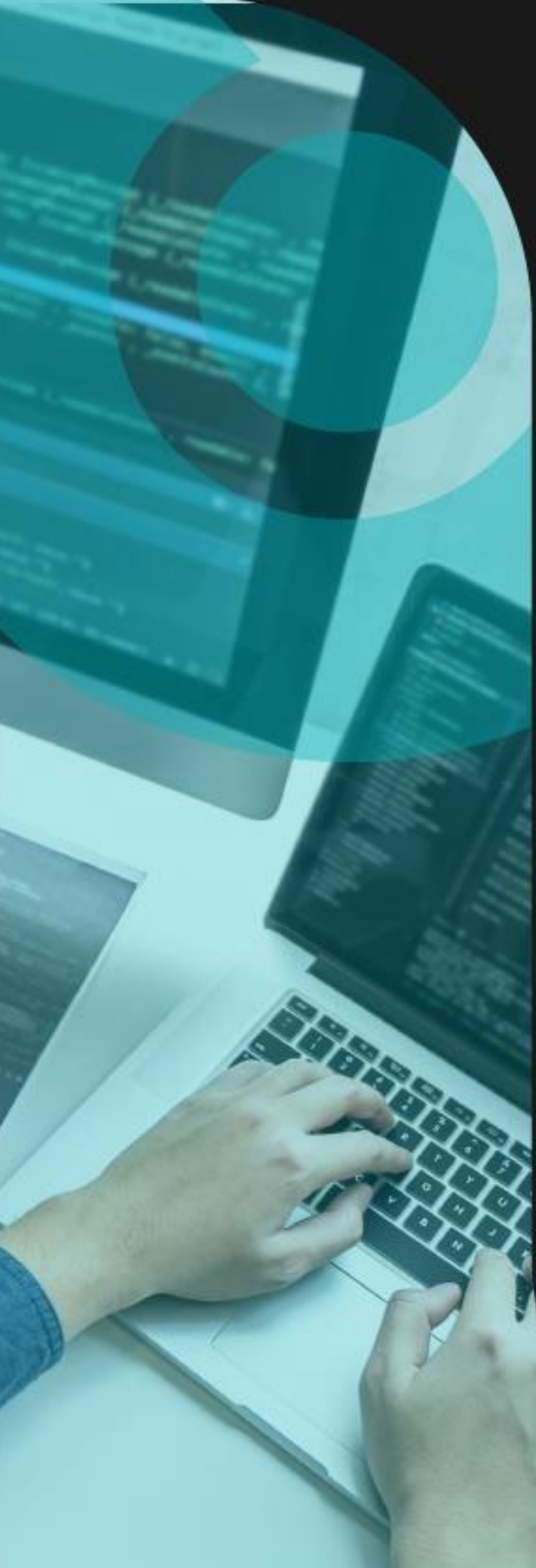


AWAKELAB

BASECAMP

Ciencia de Datos

Módulo: Aprendizaje de Máquina No Supervisado

Aprendizaje Esperado

---

3. Elaborar un modelo predictivo aplicando técnicas de Agrupamiento Jerárquico utilizando lenguaje Python para resolver un problema de clusterización

