می‌شوند .۳ .ساختار سلسله‌مراتبی : این روند ادامه می‌یابد تا . . . .

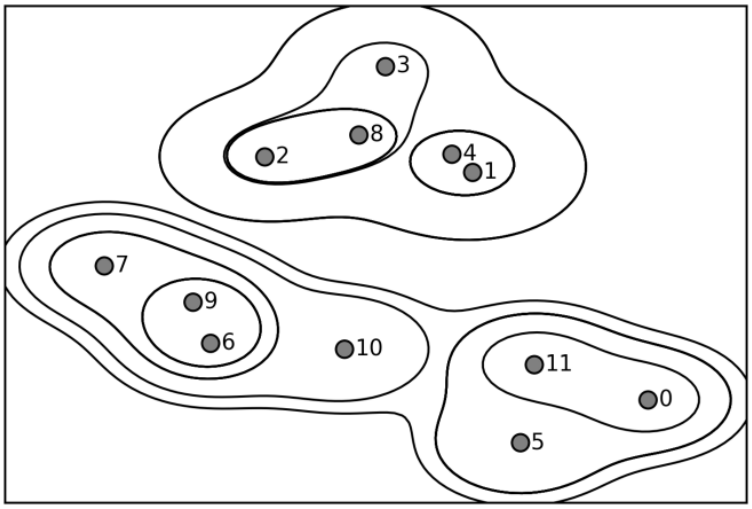
```
from sklearn.cluster import AgglomerativeClustering
clustering = AgglomerativeClustering(n_clusters=3)
labels = clustering.fit_predict(X)
```

[1] Eps-neighborhood of a point
[2] Core point
[3] Directly density-reachable
[4] Density-reachable
[5] Density-connected
[6] Noise
[7] Boarder point

تی اس ان ای (t-sne) و پی سی ای (PCA) هر دو از تکنیک‌های کاهش ابعاد هستند، اما با اهداف و روش‌های متفاوت:

- پی سی ای (PCA) یک روش خطی است که برای کاهش ابعاد داده‌های دارای ساختار خطی استفاده می‌شود. این روش داده‌ها را به محورهایی جدید تبدیل می‌کند که بیشترین واریانس را دارند، یعنی اطلاعات اصلی را حفظ می‌کند و نویز را کاهش می‌دهد.
- تی اس ان ای (t-SNE) یک روش غیرخطی است که عمدتاً برای تصویری‌سازی داده‌های پیچیده استفاده می‌شود. این روش سعی می‌کند نقاط داده‌ای مشابه را به یکدیگر نزدیک کند، در حالی که فاصله‌های بین داده‌های غیرمشابه را حفظ می‌کند، که باعث ایجاد ساختار خوشه‌ای واضح‌تر در داده‌های تصویری‌شده می‌شود.

به طور خلاصه، پی سی ای برای کاهش ابعاد کارآمد و سریع است، ولی تی اس ان ای بیشتر برای نمایش و درک خوشه‌ها در داده‌های پیچیده کاربرد دارد.



تمرین - برنامه زیر را اجرا کنید و نتایج را گزارش دهید

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

# لود آیریس
X = . . .

# Apply hierarchical clustering
linkage_matrix = linkage(X, method='ward')

# Plot dendrogram
plt.figure(figsize=(10, 5))
dendrogram(linkage_matrix)
plt.title("Dendrogram - Hierarchical Clustering")
plt.xlabel("Sample Index")
plt.ylabel("Distance")
plt.show()
```

```
# Perform clustering
clustering = AgglomerativeClustering(n_clusters=3)
labels = clustering.fit_predict(X)

# Plot clusters
plt.scatter(X[:, 0], X[:, 1], c=labels, cmap='viridis', s=50, edgecolors='k')
plt.title("Hierarchical Clustering Results")
plt.xlabel("Feature 1")
plt.ylabel("Feature 2")
plt.show()
```