ACÁMICA

¡Bienvenidas/os a Data Science!

Agenda

¿Cómo anduvieron?

Aplicación: Clustering para compresión de imágenes

Puesta en común: Entrega 5

Explicación: SVD

Break

Hands-On

Cierre

¿Cómo anduvieron?

Aplicaciones:

¿Para qué se puede usar un algoritmo de Clustering?

Puesta en Común:

Entrega 5

Repaso: Aprendizaje no supervisado

Solo datos

Llamamos Aprendizaje No Supervisado a los métodos para trabajar con datos (instancias) que no tienen asociados una etiqueta (una clase o un valor).

Los objetivos principales en Aprendizaje No Supervisado son:

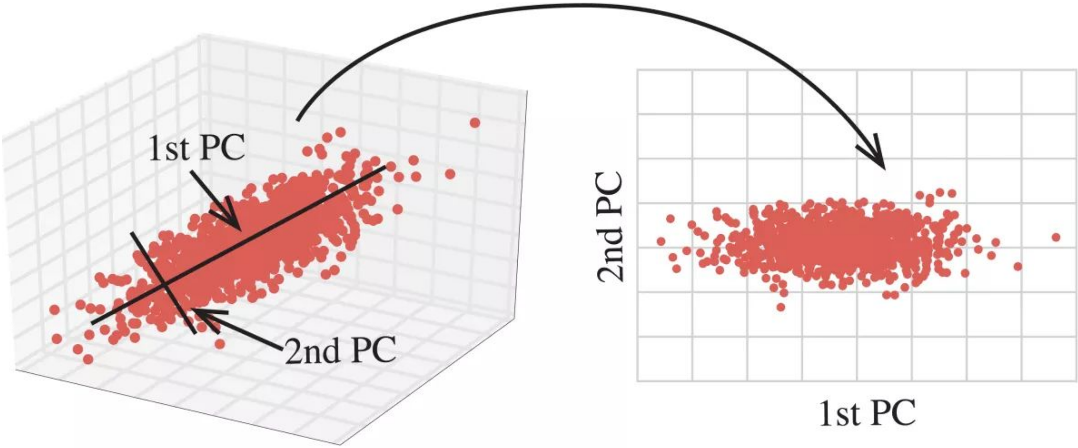
- Clustering
- Reducción de dimensionalidad

- Clustering
- Reducción de dimensionalidad

Aprendizaje No Supervisado • Reducción de la dimensionalidad

Buscamos reducir la cantidad de features de un dataset, pero reteniendo la mayor cantidad de "información" posible.

Reducción



3 features originales

2 features <u>nuevas</u>

¿Para qué sirve?

Reducir la cantidad de features en un dataset puede servir para:

- Reducir el input en un modelo de regresión o clasificación
- Compresión de archivos
- Visualización
- Detectar features relevantes en datasets
- Muchísimas mas cosas

Aprendizaje No Supervisado • Reducción de la dimensionalidad

¿Para qué sirve?

Reducir la cantidad de features en un dataset puede servir para:

- Reducir el input en un modelo de regresión o clasificación
- Compresión de archivos
- Visualización
- Detectar features relevantes en datasets
- Muchísimas mas cosas

¿Cómo se hace?

Algunos de los métodos de reducción de dimensionalidad son:

- PCA: Principal Component Analysis (usa SVD)
- MDS: Multidimensional scaling
- t-SNE: t-distributed Stochastic Neighbor Embedding
- Auto-Encoders (Se hace con Redes Neuronales)
- LDA: Linear Discriminant Analysis (si hay etiquetas de clases)

¿Para qué sirve?

Reducir la cantidad de features en un dataset puede servir para:

- Reducir el input en un modelo de regresión o clasificación
- Compresión de archivos
- Visualización
- Detectar features relevantes en datasets
- Muchísimas mas cosas

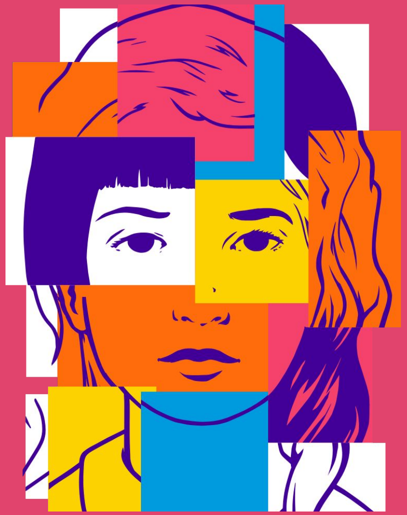
¿Cómo se hace?