---

Agrupamiento Jerárquico

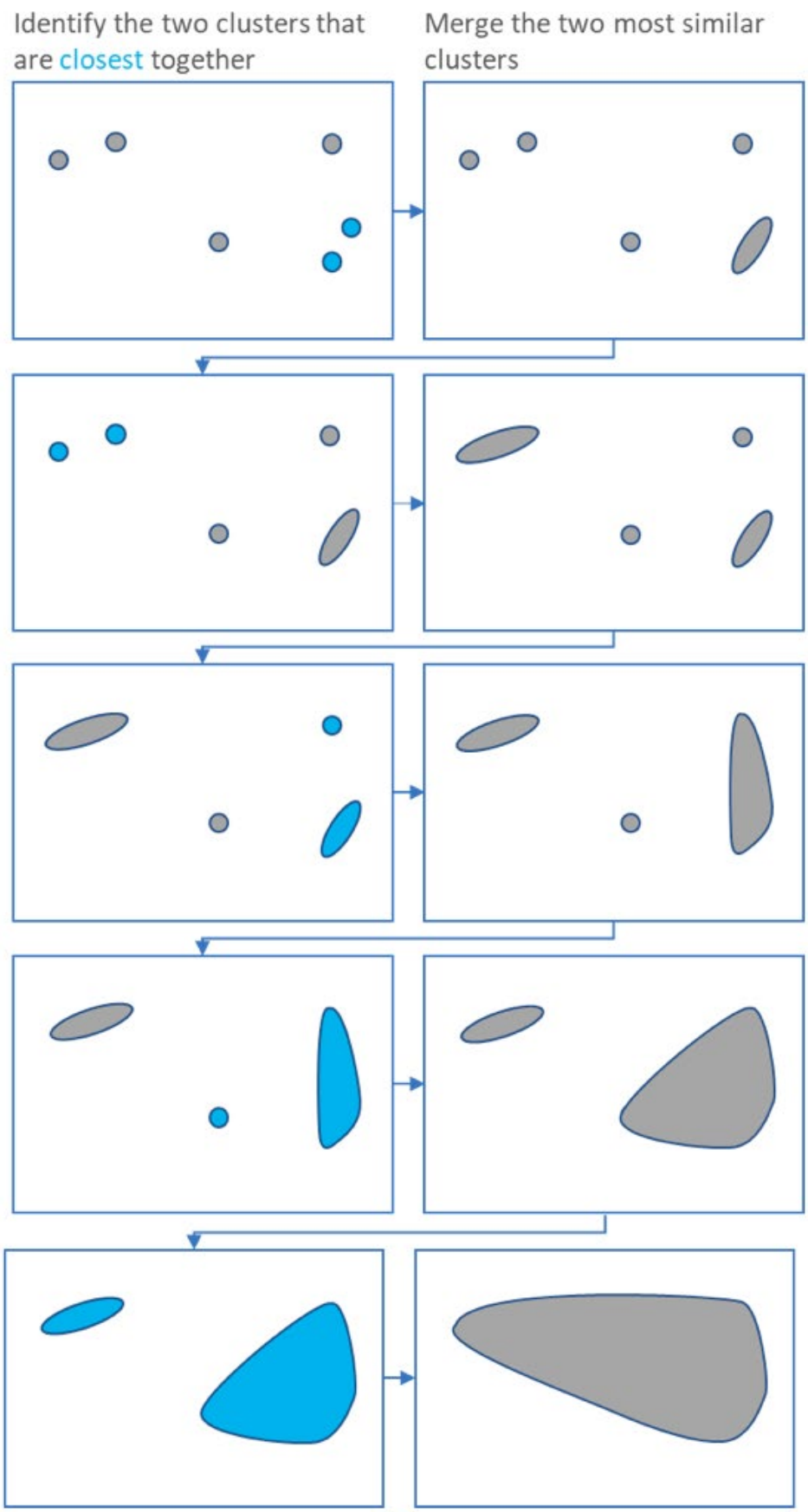
El agrupamiento jerárquico *(hierarchical clustering),* también conocido como *análisis de conglomerados jerárquicos,* es un algoritmo que agrupa objetos similares en grupos llamados *conglomerados*. El punto final es un conjunto de clústeres *,* donde cada clúster es distinto de los demás y los objetos dentro de cada clúster son muy similares entre sí.

Datos requeridos

La agrupación jerárquica se puede realizar con una *matriz de distancia* o *datos sin procesar.* Cuando se proporcionan datos sin procesar, el software calculará automáticamente una matriz de distancia en segundo plano. La siguiente matriz de distancia muestra la distancia entre seis objetos.

