KIBERNUM
IT Academy

Módulo 6
Clase 7

Aprendizaje de Máquina No Supervisado

Objetivos

- Describir características del algoritmo T-SNE
- Implementar una rutina en Python con T-SNE

T-SNE

¿Qué es T-SNE?

T-SNE (t-Distributed Stochastic Neighbor Embedding en inglés) es una **técnica de reducción de dimensionalidad** utilizada en el aprendizaje automático y la minería de datos. Su objetivo es **visualizar datos de alta dimensionalidad en un espacio de dos o tres dimensiones**, permitiendo una mejor comprensión de las relaciones entre los datos.

La técnica funciona **transformando los puntos de datos en probabilidades que representan las similitudes entre ellos**. Luego, busca una distribución de baja dimensionalidad que tenga probabilidades similares a las de alta dimensionalidad. Finalmente, representa los puntos de datos en el espacio de baja dimensionalidad para visualizarlos.

T-SNE es especialmente útil para visualizar datos complejos y encontrar patrones ocultos en grandes conjuntos de datos. Se utiliza en diversos campos, como la bioinformática, la visión por computadora y el análisis de redes sociales.

Intuición de t-SNE

El algoritmo t-SNE consiste en crear una distribución de probabilidad que represente las similitudes entre vecinos en un espacio de gran dimensión y en un espacio de menor dimensión. Por similitud, intentaremos convertir las distancias en probabilidades. Se puede conceptualizar en 3 pasos:

Paso 1

En el espacio de alta dimensionalidad, toma cada punto y calcula la similaridad con respecto a cada uno de los demás puntos.

Paso 2

Creamos un espacio dimensional más pequeño para representar nuestros datos. En éste, distribuimos los puntos de forma aleatoria y luego volvemos a calcular la similitud entre los puntos.