Algunos de los métodos de reducción de dimensionalidad son:

- **PCA: Principal Component Analysis (usa SVD)**
- MDS: Multidimensional scaling
- t-SNE: t-distributed Stochastic Neighbor Embedding
- Auto-Encoders (Se hace con Redes Neuronales)
- LDA: Linear Discriminant Analysis (si hay etiquetas de clases)

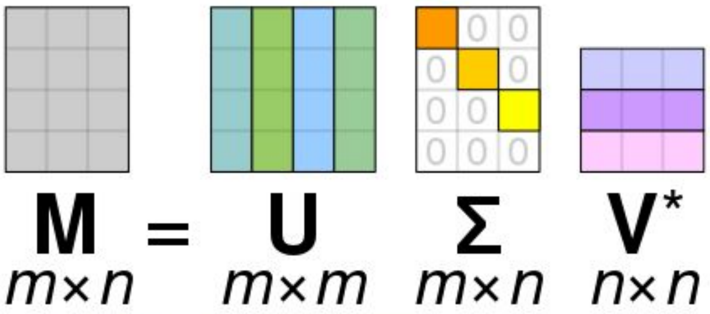
Aprendizaje No Supervisado

SVD (Singular Value Decomposition)



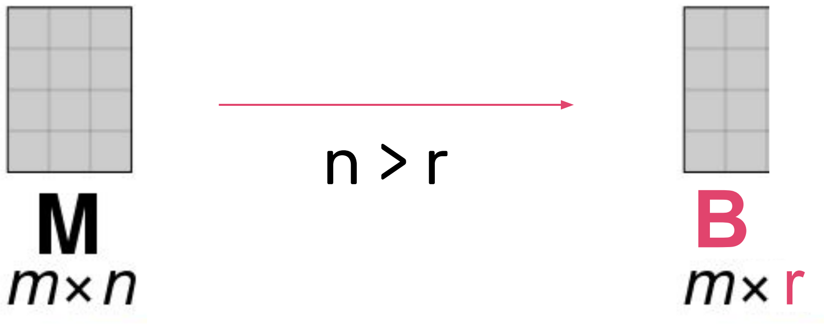
SVD • Definición

Es un método de álgebra lineal que nos permite representar cualquier matriz en términos de la multiplicación de otras 3 matrices.



SVD • ¿Para qué sirve?

Para MUCHAS COSAS. Es parte del corazón de muchos algoritmos numéricos (solución sis. lineal, pseudoinversa,etc.). En este contexto vamos a usarlo para “reducir” adecuadamente la matriz M (pasar de tener muchos features a tener menos, pero que sean buenos).



SVD • Álgebra

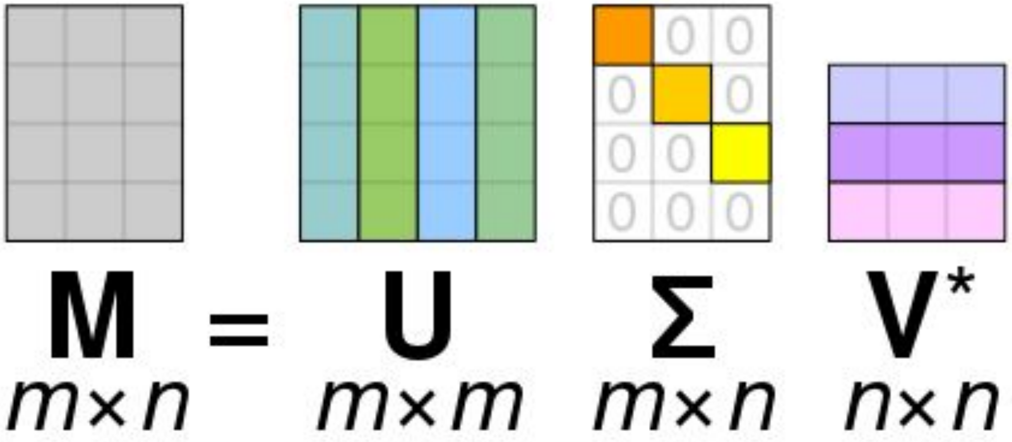
Se puede demostrar que a toda matriz M la podemos escribir como :

Matriz de Datos (m instancias, n features)

Matriz de vectores singulares por izquierda

Matriz de los <u>valores singulares</u>

Matriz de vectores singulares por derecha



Matriz Unitaria

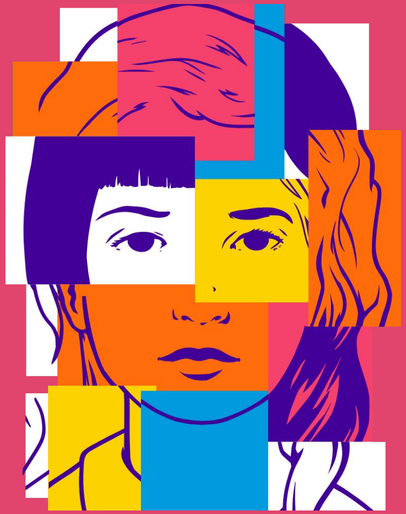
Matriz Diagonal

Matriz Unitaria

¿Y qué tiene que ver esto con todo lo que venimos hablando?

