

ACÀMICA

¡Bienvenidas/os a Data Science!



Agenda

¿Cómo anduvieron?

Actividad: Data Science en mi vida

Repaso: SVD

Explicación: PCA

Break

Hands-on training

Cierre



¿Dónde estamos?



¿Cómo anduvieron?



Actividad: Data Science en mi vida



Data Science en mi vida

¡Preparen sus charlas relámpago!

En 7 minutos con 7 slides comparte con tus compañeros:

En qué problemas estás aplicando lo aprendido en Data Science y cómo lo estás haciendo.

O bien, en qué problemas te gustaría aplicar Data Science y cómo lo harías.

¡Elige algún tema o proyecto que te interese y relaciónalo con lo aprendido!

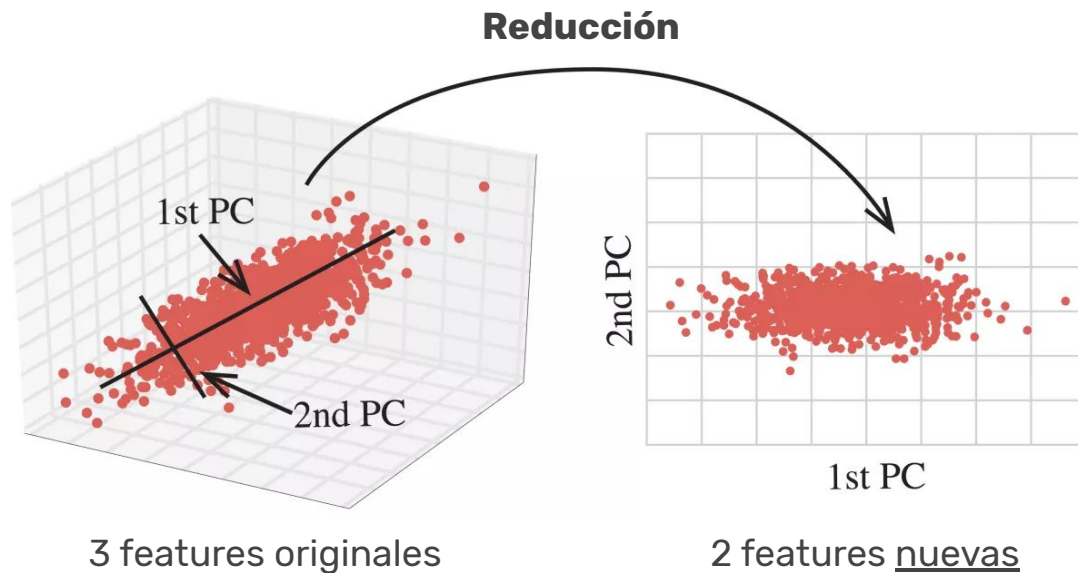
Repaso: Reducción de la dimensionalidad



- 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.



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

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)

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)

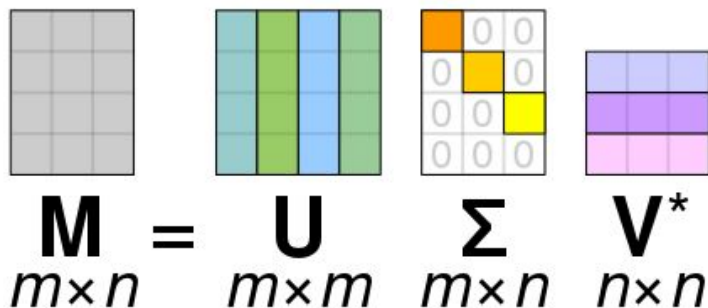
Aprendizaje No Supervisado

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.



The diagram illustrates the SVD decomposition of a matrix M into three matrices U , Σ , and V^* . Each matrix is represented by a grid of colored squares:

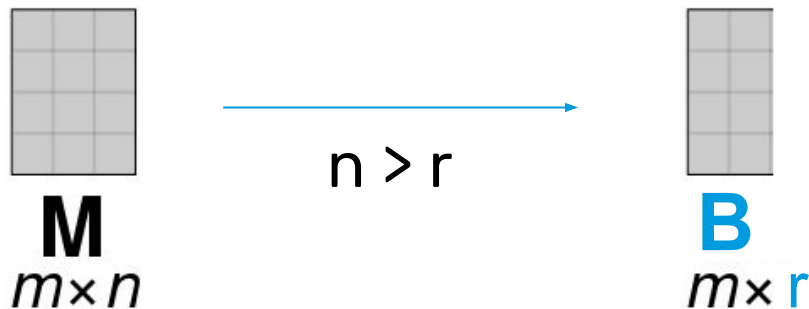
- M is a 4×4 matrix represented by a gray grid.
- U is a 4×4 matrix represented by a grid with four vertical columns of different colors: light blue, green, light blue, and green.
- Σ is a 4×4 matrix represented by a grid with three non-zero elements (orange, yellow, and yellow) on the main diagonal and zeros elsewhere.
- V^* is a 4×4 matrix represented by a grid with four horizontal rows of different colors: light blue, purple, purple, and pink.

