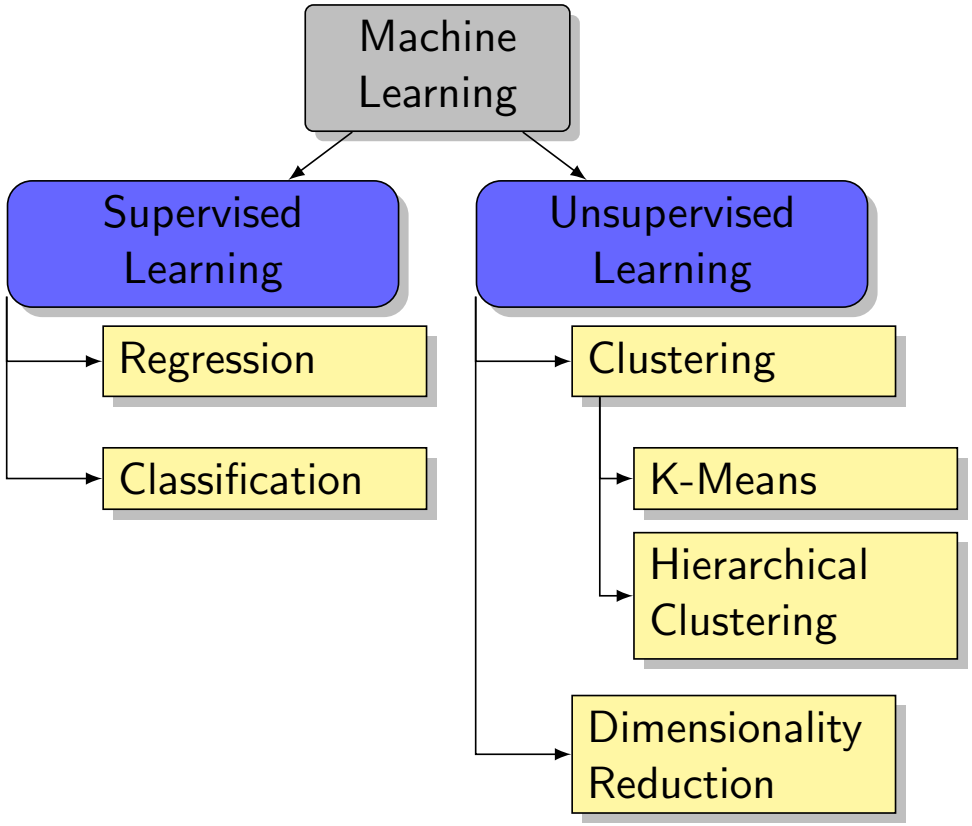
Clustering

Dr. Mauricio Toledo-Acosta

Diplomado Ciencia de Datos con Python

Table of Contents

Introducción



Clustering

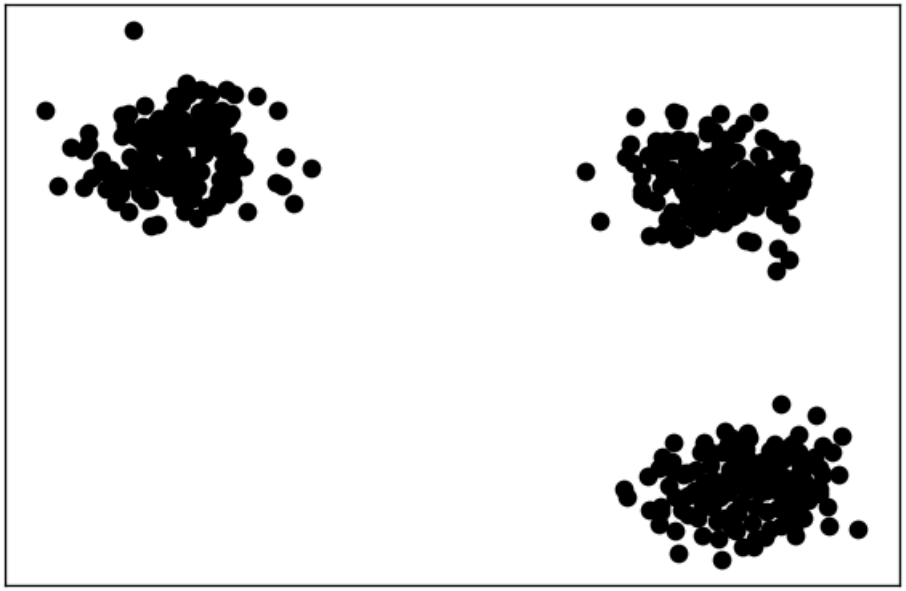
Clustering

El clustering, o agrupamiento, es la tarea que consiste en agrupar objetos de tal manera que los objetos en el mismo conjunto (cluster) son más similares entre sí que con los objetos de los otros conjuntos.

Clustering

Clustering

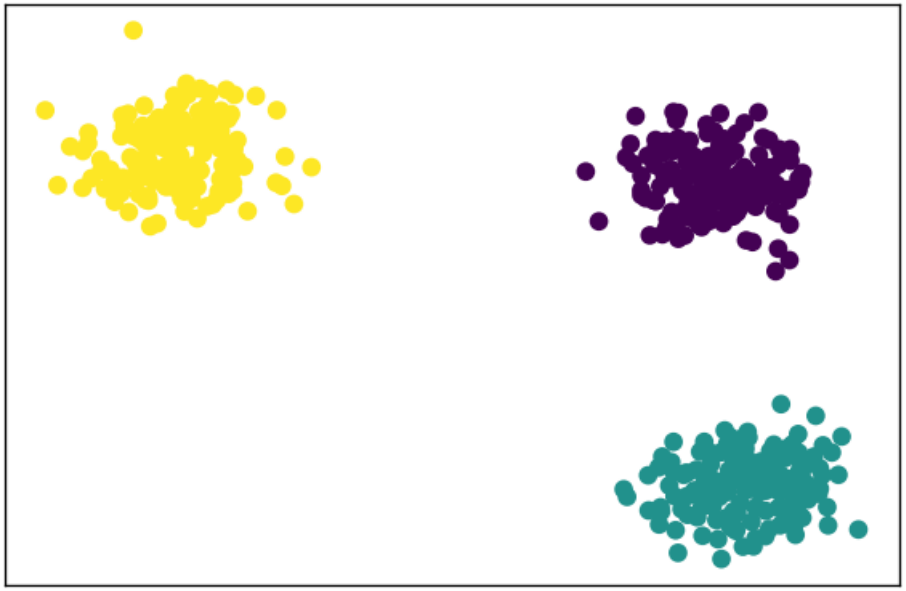
El clustering, o agrupamiento, es la tarea que consiste en agrupar objetos de tal manera que los objetos en el mismo conjunto (cluster) son más similares entre sí que con los objetos de los otros conjuntos.



Clustering

Clustering

El clustering, o agrupamiento, es la tarea que consiste en agrupar objetos de tal manera que los objetos en el mismo conjunto (cluster) son más similares entre sí que con los objetos de los otros conjuntos.



Tipos de clustering

El clustering puede realizarse usando varios algoritmos que difieren en cuanto al significado de *qué significa un cluster*:

- **Connectividad**: Cluster jerárquico basado en conectividad por distancia.

Tipos de clustering

El clustering puede realizarse usando varios algoritmos que difieren en cuanto al significado de *qué significa un cluster*:

- **Connectividad**: Cluster jerárquico basado en conectividad por distancia.
- **Centroides**: Los clusters son representados por un vector promedio.

Tipos de clustering

El clustering puede realizarse usando varios algoritmos que difieren en cuanto al significado de *qué significa un cluster*:

- **Connectividad**: Cluster jerárquico basado en conectividad por distancia.
- **Centroides**: Los clusters son representados por un vector promedio.
- **Distribución**: Los clusters son modelados usando distribuciones de probabilidad.

Tipos de clustering

El clustering puede realizarse usando varios algoritmos que difieren en cuanto al significado de *qué significa un cluster*:

- **Connectividad**: Cluster jerárquico basado en conectividad por distancia.
- **Centroides**: Los clusters son representados por un vector promedio.
- **Distribución**: Los clusters son modelados usando distribuciones de probabilidad.
- **Densidad**: Los clusters son regiones densas conectadas.

Tipos de clustering

El clustering puede realizarse usando varios algoritmos que difieren en cuanto al significado de *qué significa un cluster*:

- **Connectividad**: Cluster jerárquico basado en conectividad por distancia.
- **Centroides**: Los clusters son representados por un vector promedio.
- **Distribución**: Los clusters son modelados usando distribuciones de probabilidad.
- **Densidad**: Los clusters son regiones densas conectadas.
- **Subespacios**: Al mismo tiempo se clusterizan filas y columnas.

Tipos de clustering

El clustering puede realizarse usando varios algoritmos que difieren en cuanto al significado de *qué significa un cluster*:

- **Connectividad**: Cluster jerárquico basado en conectividad por distancia.
- **Centroides**: Los clusters son representados por un vector promedio.
- **Distribución**: Los clusters son modelados usando distribuciones de probabilidad.
- **Densidad**: Los clusters son regiones densas conectadas.
- **Subespacios**: Al mismo tiempo se clusterizan filas y columnas.
- **Grafos**: Los clusters son cliques en grafos de cercanía.

