Aprendizaje no supervisado

Fabián Villena

Aprendizaje no supervisado

Los métodos supervisados son los más utilizados y se pueden aplicar cuando tenemos características de observaciones junto con variables respuesta asociadas a cada uno de los objetos observados.

En el aprendizaje no supervisado no tenemos estas variables respuesta y sólo tenemos las características observadas de cada objeto. No nos interesa la predicción porque no tenemos etiquetas a predecir. Nos interesa encontrar patrones interesantes dentro de los datos.

El desafío del aprendizaje no supervisado

Al contrario que en los métodos supervisados, no tenemos un objetivo que sabemos que debemos alcanzar, por lo que la validación de los resultados de estos métodos es más subjetiva.

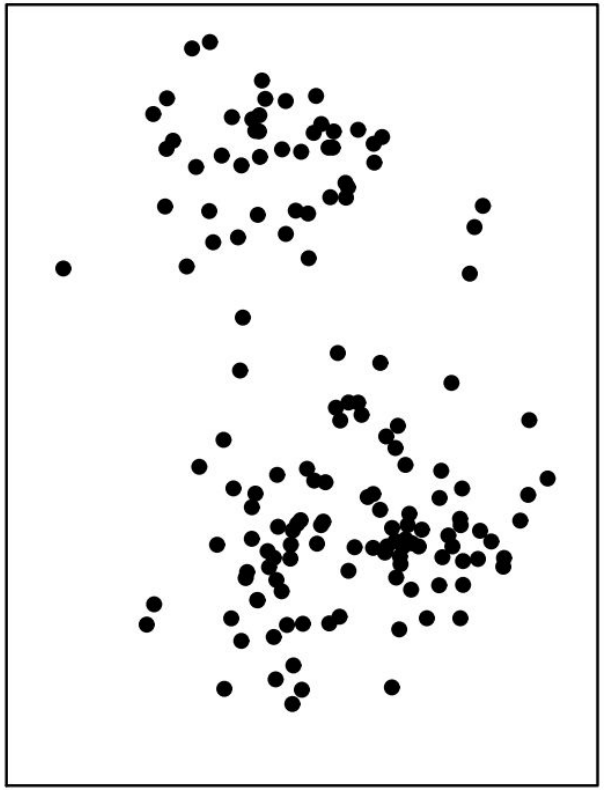
No hay forma de verificar el rendimiento de estas técnicas porque no sabemos la respuesta correcta.

Los métodos no supervisados típicamente se sitúan en la fase de análisis exploratorio.

Agrupamiento

El agrupamiento o clustering agrupa una serie de técnicas que buscan encontrar subgrupos dentro de un conjunto de datos.

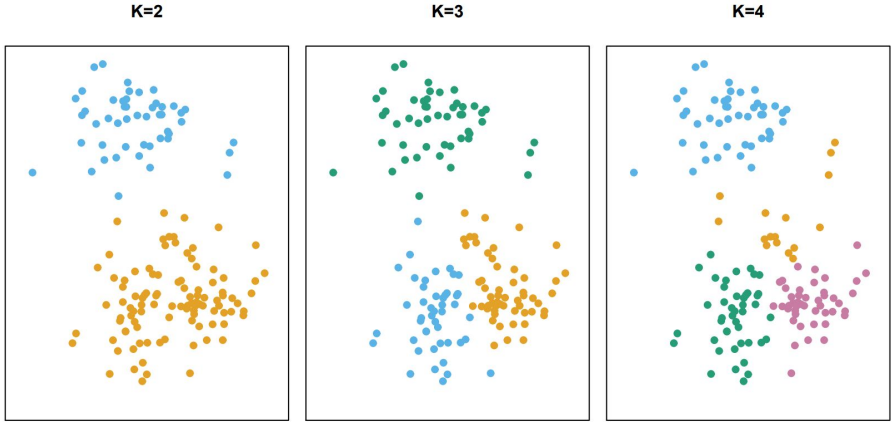
Cuando agrupamos objetos esperamos que los objetos que pertenecen a cada grupo sean muy similares entre sí, mientras que son muy distintos con los objetos de otro grupo.