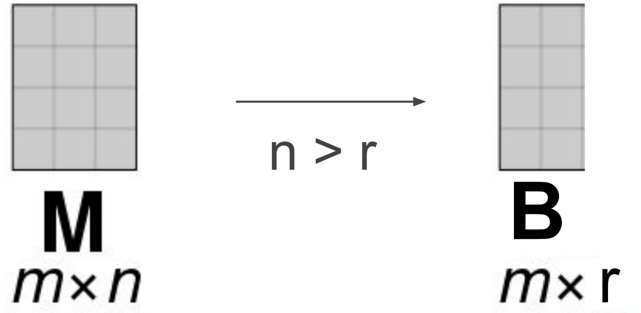
Aprendizaje No Supervisado

SVD truncado



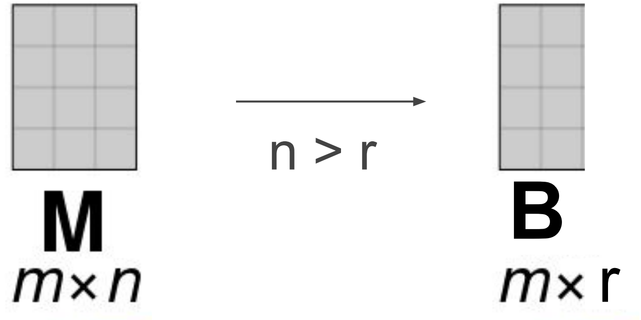
SVD **truncado**

Objetivo: queremos una nueva matriz B que reemplace a M, que tenga menos columnas (menos features).

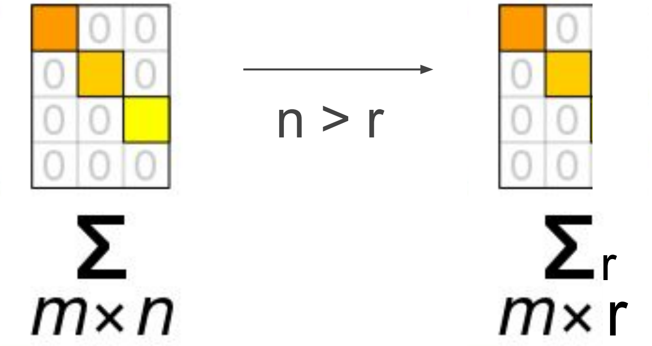


SVD truncado

Objetivo: queremos una nueva matriz B que reemplace a M, que tenga menos columnas (menos features).

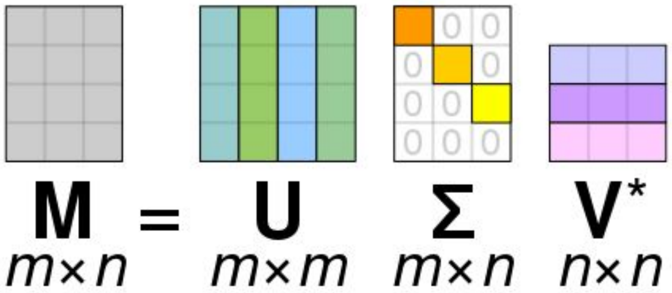


Idea de cómo lograrlo: si tomamos solo los r valores principales (elementos en la diagonal de Sigma) de valor más grande, podemos construir una matriz B que sea una "buena" reducción de M.



SVD truncado

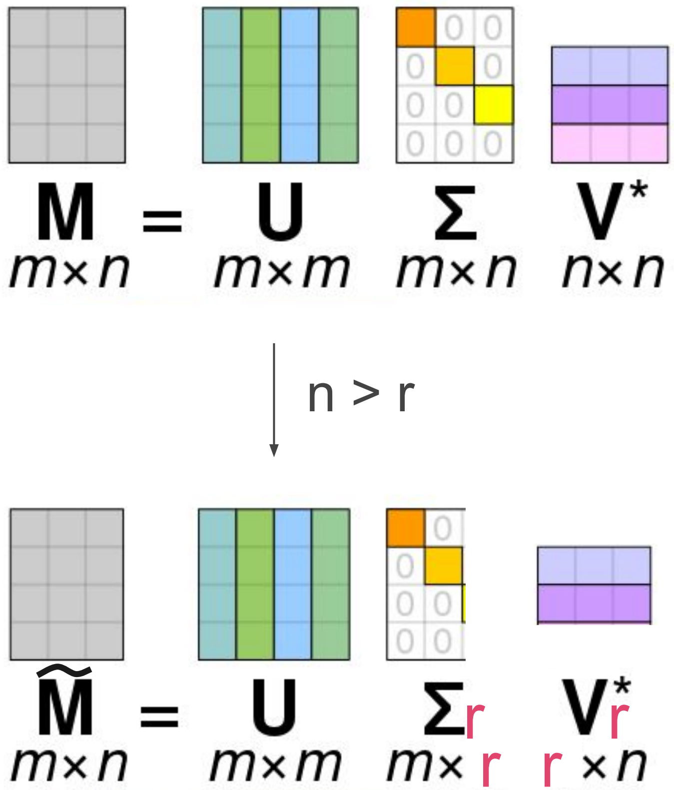
Matriz completa: es la M original, tiene toda la información.