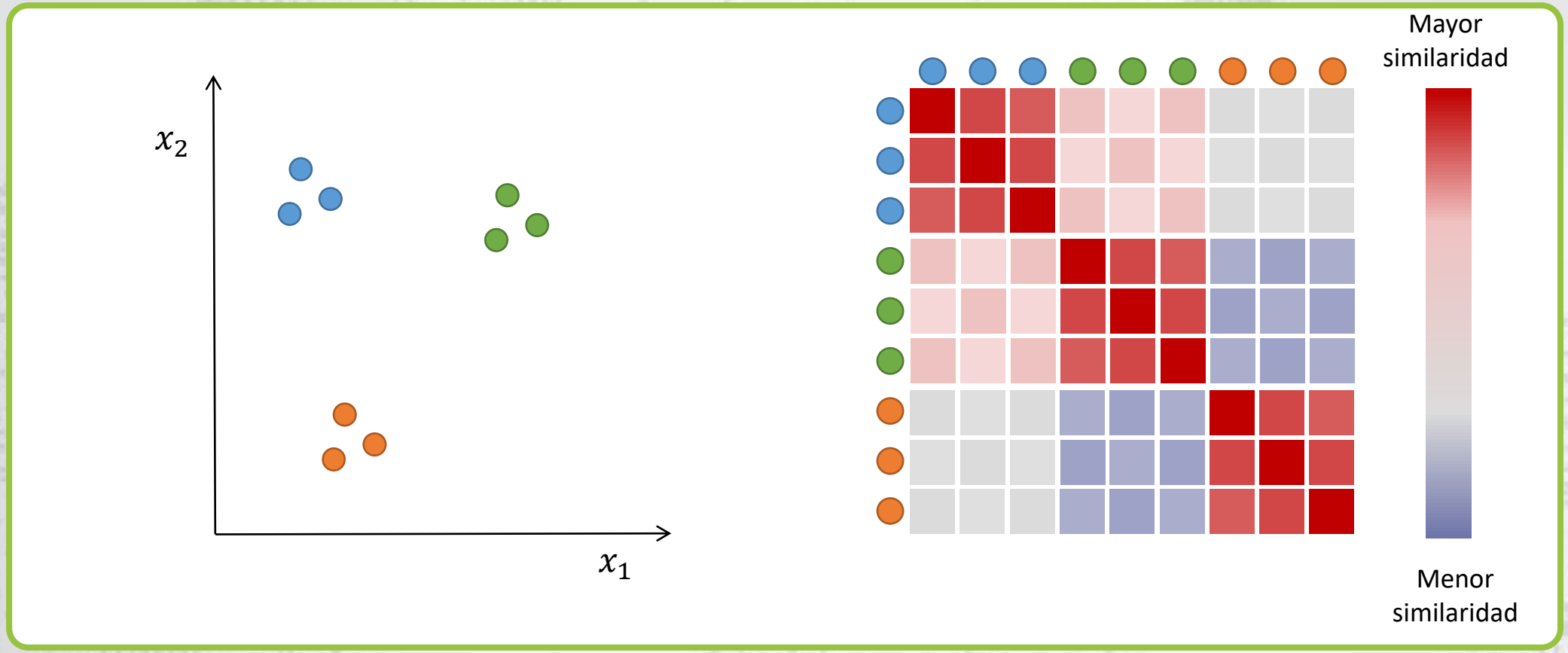
Paso 3

Ahora, ajustamos los puntos de forma de representar en el nuevo espacio las medidas de similitud lo más parecido al espacio original.

Intuición de t-SNE

PASO 1

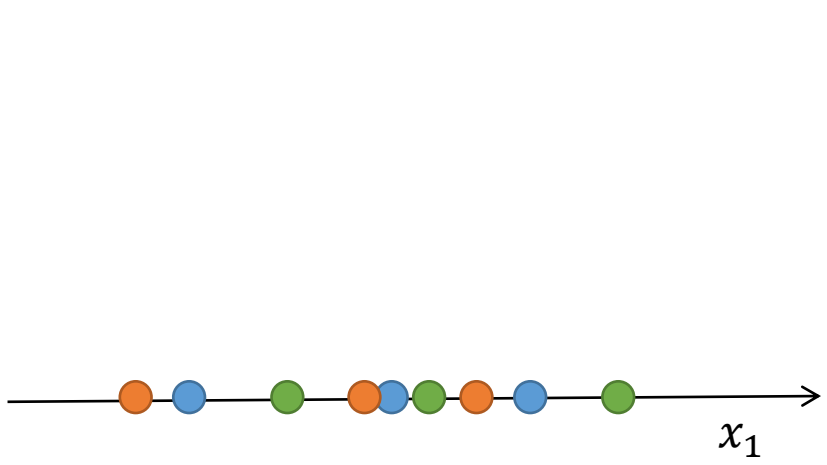
Sea el siguiente espacio dimensional, con dos dimensiones. Vamos a construir