k-Means

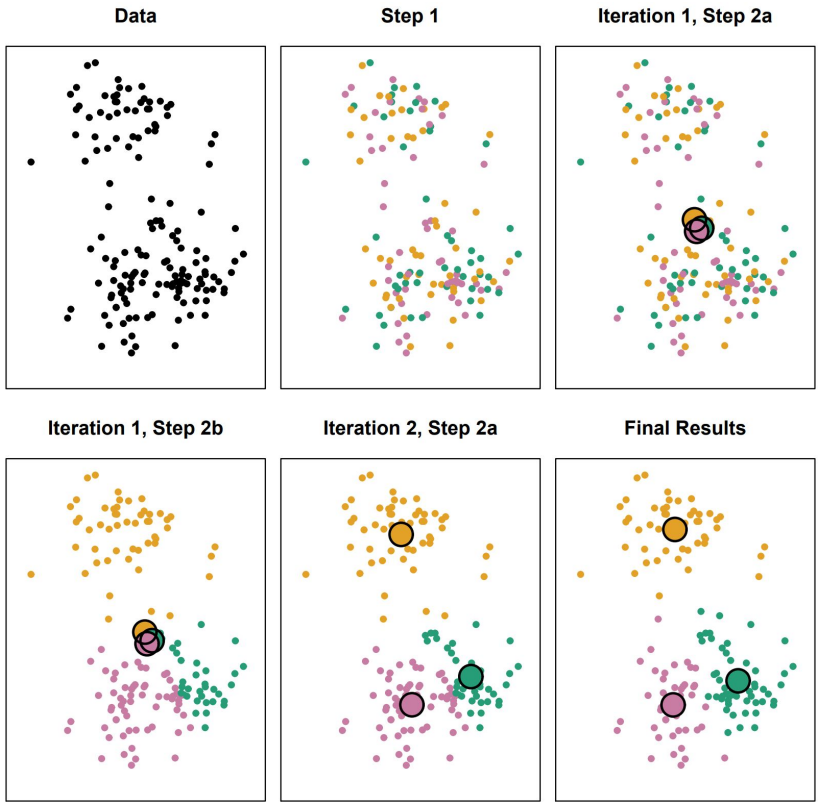
Con este algoritmo se busca particionar los objetos en una cantidad predefinida de grupos.

Para generar grupos mediante el algoritmo de k-means necesitamos primero definir la cantidad de grupos que queremos encontrar y el algoritmo le asignará exactamente un grupo a cada uno de los objetos.



El algoritmo de k-means

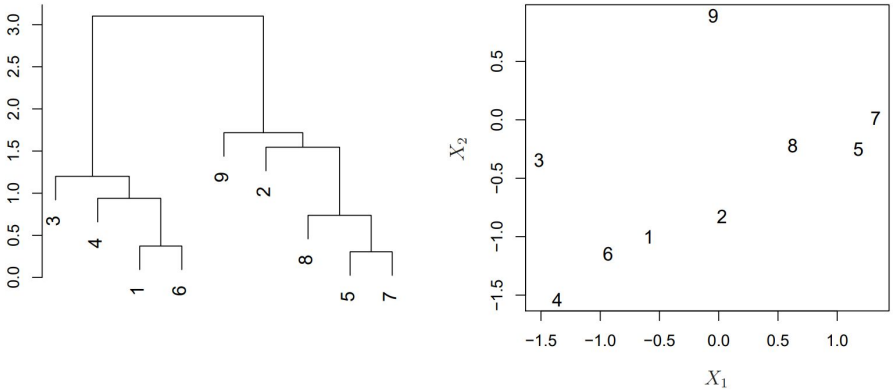
1. Asignar al azar un entero de 1 a k a cada uno de los objetos del conjunto de datos.
2. Iterar hasta que los asignamientos de grupos no cambien.
    a. Para cada uno de los k grupos, calcular el centroide.
    b. Asignar a cada uno de los objetos al grupo cuyo centroide está más cerca.