SVD **truncado**

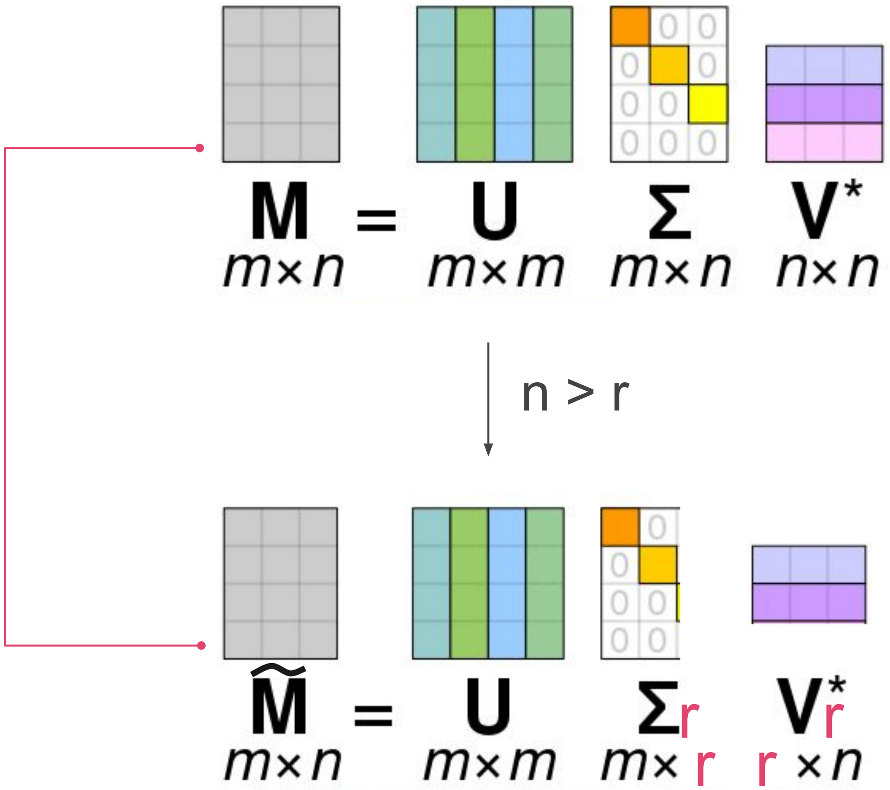
Matriz completa: es la M original, tiene toda la información.

Matriz truncada: perdimos información. Pero si tomamos un valor de r adecuado, M moño es muy parecida a M. Construimos una matriz B mas chica que M, esta es la matriz con la que vamos a trabajar.