Cómo funciona el agrupamiento jerárquico

El agrupamiento jerárquico comienza tratando cada observación como un grupo separado. Luego, ejecuta repetidamente los siguientes dos pasos: (1) identifica los dos clústeres que están más cerca entre sí y (2) fusiona los dos clústeres más similares. Este proceso iterativo continúa hasta que todos los clústeres se fusionan. Esto se ilustra en los diagramas siguientes.


Identify the two clusters that are closest together
Merge the two most similar clusters

AWAKELAB

Diferencia entre K medias y agrupamiento jerárquico

K-means es un método de análisis de conglomerados que utiliza un número predeterminado de nodos. Requiere un conocimiento previo de 'K'.
El agrupamiento jerárquico, también conocido como análisis de conglomerados jerárquicos (HCA), también es un método de análisis de conglomerados que busca construir una jerarquía de conglomerados sin tener un número fijo de conglomerados.

Las principales diferencias entre K medias y el agrupamiento jerárquico son:

Agrupación de k-medias	Agrupación jerárquica
k-medias, utilizando un número preespecificado de grupos, el método asigna registros a cada grupo para encontrar el grupo mutuamente excluyente de forma esférica en función de la distancia.	Los métodos jerárquicos pueden ser divisivos o aglomerativos.
K Significa que el agrupamiento necesitaba un conocimiento avanzado de K, es decir, no. de	En el agrupamiento jerárquico, uno puede detenerse en cualquier número de grupos, uno lo

grupos uno quiere dividir sus datos.	encuentra apropiado interpretando el dendrograma.
Se puede usar la mediana o la media como centro de un conglomerado para representar cada conglomerado.	Los métodos aglomerativos comienzan con 'n' grupos y combinan secuencialmente grupos similares hasta que solo se obtiene un grupo.
Los métodos utilizados normalmente son menos intensivos en computación y son adecuados para conjuntos de datos muy grandes.	Los métodos divisivos funcionan en la dirección opuesta, comenzando con un grupo que incluye todos los registros y los métodos jerárquicos son especialmente útiles cuando el objetivo es organizar los grupos en una jerarquía natural.
En el agrupamiento K Means, dado que uno comienza con una elección aleatoria de grupos, los resultados producidos al ejecutar el algoritmo muchas veces pueden diferir.	En el agrupamiento jerárquico, los resultados son reproducibles en el agrupamiento jerárquico

K- significa agrupar simplemente una división del conjunto de objetos de datos en subconjuntos no superpuestos (grupos) de modo que cada objeto de datos esté exactamente en un subconjunto).	Una agrupación jerárquica es un conjunto de clústeres anidados que se organizan en forma de árbol.
Se encuentra que el agrupamiento de K significa que funciona bien cuando la estructura de los agrupamientos es hiperesférica (como un círculo en 2D, una esfera en 3D).	El agrupamiento jerárquico no funciona tan bien como cuando la forma de los agrupamientos es hiperesférica.
Ventajas: 1. Se garantiza la convergencia. 2. Especializados en racimos de diferentes tamaños y formas.	**Ventajas:** 1. Facilidad de manejo de cualquier forma de semejanza o distancia. 2. En consecuencia, la aplicabilidad a cualquier tipo de atributos.

Desventajas: 1. El valor K es difícil de predecir. 2. No funcionó bien con el clúster global.	**Desventaja:** 1. La agrupación jerárquica requiere el cálculo y el almacenamiento de una matriz de distancia n×n. Para conjuntos de datos muy grandes, esto puede ser costoso y lento.

Tipos de agrupamiento jerárquico

Hay dos tipos principales de agrupación jerárquica:

- **Aglomerativo:** Inicialmente, cada objeto se considera como su propio grupo. De acuerdo con un procedimiento particular, los grupos se fusionan paso a paso hasta que queda un solo grupo. Al final del proceso de fusión de clústeres, se formará un clúster que contiene todos los elementos.