Agrupamiento jerárquico

Una desventaja de k-means es que debemos definir la cantidad de grupos que vamos a encontrar.

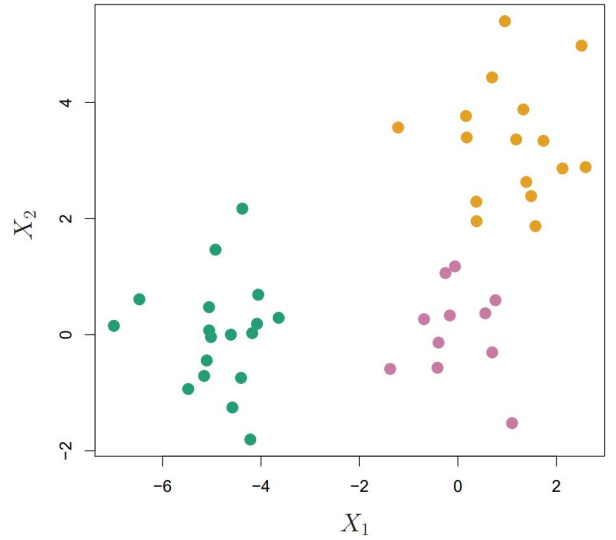
En el agrupamiento jerárquico no necesitamos definir la cantidad de grupos y con este método obtenemos una representación jerárquica de los objetos de nuestro conjunto de datos.

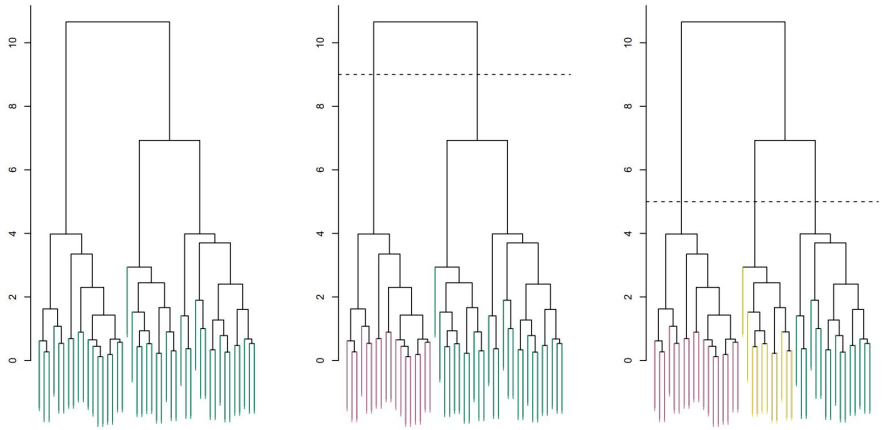


Dendrograma

Un dendrograma es una representación basada en un árbol de los objetos de un conjunto de datos.