The equation is shown as:

$$\begin{matrix} \text{Grid} & = & \text{Grid} & \text{Grid} & \text{Grid} \\ \mathbf{M} & = & \mathbf{U} & \mathbf{\Sigma} & \mathbf{V}^* \\ m \times n & & m \times m & m \times n & n \times n \end{matrix}$$

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

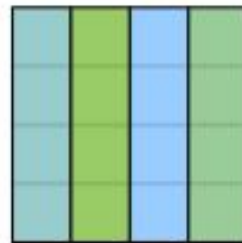
Se puede demostrar que a toda matriz M la podemos escribir como :

Matriz de Datos
(m instancias,
 n features)



$$\mathbf{M}_{m \times n}$$

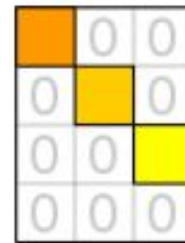
Matriz de
vectores
singulares por
izquierda



$$\mathbf{U}_{m \times m}$$

Matriz
Unitaria

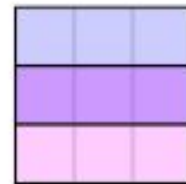
Matriz de los
valores
singulares



$$\mathbf{\Sigma}_{m \times n}$$

Matriz
Diagonal

Matriz de
vectores
singulares por
derecha



$$\mathbf{V}^*_{n \times n}$$

Matriz
Unitaria

Aprendizaje No Supervisado

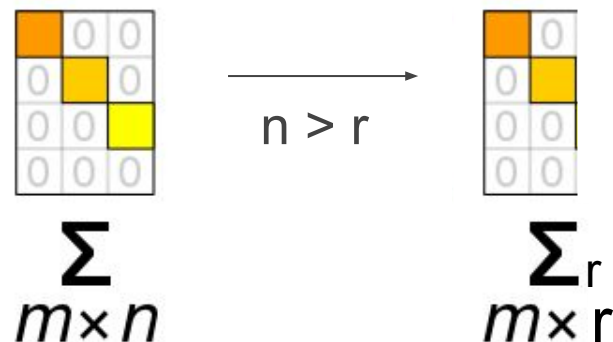
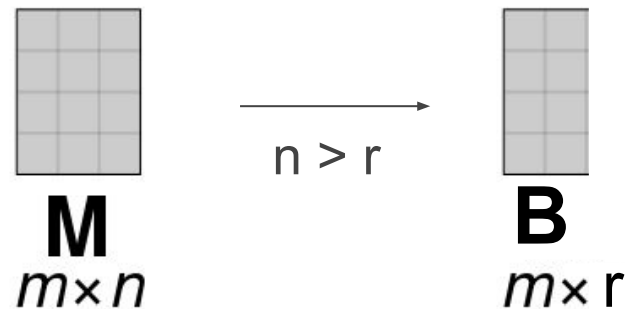
SVD truncado



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.

$$\begin{matrix} \text{4x4} & & \text{4x4} & & \text{4x4} & & \text{4x4} \\ \mathbf{M} & = & \mathbf{U} & & \mathbf{\Sigma} & & \mathbf{V}^* \\ m \times n & & m \times m & & m \times n & & n \times n \end{matrix}$$

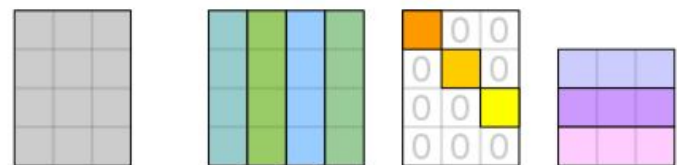
Matriz truncada: perdimos información. Pero si tomamos un valor de r adecuado, \hat{M} moño es muy parecida a M . Construimos una matriz B mas chica que M , esta es la matriz con la que vamos a trabajar.

$\downarrow n > r$

$$\begin{matrix} \text{4x4} & & \text{4x4} & & \text{4x4} & & \text{2x4} \\ \hat{\mathbf{M}} & = & \mathbf{U} & & \mathbf{\Sigma}_r & & \mathbf{V}_r^* \\ m \times n & & m \times m & & m \times r & & r \times n \end{matrix}$$

SVD truncado


Parecidas



The diagram illustrates the full SVD decomposition of a matrix M (represented by a 4x4 gray grid) into three components: U (a 4x4 matrix with 4 colored columns: teal, green, blue, and green), Σ (a 4x4 diagonal matrix with 4 colored squares: orange, yellow, yellow, and yellow), and V^* (a 4x4 matrix with 4 colored rows: light blue, light blue, light blue, and light blue).