- **Divisivo:** El método Divisivo es lo opuesto al método aglomerativo. Inicialmente, todos los objetos se consideran en un solo grupo. Luego, el proceso de división se realiza paso a paso hasta que cada objeto forma un grupo diferente. El procedimiento de división o división del grupo se lleva a cabo de acuerdo con algunos principios que establecen la distancia máxima entre objetos vecinos en el grupo.

Entre el agrupamiento aglomerativo y divisivo, el agrupamiento aglomerativo es generalmente el método preferido.

Por otro lado, el resultado principal de Hierarchical Clustering es un **dendrograma** , que muestra la relación jerárquica entre los clústeres.

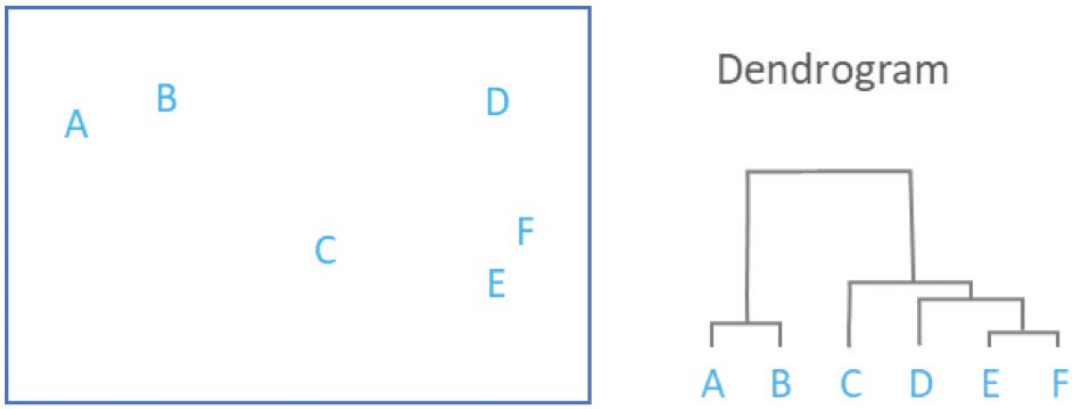
¿Qué es un Dendrograma?

Un *dendrograma* es un diagrama que muestra la relación jerárquica entre objetos. Se crea más comúnmente como una salida de la clusterización jerárquica. El uso principal de un dendrograma es encontrar la mejor manera de asignar objetos a grupos. El dendrograma a continuación muestra el agrupamiento jerárquico de seis *observaciones* que se muestran en el *diagrama* de dispersión de la izquierda.

Cómo leer un dendrograma

La clave para interpretar un dendrograma es centrarse en la altura a la que se unen dos objetos cualesquiera. En el ejemplo anterior, podemos ver que E y F son las más parecidas, ya que la altura del eslabón que las une es la más pequeña. Los siguientes dos objetos más similares son A y B.

En el dendrograma anterior, la altura del dendrograma indica el orden en que se unieron los grupos. Se puede crear un dendrograma más informativo donde las alturas reflejan la distancia entre los grupos como se muestra a continuación. En este caso, el dendrograma nos muestra que la gran diferencia entre conglomerados está entre el conglomerado de A y B versus el de C, D, E y F.



Es importante apreciar que el dendrograma es un resumen de la matriz de distancias y, como ocurre con la mayoría de los resúmenes, se pierde

información. Por ejemplo, el dendrograma sugiere que C y D están mucho más cerca entre sí que C y B, pero los datos originales (que se muestran en el diagrama de dispersión) nos muestran que esto no es cierto. Para usar un poco de jerga, un dendrograma solo es preciso cuando los datos satisfacen la *desigualdad del árbol ultramétrico,* y esto es poco probable para cualquier dato del mundo real.

La consecuencia de la pérdida de información es que los dendrogramas son más precisos en la parte inferior, mostrando qué elementos son muy similares.