una matriz de similaridad entre todos los puntos.

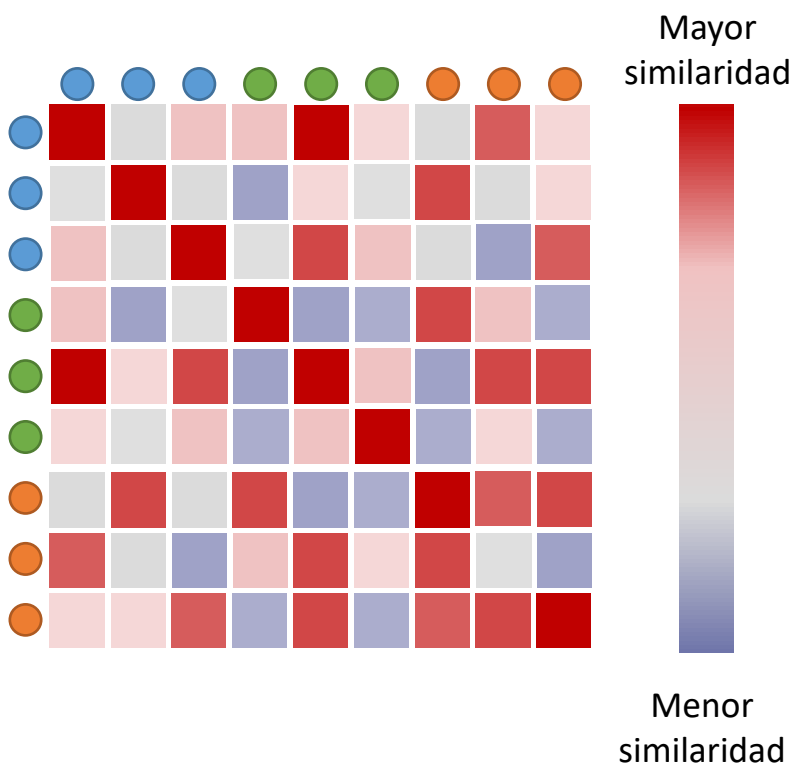


Intuición de t-SNE

PASO 2

Ahora crearemos un espacio dimensional menor y posicionaremos los puntos, de forma desordenada, y volveremos a calcular la similaridad entre los puntos, ahora, en este nuevo espacio con 1 dimensión.