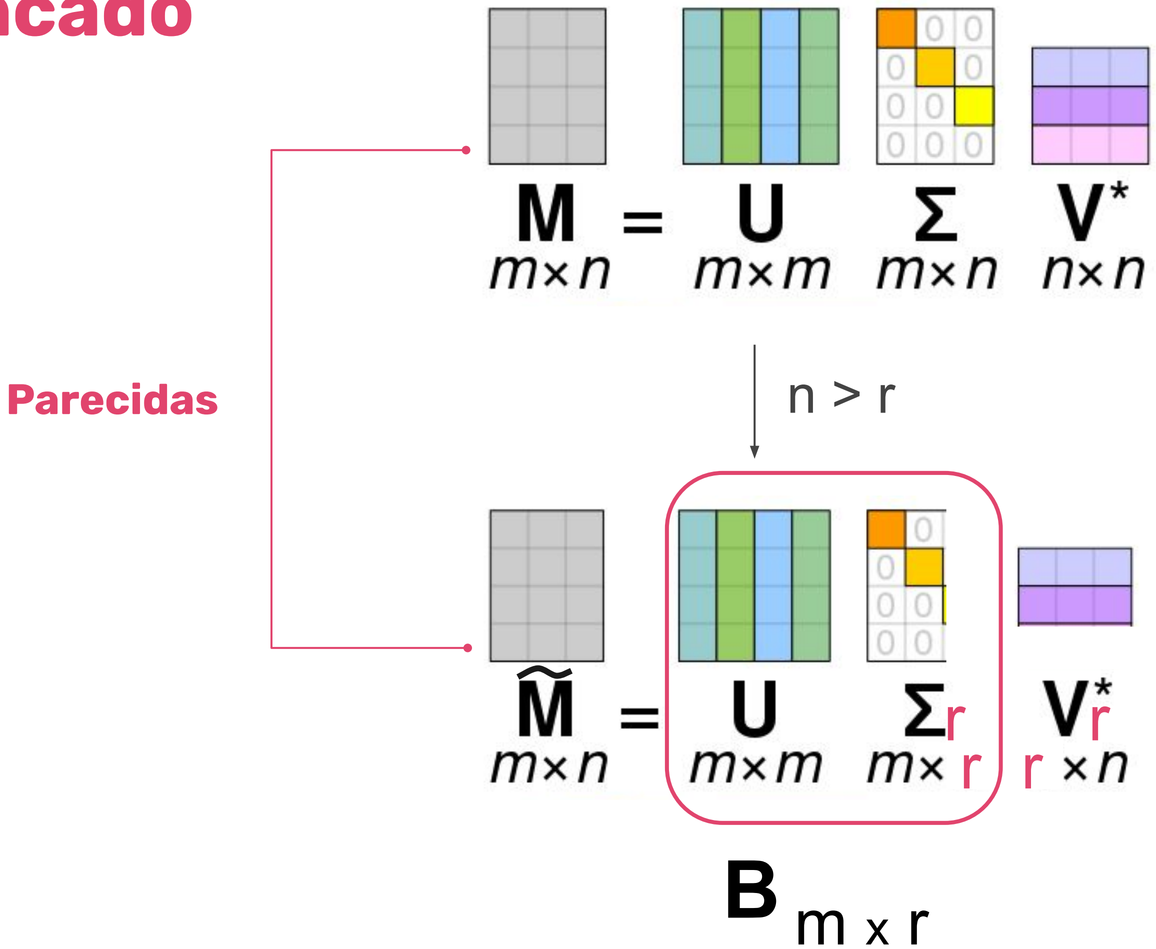


SVD **truncado**

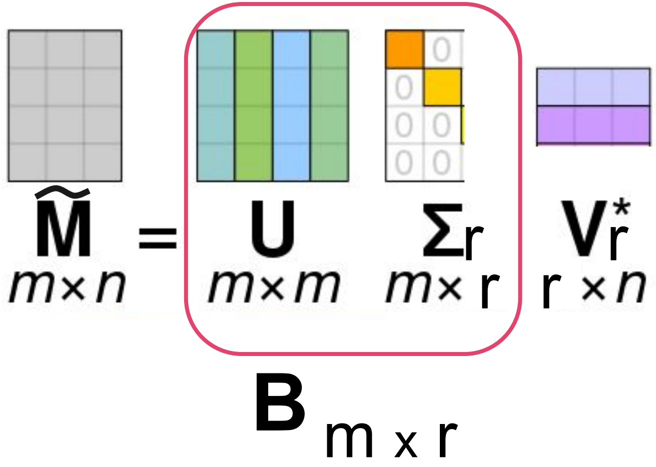
Parecidas



SVD truncado



SVD **truncado**

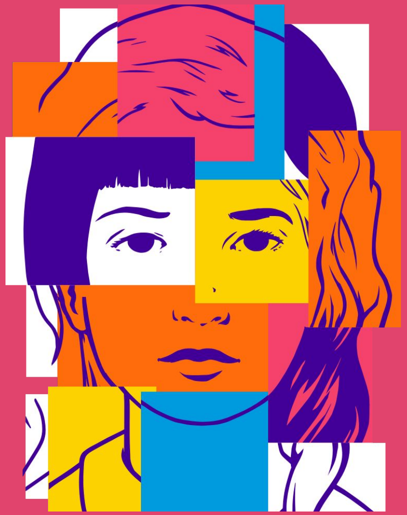


Esta matriz funciona como un diccionario para pasar del mundo de B al mundo de M.

Matriz con la que vamos a trabajar en vez de M, tiene la misma información que M moño.

Aprendizaje No Supervisado

Ejemplo conceptual SVD



SVD • Ejemplo 1

Tenemos un dataset de 7 usuarios y 5 peliculas. Cada usuario puso un valor entre 0 a 5 a cada película.

SVD • Ejemplo 1

Tenemos un dataset de 7 usuarios y 5 peliculas. Cada usuario puso un valor entre 0 a 5 a cada película.

$$
M_{7 \times 5} = \begin{array}{c} \begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Avatar} & \text{Titanic} & \text{Amelie} \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} \end{array}
$$