Medidas de distancia (similitud)

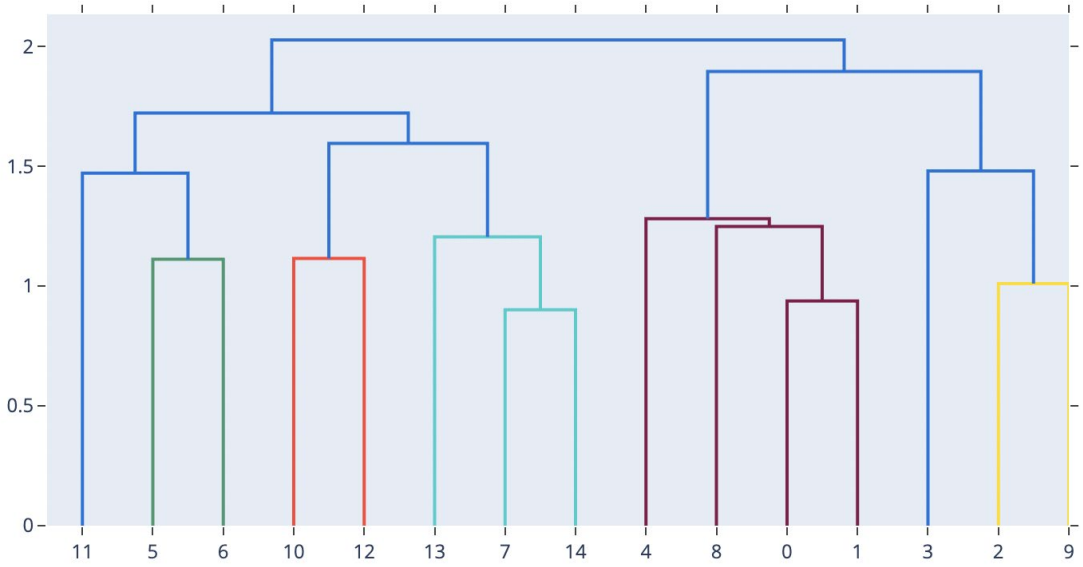
En el ejemplo anterior, la *distancia* entre dos grupos se calculó en función de la longitud de la línea recta trazada de un grupo a otro. Esto se conoce comúnmente como la *distancia euclidiana.* Se han desarrollado muchas otras *métricas de distancia.*

La elección de la métrica de distancia debe hacerse en base a preocupaciones teóricas del dominio de estudio. Es decir, una métrica de distancia necesita definir la similitud de una manera que sea sensible para el campo de estudio. Por ejemplo, si se agrupan sitios delictivos en una ciudad, la distancia entre cuadras de la ciudad puede ser adecuada. O, mejor aún, el tiempo que se tarda en viajar entre cada ubicación. Cuando no existe una justificación teórica para una alternativa, generalmente se debe preferir la euclidiana, ya que suele ser la medida apropiada de distancia en el mundo físico.

Utilizando un dendrograma

Los valores en el eje de profundidad del árbol corresponden a distancias entre grupos. Los diagramas de dendrograma se usan comúnmente en biología computacional para mostrar la agrupación de genes o muestras, a veces en el margen de los mapas de calor.

```
import plotly.figure_factory as ff

import numpy as np

np.random.seed(1)


X = np.random.rand(15, 12) # 15 samples, with 12
dimensions each

fig = ff.create_dendrogram(X)

fig.update_layout(width=800, height=500)

fig.show()
```



Etapas de agrupación jerárquica

El agrupamiento jerárquico emplea una medida de distancia/similitud para crear nuevos agrupamientos. Los pasos para el agrupamiento aglomerativo se pueden resumir de la siguiente manera:

- Paso 1: Calcule la matriz de proximidad usando una métrica de distancia particular
- Paso 2: Cada punto de datos se asigna a un clúster
- Paso 3: Fusionar los clústeres en función de una métrica de similitud entre clústeres
- Paso 4: Actualiza la matriz de distancia
- Paso 5: Repita los pasos 3 y 4 hasta que solo quede un clúster

Ventajas y desventajas

Ventajas

1. No hace falta determinar un número específico de grupos.
2. Su implementación es sencilla.
3. El dendrograma que se genera es muy útil para entender los datos.