Tipos de clustering

El clustering puede realizarse usando varios algoritmos que difieren en cuanto al significado de *qué significa un cluster*:

- **Connectividad**: Cluster jerárquico basado en conectividad por distancia.
- **Centroides**: Los clusters son representados por un vector promedio.
- **Distribución**: Los clusters son modelados usando distribuciones de probabilidad.
- **Densidad**: Los clusters son regiones densas conectadas.
- **Subespacios**: Al mismo tiempo se clusterizan filas y columnas.
- **Grafos**: Los clusters son cliques en grafos de cercanía.
- **Modelos neuronales**: Usan redes neuronales no supervisadas y pueden ser similares a uno o varios de los enfoques anteriores.

Tipos de clustering

Hay dos tipos de clustering:

- **Clustering duro**: Cada elemento pertenece a exactamente un cluster. Es decir, no hay traslapes entre pares de clústers y la unión de todos, cubren todo el conjunto de puntos.

Tipos de clustering

Hay dos tipos de clustering:

- **Clustering duro**: Cada elemento pertenece a exactamente un cluster. Es decir, no hay traslapes entre pares de clústers y la unión de todos, cubren todo el conjunto de puntos.
- **Clustering suave**: Puede haber puntos que no pertenezcan a ningún cluster, pueden traslaparse los clústers.

Tipos de clustering

Hay dos tipos de clustering:

- **Clustering duro**: Cada elemento pertenece a exactamente un cluster. Es decir, no hay traslapes entre pares de clústers y la unión de todos, cubren todo el conjunto de puntos.
- **Clustering suave**: Puede haber puntos que no pertenezcan a ningún cluster, pueden traslaparse los clústers.

Hay varios tipos de clustering:

- **Partición**: Cada elemento pertenece a un cluster o no.

Tipos de clustering

Hay dos tipos de clustering:

- **Clustering duro**: Cada elemento pertenece a exactamente un cluster. Es decir, no hay traslapes entre pares de clústers y la unión de todos, cubren todo el conjunto de puntos.
- **Clustering suave**: Puede haber puntos que no pertenezcan a ningún cluster, pueden traslaparse los clústers.

Hay varios tipos de clustering:

- **Partición**: Cada elemento pertenece a un cluster o no.
- **Partición con ruido**: Cada elemento pertenece a un cluster, hasta cierto punto.

Tipos de clustering

Hay dos tipos de clustering:

- **Clustering duro**: Cada elemento pertenece a exactamente un cluster. Es decir, no hay traslapes entre pares de clústers y la unión de todos, cubren todo el conjunto de puntos.
- **Clustering suave**: Puede haber puntos que no pertenezcan a ningún cluster, pueden traslaparse los clústers.

Hay varios tipos de clustering:

- **Partición**: Cada elemento pertenece a un cluster o no.
- **Partición con ruido**: Cada elemento pertenece a un cluster, hasta cierto punto.
- **Overlapping Clustering**: Cada elemento puede pertenecer a varios clusters.

Tipos de clustering

Hay dos tipos de clustering:

- **Clustering duro**: Cada elemento pertenece a exactamente un cluster. Es decir, no hay traslapes entre pares de clústers y la unión de todos, cubren todo el conjunto de puntos.
- **Clustering suave**: Puede haber puntos que no pertenezcan a ningún cluster, pueden traslaparse los clústers.

Hay varios tipos de clustering:

- **Partición**: Cada elemento pertenece a un cluster o no.
- **Partición con ruido**: Cada elemento pertenece a un cluster, hasta cierto punto.
- **Overlapping Clustering**: Cada elemento puede pertenecer a varios clusters.
- **Clustering de subespacios**: En varios subconjuntos de features se forman clusters.

Tipos de clustering

Hay dos tipos de clustering:

- **Clustering duro**: Cada elemento pertenece a exactamente un cluster. Es decir, no hay traslapes entre pares de clústers y la unión de todos, cubren todo el conjunto de puntos.
- **Clustering suave**: Puede haber puntos que no pertenezcan a ningún cluster, pueden traslaparse los clústers.

Hay varios tipos de clustering:

- **Partición**: Cada elemento pertenece a un cluster o no.
- **Partición con ruido**: Cada elemento pertenece a un cluster, hasta cierto punto.
- **Overlapping Clustering**: Cada elemento puede pertenecer a varios clusters.
- **Clustering de subespacios**: En varios subconjuntos de features se forman clusters.
- **Clustering jerarquico**: Los objetos que pertenecen a un cluster hijo, pertenecen también al cluster raiz.

Table of Contents

K-Means

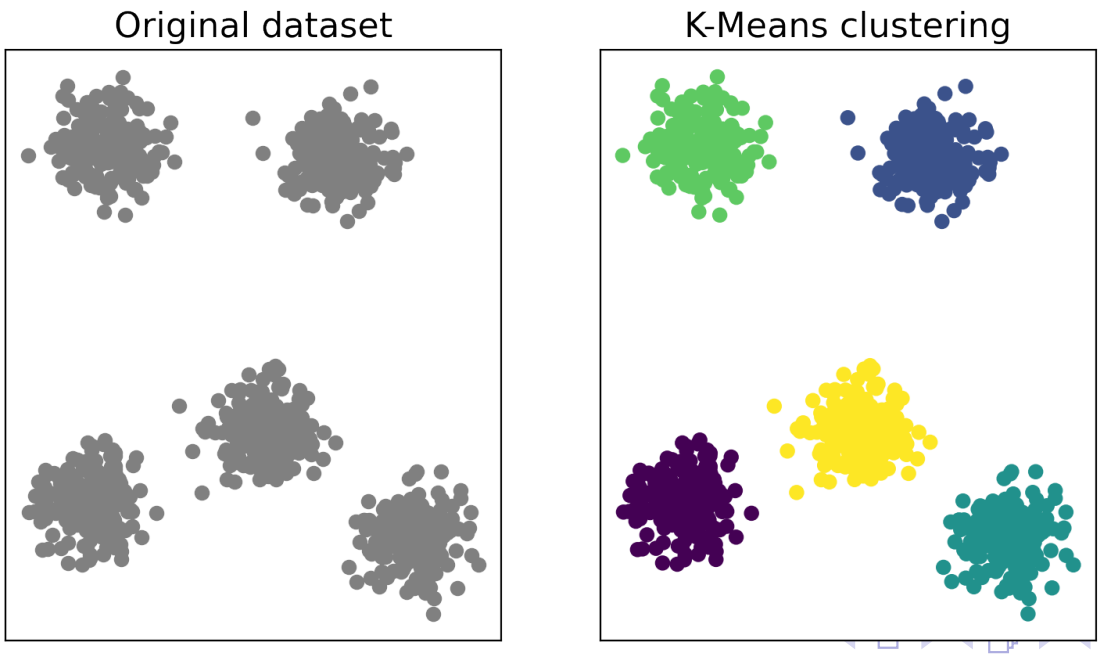
K-Means

Método de clustering que busca particionar n puntos en k clusters de manera que cada punto pertenezca al cluster cuyo centroide esté más cerca. Este centroide representa al cluster.

K-Means

K-Means

Método de clustering que busca particionar n puntos en k clusters de manera que cada punto pertenezca al cluster cuyo centroide esté más cerca. Este centroide representa al cluster.

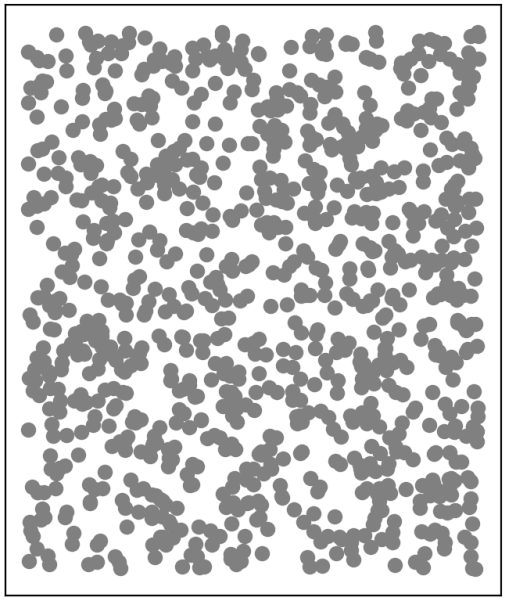


K-Means

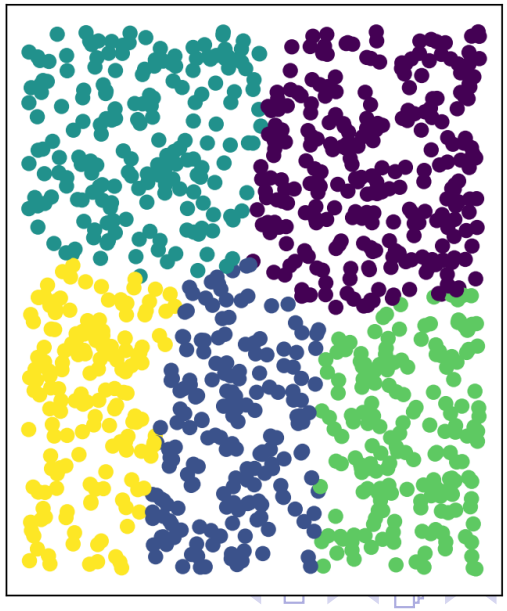
K-Means

Método de clustering que busca particionar n puntos en k clusters de manera que cada punto pertenezca al cluster cuyo centroide esté más cerca. Este centroide representa al cluster.

