

Introducción a la Ciencia de Datos

Maestría en Probabilidad y Estadística

Dr. Marco Antonio Aquino López

Centro de Investigación en Matemáticas

Agosto–Diciembre 2025



- **Adquisición & Gestión:** fuentes, bases, APIs.
- **Preprocesamiento & EDA:** limpieza, codificación, escalamiento, visualización.
- **Modelado:** ML y modelos estadísticos.
- **Validación:** métricas, incertidumbre.
- **Comunicación:** reportes, dashboards, decisiones.

Idea clave

ML es el *mecanismo de predicción*
Modelos estadísticos es el proceso por el cual se hace inferencia sobre los parámetros

Lo que ML *no* reemplaza

- **Diseño experimental:** sin buen diseño, hay sesgos irrecuperables.
- **Gestión de datos:** calidad, versionado, gobernanza.
- **Comunicación:** interpretabilidad, narrativa, ética.
- **Conocimiento de dominio:** define supuestos y utilidad de los modelos.

Resultados esperados

- **Predicción y explicación** con incertidumbre cuantificada.
- **Reproducibilidad**: trazabilidad de datos y código.
- **Impacto**: decisiones informadas en el dominio (salud, ambiente, finanzas, etc.).

Estadística, Computación y Matemáticas

- **Estadística:** inferencia, diseño, teoría de la decisión, modelos probabilísticos.
- **Computación:** algoritmos, estructuras de datos, bases, sistemas distribuidos.
- **Matemáticas:** optimización (convexa/no convexa), álgebra lineal, probabilidad, info. teórica.

Resumen

- La estadística/ML es el motor de *modelado* dentro del ciclo de Ciencia de Datos.
- Ciencia de Datos abarca desde *datos crudos* hasta *decisiones* con ética y reproducibilidad.

¿Que se considera ML?

- **Aprendizaje supervisado:** regresión, clasificación.
- **Aprendizaje no supervisado:** clustering, reducción de dimensión.
- **Aprendizaje por refuerzo:** políticas de decisión.

¿Qué es el Aprendizaje Supervisado?

- Se dispone de un conjunto de datos con pares (X_i, Y_i) :

$$\{(X_1, Y_1), \dots, (X_n, Y_n)\},$$

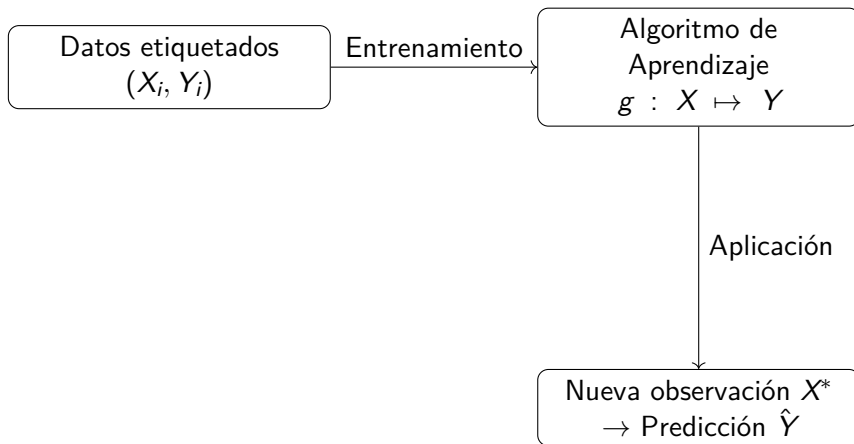
donde X_i son las variables explicativas (predictores) y Y_i es la variable respuesta (etiqueta).

- Objetivo: aprender una función g tal que $g(X)$ prediga Y con el menor error posible.
- Dos tipos de problema:
 - 1 **Clasificación:** Y es categórica (p. ej., correo *spam/no spam*).
 - 2 **Regresión:** Y es continua (p. ej., precio de una casa).

