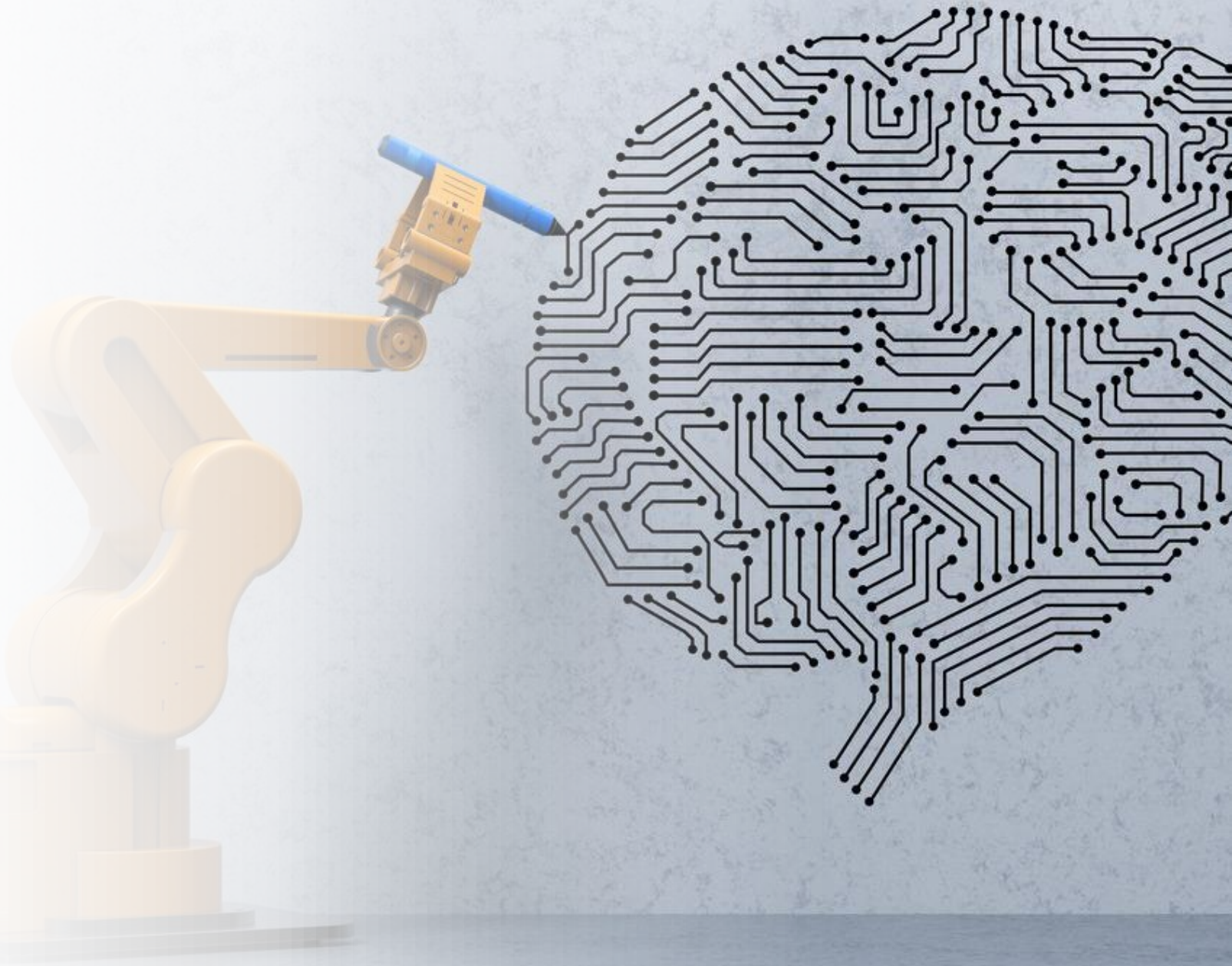
Reconocimiento de Patrones

Clase 6

Reducción de dimensionalidad II



Antes de empezar

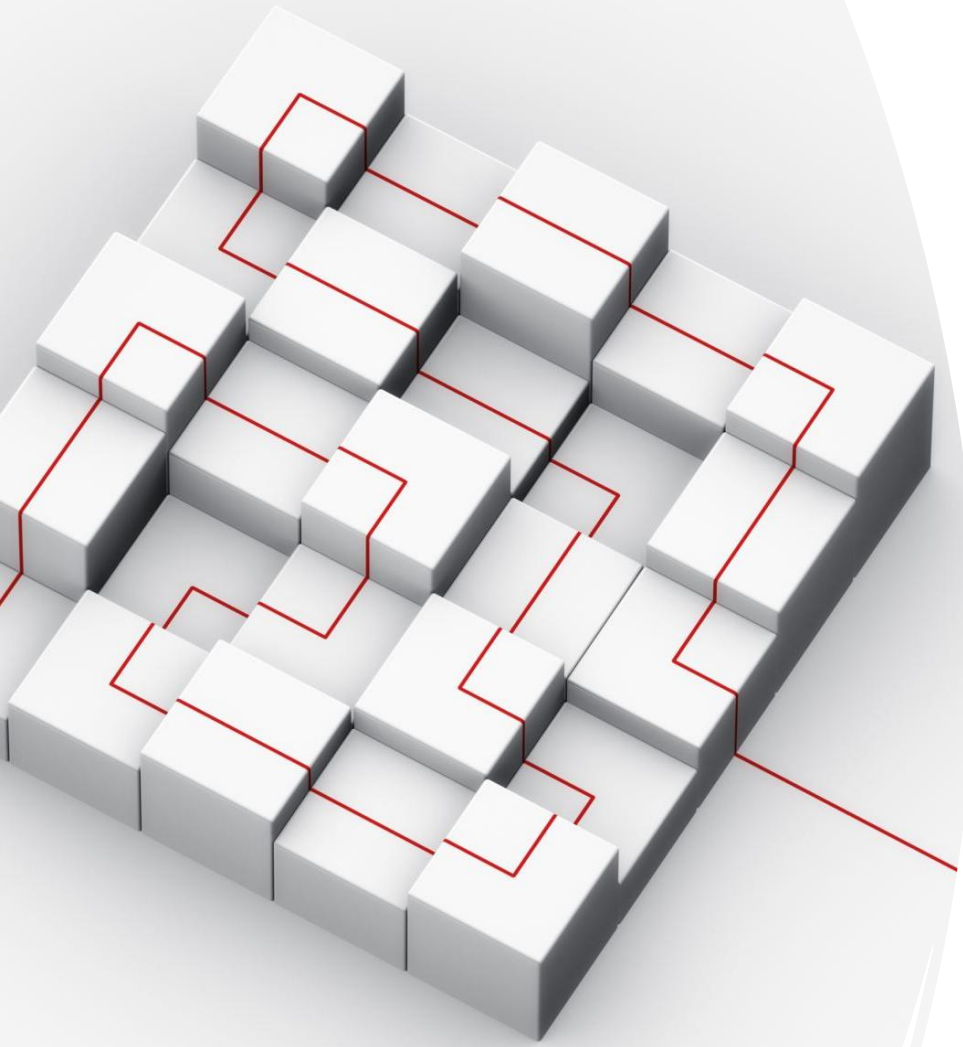
- ¿Dudas de la clase anterior?

