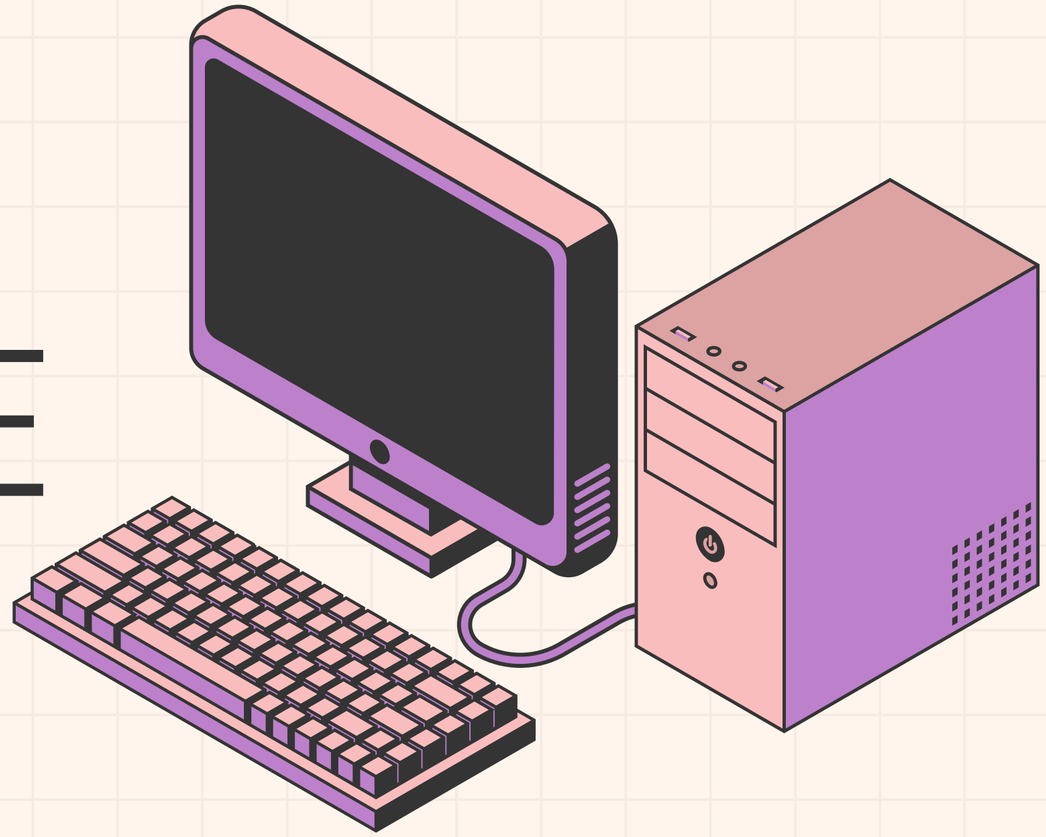
9/ABRIL/2025

ANÁLISIS DISCRIMINANTE

Sabrina Arroyo, Carolina Daniells,
Vladimir Espinosa, Estela Gil, Aitana Orduña,
Beth Pérez, Larissa Rodríguez & Oswaldo Rojano



¿QUÉ ES? 🤔

Según IBM:

- Un modelo predictivo para la pertenencia al grupo.
- El modelo está compuesto por una función discriminante (o, para más de dos grupos, un conjunto de funciones discriminantes) basada en combinaciones lineales de las variables predictoras que proporcionan la mejor discriminación posible entre los grupos.
- Las funciones se generan a partir de una muestra de casos para los que se conoce el grupo de pertenencia; posteriormente, las funciones pueden ser aplicadas a nuevos casos que dispongan de mediciones para las variables predictoras pero de los que se desconozca el grupo de pertenencia.
    - ¡Es un método SUPERVISADO!

IMPORTANCIA

Clasificación precisa:
Su propósito central es asignar correctamente nuevas observaciones a grupos predefinidos, basándose en variables conocidas.