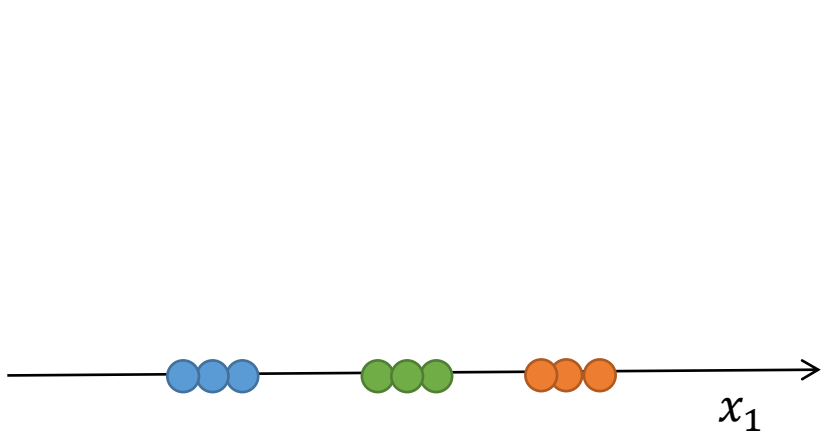


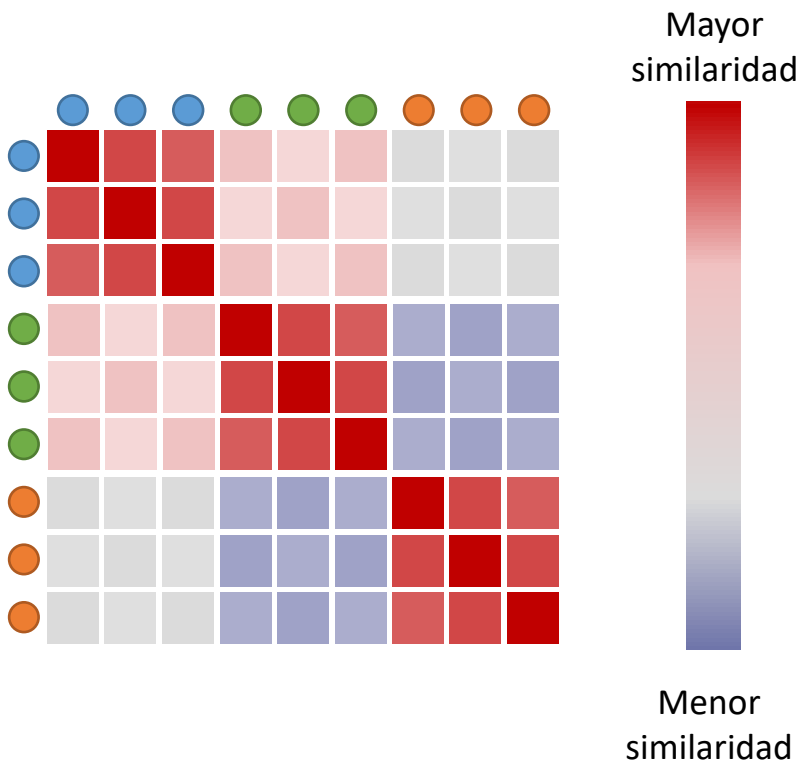


Intuición de t-SNE

PASO 3

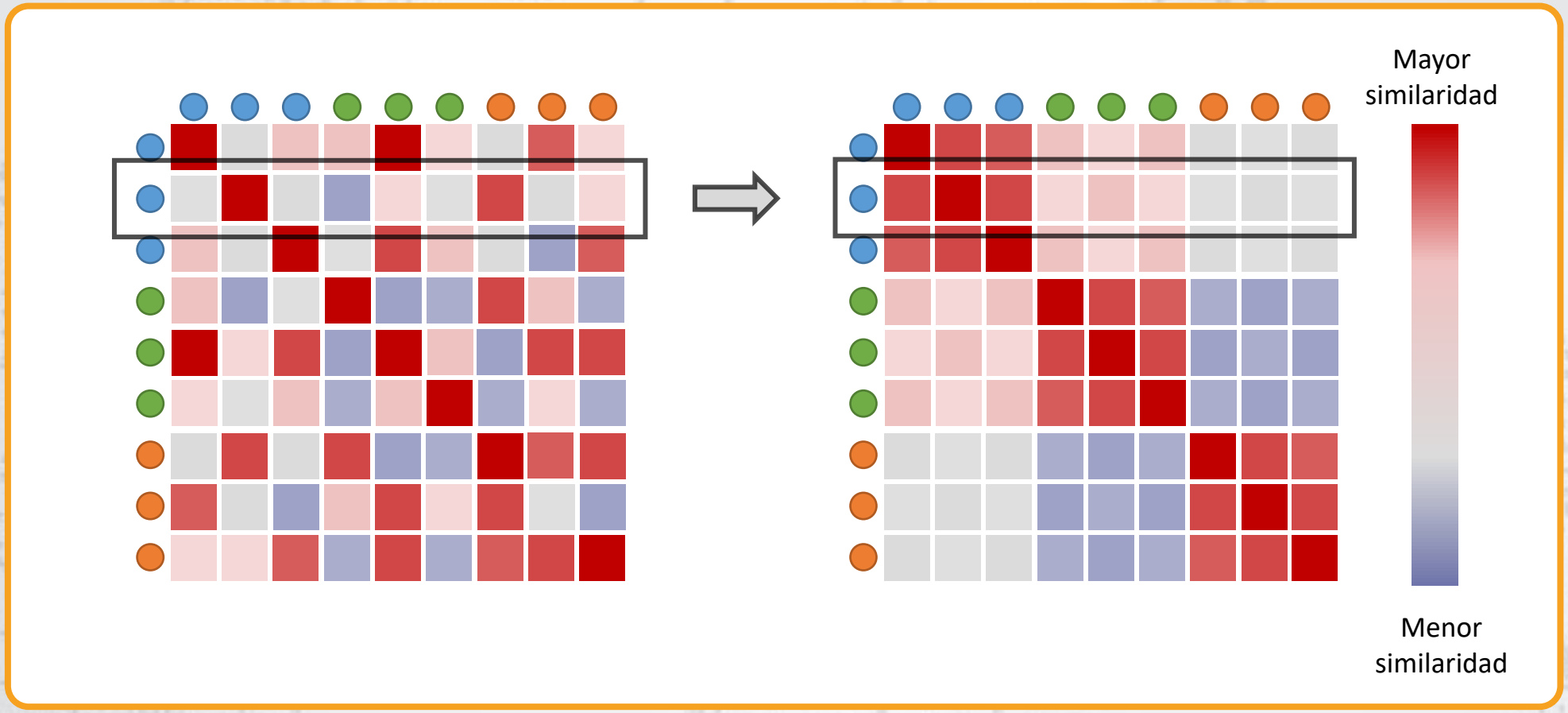
Ordenaremos los puntos en este espacio de menor dimensión con el propósito de mantener la misma matriz de similitud que en el espacio de mayor dimensionalidad. Este proceso es iterativo.