Recapitulando

- PCA es un método de reducción de dimensionalidad no supervisado.
- Es vital el central o estandarizar los datos antes de aplicar PCA:

$$X_c = X - \mu \qquad X_c = \frac{X - \mu}{\sigma}$$

- Solo utilizamos aquellos vectores propios que proporciona más información del conjunto de datos original.
- Para saber cuantos usar nos podemos apoyar de la varianza explicada mediante la regla del codo, umbral de varianza o el criterio de Kaiser.



Reconstrucción

- PCA es una transformación lineal, por lo tanto, si se puede proyectar los datos a un espacio reducido, también puede aplicarse el camino inverso.
- La reconstrucción no será perfecta a menos que se use todos los componentes.
- Esto permite ver qué tanta información conservaste después de reducir la dimensionalidad.

¿Cómo se hace?

- Recordemos:

$$X \in \mathbb{R}^{mxn}(datos\ originales)$$
$$X_c = X - \mu\ (datos\ centrados)$$
$$W_k \in \mathbb{R}^{nxk}(eigenvectores)$$
$$Z = X_c W_k \in \mathbb{R}^{mxk}$$
$$(datos\ proyectados)$$

- La reconstrucción es:

Como W es una matriz ortonormal, se puede aplicar:

$$\widehat{X_c} = Z{W_k}^T$$

(mxn)=(mxk)(kxn)

Pero como los datos estaban centrados:

$$\hat{X} = Z{W_k}^T + \mu$$

Error de reconstrucción

- Como se esta utilizando solo k direcciones, la reconstrucción es una aproximación por lo que tiene asociado un cierto error:

$$\|X - \hat{X}\|^2$$

- Este error es mínimo cuando se usan los k componentes que aportan la mayor varianza. Esta reconstrucción siempre minimiza el MSE entre todos los posibles subespacios de dimensión k.

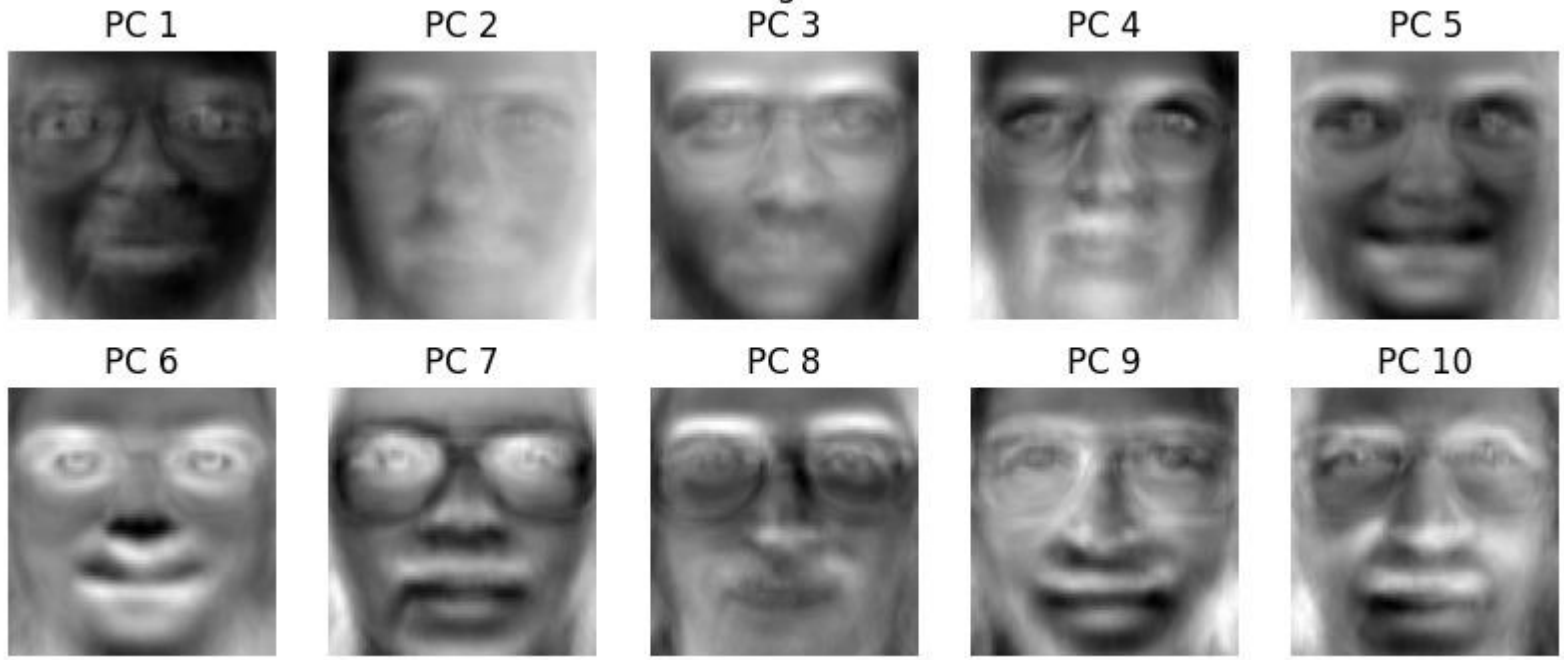
- Es **la mejor aproximación posible** de rango k.

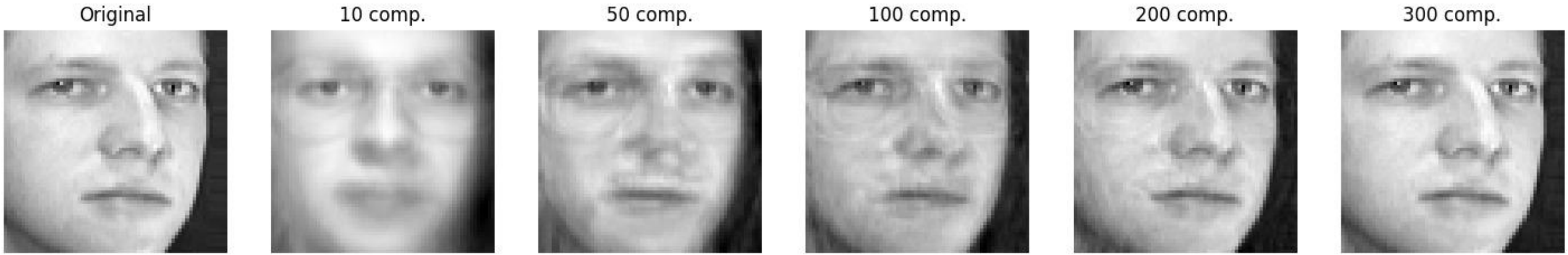
- Una forma de aplicarlo...

EigenFaces

- Son un conjunto de vectores propios que pueden ser utilizados para el reconocimiento facial.

Primeras Eigenfaces





¿Qué tan bueno es PCA?

Ventajas

- Reducción de dimensionalidad optima (MSE).
- Elimina redundancia.
- Mejora eficiencia computacional.
- Puede reducir el sobreajuste.