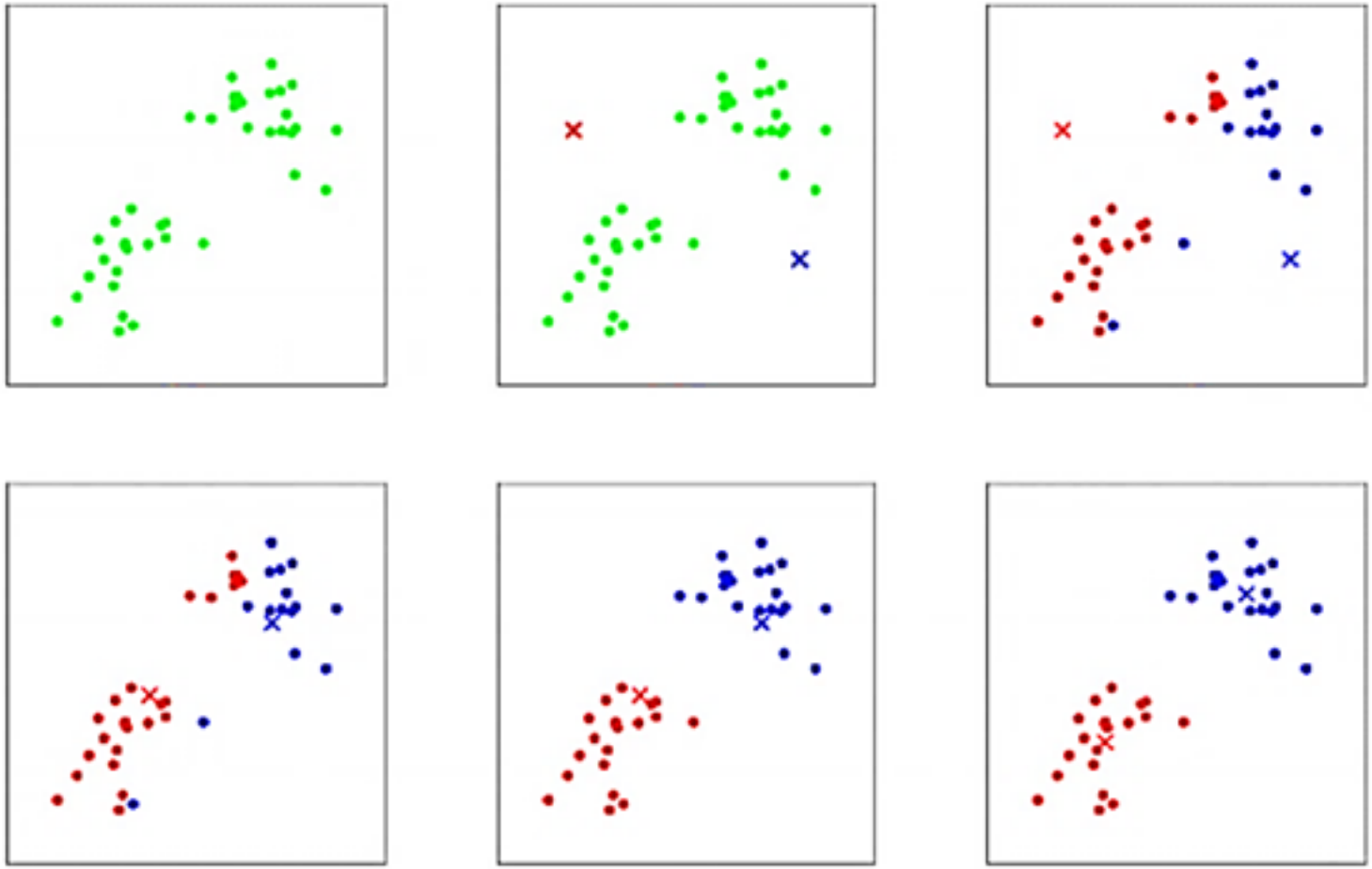
Original dataset



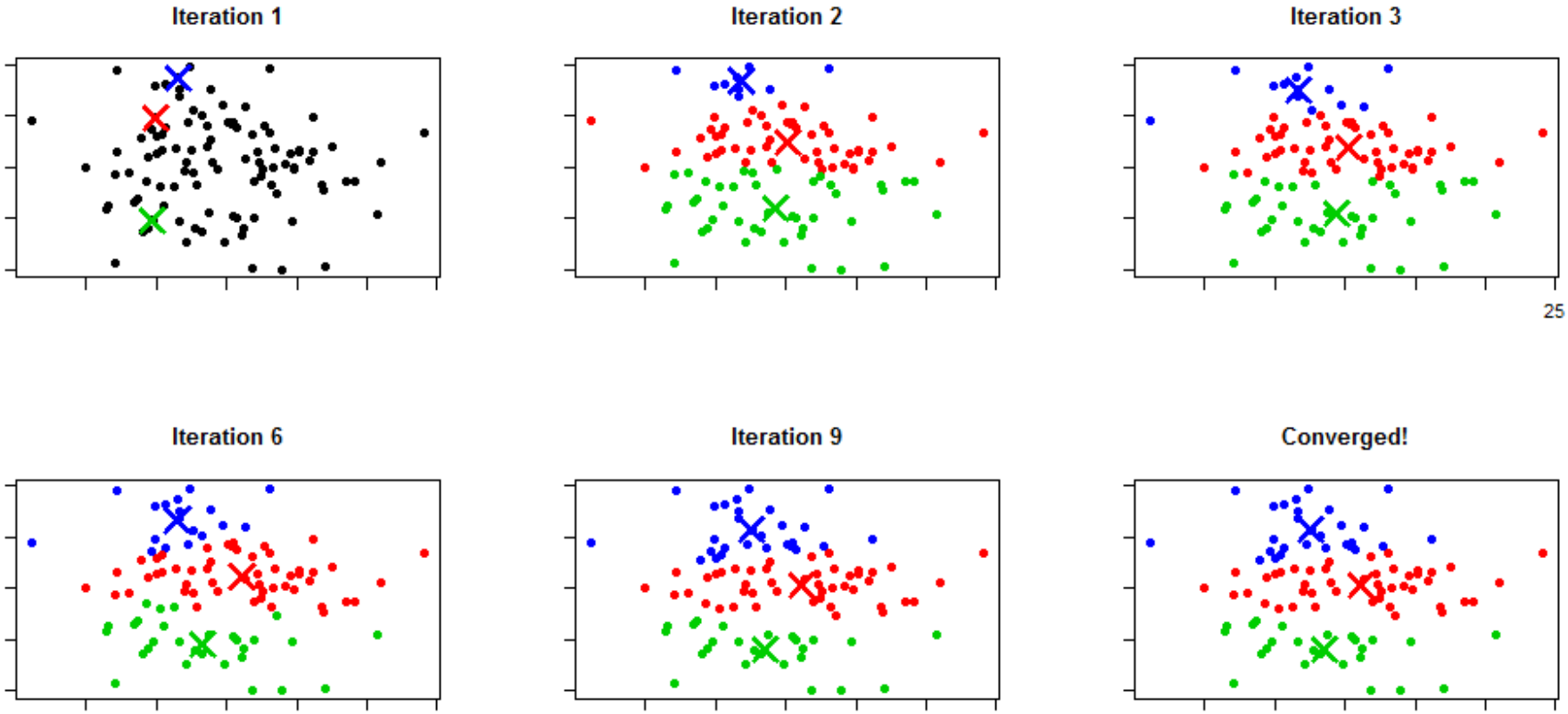
K-Means clustering



K-Means: ¿Cómo funciona?



K-Means: ¿Cómo funciona?



Ventajas

- Es rápido y eficiente.

Ventajas

- Es rápido y eficiente.
- Funciona bien en datasets grandes.

Ventajas

- Es rápido y eficiente.
- Funciona bien en datasets grandes.
- Fácil de interpretar.

Ventajas

- Es rápido y eficiente.
- Funciona bien en datasets grandes.
- Fácil de interpretar.
- Flexible a cambios de métricas.

Desventajas

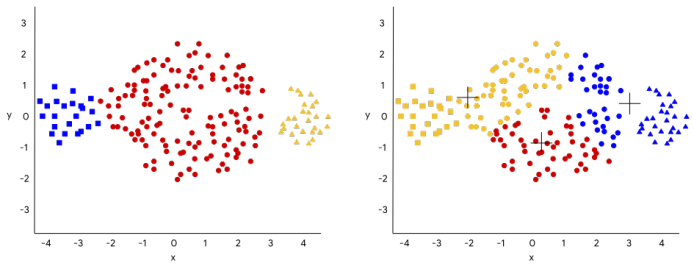
- La elección del parámetro K.

Desventajas

- La elección del parámetro K.
- Es sensible a outliers.

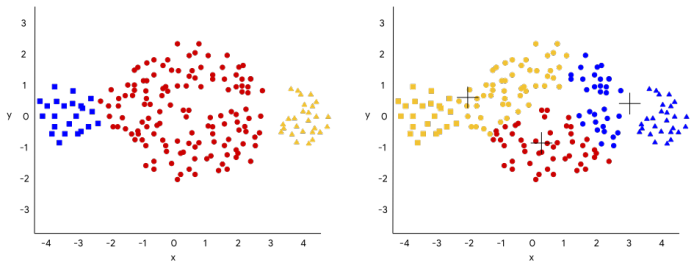
Desventajas

- La elección del parámetro K.
- Es sensible a outliers.
- Produce clusters con tamaños uniformes



Desventajas

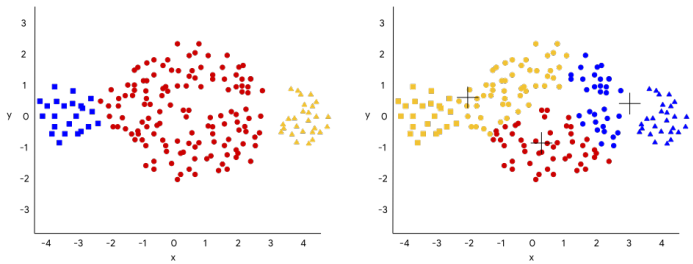
- La elección del parámetro K.
- Es sensible a outliers.
- Produce clusters con tamaños uniformes



- Sensible a la normalización.