Desventajas

1. Al agruparse de una forma, el proceso no permite realizar una corrección, esto es, no se puede alterar el paso ya creado.
2. La complejidad del tiempo para el clustering puede dar lugar a tiempos de cálculo muy largos, en comparación con algoritmos eficientes, como K-Means.
3. Si la muestra es muy grande, puede ser difícil determinar el número adecuado de clústeres por el dendrograma.

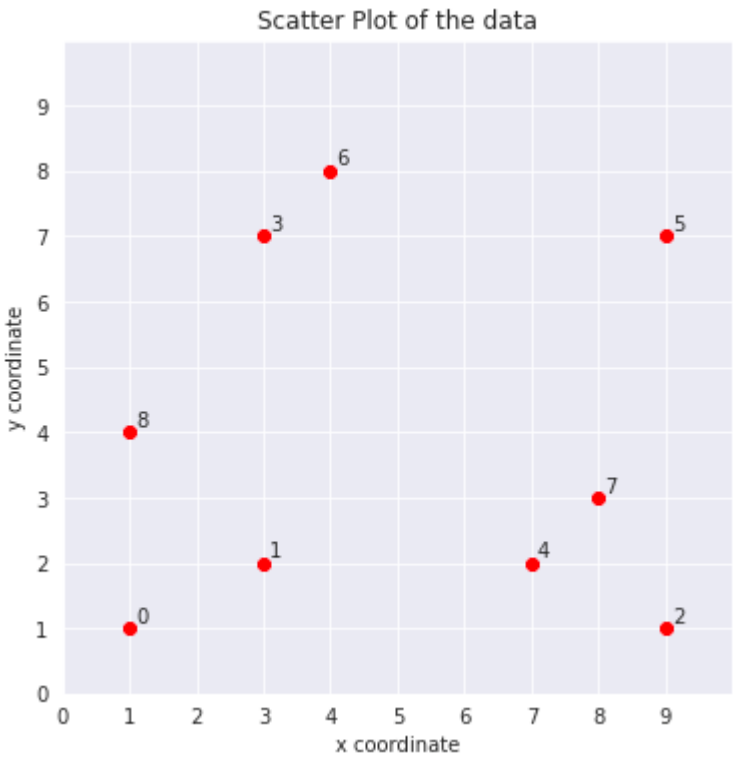
Agrupación jerárquica usando Scikit-Learn

La biblioteca Scikit-Learn tiene su propia función para el agrupamiento jerárquico aglomerativo: AgglomerativeClustering. Las opciones para calcular la distancia entre grupos incluyen **barrio, completo, promedio** y **único** . Para obtener información más específica, puede encontrar esta clase en los documentos relevantes.

```
import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

sns.set_style('dark')
```

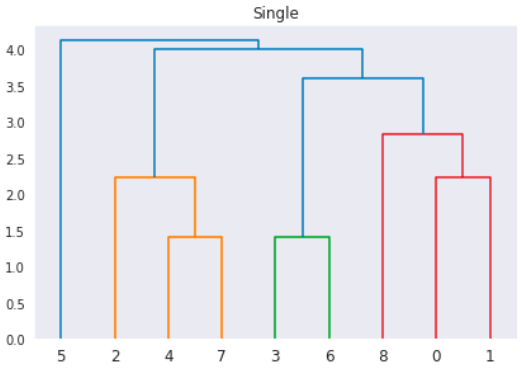
```
X1 = np.array([[1,1], [3,2], [9,1], [3,7], [7,2],
[9,7], [4,8], [8,3],[1,4]])

plt.figure(figsize=(6, 6))


plt.scatter(X1[:,0], X1[:,1], c='r')


# Create numbered labels for each point

for i in range(X1.shape[0]):

    plt.annotate(str(i), xy=(X1[i,0], X1[i,1]),
xytext=(3, 3), textcoords='offset points')
```

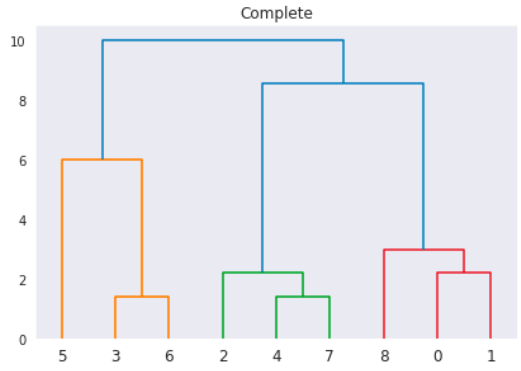
```
plt.xlabel('x coordinate')

plt.ylabel('y coordinate')

plt.title('Scatter Plot of the data')

plt.xlim([0,10]), plt.ylim([0,10])

plt.xticks(range(10)), plt.yticks(range(10))

plt.grid()


plt.show()
```

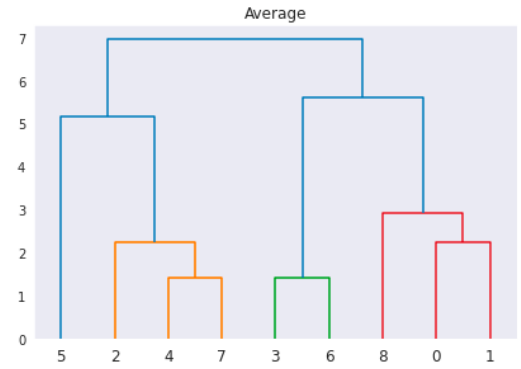


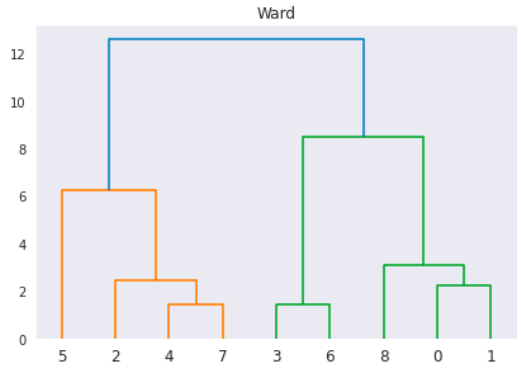

```
from scipy.cluster.hierarchy import dendrogram,
linkage
```

```
Z1 = linkage(X1, method='single', metric='euclidean')

Z2 = linkage(X1, method='complete',
metric='euclidean')

Z3 = linkage(X1, method='average',
metric='euclidean')

Z4 = linkage(X1, method='ward', metric='euclidean')


plt.figure(figsize=(15, 10))

plt.subplot(2,2,1), dendrogram(Z1),
plt.title('Single')

plt.subplot(2,2,2), dendrogram(Z2),
plt.title('Complete')

plt.subplot(2,2,3), dendrogram(Z3),
plt.title('Average')

plt.subplot(2,2,4), dendrogram(Z4), plt.title('Ward')

plt.show()
```


Single
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
5 2 4 7 3 6 8 0 1


Complete
10
8
6
4
2
0
5 3 6 2 4 7 8 0 1


Average
7
6
5
4
3
2
1
0
5 2 4 7 3 6 8 0 1


Ward
12
10
8
6
4
2
0
5 2 4 7 3 6 8 0 1

Referencias

[1] Agrupamiento jerárquico
https://estrategiastrading.com/clustering-jerarquico/

[2] Opciones de Dendrogramas en py
https://plotly.com/python/dendrogram/

[3] Agrupación de Cluster
https://www.statdeveloper.com/agrupacion-en-cluster-jerarquica/

Material Complementario

[1] Clusterización jerárquica
https://www.youtube.com/watch?v=T76paW6fJBI&t=2s

[2] Dendrogramas paso a paso
https://www.youtube.com/watch?v=rZQqhLtzHR8