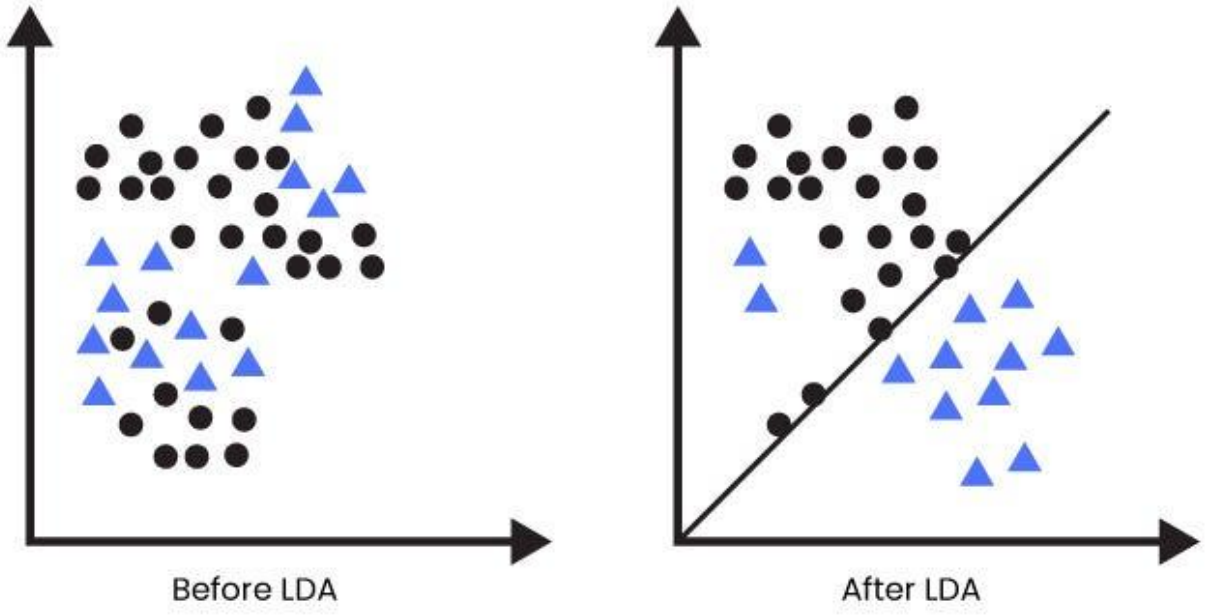
Desventajas

- Sensible a escala y outliers
- Solo captura estructura lineal
- Es un algoritmo no supervisado.
- No es robusto a ruido estructurado.

NO GARANTIZA MEJOR CLASIFICACIÓN.

Linear Discriminant Analysis (LDA)

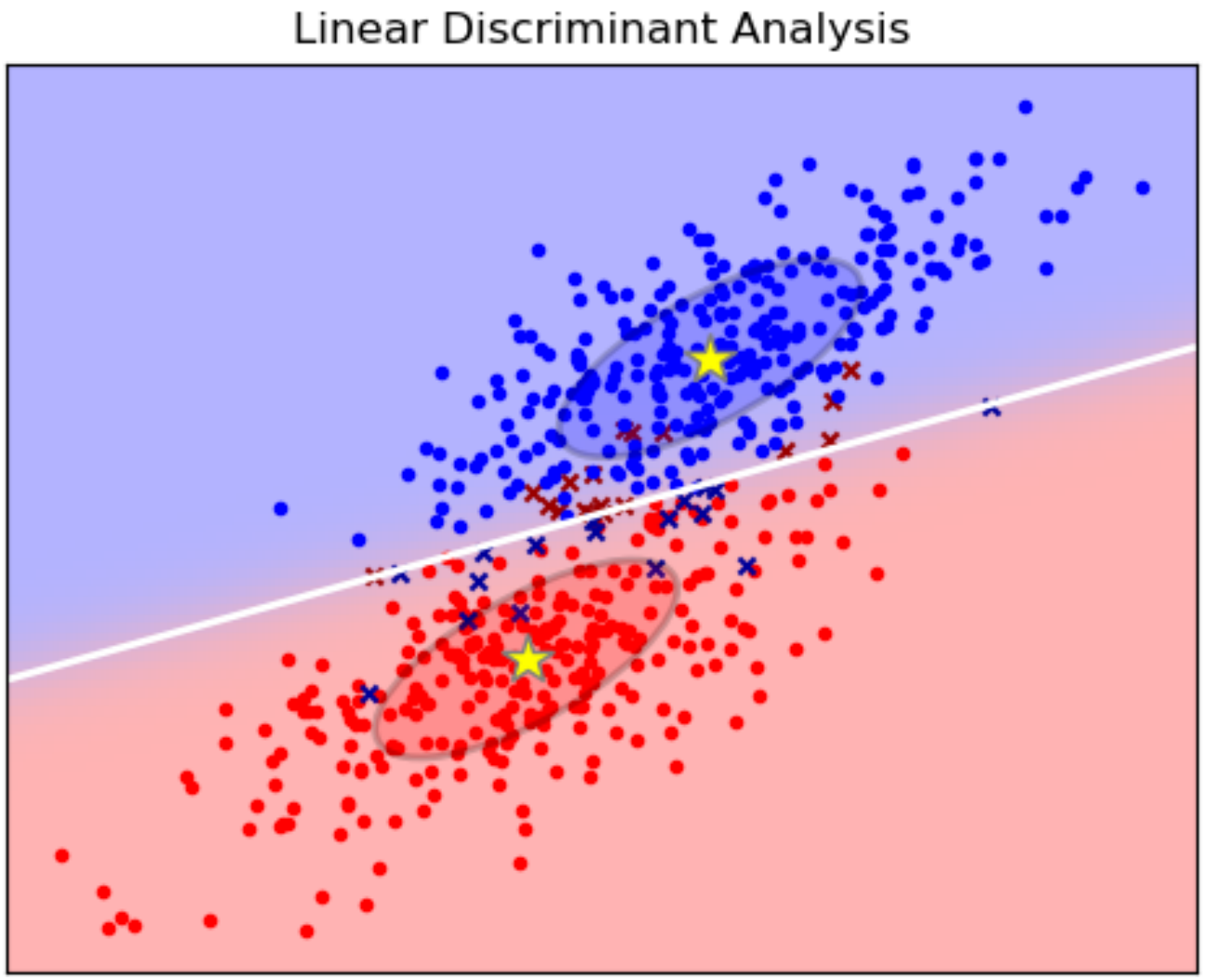
- Es una técnica de reducción de dimensionalidad supervisada, utilizada principalmente para la selección de características y para reducir la dimensionalidad de un conjunto de datos preservando al mismo tiempo la separabilidad de clases.