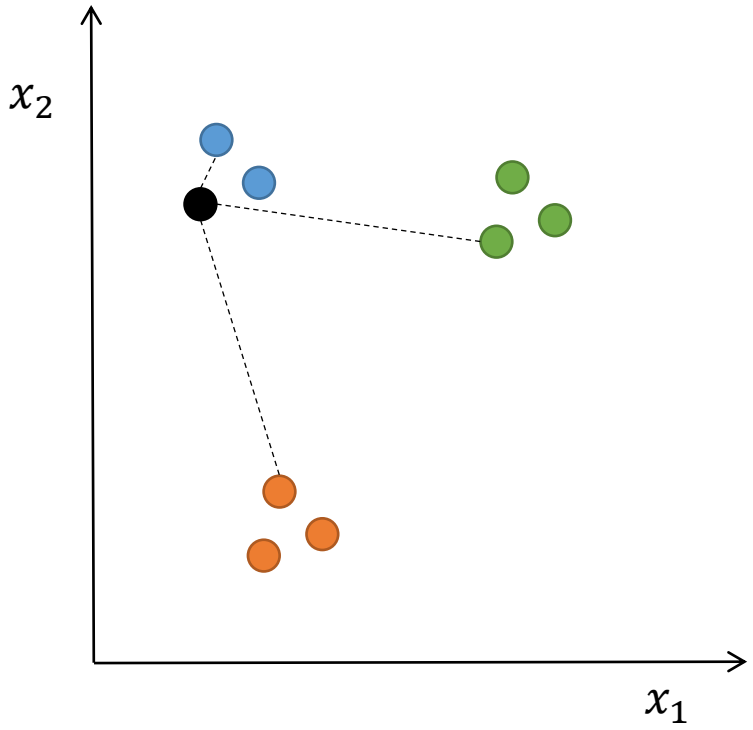
Intuición de t-SNE

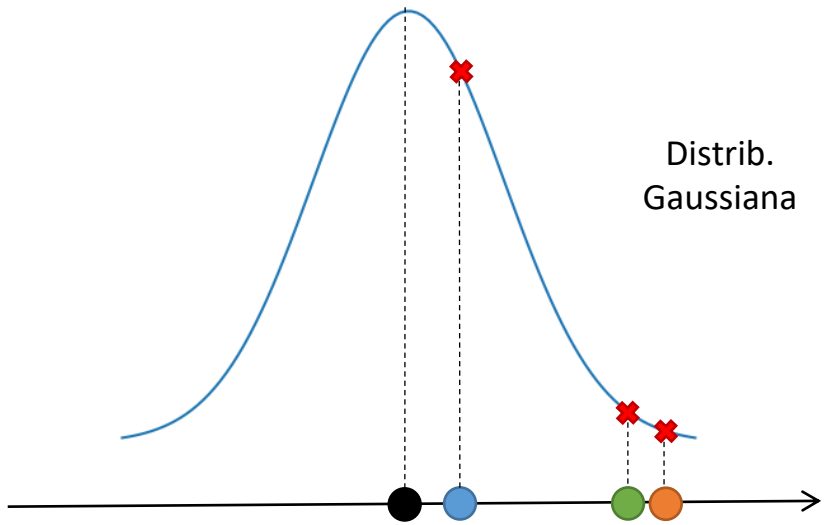
PASO 3
Es decir, cada punto se distribuye en el nuevo espacio hasta que las nueva similaridades se acerquen a las originales.



Intuición de t-SNE

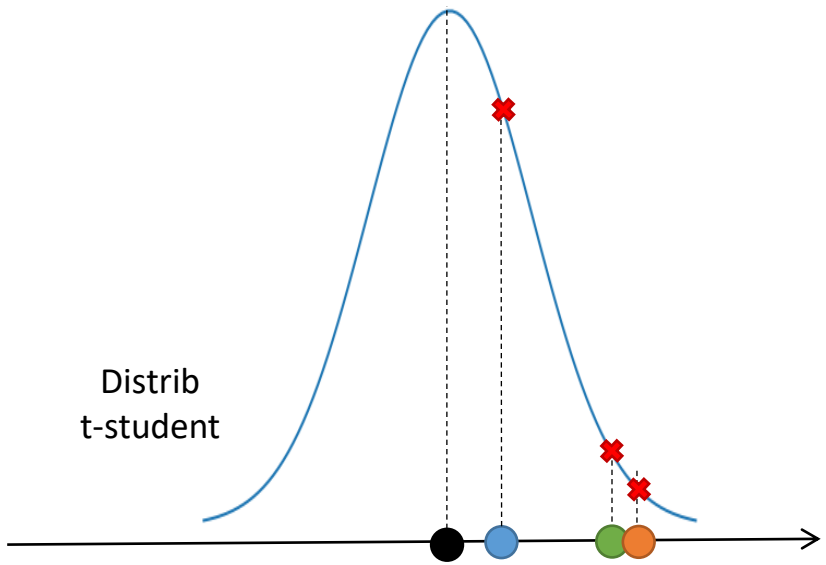
Para el cálculo de la similaridad en el paso 1, se utiliza una distribución gausiana. La desviación estándar se define por un valor llamado **perplejidad**, que corresponde al número de vecinos alrededor de cada punto. Este valor lo establece el usuario de antemano y permite estimar la desviación estándar de las distribuciones gaussianas definidas para cada punto xi. Cuanto mayor es la perplejidad, mayor es la variación.



