La descomposición en valores singulares. Revisión del método

La descomposición en valores singulares (SVD, por sus siglas en inglés, *Singular Value Decomposition*) permite una representación de baja dimensión con base en un conjunto de componentes latentes que expresan las relaciones entre los datos. Al igual que el PCA, transforma variables correlacionadas en un conjunto de variables no correlacionadas. Sin embargo, es más general que el PCA, ya que proporciona dos conjuntos base de vectores, tanto para las columnas como para las filas de la matriz de datos. Entonces, dada una matriz $\boldsymbol{X}$ de dimensión $m \times n$, una descomposición en valores singulares de $\boldsymbol{X}$ es una factorización del tipo:

$$\boldsymbol{X} = \boldsymbol{U\Sigma V'}$$

En la cual,

$\boldsymbol{U} = matriz\ ortonormal\ de\ mxm\ (vectores\ columna\ ortonormales)$. Las columnas de esta matriz son los autovectores de la matriz $\boldsymbol{XX'}$ (vectores singulares izquierdos).

$\boldsymbol{V} = matriz\ ortonormal\ de\ nxn\ (vectores\ columna\ ortonormales)$. Las columnas de esta matriz son los autovectores de la matriz $\boldsymbol{X'X}$ (vectores singulares derechos).

$\boldsymbol{\Sigma} = matriz\ diagonal\ de\ mxn$. Los valores singulares de $\boldsymbol{X}$ son los σ_i y son iguales a las raíces cuadradas de los autovalores de $\boldsymbol{X'X}$ (o de $\boldsymbol{XX'}$), es decir: $\sigma_i = \sqrt{\lambda_i}$

Los valores singulares son no negativos y están ordenados en orden descendente en la matriz $\boldsymbol{\Sigma}$. Por otra parte, el número de valores singulares diferentes de cero constituye el rango r de la matriz $\boldsymbol{X}$. Además, si los datos están centrados alrededor de la media, SVD y PCA suministran la misma transformación.

¿Cómo utilizar SVD para reducción de la dimensionalidad?

SVD produce una representación exacta de cualquier matriz, en la cual $\boldsymbol{U}$ representa una base en $\mathbb{R}^m$ (para las filas de $\boldsymbol{X}'$, $espacio\ columna\ de\ \boldsymbol{X}$) y $\boldsymbol{V}$ representa una base en $\mathbb{R}^n$ (para las filas de

Universidad de los Andes | Vigilada Mineducación Reconocimiento como
Universidad: Decreto 1297 del 30 de mayo de 1964.
Reconocimiento personería jurídica: Resolución 28 del 23 de febrero de 1949 Minjusticia.

$\boldsymbol{X}$, $espacio\ fila\ de\ \boldsymbol{X}$). En esta representación SVD define un número r de componentes latentes en la matriz de datos (Fig. 1).