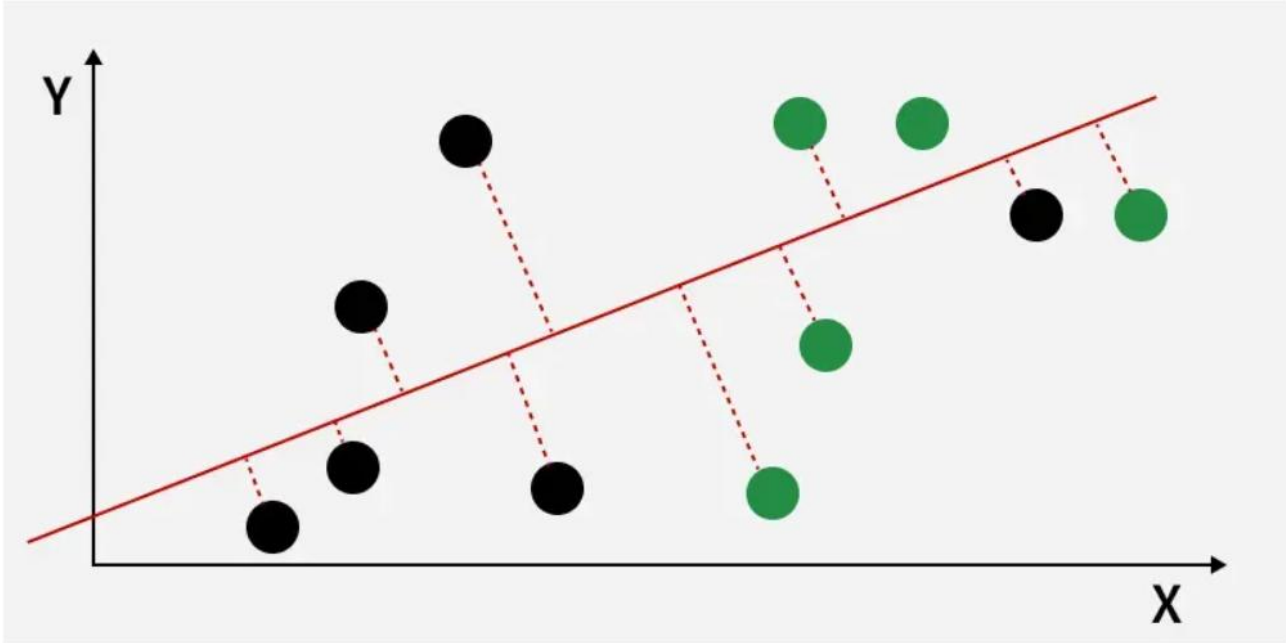
Su principal objetivo es maximizar la separabilidad entre diferentes clases en un problema de clasificación.

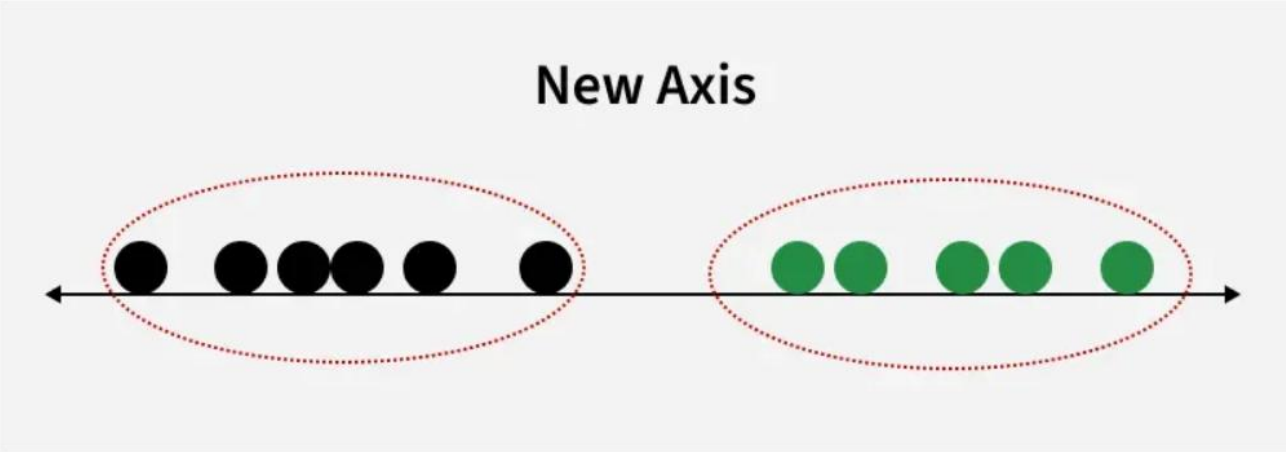
Funciona identificando una combinación lineal de características que separa o caracteriza dos o más clases de objetos o eventos.

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + a_3\mathbf{v}_3 + \cdots + a_n\mathbf{v}_n$$

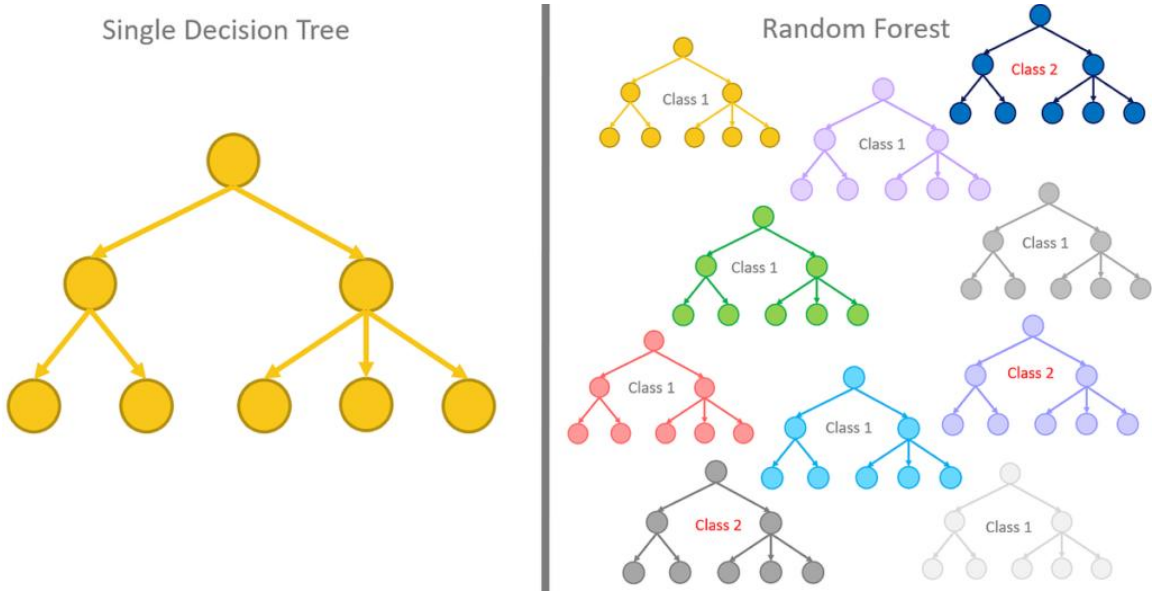


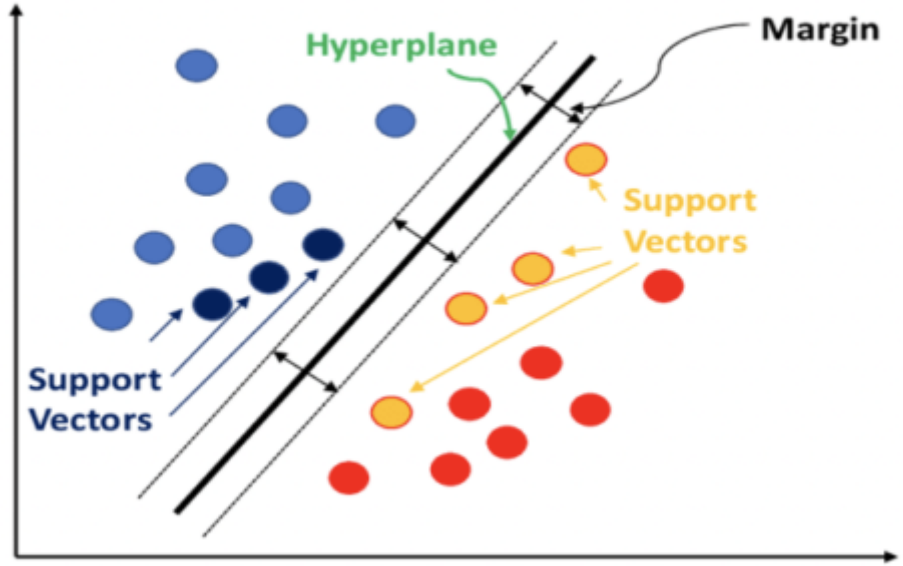
Hace esto proyectando datos con dos o más dimensiones en una dimensión para que puedan clasificarse más fácilmente.



