

REDUCCIÓN DE DIMENSIONES

Organización de datos, Argerich. FIUBA
Fecha: 23-06-22

INTRODUCCIÓN

La reducción de dimensiones se encarga de buscar una representación más compacta de un set de datos no solo para optimizar el uso de recursos sino también para otros usos: visualización de datos, compresión de imágenes, feature engineering, etc.

SVD

SVD

Surge de realizar una descomposición de una matriz. La SVD está íntimamente relacionada con el hecho de diagonalizar una matriz simétrica.

DESCOMPOSICIÓN DE MATRICES

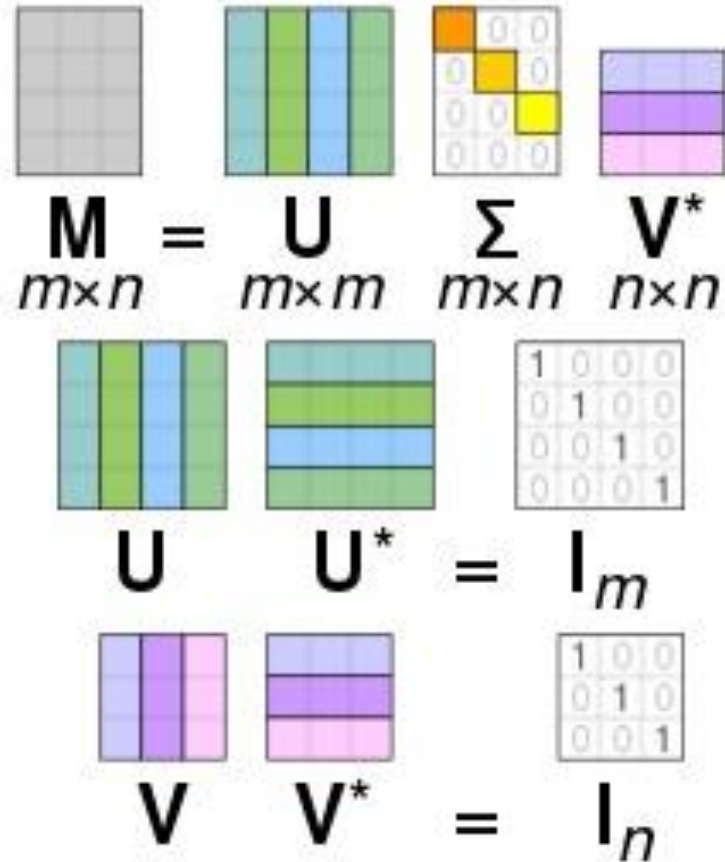
$$M = U \Sigma V^*$$

Las columnas de **U** son los autovectores de M por M traspuesto.

Sigma es la matriz más importante, es una matriz diagonal y sus valores son los llamados valores singulares, que son la raíz de los autovalores de M traspuesto por M , ordenado decrecientemente.

Las columnas de **V** son los autovectores de M traspuesto por M .

Objetivo: queremos encontrar una matriz de menor rango que la original que sea lo más parecida posible a nuestra matriz original



PCA

PCA

Para muchas aplicaciones es de utilidad **tener la posibilidad de visualizar la información, en particular en 2 o 3 dimensiones**. Uno de los usos que se le puede dar a PCA es reducir dimensiones para visualizar datos.

Es un método lineal,
utiliza planos.

Objetivo: encontrar las
componentes que maximicen
la varianza de los datos.

La varianza hace referencia
a la dispersión del
conjunto de datos respecto
de la media. Utiliza la
matriz de covarianza.

SVD vs PCA

La SVD es en general más eficiente y numéricamente más estable que PCA dado que no hace falta construir la matriz de covarianza y es por este motivo que como método genérico lineal para reducción de dimensiones se prefiere siempre usar la SVD.

MDS

MDS

Este algoritmo busca resolver el siguiente problema: Dada una matriz con las distancias entre ciudades que se desconocen, se quieren encontrar coordenadas en el plano para estas ciudades de forma tal que las distancias entre las mismas se aproximen lo más posible a las distancias de nuestra matriz.

T-SNE

T-SNE

Es un algoritmo para la representación de datos en dos y tres dimensiones. Es cuadrático con dimensiones.

Objetivo: a intentar que los puntos que estaban cerca en n dimensiones se mantengan cercanos en k dimensiones.

¿CUANDO LO USAMOS?

Cuando queremos ver qué está pasando con todas las dimensiones al mismo tiempo.

UMAP

UMAP

Es una técnica de aprendizaje para la reducción de dimensiones.

UMAP se construye a partir de un marco teórico basado en geometría y topología algebraica.

Se puede aplicar a cualquier cantidad de dimensiones.

Soporta nuevos puntos, se puede usar para predecir.