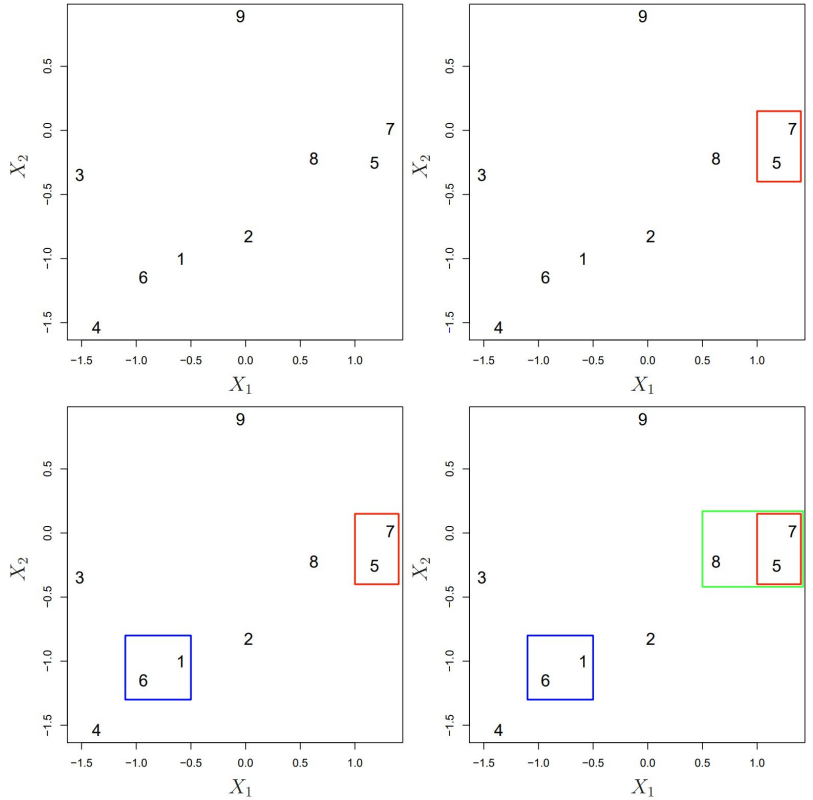
Cada hoja del dendrograma representa un objeto y a medida que subimos el en dendrograma estas hojas se fusionan en ramas, correspondiendo a objetos similares entre sí.





El algoritmo de agrupamiento jerárquico

1. Comienza con las observaciones y una medida de todas las combinaciones de pares de disimilaridades. Trata cada observación como su propio grupo.
2. Por cada objeto:
   a. Examina cada disimilaridad entre los grupos e identifica el pat de grupos que sean más similares. Fusiona esos grupos. La disimilaridad entre esos grupos indica la altura en el dendrograma donde ocurre la fusión.
   b. Calcula las nuevas disimilaridades entre los grupos remanentes



Medidas de disimilaridad y enlace

La medida de disimilaridad típicamente es la distancia euclidiana entre los puntos, pero podemos utilizar otra métrica de distancia.

Para medir la disimilaridad entre dos objetos, es simple porque simplemente se mide entre esos dos puntos, pero cuando queremos medir la disimilaridad entre dos grupos, debemos decidir cómo la mediremos, estos distintos métodos se llaman enlaces.

Las decisiones antes de hacer agrupamiento

- Decidir si las características de mi conjunto de datos deben ser normalizadas de alguna manera, porque algunos métodos son sensibles a la escala de las características.
- En el agrupamiento jerárquico:
  - Decidir qué medida de disimilaridad se va a usar.
  - Decidir el enlace que se usará.
  - Dónde debemos cortar el dendrograma para obtener los grupos.
- En k-means debemos decidir cuántos grupos queremos encontrar en los datos.