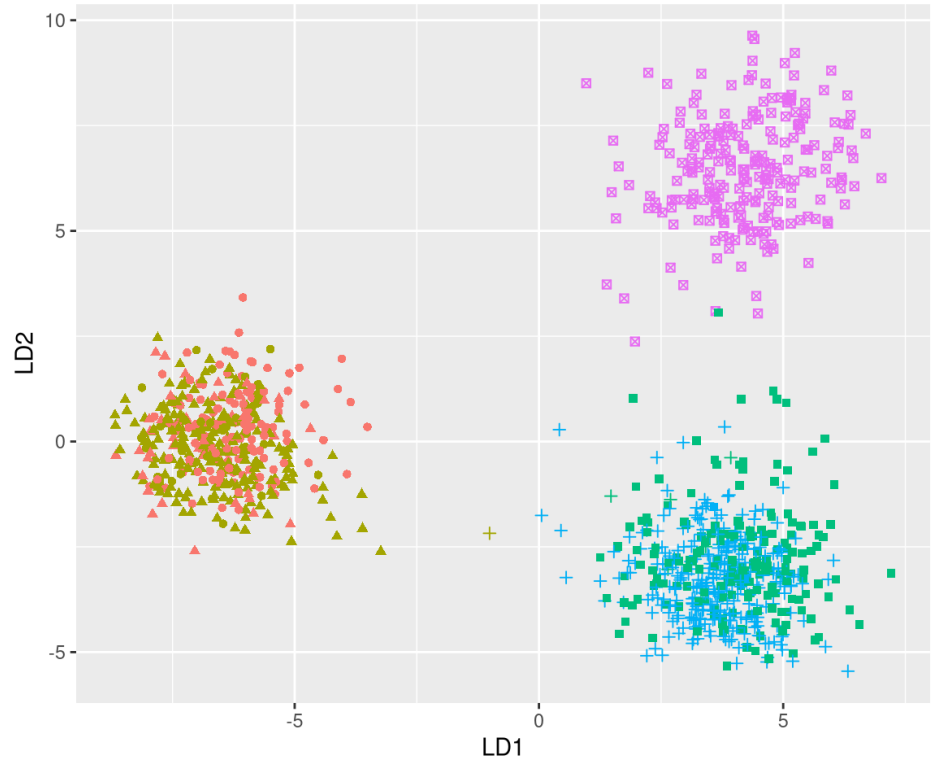


Es utilizado comúnmente para mejorar el funcionamiento de otros algoritmos de clasificación como arboles de decisión, bosque aleatorio o las máquinas de vectores de soporte (SVM).





¿Qué es lo que busca?



- La separación implica:

1. maximizar la distancia entre las medias proyectadas
2. minimizar la varianza proyectada dentro de las clases.

- En otras palabras:

LDA busca una proyección lineal donde: las clases estén lo **más separadas posible** y los datos dentro de cada clase estén **lo más compactos posible.**

Supuestos de LDA

- Distribución gaussiana. Los datos de cada clase deben de seguir una distribución normal.
- Cada clase tiene la misma matriz de covarianza
- El conjunto de datos es linealmente separable. Se puede dibujar una línea recta que separe los datos.

Algunas definiciones...

- Nuestros datos los podemos representar como una matriz:

$X \in \mathbb{R}^{Nxd}$

c: clases

μ_i: media de la clase i

μ: media global

- Donde estas pueden calcularse como:

$$\mu_i = \frac{1}{n_i} \sum_{x \in c_i} x$$

$$\mu = \frac{1}{n} \sum_{x \in c_i} n_i \, \mu_i = \frac{1}{n} \sum_{k=1}^{n} x_k$$

- Matriz de dispersión entre clases (Between-class scatter, S_B):

Mide qué tan separadas están las medias de cada clase respecto a la media global.

$$S_B = \sum_{i=1}^{C} n_i(\mu_i - \mu)\,(\mu_i - \mu)^T$$

- Matriz de dispersión dentro de clases (Within-class scatter, S_w):

Mide la varianza interna de cada clase, es decir, qué tan dispersos están los datos dentro de cada clase.

$$S_W = \sum_{i=1}^{c} \sum_{x \in C_i} (x - \mu_i)\,(x - \mu_i)^T$$

Coeficiente de Rayleigh

LDA busca maximizar este coeficiente:

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

Donde w es el vector de proyección.

Interpretación:
Numerador → separación entre clases
Denominador → dispersión interna

Para maximizar este coeficiente se busca que el numerador sea grande y el denominador sea pequeño.

¿Cómo se calcula LDA?

- Antes de considerar aplicar LDA:

1. Tener etiquetas: (X, y)

2. De preferencia normalizar los datos: $x' = \frac{x-\mu}{\sigma}$

3. Verificar singularidad: d>n

1. Calcular las medias: la media para cada clase i y la media global μ_i, μ

2. Construir las matrices S_B y S_W : Una S_W grande relativa a S_B indica clases muy solapadas y un problema difícil de separar. Verificar también que S_W sea invertible

$$S_B = \sum_{i=1}^{C} n_i (\mu_i - \mu)\,(\mu_i - \mu)^T$$

$$S_W = \sum_{i=1}^{c} \sum_{x \in C_i} (x - \mu_i)\,(x - \mu_i)^T$$

3. Resolver el problema de autovalores. Recordemos que:

$$A\,w = \lambda\,w$$

Para este caso nuestra ecuación a resolver es:

$$S_B w = \lambda S_W w$$

O su equivalente: $S_W^{-1} S_B w = \lambda w$

S_W^{-1} **tiene que existir**, de lo contrario el problema se vuelve mal condicionado y hay que tomar otros enfoques.

Si S_W^{-1} existe, entonces $A = S_W^{-1} S_B$ y se establece la ecuación: $A\,w = \lambda\,w$

A diferencia de PCA, LDA proporciona K-1 vectores propios.

4. Seleccionar los discriminantes lineales

Los autovectores asociados a los autovalores más grandes son los ejes discriminantes. Para K clases, hay como máximo: **min(K-1, d)**

$$W = [w_1, \cdots, w_k]$$

6. Proyectar los datos sobre los discriminante seleccionados

Los datos ahora viven en un espacio de menor dimensión máximamente discriminante

$$Z = XW, \qquad Z \in \mathbb{R}^{Nxk}, \qquad k < d$$

¡¡OJO!!