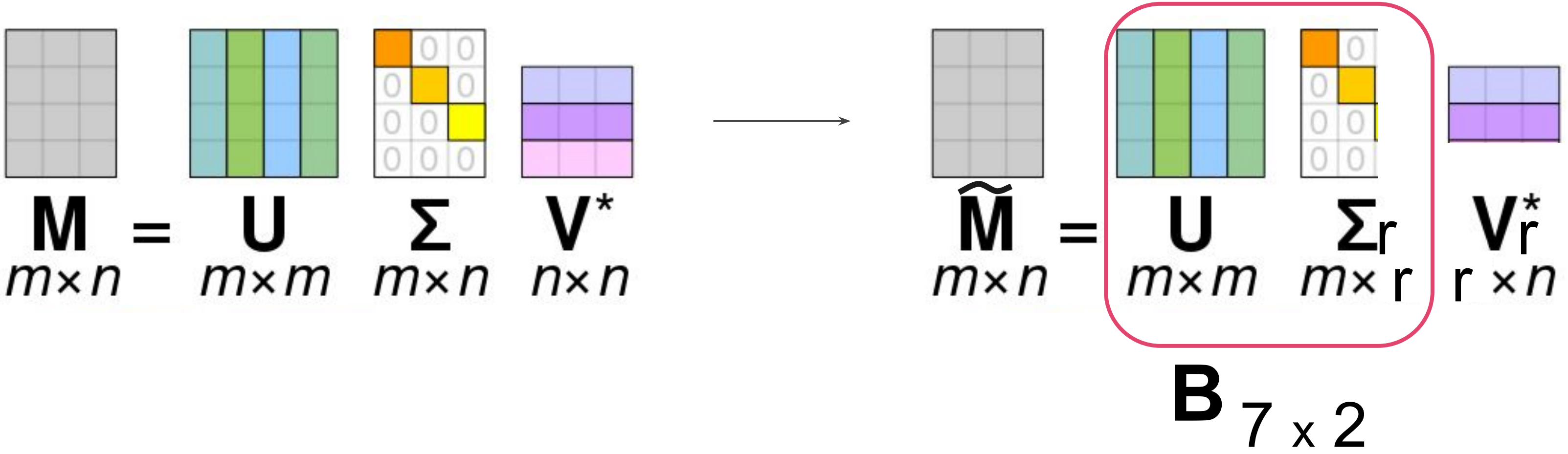
SVD • Ejemplo 1

Tenemos un dataset de 7 usuarios y 5 peliculas. Cada usuario puso un valor entre 0 a 5 a cada película.

$$M_{7 \times 5} = \begin{array}{c} \begin{matrix} \text{Matrix} & \text{Alien} & \text{Avatar} & \text{Titanic} & \text{Amelie} \end{matrix} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} \end{array} \begin{array}{l} \text{Usuario 1} \\ \text{Usuario 2} \\ \dots \end{array} \begin{array}{l} \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} \text{Ciencia Ficción} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \text{Románticas} \end{array}$$

SVD • Ejemplo 1

Buscamos una matriz B más con menos columnas que M. Proponemos usar un valor de r =2 es decir que B será de 7 x 2. Veamos como quedaría:



Esta vez usaremos solo los 2 valores singulares más grandes de SIgma.

SVD • Ejemplo 1

$$U_r = \begin{bmatrix} 0.13 & 0.02 \\ 0.41 & 0.07 \\ 0.55 & 0.09 \\ 0.68 & 0.11 \\ 0.15 & -0.59 \\ 0.07 & -0.73 \\ 0.07 & -0.29 \end{bmatrix} \quad \Sigma_r = \begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix} \quad V_r^* = \begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & -0.12 & -0.69 & -0.69 \end{bmatrix}$$

SVD • Ejemplo 1

$$
U_r = \begin{bmatrix} 0.13 & 0.02 \\ 0.41 & 0.07 \\ 0.55 & 0.09 \\ 0.68 & 0.11 \\ 0.15 & -0.59 \\ 0.07 & -0.73 \\ 0.07 & -0.29 \end{bmatrix}
\quad
\Sigma_r = \begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix}
\quad
V_r^* = \begin{array}{c}
\begin{array}{ccccc} \overbrace{\text{Matrix} \quad \text{Alien} \quad \text{Avatar}}^{\text{Ciencia Ficción}} & \overbrace{\text{Titanic} \quad \text{Amelie}}^{\text{Románticas}} \end{array} \\
\begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & -0.12 & -0.69 & -0.69 \end{bmatrix}
\end{array}
$$

Pesos: X Z

SVD • Ejemplo 1

$$U_r = \begin{bmatrix} 0.13 & 0.02 \\ 0.41 & 0.07 \\ 0.55 & 0.09 \\ 0.68 & 0.11 \\ 0.15 & -0.59 \\ 0.07 & -0.73 \\ 0.07 & -0.29 \end{bmatrix} \quad \Sigma_r = \begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix}$$

$$\begin{array}{r c} & \begin{array}{ccccc} \multicolumn{3}{c}{\overbrace{\hphantom{\text{Matrix} \quad \text{Alien} \quad \text{Avatar}}}^{\text{Ciencia Ficción}}} & \multicolumn{2}{c}{\overbrace{\hphantom{\text{Titanic} \quad \text{Amelie}}}^{\text{Románticas}}} \\ \text{Matrix} & \text{Alien} & \text{Avatar} & \text{Titanic} & \text{Amelie} \end{array} \\ V^*_r = & \begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & -0.12 & -0.69 & -0.69 \end{bmatrix} \end{array}$$

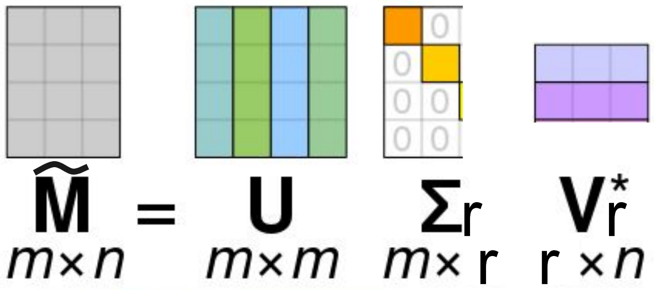
Pesos: X Z

- Ahora cada Usuario estará identificado por dos features X y Z. Notemos que los primeros 4 usuarios tienen un valor alto de X y bajo de Z. En los otros 3, se da al revés.
- Los features encontrados corresponden a los géneros.