$$\begin{matrix} \text{4x4} & & \text{4x4} & & \text{4x4} & & \text{4x4} \\ \mathbf{M} & = & \mathbf{U} & & \mathbf{\Sigma} & & \mathbf{V}^* \\ m \times n & & m \times m & & m \times n & & n \times n \end{matrix}$$

$n > r$



The diagram illustrates the truncated SVD decomposition of a matrix M (represented by a 4x4 gray grid) into three components: U (a 4x4 matrix with 4 colored columns: teal, green, blue, and green), Σ_r (a 4x4 diagonal matrix with 4 colored squares: orange, yellow, yellow, and yellow), and V_r^* (a 4x4 matrix with 4 colored rows: light blue, light blue, light blue, and light blue).

$$\begin{matrix} \text{4x4} & & \text{4x4} & & \text{4x4} & & \text{4x4} \\ \mathbf{\tilde{M}} & = & \mathbf{U} & & \mathbf{\Sigma}_r & & \mathbf{V}_r^* \\ m \times n & & m \times m & & m \times r & & r \times n \end{matrix}$$

SVD truncado

$$\hat{\mathbf{M}}_{m \times n} = \mathbf{U}_{m \times m} \mathbf{\Sigma}_r_{m \times r} \mathbf{V}_r^*_{r \times n}$$

B
 $m \times r$



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

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

SVD • Hiperparámetro r

¿Cómo podríamos elegir el valor de r?