- Si Sw es singular significa que su determinante es 0 y por lo tanto no tiene inversa.
- Si es singular se puede:

1. Aplicar primero PCA y luego LDA
2. Hacer una regularización a la matriz: $S_w^{(\alpha)} = (1-\alpha)S_w + \alpha I$
3. Pseudoinversa: mediante SVD

Ejemplo 1

Se tienen $n = 8$ observaciones, $p = 2$ variables (x_1, x_2) y $K = 2$ clases ($n_k = 4$ por clase):

Clase 1		Clase 2	
x_1	x_2	x_1	x_2
2	3	7	6
3	4	8	7
3	3	8	8
4	5	9	9

$$\mu_i = \frac{1}{n_i} \sum_{x \in c_i} x$$

$$\mu = \frac{1}{n} \sum_{x \in c_i} n_i \, \mu_i = \frac{1}{n} \sum_{k=1}^{n} x_k$$

Clase 1 $(n_1 = 4)$:

$$\boldsymbol{\mu}_1 = \frac{1}{4} \begin{pmatrix} 2+3+3+4 \\ 3+4+3+5 \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 12 \\ 15 \end{pmatrix} = \begin{pmatrix} 3.00 \\ 3.75 \end{pmatrix}$$

Clase 2 $(n_2 = 4)$:

$$\boldsymbol{\mu}_2 = \frac{1}{4} \begin{pmatrix} 7+8+8+9 \\ 6+7+8+9 \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 32 \\ 30 \end{pmatrix} = \begin{pmatrix} 8.00 \\ 7.50 \end{pmatrix}$$

$$\boldsymbol{\mu} = \frac{1}{8}(4 \cdot \boldsymbol{\mu}_1 + 4 \cdot \boldsymbol{\mu}_2) = \frac{1}{2} \left[\begin{pmatrix} 3.00 \\ 3.75 \end{pmatrix} + \begin{pmatrix} 8.00 \\ 7.50 \end{pmatrix} \right] = \begin{pmatrix} 5.50 \\ 5.625 \end{pmatrix}$$

$$S_B = \sum_{i=1}^{C} n_i(\mu_i - \mu)\,(\mu_i - \mu)^T$$

Desviaciones de cada media de clase respecto a $\boldsymbol{\mu} = (5.50,\ 5.625)^\top$:

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}) = \begin{pmatrix} 3.00 - 5.50 \\ 3.75 - 5.625 \end{pmatrix} = \begin{pmatrix} -2.50 \\ -1.875 \end{pmatrix} \qquad (\boldsymbol{\mu}_2 - \boldsymbol{\mu}) = \begin{pmatrix} 8.00 - 5.50 \\ 7.50 - 5.625 \end{pmatrix} = \begin{pmatrix} 2.50 \\ 1.875 \end{pmatrix}$$

Contribución de cada clase $(n_k = 4)$:

$$4\,(\boldsymbol{\mu}_1 - \boldsymbol{\mu})(\boldsymbol{\mu}_1 - \boldsymbol{\mu})^\top = 4 \begin{pmatrix} -2.50 \\ -1.875 \end{pmatrix} \begin{pmatrix} -2.50 & -1.875 \end{pmatrix} = 4 \begin{pmatrix} 6.25 & 4.6875 \\ 4.6875 & 3.5156 \end{pmatrix} = \begin{pmatrix} 25.00 & 18.75 \\ 18.75 & 14.063 \end{pmatrix}$$

$$4\,(\boldsymbol{\mu}_2 - \boldsymbol{\mu})(\boldsymbol{\mu}_2 - \boldsymbol{\mu})^\top = 4 \begin{pmatrix} 2.50 \\ 1.875 \end{pmatrix} \begin{pmatrix} 2.50 & 1.875 \end{pmatrix} = 4 \begin{pmatrix} 6.25 & 4.6875 \\ 4.6875 & 3.5156 \end{pmatrix} = \begin{pmatrix} 25.00 & 18.75 \\ 18.75 & 14.063 \end{pmatrix}$$

Sumando:

$$\boxed{S_B = \begin{pmatrix} 25.00 & 18.75 \\ 18.75 & 14.063 \end{pmatrix} + \begin{pmatrix} 25.00 & 18.75 \\ 18.75 & 14.063 \end{pmatrix} = \begin{pmatrix} 50.00 & 37.50 \\ 37.50 & 28.125 \end{pmatrix}}$$

$$S_W = \sum_{i=1}^{c} \sum_{x \in C_i} (x - \mu_i)\,(x - \mu_i)^T$$

Contribución de la Clase 1 ($\mu_1 = (3.00,\ 3.75)^\top$)

Las desviaciones respecto a $\boldsymbol{\mu}_1$ son:

$$\mathbf{x}^{(1)} - \boldsymbol{\mu}_1 = \begin{pmatrix} -1.00 \\ -0.75 \end{pmatrix}, \quad \mathbf{x}^{(2)} - \boldsymbol{\mu}_1 = \begin{pmatrix} 0.00 \\ 0.25 \end{pmatrix}, \quad \mathbf{x}^{(3)} - \boldsymbol{\mu}_1 = \begin{pmatrix} 0.00 \\ -0.75 \end{pmatrix}, \quad \mathbf{x}^{(4)} - \boldsymbol{\mu}_1 = \begin{pmatrix} 1.00 \\ 1.25 \end{pmatrix}$$

Sumando los productos exteriores:

$$S_1 = \begin{pmatrix} 1.00 & 0.75 \\ 0.75 & 0.5625 \end{pmatrix} + \begin{pmatrix} 0.00 & 0.00 \\ 0.00 & 0.0625 \end{pmatrix} + \begin{pmatrix} 0.00 & 0.00 \\ 0.00 & 0.5625 \end{pmatrix} + \begin{pmatrix} 1.00 & 1.25 \\ 1.25 & 1.5625 \end{pmatrix} = \begin{pmatrix} 2.00 & 2.00 \\ 2.00 & 2.75 \end{pmatrix}$$

Contribución de la Clase 2 ($\mu_2 = (8.00,\ 7.50)^\top$)

$$\mathbf{x}^{(5)} - \boldsymbol{\mu}_2 = \begin{pmatrix} -1.00 \\ -1.50 \end{pmatrix}, \quad \mathbf{x}^{(6)} - \boldsymbol{\mu}_2 = \begin{pmatrix} 0.00 \\ -0.50 \end{pmatrix}, \quad \mathbf{x}^{(7)} - \boldsymbol{\mu}_2 = \begin{pmatrix} 0.00 \\ 0.50 \end{pmatrix}, \quad \mathbf{x}^{(8)} - \boldsymbol{\mu}_2 = \begin{pmatrix} 1.00 \\ 1.50 \end{pmatrix} \qquad S_2 = \begin{pmatrix} 2.00 & 3.00 \\ 3.00 & 5.00 \end{pmatrix}$$

Suma total

$$\boxed{S_W = S_1 + S_2 = \begin{pmatrix} 2.00 & 2.00 \\ 2.00 & 2.75 \end{pmatrix} + \begin{pmatrix} 2.00 & 3.00 \\ 3.00 & 5.00 \end{pmatrix} = \begin{pmatrix} 4.00 & 5.00 \\ 5.00 & 7.75 \end{pmatrix}}$$

Calcular $S_W^{-1} S_B$

Inversa de S_W

Para una matriz 2×2: $A^{-1} = \dfrac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

$$\det(S_W) = (4.00)(7.75) - (5.00)(5.00) = 31.00 - 25.00 = 6.00$$

$$S_W^{-1} = \frac{1}{6.00} \begin{pmatrix} 7.75 & -5.00 \\ -5.00 & 4.00 \end{pmatrix} = \begin{pmatrix} 1.2917 & -0.8333 \\ -0.8333 & 0.6667 \end{pmatrix}$$

Verificación: $S_W^{-1} S_W = I$

$$\begin{pmatrix} 1.2917 & -0.8333 \\ -0.8333 & 0.6667 \end{pmatrix} \begin{pmatrix} 4.00 & 5.00 \\ 5.00 & 7.75 \end{pmatrix} = \begin{pmatrix} 5.167 - 4.167 & 6.458 - 6.458 \\ -3.333 + 3.333 & -4.167 + 5.167 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \checkmark$$