SVD • Ejemplo 1

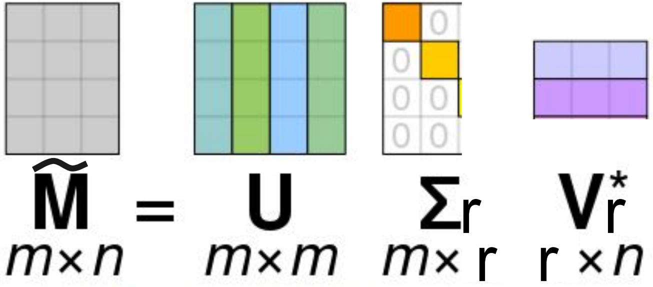
Pasamos de identificar a cada usuario con un puntaje al género de las películas en lugar de a las películas en sí, pasamos de 5 a 2 features.

Cuanta información perdemos por usar B en lugar de M?



SVD • Ejemplo 1

Pasamos de identificar a cada usuario con un puntaje al género de las películas en lugar de a las películas en sí, pasamos de 5 a 2 features.



Cuanta información perdemos por usar B en lugar de M?

$$M = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix}$$

$$\widetilde{M} = \begin{bmatrix} 0.92 & 0.95 & 0.92 & 0.01 & 0.01 \\ 2.91 & 3.01 & 2.91 & -0.01 & -0.01 \\ 3.90 & 4.04 & 3.90 & 0.01 & 0.01 \\ 4.82 & 5.00 & 4.82 & 0.03 & 0.03 \\ 0.70 & 0.53 & 0.70 & 4.11 & 4.11 \\ -0.69 & 1.34 & -0.69 & 4.78 & 4.78 \\ 0.32 & 0.23 & 0.32 & 2.01 & 2.01 \end{bmatrix}$$

Estamos muy cerca!!

Hiperparámetro r

SVD • Hiperparámetro r

¿Cómo podríamos elegir el valor de r?

SVD • Hiperparámetro r

¿Cómo podríamos elegir el valor de r?

Una posibilidad es **mirar la distancia entre M y M moño.**

$$\| M - \widetilde{M} \|_F = \sqrt{\sum_{ij} (M_{ij} - \widetilde{M}_{ij})^2}$$

El método de SVD nos GARANTIZA que elegimos los mejores r vectores (combinaciones de features) para minimizar esta norma!

SVD • Hiperparámetro r

¿Cómo podríamos elegir el valor de r?

Una posibilidad es **mirar la distancia entre M y M moño.**

$$\| M - \widetilde{M} \|_F = \sqrt{\sum_{ij} (M_{ij} - \widetilde{M}_{ij})^2}$$

El método de SVD nos GARANTIZA que elegimos los mejores r vectores (combinaciones de features) para minimizar esta norma!

SVD • Hiperparámetro r

¿Cómo podríamos elegir el valor de r?

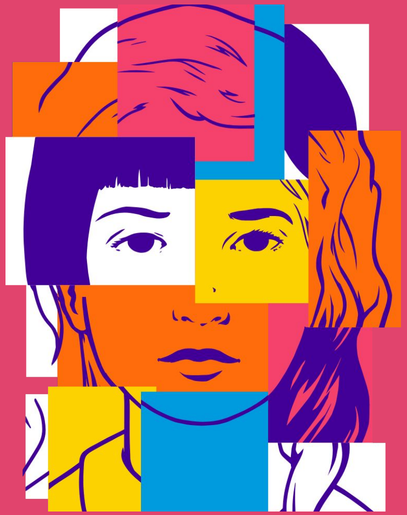
Una posibilidad es **mirar la distancia entre M y M moño.**

Otra posibilidad es **tener algún criterio sobre el peso relativo** de los valores singulares seleccionados respecto a la suma de todos. (Es más costoso, hay que calcular todos los valores singulares)

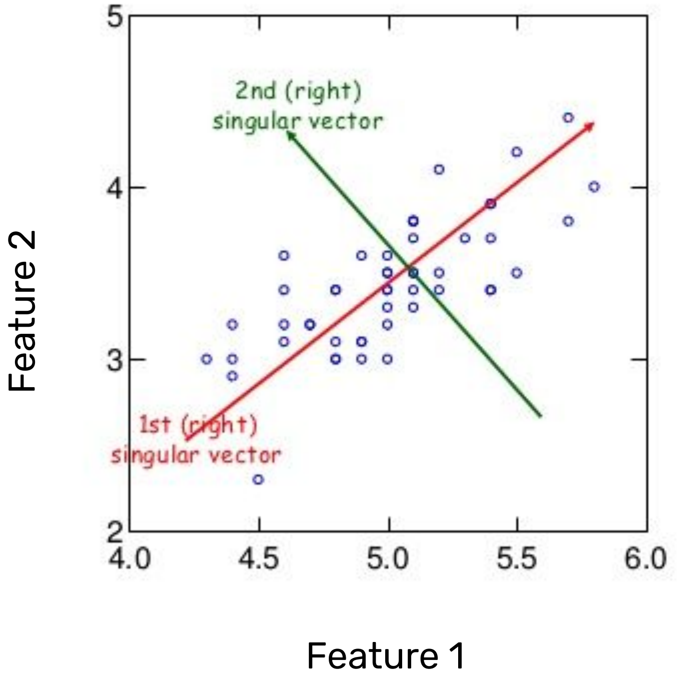
¿Y no hay algo un poco más visual?