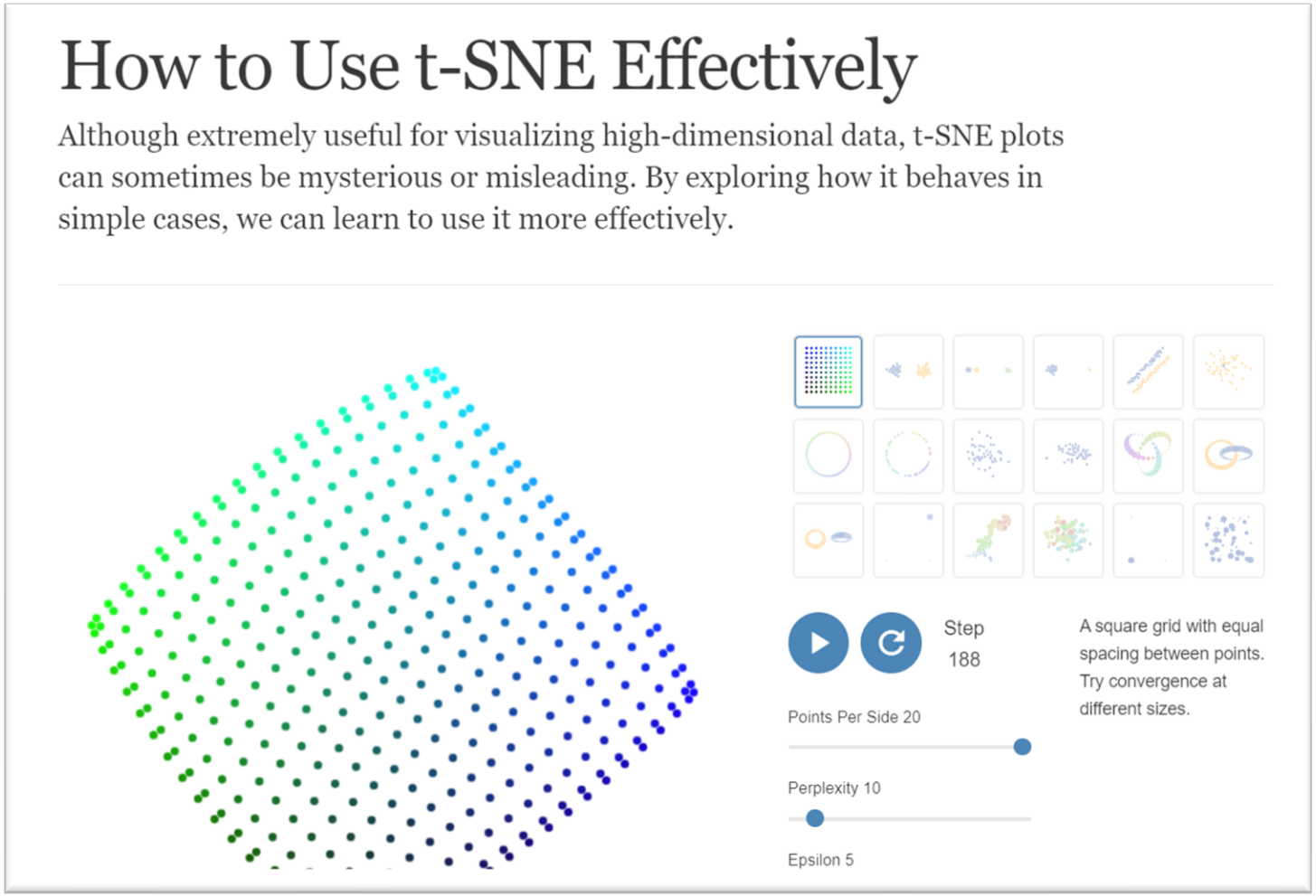


Intuición de t-SNE

Para el cálculo de la similaridad en el paso 3, se utiliza una distribución t-student.



Recursos Complementarios



https://distill.pub/2016/misread-tsne/

Playground

https://projector.tensorflow.org/

Ventajas

1. t-SNE **es muy efectivo para visualizar datos de alta dimensionalidad en dos o tres dimensiones**. Puede ayudar a encontrar patrones y estructuras en los datos que de otro modo serían difíciles de detectar.

2. t-SNE **es capaz de preservar las relaciones locales entre los datos**, lo que significa que los puntos cercanos en el espacio de alta dimensión seguirán siendo cercanos en el espacio de baja dimensión.

3. t-SNE es **capaz de encontrar agrupaciones naturales** de datos y separarlas en regiones distintas en la visualización de baja dimensión.

Desventajas

1. t-SNE **es computacionalmente costoso** y puede tardar mucho tiempo en ejecutarse en grandes conjuntos de datos.

2. t-SNE **no siempre preserva las distancias globales entre los datos**, lo que significa que los puntos que están muy separados en la alta dimensión pueden aparecer más cerca en la visualización de baja dimensión.

3. t-SNE **puede ser sensible a la elección de parámetros**, como la tasa de aprendizaje y la perplejidad, y puede producir resultados diferentes si se ejecuta varias veces con diferentes configuraciones de parámetros.

Dudas y consultas

Gracias

KIBERNUM