Producto $A = S_W^{-1} S_B$

$$A = S_W^{-1} S_B = \begin{pmatrix} 1.2917 & -0.8333 \\ -0.8333 & 0.6667 \end{pmatrix} \begin{pmatrix} 50.00 & 37.50 \\ 37.50 & 28.125 \end{pmatrix}$$

Calculando cada entrada:

$$A_{11} = (1.2917)(50.00) + (-0.8333)(37.50) = 64.583 - 31.249 = 33.334$$
$$A_{12} = (1.2917)(37.50) + (-0.8333)(28.125) = 48.439 - 23.436 = 25.003$$
$$A_{21} = (-0.8333)(50.00) + (0.6667)(37.50) = -41.665 + 25.001 = -16.664$$
$$A_{22} = (-0.8333)(37.50) + (0.6667)(28.125) = -31.249 + 18.751 = -12.498$$

$$\boxed{A = S_W^{-1} S_B \approx \begin{pmatrix} 33.33 & 25.00 \\ -16.67 & -12.50 \end{pmatrix}}$$

$$\det(A - \lambda I) = 0$$

$$\det \begin{pmatrix} 33.33 - \lambda & 25.00 \\ -16.67 & -12.50 - \lambda \end{pmatrix} = (33.33 - \lambda)(-12.50 - \lambda) - (25.00)(-16.67) = 0$$

Expandiendo el primer término:

$$(33.33)(-12.50) + (33.33)(-\lambda) + (-\lambda)(-12.50) + \lambda^2 + 416.75 = 0$$

$$-416.63 - 33.33\lambda + 12.50\lambda + \lambda^2 + 416.75 = 0$$

$$\lambda^2 - 20.83\lambda + 0.12 = 0$$

$$\boxed{\lambda_1 \approx 20.83} \qquad \lambda_2 \approx 0.005 \approx 0$$

Autovector principal ($\lambda_1 \approx 20.83$)

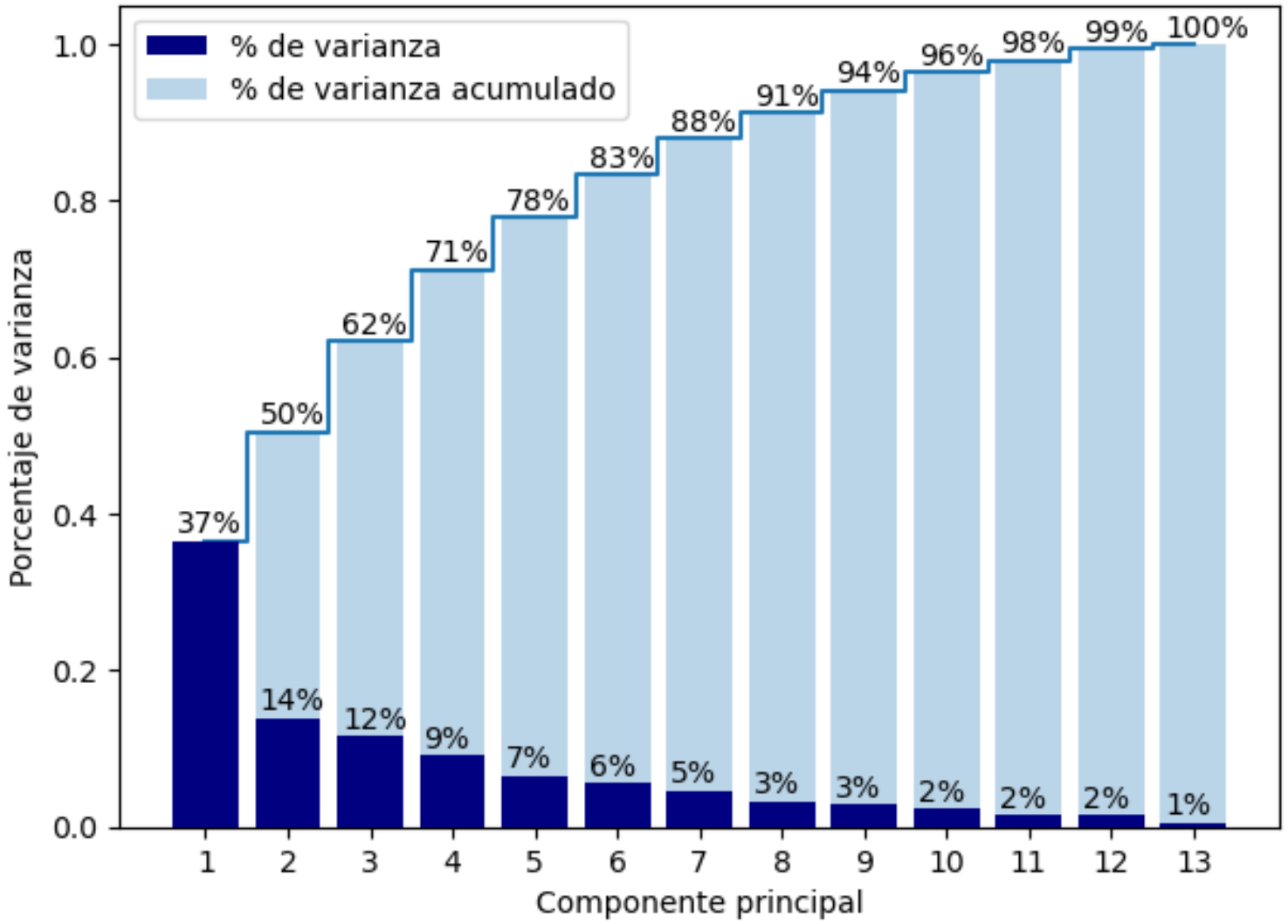
Resolviendo $(A - \lambda_1 I)\,\mathbf{w} = \mathbf{0}$:

$$\begin{pmatrix} 33.33 - 20.83 & 25.00 \\ -16.67 & -12.50 - 20.83 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 12.50 & 25.00 \\ -16.67 & -33.33 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \mathbf{0}$$

De la primera fila: $12.50\,w_1 + 25.00\,w_2 = 0 \;\Rightarrow\; w_1 = -2\,w_2$
Tomando $w_2 = -1$: $\mathbf{w} = (2,\ 1)^\top$. Normalizando:

$$\|\mathbf{w}\| = \sqrt{2^2 + 1^2} = \sqrt{5} \approx 2.236$$

$$\boxed{\mathbf{w}_1 = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 0.894 \\ 0.447 \end{pmatrix} \quad (\mathrm{LD}_1)}$$



¿Cuantas k escoger?

Varianza Explicada

¿Vale la pena usar LDA?

Ventajas

- Reducción de dimensionalidad supervisada
- Interpretabilidad
- Computacionalmente eficiente
- También puede funcionar como clasificador.

Desventajas

- Supone distribución gaussiana
- Supone covarianza compartida
- No captura relaciones no lineales
- Limite de min(K-1, d) discriminantes

Otros métodos...