Aprendizaje No Supervisado

Representación gráfica SVD

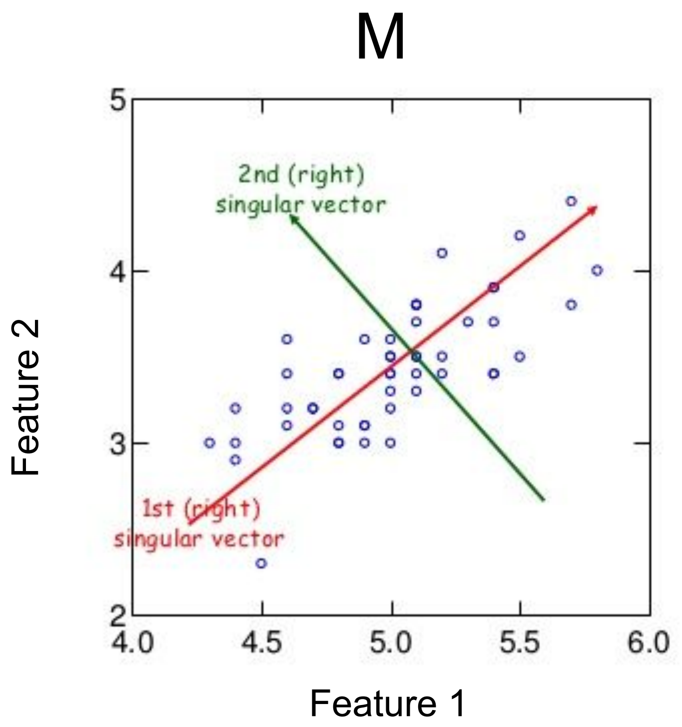


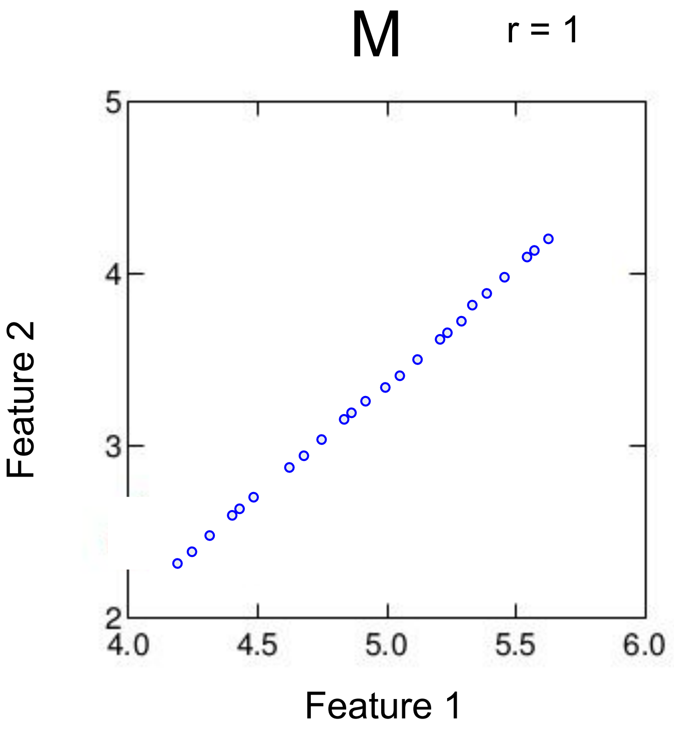
SVD • Representación gráfica



- El espacio original tiene 2 coordenadas, 2 features. Esto sirve para definir la posición de todas las instancias del dataset (cada punto azul).
- SVD nos da dos nuevos vectores, el 1er y 2do vector singular. Si usamos ambos como coordenadas, podemos definir perfecto la posición de cada punto.
- Veamos qué pasa si ahora sólo usamos el primer vector singular para definir los puntos.

SVD • Representación gráfica






¡BREAK!
Ph. Credit: Drew Coffmann

Hands-on training

Hands-on training

DS_Encuentro_40_SVD.ipynb

Para la próxima

1. Terminar de ver los videos de Reducción de Dimensionalidad.
2. Completar los notebooks de hoy y atrasados.
3. Si están leyendo sobre PCA, pueden jugar con esta página: http://setosa.io/ev/principal-component-analysis/.
4. Preparar el relato “Data Science en mi vida”.

ACÁMICA