Amplia aplicabilidad:
Su uso se extiende a numerosos campos, como la medicina (diagnóstico clínico), finanzas (evaluación de riesgo crediticio), marketing (segmentación de mercados), ciencias sociales (clasificación de perfiles psicológicos) y ciencia forense (identificación de patrones criminales), entre otros.

Interpretación de datos:
Más allá de clasificar, permite comprender qué variables son más relevantes para distinguir entre grupos, ayudando a identificar factores clave.

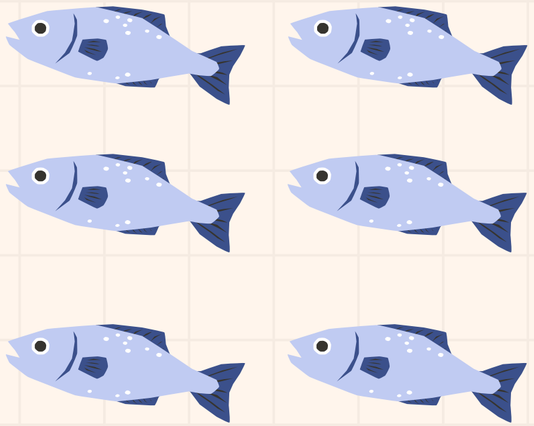
Compatibilidad con otros métodos:
Puede complementarse con técnicas como PCA, regresión logística o redes neuronales, formando parte de sistemas más complejos de análisis predictivo.

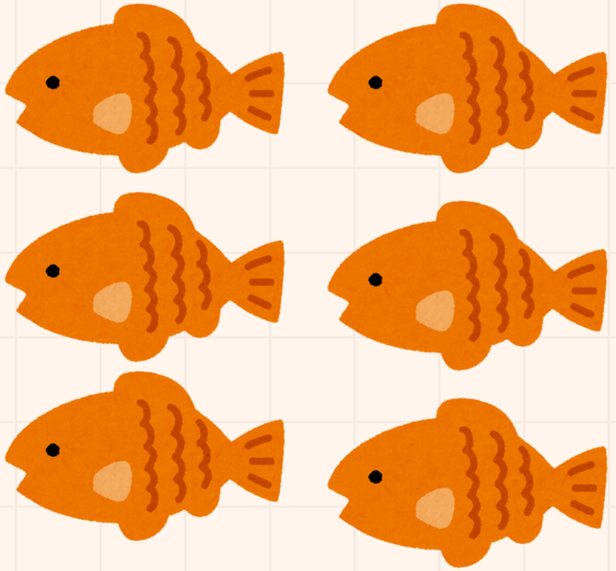
Flexibilidad metodológica:
Puede adaptarse a diferentes tipos de datos y situaciones (tipos).

TIPOS
DE ANÁLISIS DISCRIMINANTES

LDA

- Supone que las clases tienen la **misma matriz de covarianza** (“forma” y “dispersión”).
- Separa a los grupos mediante líneas rectas **(fronteras lineales).**
- Ideal cuando: los datos está bien separados y las **distribuciones son SIMILARES.**

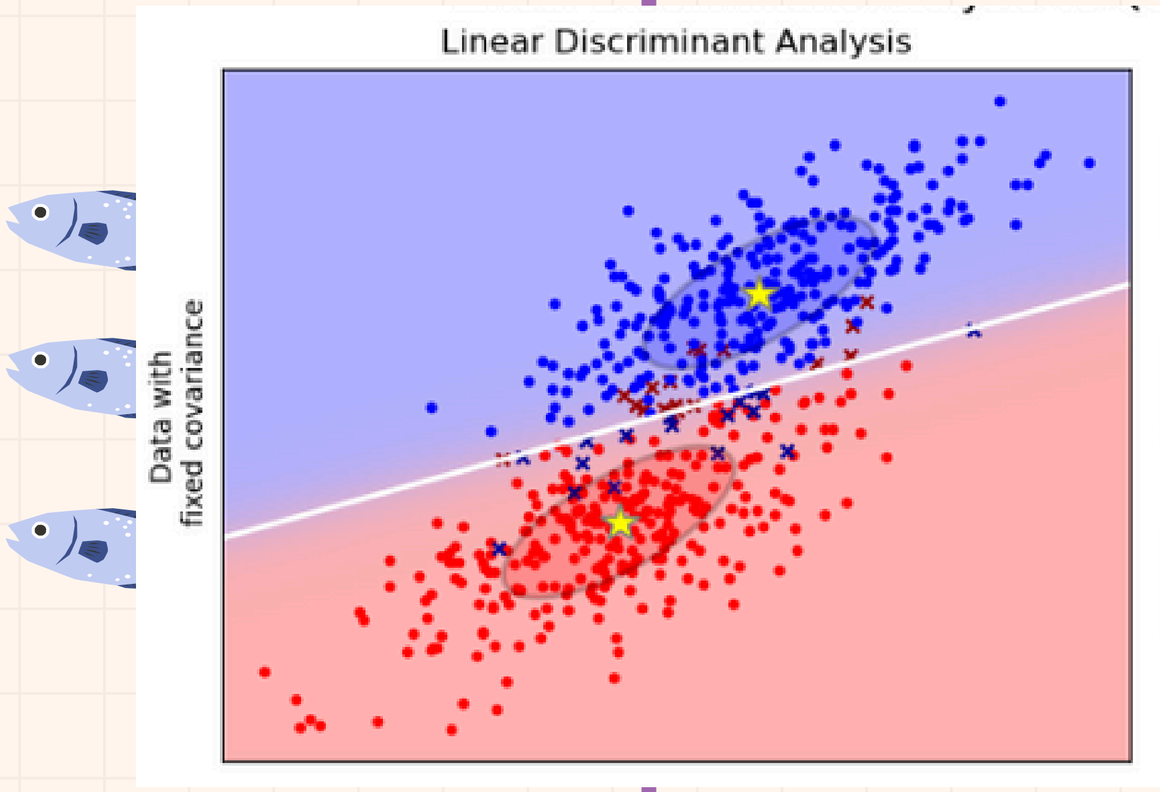




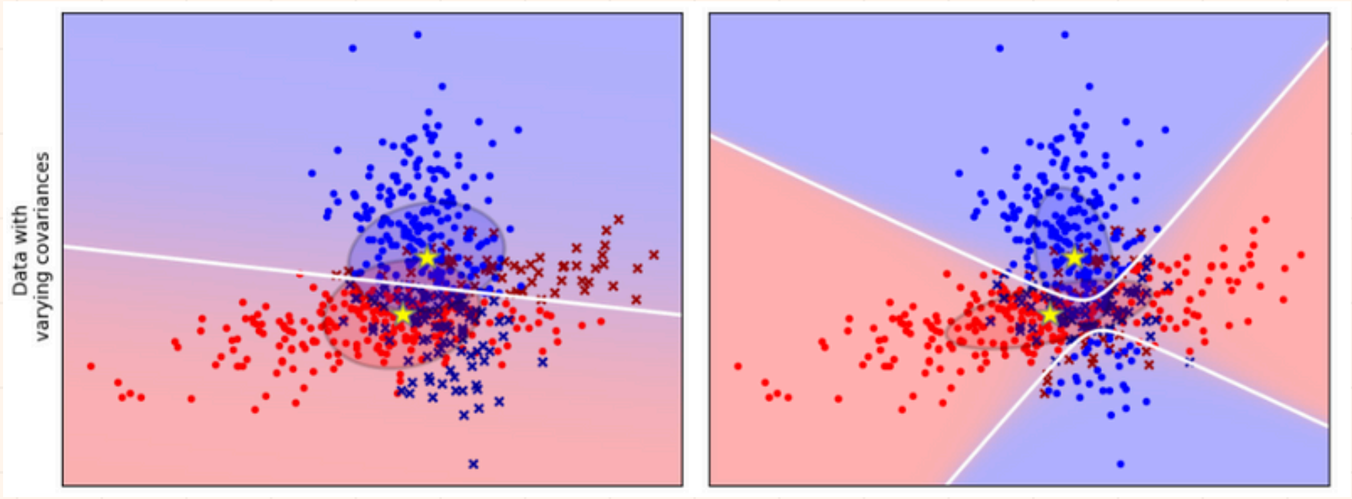
TIPOS
DE ANÁLISIS DISCRIMINANTES

LDA

- Supone que las clases tienen la **misma matriz de covarianza** (“forma” y “dispersión”).
- Separa a los grupos mediante líneas rectas **(fronteras lineales).**
- Ideal cuando: los datos está bien separados y las **distribuciones son SIMILARES.**



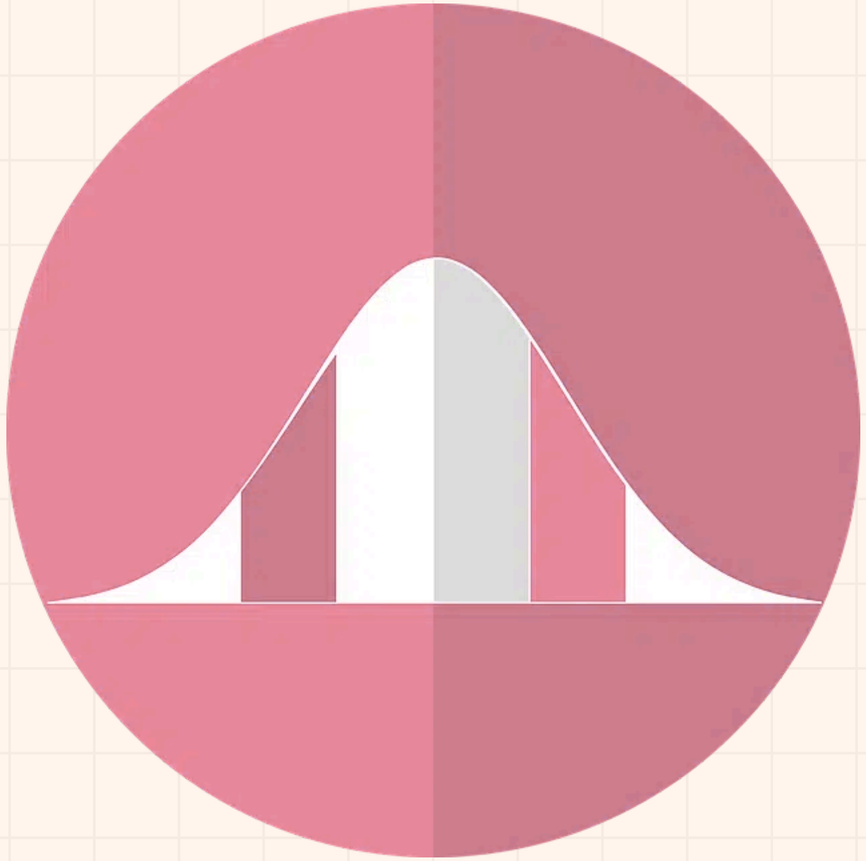
TIPOS
DE ANÁLISIS DISCRIMINANTES



QDA

- **No asume** que los grupos tienen la misma matriz de covarianza.
- Permite **fronteras** de decisión **curvas (¡++ flexibilidad!).**
- Ideal cuando: los datos están un poco más “enredados”

FUNDAMENTOS MATEMÁTICOS



OBJETIVOS

Encontrar una línea o plano que **separe** de la mejor manera posible a los diferentes grupos, basándose en sus características

¿Cómo lo logra?

Maximizar la separación entre clases

Minimizar la variabilidad dentro de cada clase

Busca un nuevo espacio donde la distancia entre los **centros** de los grupos sea máxima, y la **dispersión** dentro de cada grupo sea mínima

STUDIO SHODWE

ENTONCES... **¿CENTROS?**

Son **vectores** formados con las **medias** de las variables que definen a los datos de cada clase:

Imagina 2 grupos:

- Px sanos
- Px con enfermedad

Cada grupo forma una **nube de puntos** en el espacio de las características.

EL CENTRO DE CADA NUBE ES... EL VECTOR DE MEDIAS DE ESA CLASE => EL PROMEDIO DE CADA VARIABLE

¿Para qué sirve?

Medir qué tan lejos están los grupos entre sí

¿QUÉ PASA SI TENEMOS **+ DE 2** CLASES?

1. **CALCULA LOS CENTROS DE CADA GRUPO:**
   - A. **CENTRO A**
   - B. **CENTRO B**
   - C. **CENTRO C**

2. SE ANALIZA CÓMO ESTÁN DISTRIBUIDOS ENTRE ELLOS

¡PUEDE ENCONTRAR VARIAS DIRECCIONES en las que se maximice la separación entre todos los centros al mismo tiempo

STUDIO
SHODWE

DISPERSIÓN **INTERNA**

- Mide qué tan **dispersos** están los puntos alrededor de su centro => matriz de covarianza dentro de las clases
- Refleja cuánta variación hay dentro de cada clase

Poca dispersión interna = datos de una clase **+ juntos**

DISPERSIÓN **EXTERNA**

- Calcula qué tan lejanos están los centros entre sí => matriz de covarianza **entre** clases
- Refleja cuánta diferencia hay entre grupos

Mucha separación entre centros = los grupos son bien **distintos**

DIRECCIÓN **ÓPTIMA**

- LDA busca una **combinación lineal** de las variables para que:
  - Los centros proyectados estén lo más separados posible
  - La dispersión proyectada dentro de las clases sea la menor posible

$$\frac{\text{VARIANZA ENTRE CLASES}}{\text{VARIANZA DENTRO DE CLASES}}$$

maximizar esta razón

crietrio de Fisher

CLADIFICADORES Y FUNCIONES DISCRIMINANTES

UN clasificador se define como:

$$c(x) = \arg\max_{c} \; g_c(x)$$

donde cada clase c se define su funcion discriminante g_c.
El grado de pertenencia del objeto x a la clase c es $g_c(x)$.
c(x) es la clase a la que el objeto x pertenece en mayor grado.

El Clasificador de Bayes se obtiene como:

$$c(x) = \arg\max_{c} \; p(c \mid x)$$

Para cada punto x, calculamos δk(x) para cada clase k. El valor de δk(x) es como un "puntaje de pertenencia"

LDA se basa en la proyección de los datos a un espacio de menor dimensión, de manera que las clases sean lo más separadas posible. En este contexto, se buscan funciones discriminantes que maximicen la distancia entre las medias de las clases y minimicen la dispersión dentro de cada clase.

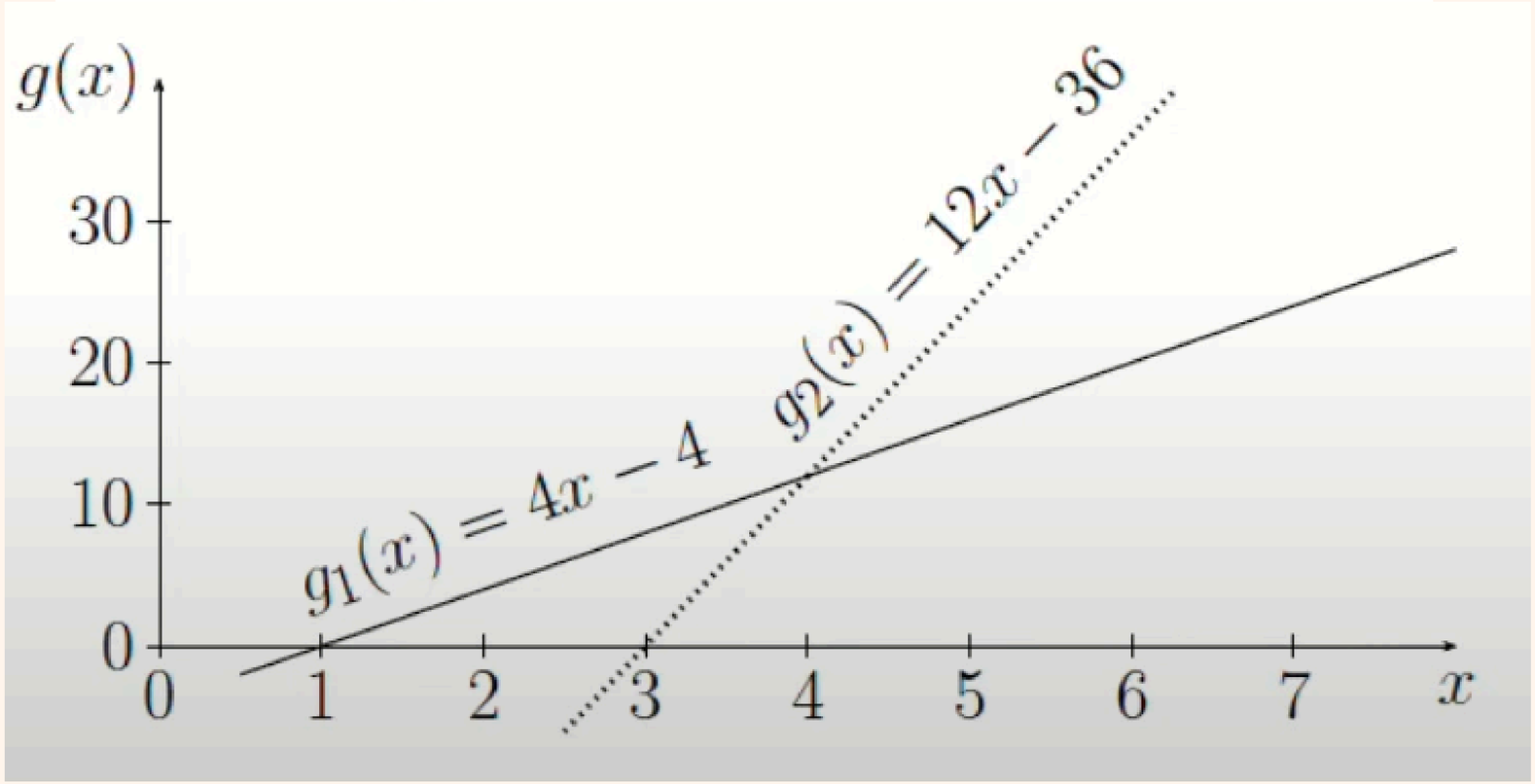
La dirección óptima para la proyección se obtiene resolviendo el problema de maximizar la razón de la dispersión entre clases a la dispersión dentro de las clases: calculo de autovalores y autovectores.

La idea es encontrar una dirección (o función) en el espacio de las variables que maximice la separación entre las categoría

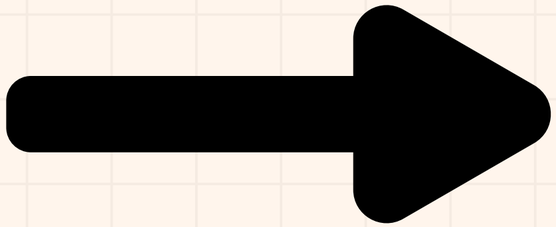
CLASIFICADOR LINEAL

Se define en terminos de funciones discriminantes lineales:

$$g_c(\boldsymbol{x}) = \sum_d w_{cd}\, x_d + w_{c0} = \boldsymbol{w}_c^t\, \boldsymbol{x} + w_{c0}$$



PROYECCIÓN DE DATOS



CLASIFICACIÓN

Una vez que se encuentra el vector w, **proyecta** los datos originales a esa **nueva dirección**:

- Si había datos en 4 dimensiones, ahora cada dato es un número en una línea (su proyección).
- Esa línea fue elegida porque ahí los grupos se ven más separados.

Ahora que los datos están proyectados, se puede usar una frontera simple (como un umbral) para decir

- Si la proyección cae a la **izquierda** → clase A
- Si cae a la **derecha** → clase B

¡¡LIMITACIONES!!

- Necesita saber de antemano los grupos.
  - **No funciona si los grupos son desconocidos** (para eso usarías clustering).
- Supone que los datos siguen una distribución normal.
  - Si tus **datos están muy desbalanceados o que no tengan distribución normal, puede fallar.**
- **Sensibilidad** a valores extremos
  - Un solo dato muy raro puede mover la frontera entre grupos.
- No es bueno si las **variables se parecen mucho entre sí.**

GRACIAS