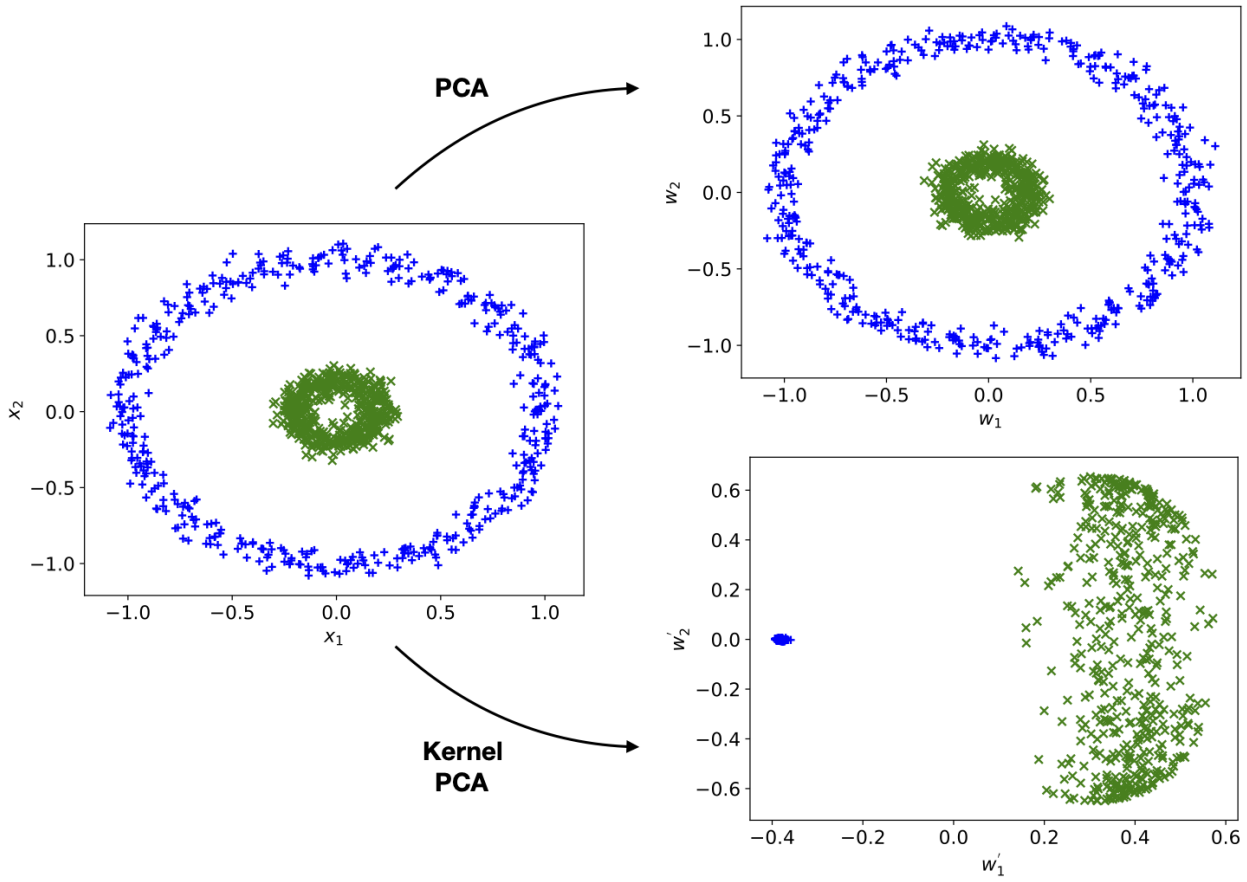
t-SNE (t-Distributed Stochastic Neighbor Embedding)

Método no lineal de reducción de dimensionalidad usado principalmente para visualización (2D o 3D).

Mide similitudes en el espacio original y el reducido usando una distribución Gaussiana y t de Student respectivamente. Minimiza la divergencia KL entre ambas distribuciones.

Kernel PCA

- Extensión no lineal de PCA usando el truco del kernel.
- Se mapean los datos a un espacio de mayor dimensionalidad usando una funcion kernel, en ese espacio se aplica PCA
- Todo se calcula implícitamente mediante la matriz de kernel $K(x_i, x_j)$, nunca se computa el mapeo explícitamente.

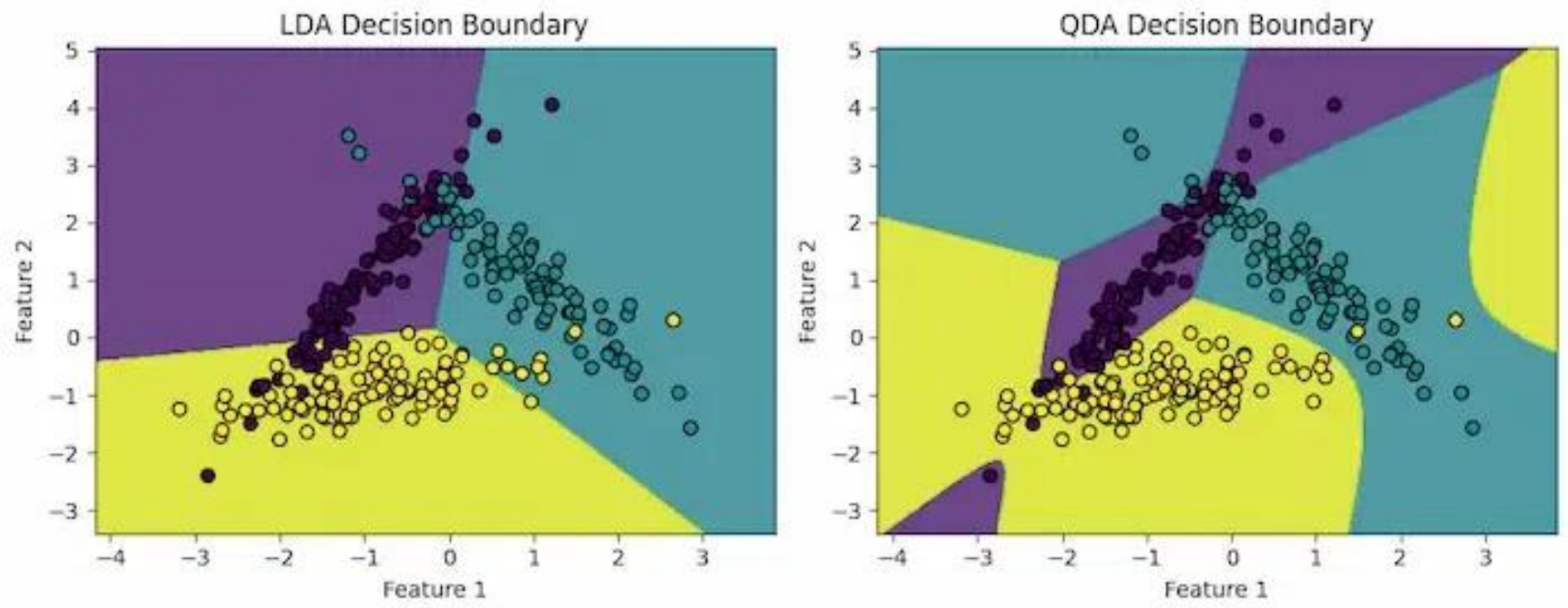


Quadratic LDA

Extensión de LDA que permite fronteras cuadráticas entre clases.

No asume una matriz de covarianza igual para cada clase, sino que cada clase tiene una matriz distinta.

Cada clase tiene su propia distribución gaussiana con su propia forma y orientación.



Practica 1. PCA

1. Dataset sintético
2. Dataset Iris
3. Dataset Wine
4. Comparación.
5. Eigenfaces

Para la próxima clase...

- Evaluación de modelos