Técnicas Estadísticas que conforman el Aprendizaje Supervisado

- **Clasificación óptima y regla de Bayes** (*clasificación*).
- **Regresión lineal y logística** (*regresión / clasificación binaria*).
- **Discriminantes lineales**: LDA, QDA, criterio de Fisher (*clasificación*).
- **k-Vecinos más cercanos (k-NN)** (*clasificación y regresión*).
- **Árboles de decisión y bosques aleatorios** (*clasificación y regresión*).

Esquema del Aprendizaje Supervisado



¿Qué es la clasificación?

Preguntas naturales:

- ¿Qué son los objetos a clasificar?
- ¿Cómo definimos una **regla de clasificación**?
- ¿Cómo medir si la clasificación es buena o mala?
- ¿Cómo optimizar la regla?

Regla de clasificación

- Los datos son vectores $X \in \mathbb{R}^p$.
- Cada dato tiene una etiqueta $Y \in \{1, \dots, K\}$.
- Una **regla de clasificación** es

$$g: \mathbb{R}^p \rightarrow \{1, \dots, K\}, \quad x \mapsto g(x).$$

- Induce una partición:

$$\mathbb{R}^p = R_1 \cup \dots \cup R_K, \quad R_k = \{x: g(x) = k\}.$$

Clasificación óptima

Cada sub-conjunto de datos $C_i = \{X_j : Y_j = i\}$ se llama *población i* o *cluster i*. Denotaremos $\{X \rightarrow R_i\}$ el evento que el punto X sea categorizado en R_i .

Suponemos que los datos de la población C_i tienen una cierta distribución subyacente de densidad f_i y que tiene una probabilidad de inclusión

$$\pi_i = P(X \rightarrow C_i).$$

Sea E el evento de hacer un error de clasificación. Entonces

$$\begin{aligned} P(E) &= \sum_{i=1}^K P(E \cap \{X \rightarrow C_i\}) = \sum_{i=1}^K P(X \rightarrow C_i) P(E \mid \{X \rightarrow C_i\}) \\ &= \sum_{i=1}^K \pi_i P(E \mid \{X \rightarrow C_i\}) = \sum_{i=1}^K \pi_i \left(1 - P(X \in R_i \mid \{X \rightarrow C_i\})\right) \\ &= 1 - \sum_{i=1}^K \pi_i \int_{R_i} f_i(x) dx. \end{aligned}$$

Clasificación óptima

Definimos, para cada i ,

$$q_i(x) = P(X \rightarrow C_i \mid X = x).$$

Por la regla de Bayes,

$$q_i(x) = \frac{\pi_i P(X = x \mid X \rightarrow C_i)}{P(X = x)} = \frac{\pi_i f_i(x)}{\sum_{j=1}^K \pi_j f_j(x)}.$$

Además, $P(X = x) = \sum_{j=1}^K \pi_j f_j(x)$, de tal manera que

$$P(\text{NoErr}(g)) = \sum_{i=1}^K \int_{R_i} q_i(x) P(X = x) dx = \int_{\mathbb{R}^p} P(X = x) \left(\sum_{i=1}^K \mathbf{1}_{\{x \in R_i\}} q_i(x) \right) dx.$$

Clasificación óptima

Vimos que la probabilidad de no hacer errores en una regla de clasificación g está dada por

$$P(\text{NoErr}(g)) = \int P(X=x) h_g(x) dx, \quad h_g(x) = \sum_{i=1}^K \mathbf{1}_{\{x \in R_i\}} q_i(x).$$

Consideramos la regla g^* dada por las regiones

$$R_i^* = \{x \in \mathbb{R}^p : \forall j, q_i(x) \geq q_j(x)\}.$$

Sea g cualquier otra regla. Para un x dado, existe j tal que $x \in R_j$ y m tal que $x \in R_m^*$. Por definición de g^* ,

$$q_m(x) \geq q_j(x) \Rightarrow h_g(x) = \sum_{i=1}^K \mathbf{1}_{\{x \in R_i\}} q_i(x) = q_j(x) \leq q_m(x) = \sum_{i=1}^K \mathbf{1}_{\{x \in R_i^*\}} q_i(x) = h_{g^*}(x).$$

Clasificación óptima

Entonces,

$$P(\text{NoErr}(g)) = \int P(X=x) h_g(x) dx \leq \int P(X=x) h_{g^*}(x) dx = P(\text{NoErr}(g^*)).$$

La regla g^* es óptima para el criterio “probabilidad de errores”.