Desventajas

- La elección del parámetro K.
- Es sensible a outliers.
- Produce clusters con tamaños uniformes



- Sensible a la normalización.
- Sensibilidad al número de dimensiones.

Detalles

Table of Contents

Hierarchical Clustering

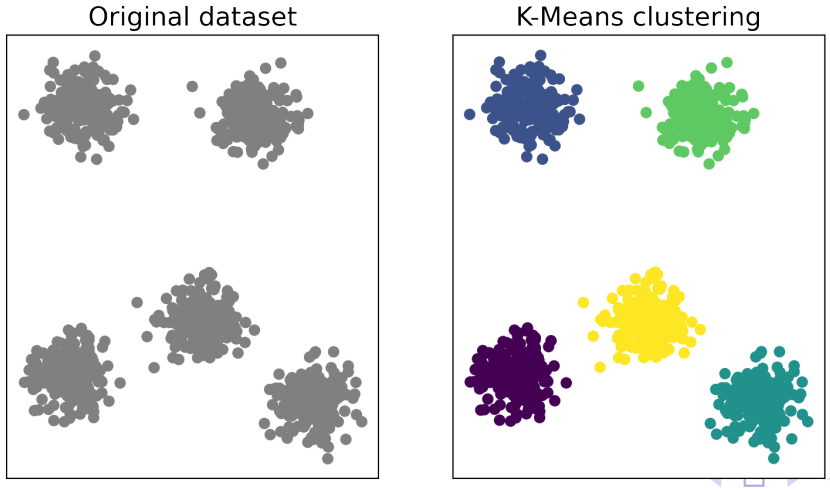
Hierarchical Clustering

La agrupación jerárquica es una familia general de algoritmos de agrupación que crean agrupaciones anidadas fusionándolas o dividiéndolas sucesivamente. Esta jerarquía de grupos se representa como un árbol (o dendrograma). La raíz del árbol es el grupo único que reúne todas las muestras, siendo las hojas los grupos con una sola muestra.

Hierarchical Clustering

Hierarchical Clustering

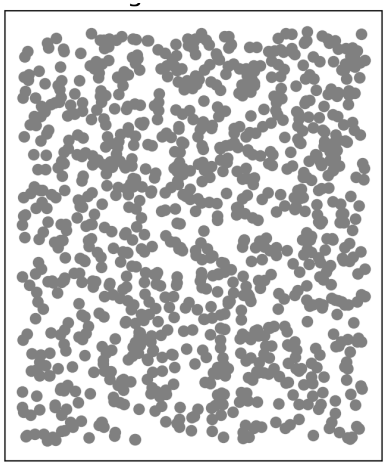
La agrupación jerárquica es una familia general de algoritmos de agrupación que crean agrupaciones anidadas fusionándolas o dividiéndolas sucesivamente. Esta jerarquía de grupos se representa como un árbol (o dendrograma). La raíz del árbol es el grupo único que reúne todas las muestras, siendo las hojas los grupos con una sola muestra.

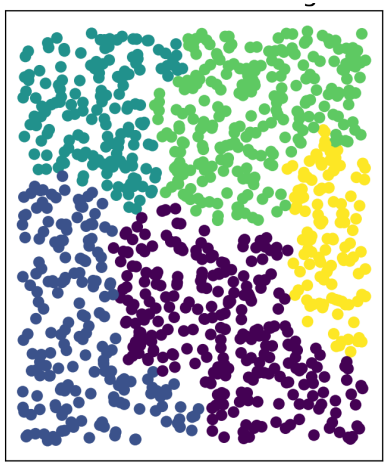


Hierarchical Clustering

Hierarchical Clustering

La agrupación jerárquica es una familia general de algoritmos de agrupación que crean agrupaciones anidadas fusionándolas o dividiéndolas sucesivamente. Esta jerarquía de grupos se representa como un árbol (o dendrograma). La raíz del árbol es el grupo único que reúne todas las muestras, siendo las hojas los grupos con una sola muestra.





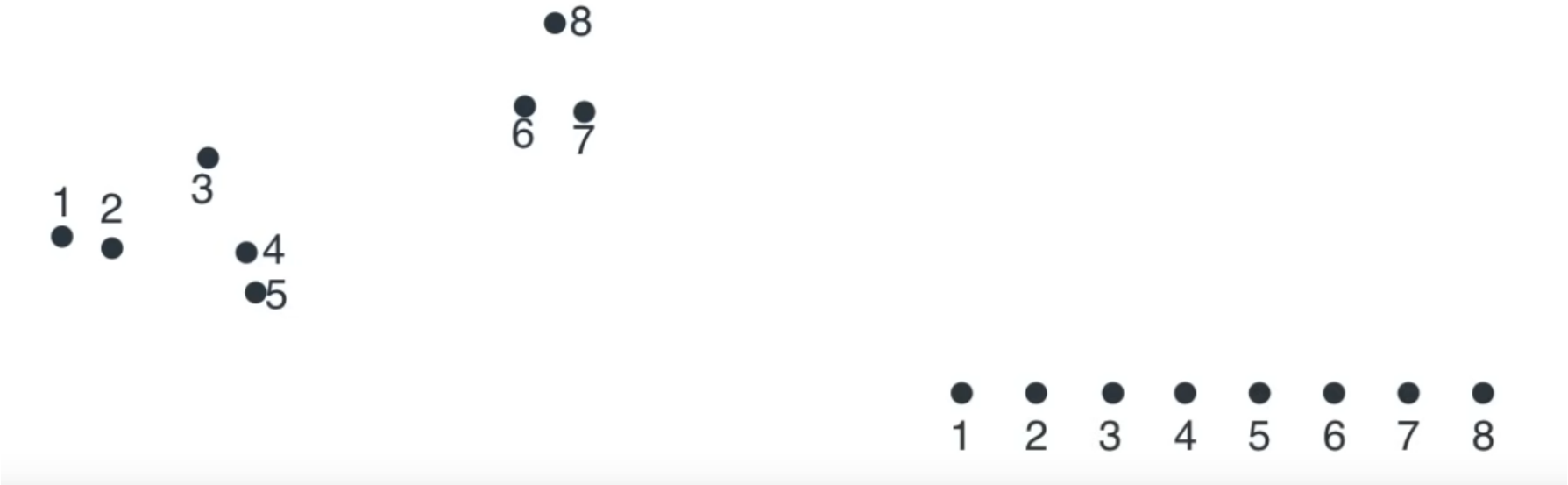
Hierarchical Clustering

En la implementación de `scikit-learn`, los criterios de vinculación (linkage) determinan la métrica utilizada para la estrategia de fusión:

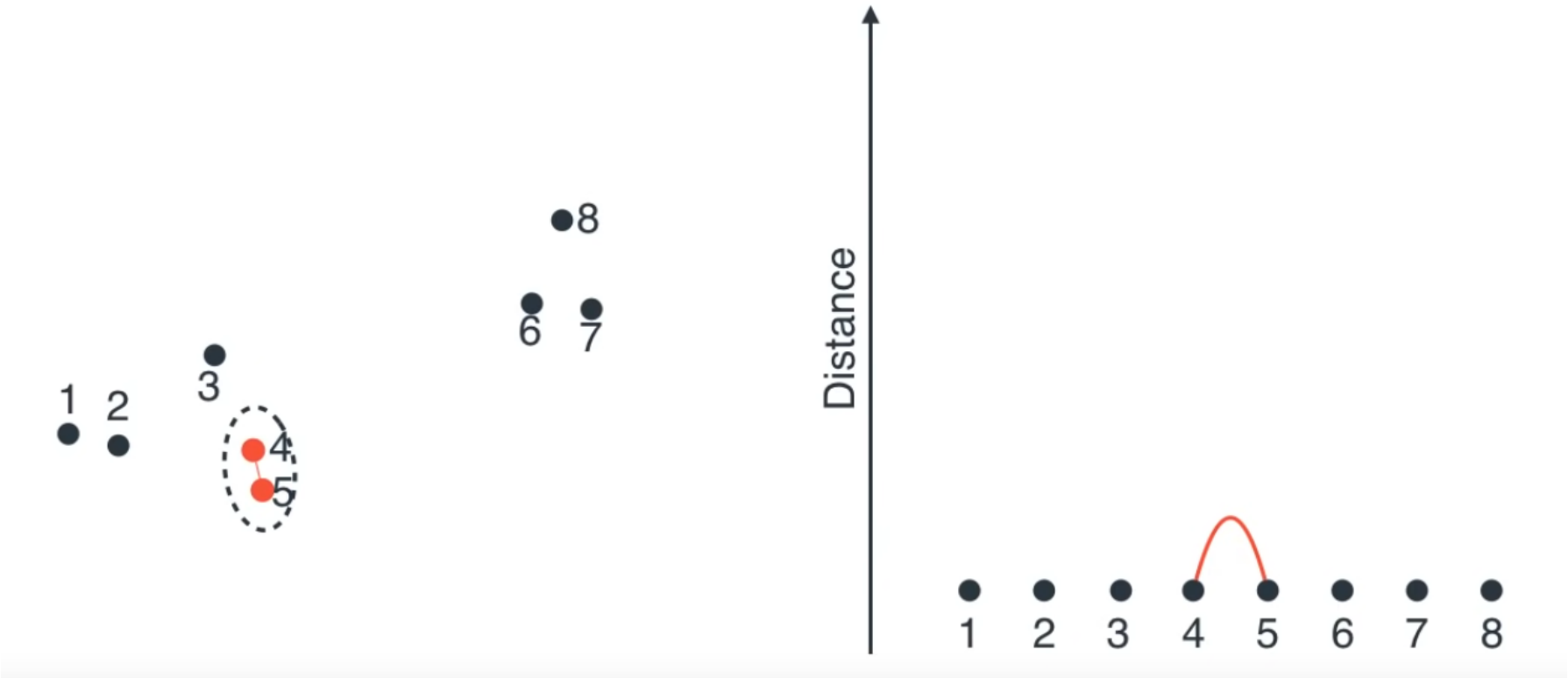
- `Ward` minimiza la suma de las diferencias al cuadrado dentro de todos los grupos, es decir, minimiza la varianza.
- `Complete Linkage` minimiza la distancia máxima entre observaciones de pares de grupos.
- `Average linkage` minimiza el promedio de las distancias entre todas las observaciones de pares de grupos.
- `Single linkage` minimiza la distancia entre las observaciones más cercanas de pares de grupos.