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

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

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

Full-Rank Dog



Rank 200 Dog



Rank 100 Dog



Rank 50 Dog



Rank 30 Dog



Rank 20 Dog



Rank 10 Dog



Rank 3 Dog

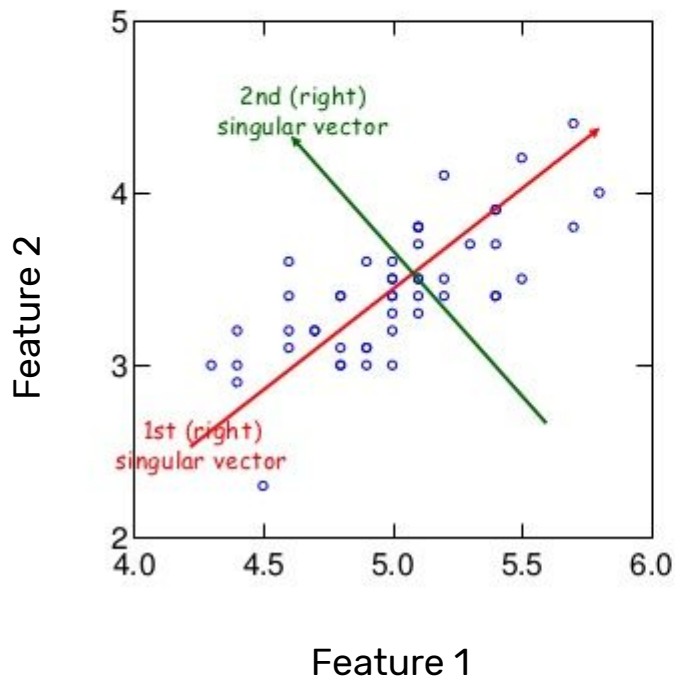


Aprendizaje No Supervisado

Repaso: Representación gráfica SVD

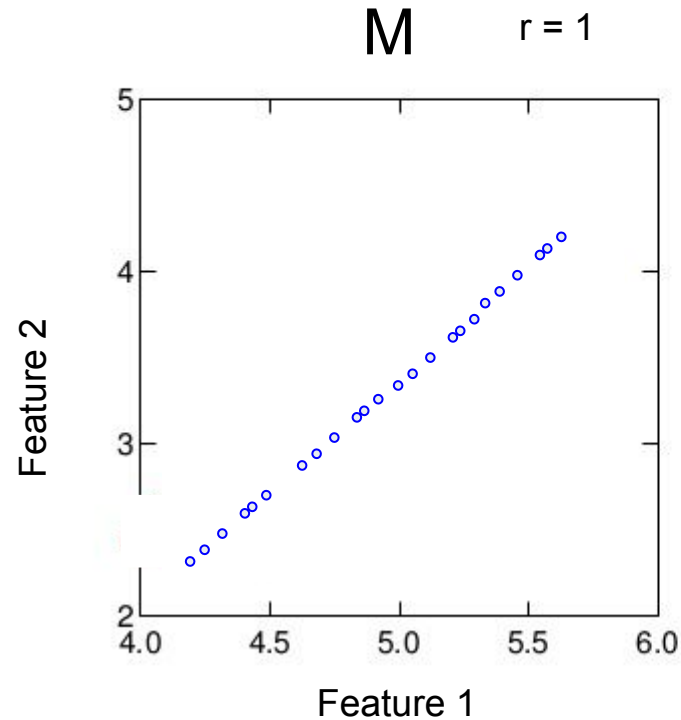
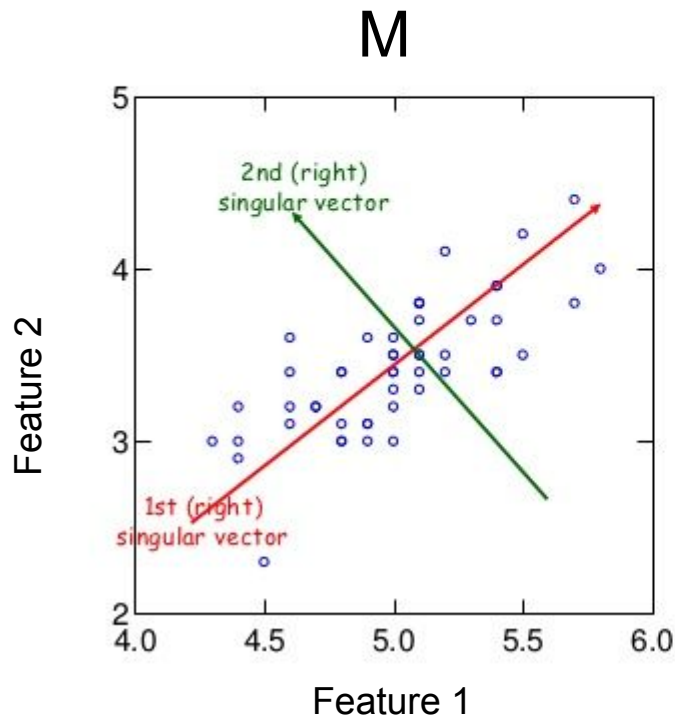


SVD • Representación gráfica



- 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





Aprendizaje No Supervisado

PCA (Principal Component Analysis)



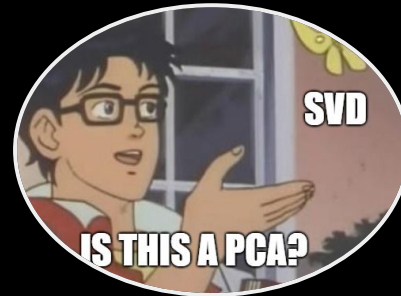
“Por lo que vimos en clase y lo que vi en los videos, SVD y PCA parecen ser lo mismo...”

Anónimo



“Casi, pero no, no son lo mismo.”

(otro) Anónimo



PCA • Definición

PCA es el método de reducción de dimensionalidad más utilizado.



¿Y PCA es muy distinto a usar
un SVD truncado?



PCA • Definición

PCA y SVD truncado son casi iguales, solo que existe una diferencia:

PCA = Centrar datos + SVD truncado



PCA • Definición

PCA y SVD truncado son casi iguales, solo que existe una diferencia:

PCA = Centrar datos + SVD truncado



Debemos sustraer la media de cada columna de Features antes de aplicar SVD truncado.

¿Por qué PCA es tan relevante?
¿Por qué no se lo ve solo como
un caso particular de SVD?



PCA • Importancia

- Matemáticamente, se puede llegar por otro camino (Matriz de covarianza)
- Tiene una interpretación muy intuitiva:

**Componentes
Principales**



**Direcciones de
máxima varianza**

La primer componente principal está en la dirección donde los datos presentan varianza máxima.

La segunda componente principal está la segunda dirección en términos de la varianza, y así sucesivamente.



¿Cuándo uso PCA y cuándo SVD?



Comparación • PCA vs. SVD

Para muchos casos, el resultado de usar uno u otro método va a ser muy parecido. Dependiendo del problema, puede que restar la media y trabajar con la distancia de cada feature a la media en el dataset sea mejor (si no saben, lo más común es usar PCA).

Analogías

PCA	SVD
Numero de componentes	Rango R
Componentes principales	Vectores singulares por derecha
Autovalores	Valores singulares
Maximiza Varianza	Minimiza Distancia

A close-up photograph of a white ceramic cup filled with a latte. The surface of the milk is decorated with intricate latte art, featuring a central heart shape surrounded by concentric, wavy lines. The cup is placed on a matching white saucer. In the background, a white napkin and a silver spoon are visible, though they are out of focus. The overall lighting is soft and even, highlighting the textures of the coffee and the smooth surface of the cup.

¡BREAK!



Hands-on training



Hands-on training



DS_Encuentro_39_PCA.ipynb

Para la próxima

1. Ver los videos de la plataforma “Aprendizaje No Supervisado: Sistemas de recomendación”.
2. Completar el notebook de hoy y atrasados.
3. Bajar el siguiente dataset
<https://www.kaggle.com/netflix-inc/netflix-prize-data>

ACÀMICA