Cálculo de R_i^* : encontrar R_i^* equivale a comparar los valores $\pi_1 f_1(x), \dots, \pi_K f_K(x)$. Si conocemos $\{\pi_i, f_i\}$, el algoritmo es:

- ➊ **Input:** x
- ➋ Calcular $\pi_1 f_1(x), \dots, \pi_K f_K(x)$
- ➌ Calcular $k = \arg \max_i \pi_i f_i(x)$
- ➍ $k \rightarrow Y$
- ➎ **Output:** Y

Clasificación óptima de poblaciones normales

Si tenemos K poblaciones *normales* de proporciones π_i **iguales**,

$$f_i(x) = \frac{1}{(2\pi)^{p/2} |\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_i)^\top \Sigma_i^{-1}(x - \mu_i)\right).$$

Las **fronteras** de las regiones R_i^* son de la forma $f_i(x) = f_j(x)$ sujeto a $\forall k \neq i, j, f_k(x) \leq f_i(x)$.

Error de Bayes

En general, una regla óptima g^* satisface

$$g^* = \arg \min_{g: \mathbb{R}^p \rightarrow \{1, \dots, K\}} P(g(X) \neq Y).$$

La cantidad $L(g) = P(g(X) \neq Y)$ es la misma utilizada antes. El número real $L(g^*)$ se llama **error de Bayes**.

El error de Bayes es una *noción teórica* y no se puede calcular en la práctica salvo en casos particulares:

La distribución de (X, Y) es generalmente desconocida.

Un clasificador real se construye con datos $(X_1, Y_1), \dots, (X_n, Y_n)$. Denotamos g_n a dicho clasificador; su error teórico es $L(g_n)$ y su error *estimado* es

$$L_n(g_n) = P(g(X) \neq Y \mid \text{“datos”}).$$

Riesgo condicional y regla de decisión

Contexto de dos clases $Y \in \{0, 1\}$. Denotamos

$$\eta(x) = P(Y = 1 \mid X = x).$$

En este contexto, el riesgo esperado es

$$L(g) = \Pr(g(X) \neq Y).$$

Podemos descomponerlo como

$$L(g) = \mathbb{E}[\Pr(g(X) \neq Y \mid X)].$$

Riesgo condicional y regla de decisión

Contexto de dos clases $Y \in \{0, 1\}$. Denotamos

$$\eta(x) = P(Y = 1 \mid X = x).$$

En este contexto, el riesgo esperado es

$$L(g) = \Pr(g(X) \neq Y).$$

Podemos descomponerlo como

$$L(g) = \mathbb{E}[\Pr(g(X) \neq Y \mid X)].$$

Riesgo condicional: dado $X = x$:

$$\Pr(\text{error} \mid X = x, g(x) = 1) = \Pr(Y = 0 \mid X = x) = 1 - \eta(x),$$

$$\Pr(\text{error} \mid X = x, g(x) = 0) = \Pr(Y = 1 \mid X = x) = \eta(x).$$

Riesgo condicional y regla de decisión

Contexto de dos clases $Y \in \{0, 1\}$. Denotamos

$$\eta(x) = P(Y = 1 \mid X = x).$$

En este contexto, el riesgo esperado es

$$L(g) = \Pr(g(X) \neq Y).$$

Podemos descomponerlo como

$$L(g) = \mathbb{E}[\Pr(g(X) \neq Y \mid X)].$$

Riesgo condicional: dado $X = x$:

$$\Pr(\text{error} \mid X = x, g(x) = 1) = \Pr(Y = 0 \mid X = x) = 1 - \eta(x),$$

$$\Pr(\text{error} \mid X = x, g(x) = 0) = \Pr(Y = 1 \mid X = x) = \eta(x).$$

La **regla de Bayes** consiste en elegir, para cada x , la clase que minimiza la probabilidad de error condicional:

$$g^*(x) = \arg \min_{y \in \{0,1\}} \Pr(\text{error} \mid X = x, g(x) = y).$$

De Bayes \rightarrow Naive Bayes

El clasificador óptimo es

$$g^*(x) = \arg \max_k \pi_k f_k(x), \quad x = (x_1, \dots, x_p).$$

Por la *chain rule*:

$$\begin{aligned} f_k(x) &= P(X_1 = x_1 \mid Y = k) P(X_2 = x_2 \mid X_1 = x_1, Y = k) \cdots \\ &\quad \times P(X_p = x_p \mid X_1 = x_1, \dots, X_{p-1} = x_{p-1}, Y = k). \end{aligned}$$

De Bayes → Naive Bayes

El clasificador óptimo es

$$g^*(x) = \arg \max_k \pi_k f_k(x), \quad x = (x_1, \dots, x_p).$$

Por la *chain rule*:

$$\begin{aligned} f_k(x) &= P(X_1 = x_1 \mid Y = k) P(X_2 = x_2 \mid X_1 = x_1, Y = k) \cdots \\ &\quad \times P(X_p = x_p \mid X_1 = x_1, \dots, X_{p-1} = x_{p-1}, Y = k). \end{aligned}$$

Supuesto Naive (independencia condicional):

$$X_1, \dots, X_p \perp\!\!\!\perp \text{ dado } Y = k \implies f_k(x) = \prod_{j=1}^p f_{k,j}(x_j).$$

De Bayes → Naive Bayes

El clasificador óptimo es

$$g^*(x) = \arg \max_k \pi_k f_k(x), \quad x = (x_1, \dots, x_p).$$

Por la *chain rule*:

$$\begin{aligned} f_k(x) &= P(X_1 = x_1 \mid Y = k) P(X_2 = x_2 \mid X_1 = x_1, Y = k) \cdots \\ &\quad \times P(X_p = x_p \mid X_1 = x_1, \dots, X_{p-1} = x_{p-1}, Y = k). \end{aligned}$$

Supuesto Naive (independencia condicional):

$$X_1, \dots, X_p \perp\!\!\!\perp \text{ dado } Y = k \implies f_k(x) = \prod_{j=1}^p f_{k,j}(x_j).$$

Regla de decisión (en log para evitar underflow):

$$g_{\text{NB}}(x) = \arg \max_k \left[\log \pi_k + \sum_{j=1}^p \log f_{k,j}(x_j) \right]$$

Estimación de $f_{k,j}$: variables **continuas**

Gaussian Naive Bayes (una normal por atributo y clase):

$$x_j \mid Y = k \sim \mathcal{N}(\mu_{kj}, \sigma_{kj}^2), \quad f_{k,j}(x_j) = \frac{1}{\sqrt{2\pi\sigma_{kj}^2}} \exp\left(-\frac{(x_j - \mu_{kj})^2}{2\sigma_{kj}^2}\right).$$

Estimadores MLE con los datos de la clase k :

$$\hat{\mu}_{kj} = \frac{1}{n_k} \sum_{i:y_i=k} x_{ij}, \quad \hat{\sigma}_{kj}^2 = \frac{1}{n_k} \sum_{i:y_i=k} (x_{ij} - \hat{\mu}_{kj})^2.$$

Implementación práctica

Entrenamiento

- 1 Estimar $\hat{\pi}_k$ para cada clase k .
- 2 Para cada atributo j y clase k : estimar $f_{k,j}$ (Bernoulli/Multinomial/Gaussian).