Además, hay que especificar el número de clusters o un umbral de distancia máxima.

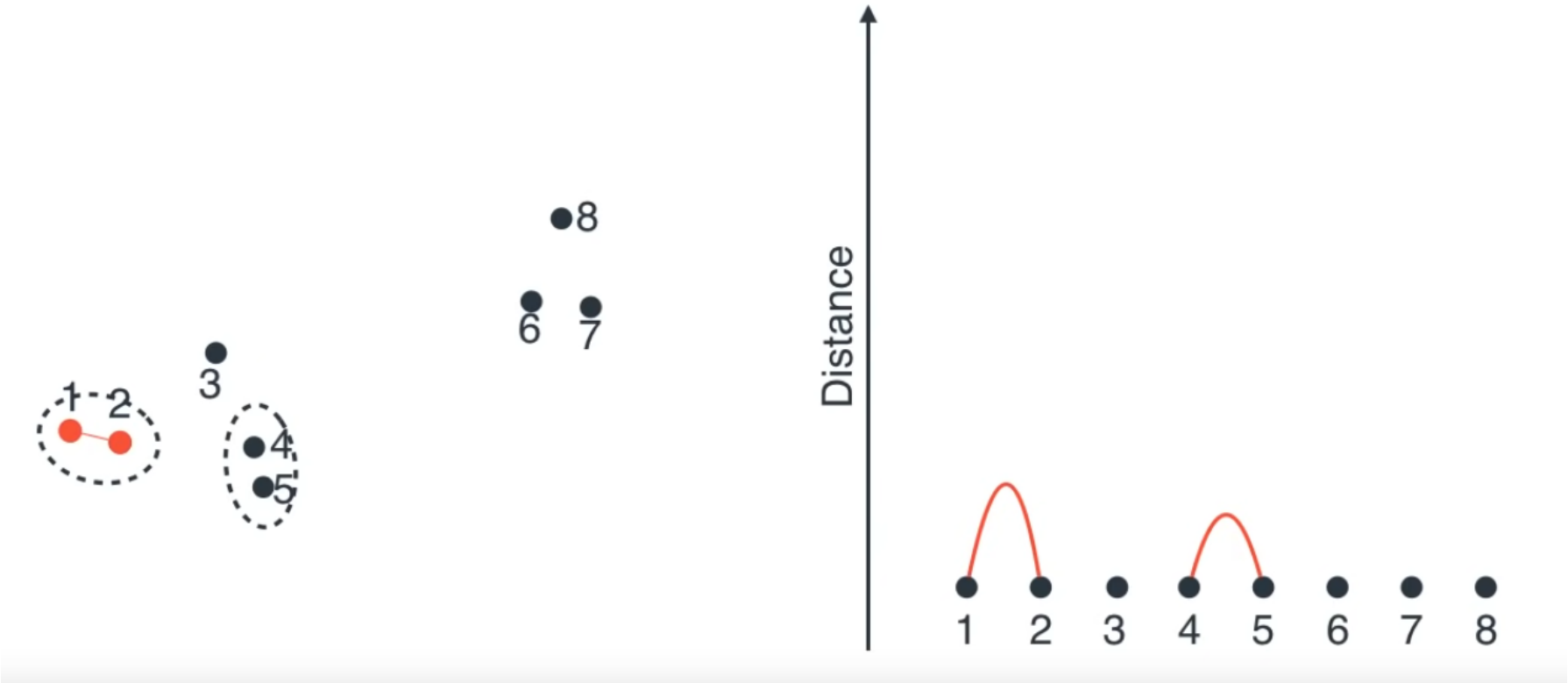
Agglomerative Clustering



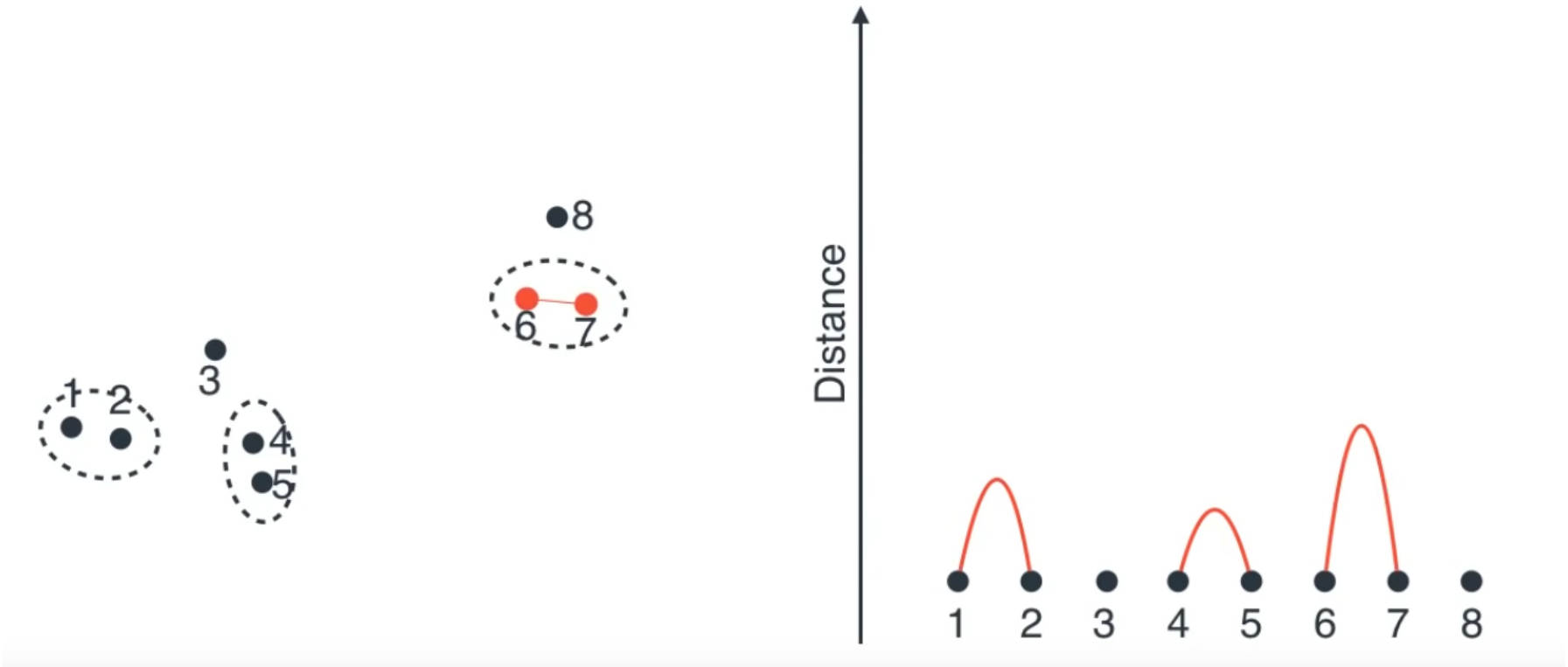
Agglomerative Clustering



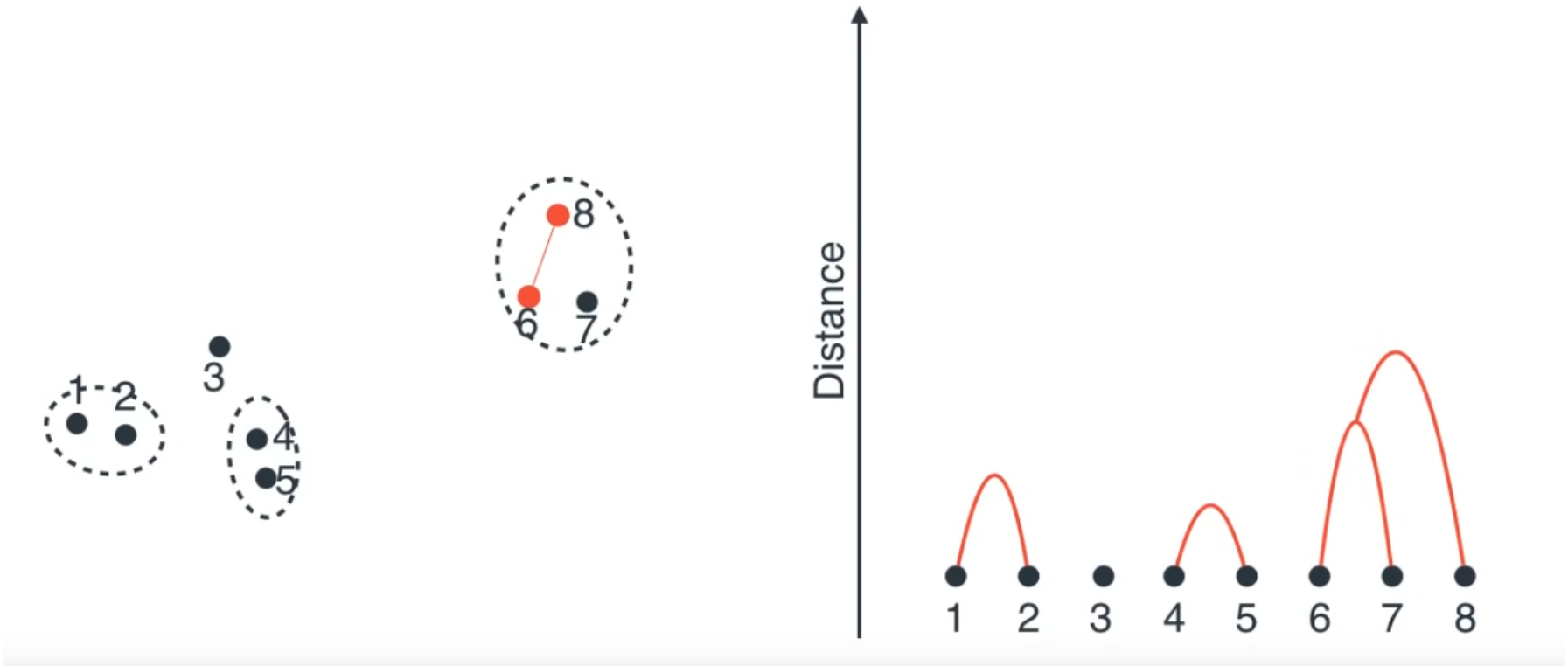
Agglomerative Clustering



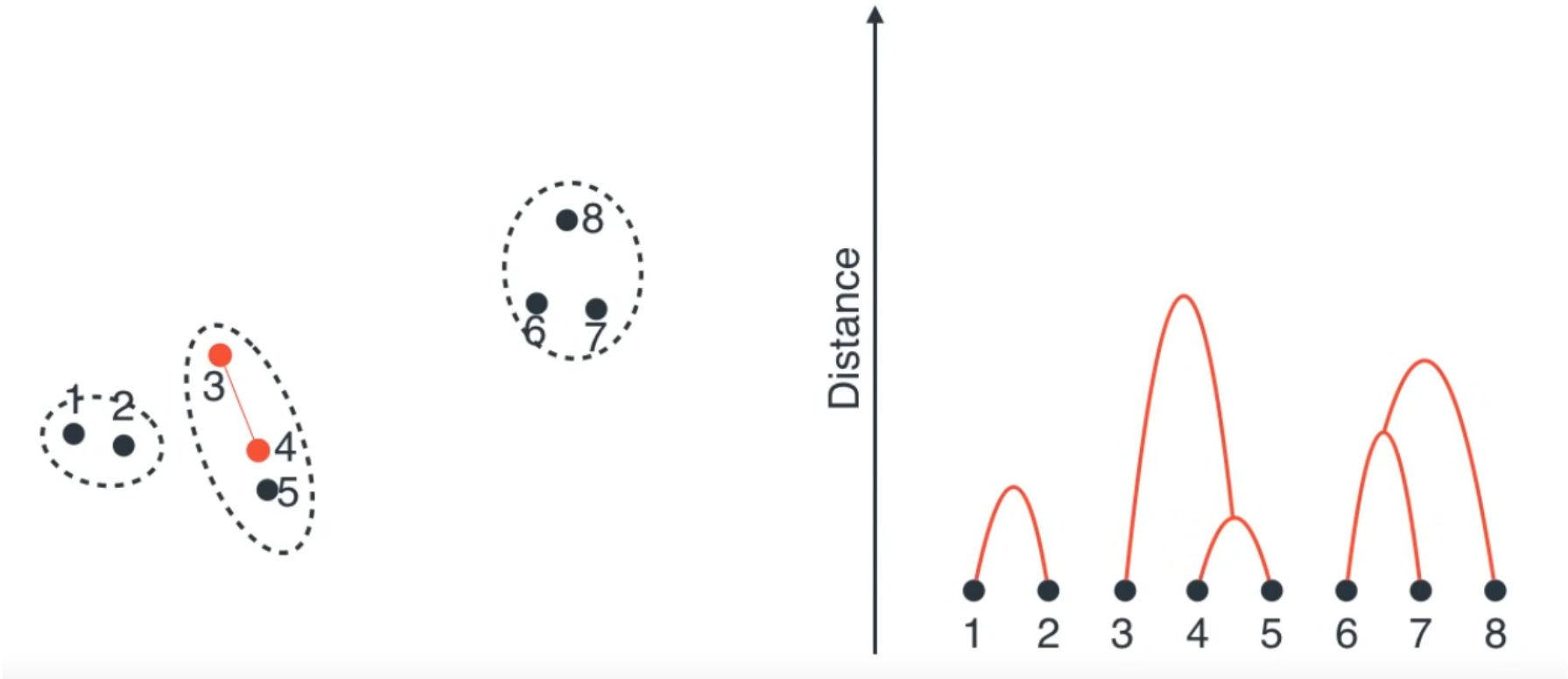
Agglomerative Clustering



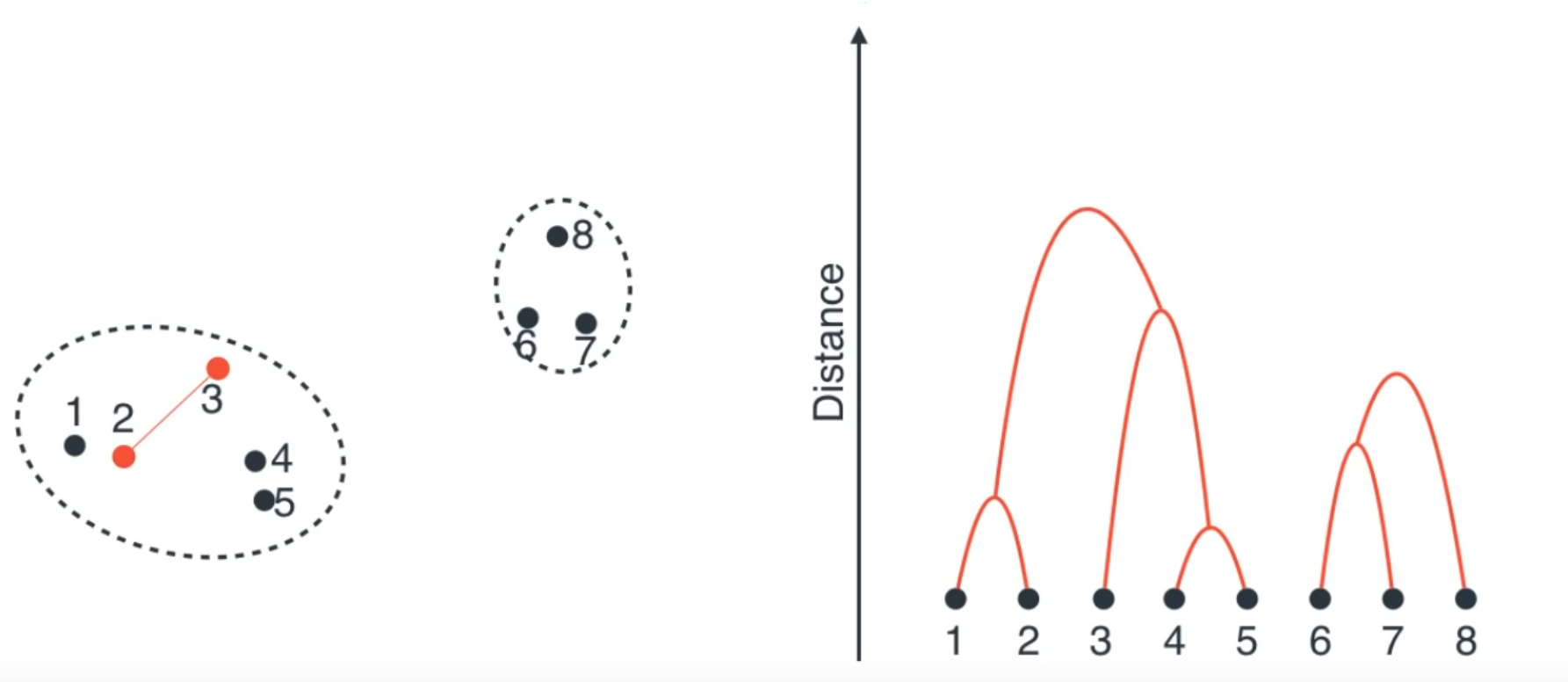
Agglomerative Clustering



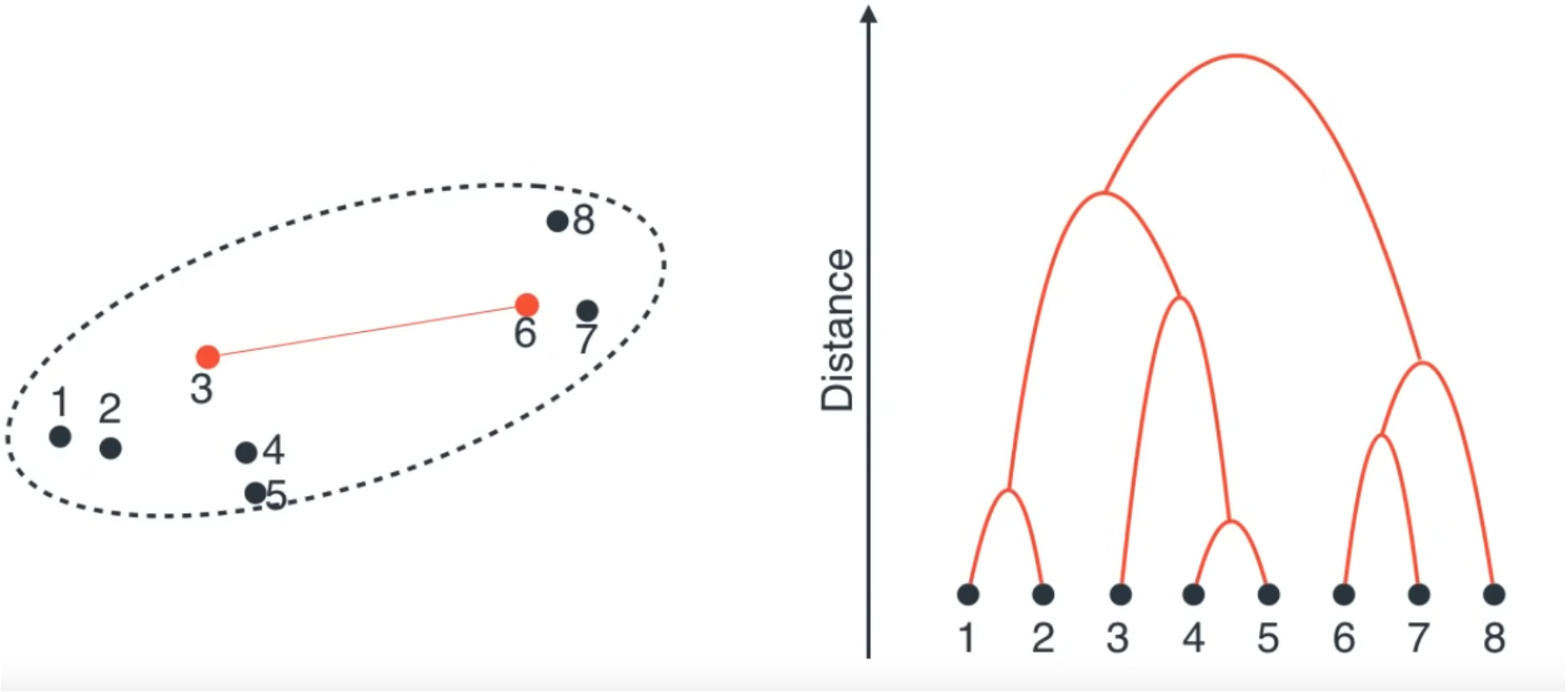
Agglomerative Clustering



Agglomerative Clustering



Agglomerative Clustering



Ventajas y Desventajas

Ventajas

- Su estructura de jerarquía ofrece más información que la simple lista de clusters.
- Fácil de implementar e interpretar.

Desventajas

- Susceptible a outliers.
- No es apto para datasets muy grandes.
- El orden de los datos impactan el resultado final.

Más algoritmos de clustering

- **K-Means**: Particiona los datos en k clústeres minimizando la varianza intra-clúster.
- **Clustering jerárquico**: Construye una jerarquía de clústeres mediante fusiones o divisiones sucesivas.
- **DBSCAN**: Agrupa puntos densamente conectados y detecta automáticam ruido (outliers).