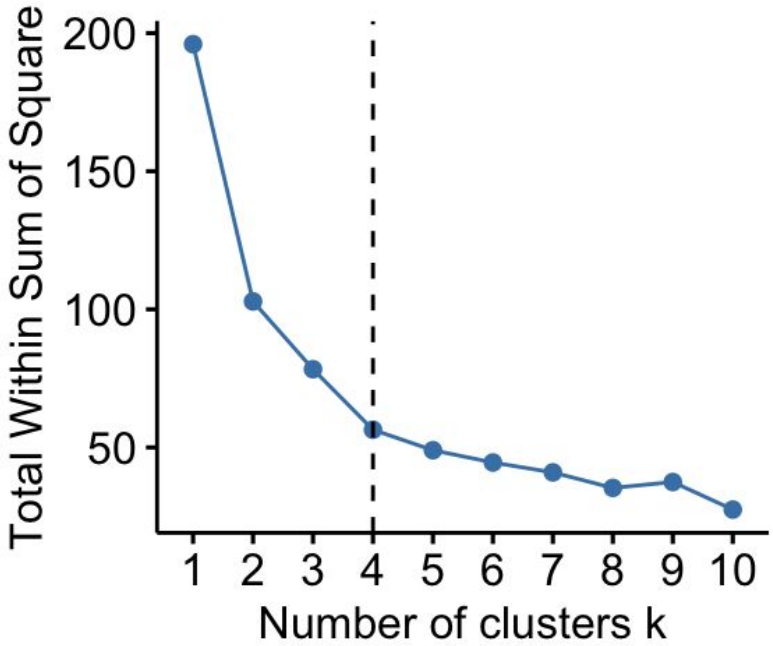
Validando los grupos obtenidos

Siempre encontraremos subgrupos en un conjunto de datos. Pero nosotros realmente queremos saber si los grupos encontrados realmente representan los subgrupos verdaderos del conjunto de datos. Esta es una pregunta difícil de responder, pero existen algunas técnicas que intentan describir la calidad del agrupamiento que se realiza.

El método del codo

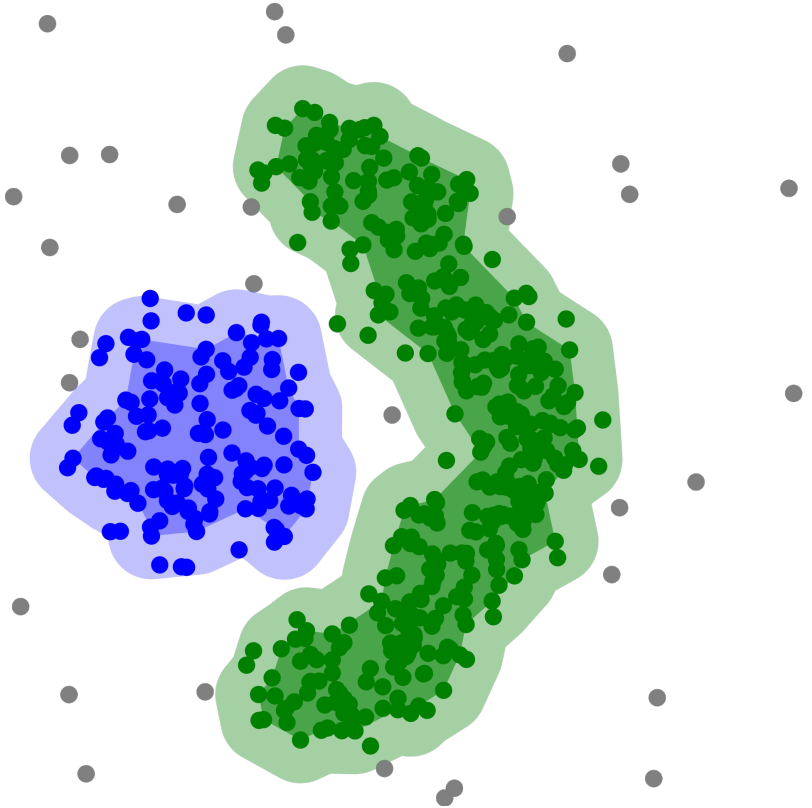
La disimilaridad dentro de los grupos disminuye a medida que voy aumentando el valor de k.

Se cree que el valor de disimilaridad decrece rápidamente en función del k hasta que se llega al k natural del conjunto de datos y después ralentiza su crecimiento a medida que avanzamos en el valor de k.



DBSCAN

Es un algoritmo de agrupamiento que dado un conjunto de puntos en un espacio, agrupa los puntos que están aglomerados en un lugar, marcando como atípicos los puntos que están solos en algún lugar de baja dimensión de puntos.