Predicción para un x nuevo

$$\text{score}_k(x) = \log \hat{\pi}_k + \sum_{j=1}^p \log \hat{f}_{k,j}(x_j), \quad \hat{y} = \arg \max_k \text{score}_k(x).$$

Detalles numéricos & prácticos

- Trabajar en log para evitar underflow.
- Suavizado de Laplace ($\alpha > 0$) evita probabilidades nulas.

Ventajas y límites (cuándo usarlo)

Ventajas

- Muy rápido y simple; escala bien a alta dimensión y datos dispersos.
- Mezcla natural de variables discretas y continuas.
- Sorprendentemente competitivo como línea base (texto, filtrado de spam, etc.).

Limitaciones

- El supuesto de independencia condicional puede ser falso (atributos correlacionados).
- Puede estar mal calibrado en probabilidades; considerar *calibration* posterior.
- Decisiones sesgadas si hay desbalance severo sin ajustar π_k o umbrales.

Cuándo funciona bien

- Como base rápida para comparar con modelos más complejos.

Caso Gaussiano: Discriminante Lineal

- Supongamos: $X|Y = k \sim N_p(\mu_k, \Sigma)$, con $k = 1, 2$.
- La densidad conjunta es $f_k(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\{-\frac{1}{2}(x - \mu_k)^\top \Sigma^{-1}(x - \mu_k)\}$.
- La regla de Bayes compara $\pi_2 f_2(x)$ y $\pi_1 f_1(x)$.

Caso Gaussiano: Discriminante Lineal

- Supongamos: $X|Y = k \sim N_p(\mu_k, \Sigma)$, con $k = 1, 2$.
- La densidad conjunta es $f_k(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\{-\frac{1}{2}(x - \mu_k)^\top \Sigma^{-1}(x - \mu_k)\}$.
- La regla de Bayes compara $\pi_2 f_2(x)$ y $\pi_1 f_1(x)$.

Regla de Decisión

Asignar x a la clase 2 si

$$(\mu_2 - \mu_1)^\top \Sigma^{-1} x > \log \frac{\pi_1}{\pi_2} + \frac{1}{2}(\mu_2 + \mu_1)^\top \Sigma^{-1}(\mu_2 - \mu_1).$$

Regla de Decisión para Dos Poblaciones Normales (LDA)

Sea $X|Y = k \sim N_p(\mu_k, \Sigma)$ con Σ común y prior π_k . Asignar x a la clase 2 si:

$$\pi_2 f_2(x) > \pi_1 f_1(x)$$

es decir

$$\frac{\pi_2}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_2)^T \Sigma^{-1} (x-\mu_2)} > \frac{\pi_1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)}$$

Regla de Decisión para Dos Poblaciones Normales (LDA)

Sea $X|Y = k \sim N_p(\mu_k, \Sigma)$ con Σ común y prior π_k . Asignar x a la clase 2 si:

$$\pi_2 f_2(x) > \pi_1 f_1(x)$$

es decir

$$\frac{\pi_2}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_2)^T \Sigma^{-1} (x-\mu_2)} > \frac{\pi_1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)}$$

Tomando logaritmos y desarrollando términos cuadráticos:

$$-\frac{1}{2}(x-\mu_2)^T \Sigma^{-1} (x-\mu_2) > \log \frac{\pi_1}{\pi_2} - \frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)$$

Regla de Decisión para Dos Poblaciones Normales (LDA)

Sea $X|Y = k \sim N_p(\mu_k, \Sigma)$ con Σ común y prior π_k . Asignar x a la clase 2 si:

$$\pi_2 f_2(x) > \pi_1 f_1(x)$$

es decir

$$\frac{\pi_2}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_2)^T \Sigma^{-1} (x-\mu_2)} > \frac{\pi_1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)}$$