- **Mean Shift**: Encuentra clústeres desplazando iterativamente los puntos hacia zonas de mayor densidad.
- **Gaussian Mixture Models (GMM)**: Supone que los datos provienen de una mezcla de distribuciones normales.
- **Spectral Clustering**: Utiliza los valores propios de una matriz de similitud para reducir dimensionalidad antes de aplicar clustering.

Table of Contents

Silhouette score (score de silueta)

Silhoutte

La silueta es un método de interpretación y validación de la coherencia dentro de un cluster. El valor de la silueta es una medida de cuán similar es un objeto a su propio cluster (cohesión) en comparación con otros cluster (separación).

- La silueta va de -1 a 1.

Silhouette score (score de silueta)

Silhoutte

La silueta es un método de interpretación y validación de la coherencia dentro de un cluster. El valor de la silueta es una medida de cuán similar es un objeto a su propio cluster (cohesión) en comparación con otros cluster (separación).

- La silueta va de -1 a 1.
- Un valor alto indica que el objeto está bien emparejado con su propio cluster y mal emparejado con los clusters vecinos.

Silhouette score (score de silueta)

Silhoutte

La silueta es un método de interpretación y validación de la coherencia dentro de un cluster. El valor de la silueta es una medida de cuán similar es un objeto a su propio cluster (cohesión) en comparación con otros cluster (separación).

- La silueta va de -1 a 1.
- Un valor alto indica que el objeto está bien emparejado con su propio cluster y mal emparejado con los clusters vecinos.
- La silueta puede ser calculada con cualquier distancia (euclidiana, Manhattan, angular, etc.).

Silhoutte score

El score de silueta para un dato x_i es s_i dado por:

$$\textbf{Cohesion: } a_i = \frac{1}{|C_I| - 1} \sum_{\substack{j \in C_I \\ j \neq i}} d(i,j)$$

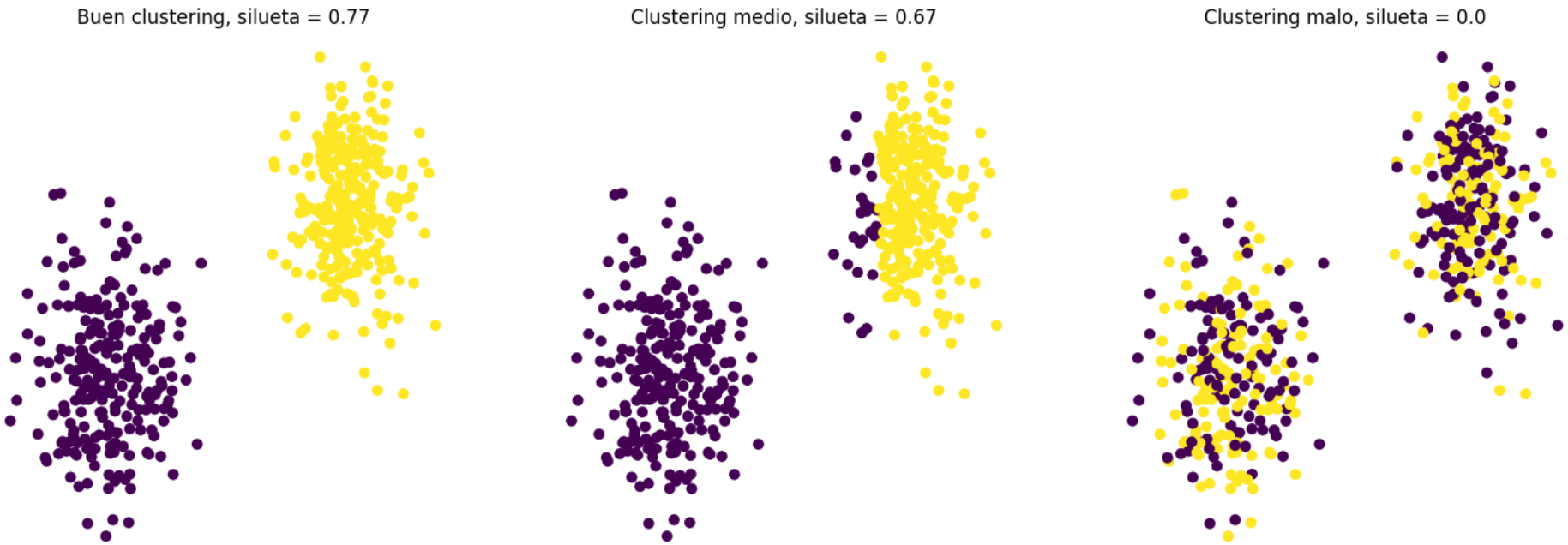
$$\textbf{Separación: } b_i = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j)$$

$$\textbf{Diferencia: } \mathbf{s_i} = \begin{cases} 1 - \frac{a_i}{b_i}, & a_i < b_i \\ 0, & a_i = b_i \\ \frac{b_i}{a_i} - 1, & b_i < a_i. \end{cases}$$

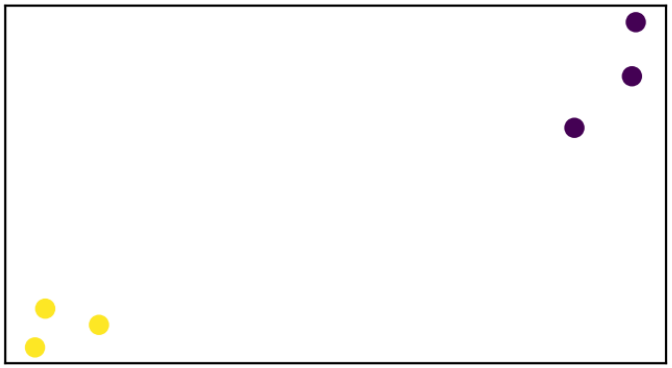
Silhoutte score

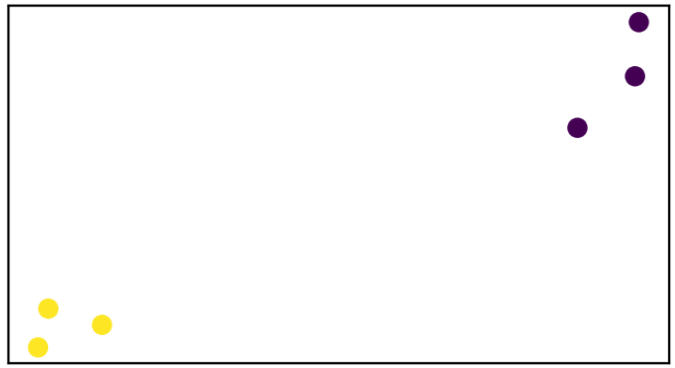
El valor de la silueta para el conjunto de datos $X = \{x_1, ..., x_N\}$ es el promedio