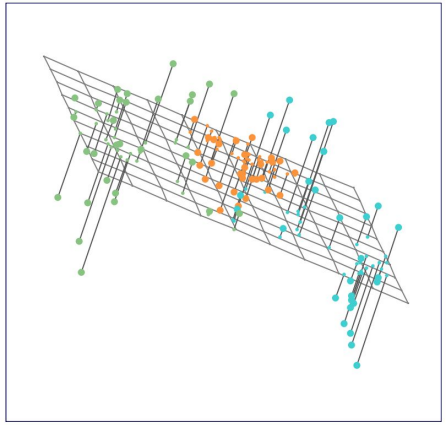


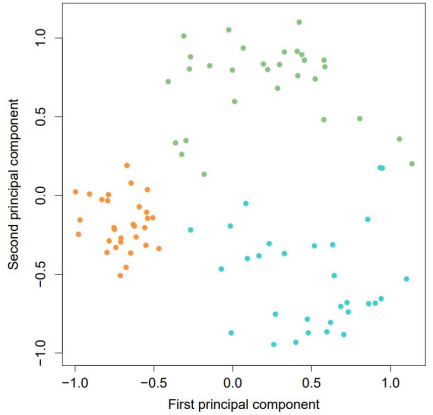
Reducción de dimensionalidad

La reducción de dimensionalidad es la transformación de los datos desde un espacio de altas dimensiones hacia un espacio de más bajas dimensiones tal que la representación en bajas dimensiones retiene las propiedades del espacio original en altas dimensiones. Una de las aplicaciones más utilizada de estos métodos es la visualización en dos dimensiones de datos multidimensionales.

Análisis de componentes principales

Cuando nos enfrentamos a un gran conjunto de variables correlacionadas, el análisis de componentes principales nos permite resumir este conjunto de características con un conjunto más pequeño que contiene sólo las más representativas que colectivamente explican la mayor variabilidad del conjunto de datos.





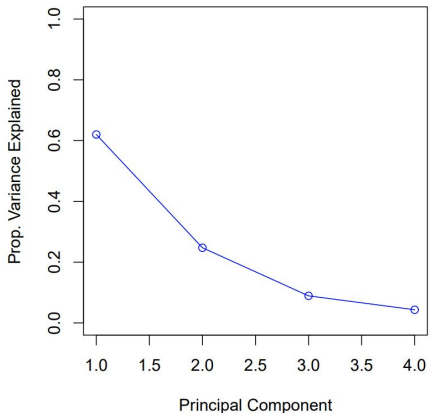
Componentes principales

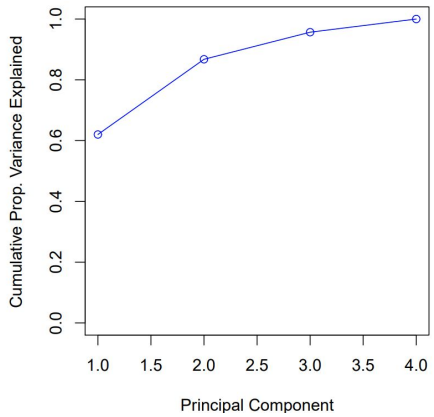
Los componentes principales son combinaciones lineales de las características del espacio original. Estas variables sintéticas que se calculan mediante el análisis de componentes principales resumen las direcciones en donde el espacio de características varían en mayor manera.

Varianza explicada

Cada uno de los componentes principales calculados por el análisis de componentes principales explica una porción de toda la varianza del conjunto de datos.

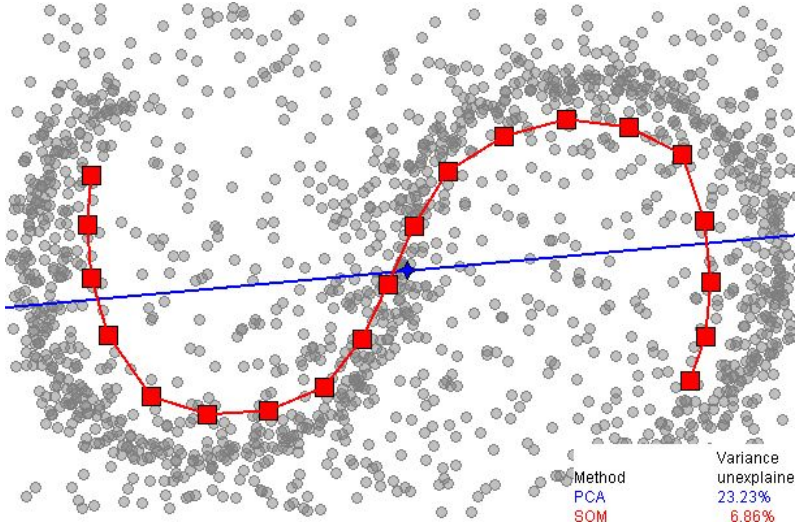
Podemos seleccionar tantos componentes principales para representar nuestros datos como varianza queramos explicar.





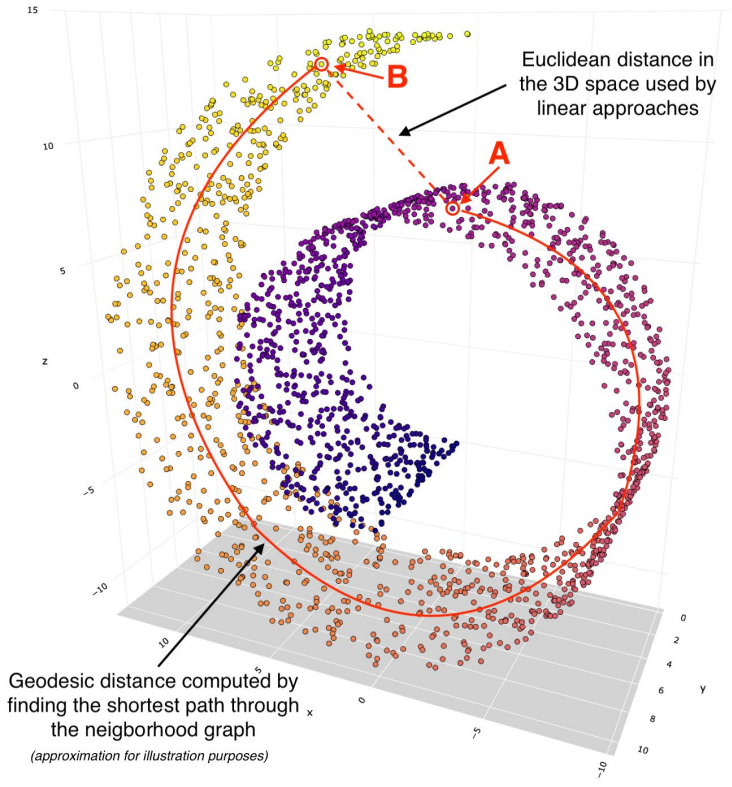
Reducción de dimensionalidad no lineal

La reducción de dimensionalidad no lineal se refiere a un conjunto de técnicas que buscan proyectar datos altamente dimensionales en un espacio de bajas dimensiones con el objetivo de visualizar los datos en bajas dimensiones o aprender el mapeo entre el espacio de alta y baja dimensionalidad.



Isomap

El algoritmo provee un método simple para estimar la geometría intrínseca de los datos, basado en un estimativo de cada punto y sus vecinos. Isomap es altamente eficiente y generalmente aplicable a un gran número de conjuntos de datos y dimensionalidades.



t-SNE

t-distributed stochastic neighbor embedding es un método estadístico para visualizar datos altamente dimensionales. Este método modela cada objeto de altas dimensiones en un espacio de dos o tres dimensiones de manera que objetos similares son modelados cerca y objetos disimilares son modelados lejanos con alta probabilidad.