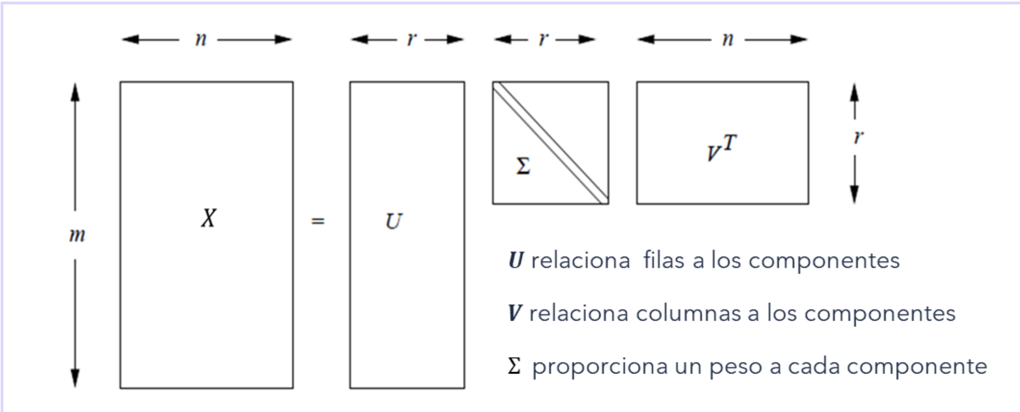
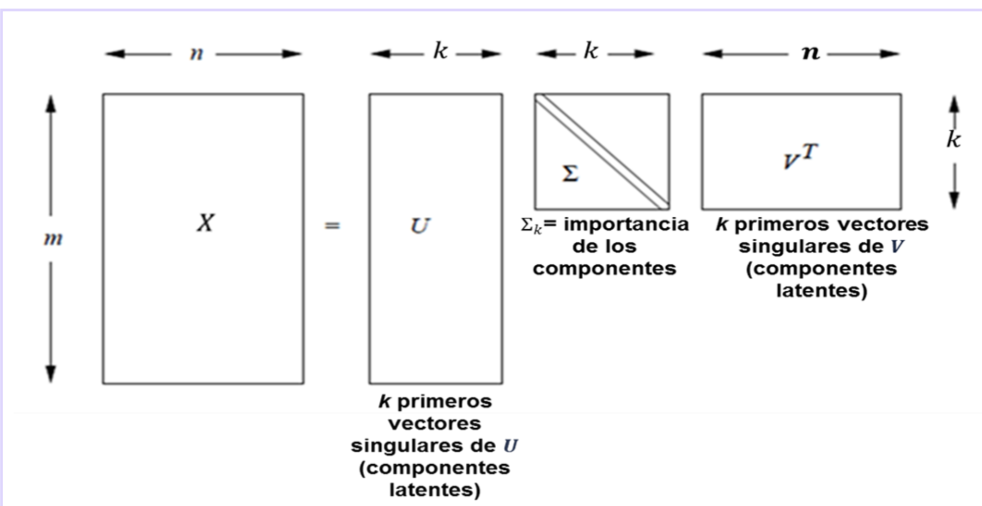


Fig. 1. Factorización SVD de una matriz.

La representación que proporciona SVD facilita la eliminación de los componentes menos importantes (aquellos con valores σ_i muy pequeños), produciendo así una representación aproximada con cualquier número deseado de dimensiones k. Este método se conoce como SVD reducida (Fig. 2).



Fig, 2. Reducción de la dimensionalidad con SVD.

Así, la SVD reducida expresa los datos en términos de k componentes latentes. La representación reducida en un sistema de base k del conjunto de datos sería:

$$\boldsymbol{X}^* = \boldsymbol{X}\boldsymbol{V_k}$$

$$X^* = XV_k = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1k} \\ v_{21} & v_{22} & \cdots & v_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nk} \end{bmatrix} = \boxed{\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1k} \\ c_{21} & c_{22} & \cdots & c_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mk} \end{bmatrix}}$$

Cada una de las m filas de $\boldsymbol{X}^*$ representa un dato original transformado en el nuevo sistema de coordenadas.

Aplicaciones de SVD

La descomposición en valores singulares es una herramienta poderosa con una amplia gama de aplicaciones en diversas disciplinas como, por ejemplo:

- Análisis de textos. En este contexto la SVD se utiliza para descomponer matrices término-documento en sus componentes principales (donde las filas son palabras y las columnas son documentos). Esto permite identificar tópicos latentes y la distribución de términos asociados a cada uno de estos, lo que facilita la comprensión de la estructura temática de un conjunto de documentos.
- Compresión de imágenes. La SVD se utiliza para representar la imagen original mediante una combinación lineal de componentes singulares. Al conservar solo las componentes más significativas se logra una compresión eficiente de la imagen con poca pérdida de información.
- Sistemas de recomendación. En el filtrado colaborativo se utiliza SVD para factorizar matrices de interacciones usuario-elemento, donde las filas son usuarios, las columnas son elementos (productos, películas, entre otros) y las celdas contienen las calificaciones que los usuarios han dado a los elementos. A partir de esta, se pueden encontrar las preferencias latentes de los usuarios y las características latentes de los elementos, mejorando así la calidad de las recomendaciones.

Bibliografía

"Singular Value Decomposition: Image Processing, Natural Language Processing, and social media". En: Nelson, H. (2022). Essential Math for AI. Next-Level Mathematics for Efficient and Successful AI Systems. Capítulo 6. O'Reilly.