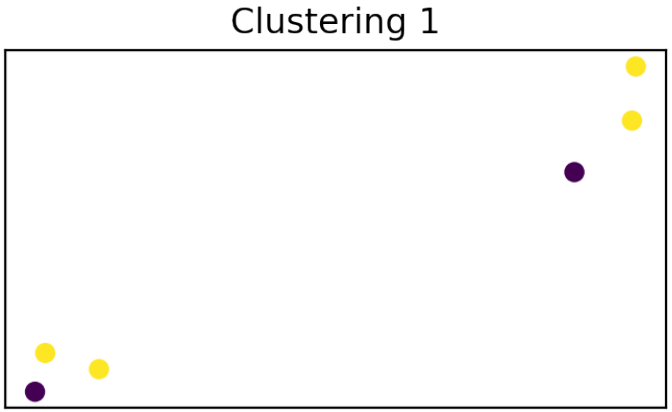
$$s(X) = \frac{1}{N} \sum_{i=1}^{N} s_i.$$

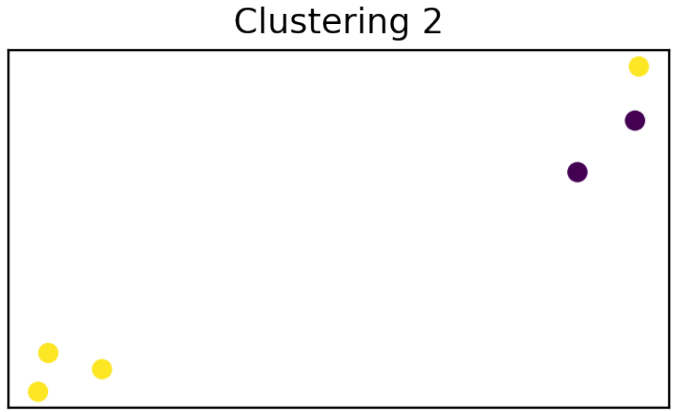


Adjusted Mutual Information









	T Cluster amarillo	T Cluster morado
Cluster amarillo	2	2
Cluster morado	1	1

	T Cluster amarillo	T Cluster morado
Cluster amarillo	3	1
Cluster morado	0	2

Adjusted Mutual Information

La Información Mutua Ajustada mide que tanta información comparten dos clusterings en términos de los elementos que comparten, es decir, del tamaño de la intersección. Suele usarse para **comparar un clustering ground truth contra uno que hemos obtenido**.

- Es un valor entre 0 y 1.

Adjusted Mutual Information

La Información Mutua Ajustada mide que tanta información comparten dos clusterings en términos de los elementos que comparten, es decir, del tamaño de la intersección. Suele usarse para **comparar un clustering ground truth contra uno que hemos obtenido**.

- Es un valor entre 0 y 1.
- Entre mayor es el valor, más similar es el clustering obtenido con el verdadero.

Adjusted Mutual Information

La Información Mutua Ajustada mide que tanta información comparten dos clusterings en términos de los elementos que comparten, es decir, del tamaño de la intersección. Suele usarse para **comparar un clustering ground truth contra uno que hemos obtenido**.

- Es un valor entre 0 y 1.
- Entre mayor es el valor, más similar es el clustering obtenido con el verdadero.
- No cambia si permutamos las etiquetas de los clusters

$$(0, 1, 0, 0, 1) \longleftrightarrow (1, 0, 1, 1, 0)$$

Adjusted Mutual Information

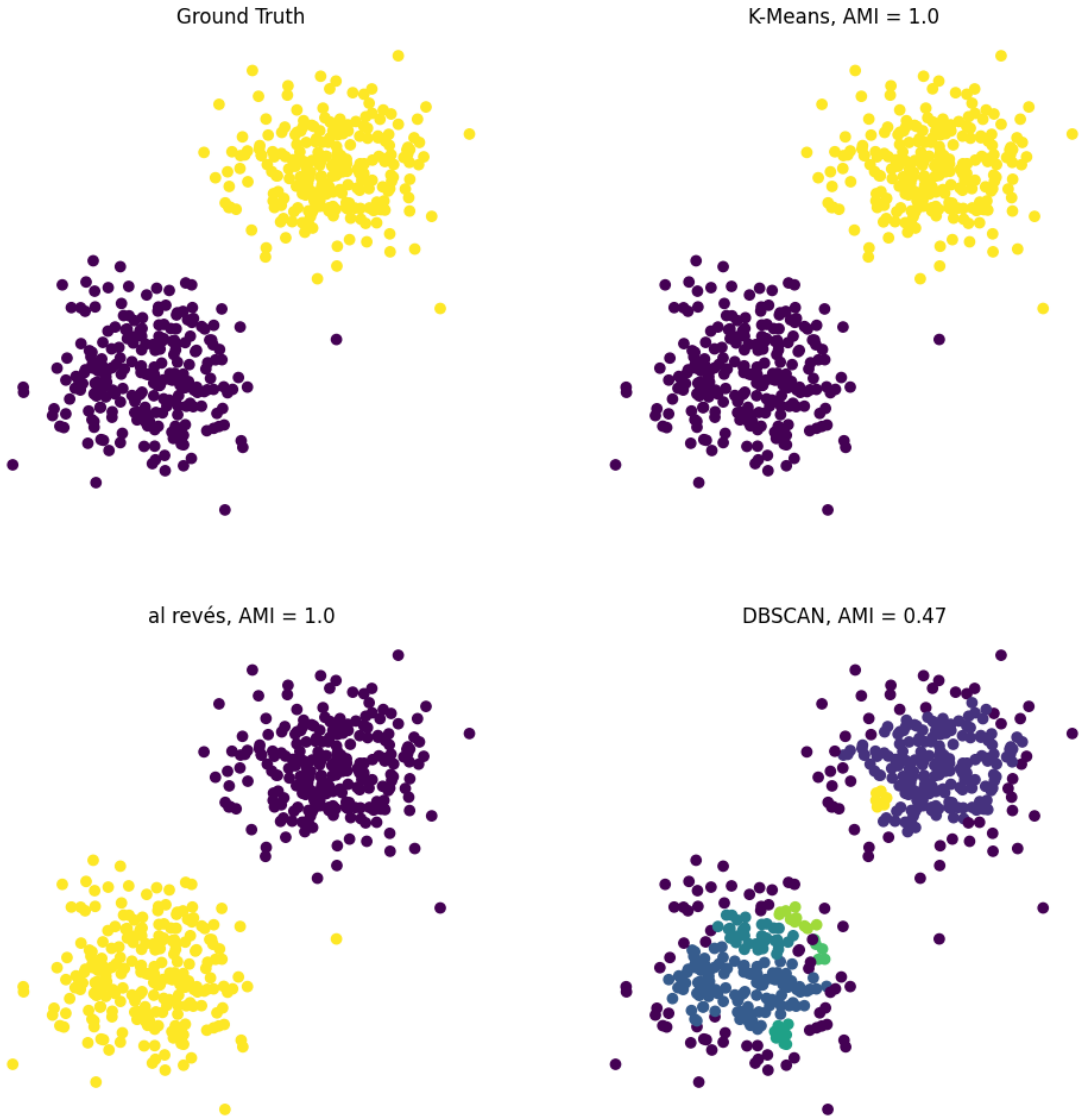
La Información Mutua Ajustada mide que tanta información comparten dos clusterings en términos de los elementos que comparten, es decir, del tamaño de la intersección. Suele usarse para **comparar un clustering ground truth contra uno que hemos obtenido**.

- Es un valor entre 0 y 1.
- Entre mayor es el valor, más similar es el clustering obtenido con el verdadero.
- No cambia si permutamos las etiquetas de los clusters

$$(0, 1, 0, 0, 1) \longleftrightarrow (1, 0, 1, 1, 0)$$

- Es simétrica.

Ejemplos de AMI



Métricas de evaluación para clustering

- **Índice de silueta**: Mide qué tan similar es un punto a su propio clúster frente a otros clústeres.
- **Adjusted Mutual Information (AMI)**: Mide el acuerdo entre dos agrupamientos, ajustado por azar.
- **Rand Index / Adjusted Rand Index (ARI)**: Evalúa la similitud entre dos asignaciones de clúster, considerando pares de elementos.
- **Davies–Bouldin index**: Mide la compacidad intra-clúster y la separación inter-clúster (valores bajos son mejores).
- **Calinski-Harabasz index**: Relación entre la dispersión entre clústeres y dentro de los clústeres (valores altos indican mejor agrupación).
- **Homogeneity, Completeness, V-measure**: Métricas externas que comparan un clustering con clases reales (si están disponibles).

Table of Contents

Maldición de la dimensionalidad