Tomando logaritmos y desarrollando términos cuadráticos:

$$-\frac{1}{2}(x-\mu_2)^T \Sigma^{-1} (x-\mu_2) > \log \frac{\pi_1}{\pi_2} - \frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)$$

Simplificando:

$$(\mu_2 - \mu_1)^T \Sigma^{-1} x > \log \frac{\pi_1}{\pi_2} + \frac{1}{2}(\mu_2^T \Sigma^{-1} \mu_2 - \mu_1^T \Sigma^{-1} \mu_1)$$

Regla de Decisión para Dos Poblaciones Normales (LDA)

Sea $X|Y = k \sim N_p(\mu_k, \Sigma)$ con Σ común y prior π_k . Asignar x a la clase 2 si:

$$\pi_2 f_2(x) > \pi_1 f_1(x)$$

es decir

$$\frac{\pi_2}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2)} > \frac{\pi_1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)}$$

Tomando logaritmos y desarrollando términos cuadráticos:

$$-\frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2) > \log \frac{\pi_1}{\pi_2} - \frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)$$

Simplificando:

$$(\mu_2 - \mu_1)^T \Sigma^{-1} x > \log \frac{\pi_1}{\pi_2} + \frac{1}{2}(\mu_2^T \Sigma^{-1} \mu_2 - \mu_1^T \Sigma^{-1} \mu_1)$$

Finalmente, se obtiene una regla lineal:

$$w^T x > w_0, \quad w = \Sigma^{-1}(\mu_2 - \mu_1), \quad w_0 = \log \frac{\pi_1}{\pi_2} + \frac{1}{2}(\mu_2^T \Sigma^{-1} \mu_2 - \mu_1^T \Sigma^{-1} \mu_1)$$

Caso Gaussiano: Discriminante Lineal

Regla de Decisión

Asignar x a la clase 2 si

$$(\mu_2 - \mu_1)^\top \Sigma^{-1} x > \log \frac{\pi_1}{\pi_2} + \frac{1}{2}(\mu_2 + \mu_1)^\top \Sigma^{-1}(\mu_2 - \mu_1).$$

- Definimos:

$$w = \Sigma^{-1}(\mu_2 - \mu_1), \quad w_0 = \log \frac{\pi_1}{\pi_2} + \frac{1}{2}(\mu_2 + \mu_1)^\top \Sigma^{-1}(\mu_2 - \mu_1)$$

- Clasificador:

$$g(x) = \begin{cases} 2, & w^\top x > w_0 \\ 1, & w^\top x \leq w_0. \end{cases}$$

- La frontera de decisión es un **hiperplano**: $\{x : w^\top x = w_0\}$.

Interpretación Geométrica

- Los datos se proyectan sobre la dirección w .
- La clasificación depende de si x cae a la izquierda o derecha de la frontera.

Motivación del Criterio de Fisher

- Consideremos dos poblaciones con medias \hat{m}_0, \hat{m}_1 y matrices de covarianza S_0, S_1 .
- Queremos proyectar los datos $X \in \mathbb{R}^p$ sobre una dirección a y luego clasificar en 1D.
- Idea: elegir a de manera que las proyecciones $z_i = a^T X_i$ estén lo más separadas posible entre clases.

Definición Formal del Criterio de Fisher

Criterio

$$J(a) = \frac{(a^T(\hat{m}_1 - \hat{m}_0))^2}{a^T S_w a} \quad \text{donde } S_w = S_0 + S_1 \text{ es la matriz de dispersión intra-clase}$$

- El numerador representa la **separación entre clases** en la dirección a .
- El denominador representa la **variabilidad dentro de las clases** en la dirección a .
- Queremos maximizar $J(a)$ para encontrar la dirección de proyección óptima.

Construcción de $J(a)$: de datos p -dimensionales a 1D

Paso 1. Proyección y medias proyectadas. Dado $a \in \mathbb{R}^p$, definimos $z = a^\top x$. Las medias de clase proyectadas son $\bar{z}_k = a^\top \hat{m}_k$ ($k = 0, 1$) y la media global $\bar{z} = a^\top \hat{m}$.

Paso 2. Dispersión entre clases en 1D. La separación de medias proyectadas es $(\bar{z}_1 - \bar{z}_0)^2 = (a^\top (\hat{m}_1 - \hat{m}_0))^2$.

Paso 3. Dispersión intra-clase en 1D. La varianza total dentro de clases tras proyectar es

$$s_w^2(a) = \sum_{k \in \{0,1\}} \sum_{i \in C_k} (a^\top x_i - a^\top \hat{m}_k)^2 = a^\top \underbrace{\left(\sum_k \sum_{i \in C_k} (x_i - \hat{m}_k)(x_i - \hat{m}_k)^\top \right)}_{S_w} a = a^\top S_w a.$$

Paso 4. Criterio adimensional (cociente de Rayleigh) agregué una diapositiva de biblio. Para comparar “separación/variabilidad” en la misma escala:

$$J(a) = \frac{(a^\top (\hat{m}_1 - \hat{m}_0))^2}{a^\top S_w a}.$$

Observaciones.

- $J(a)$ es invariante a re-escalamientos de a (si $a \leftarrow c a$, numerador y denominador se multiplican por c^2).
- Maximizar $J(a) \iff$ resolver $\max (a^\top \Delta m)^2$ s.a. $a^\top S_w a = 1$ con $\Delta m = \hat{m}_1 - \hat{m}_0$.

Optimización de $J(a)$

- Como $J(a)$ es invariante a la escala de a , podemos imponer la restricción $a^T S_w a = 1$.
- Entonces el problema es:

$$\max_a (a^T (\hat{m}_1 - \hat{m}_0))^2 \quad \text{sujeto a } a^T S_w a = 1$$

- Resolviendo con multiplicadores de Lagrange, la condición de primer orden es:

$$S_w a = \lambda (\hat{m}_1 - \hat{m}_0)$$

- Solución:

$$a^* = S_w^{-1} (\hat{m}_1 - \hat{m}_0)$$

Interpretación de a^*

- a^* es la dirección que maximiza la razón “varianza entre clases / varianza dentro de clases”.
- En el espacio proyectado $z = a^{*T}x$ obtenemos una variable 1D con máxima separabilidad.
- Luego podemos usar un umbral z_0 (por ejemplo, el punto medio entre medias proyectadas) para clasificar.

Discriminante Cuadrático (QDA)

- Si $X|Y = k \sim N_p(\mu_k, \Sigma_k)$ con matrices de covarianza **diferentes**, la regla de Bayes se vuelve cuadrática.
- La función discriminante para la clase k es:

$$\delta_k(x) = -\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x - \mu_k)^\top \Sigma_k^{-1} (x - \mu_k) + \log \pi_k$$

- Clasificamos x en la clase con mayor $\delta_k(x)$.

Discriminante Cuadrático (QDA) - Detalles Formales

Supongamos $X|Y = k \sim N_p(\mu_k, \Sigma_k)$ con matrices de covarianza **distintas**. La densidad de clase k es:

$$f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} \exp \left[-\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right]$$

La regla de Bayes asigna x a la clase k que maximiza $\pi_k f_k(x)$. Tomando logaritmos:







$$\log \pi_k - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)$$

Definimos la función discriminante cuadrática:

$$\delta_k(x) = -\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \log \pi_k$$

- La frontera de decisión $\delta_1(x) = \delta_2(x)$ es una **cuádrica** en x (elipsoide, parábola, hipérbola según el caso).
- Si $\Sigma_1 = \Sigma_2$, se recupera el discriminante lineal de LDA.

Bibliografía

-  Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, 2nd Ed.
-  Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
-  Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.
-  Friedman, J. (1989). Regularized Discriminant Analysis. *Journal of the American Statistical Association*, 84(405), 165–175.
-  Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. Wiley.
-  Venables, W. N., Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer.