Introducción Matemática al Análisis Multivariante y Aprendizaje Automático

Máster en Análisis de Datos para la Inteligencia de Negocios

30 de enero de 2026

1. Marco probabilístico del aprendizaje

Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad. Sean variables aleatorias

$$X : \Omega \to \mathcal{X} \subseteq \mathbb{R}^p \qquad Y : \Omega \to \mathcal{Y}$$

de las que observamos una muestra aleatoria simple $\mathcal{D}_n$, o sea, con componentes D_i independientes y distribuidos idénticamente

$$\mathcal{D}_n = (D_i)_{i=1}^n = (X_i, Y_i)_{i=1}^n$$

Sea $\mathcal{F}$ una familia de funciones predictivas $f : \mathcal{X} \to \mathcal{Y}$ y sea $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ una función de pérdida.

1.1. Riesgo poblacional y empírico

$$R(f) = \mathbb{E}[L(Y, f(X))], \qquad \hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^n L(Y_i, f(X_i)).$$

El problema fundamental del aprendizaje es

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \hat{R}_n(f).$$

2. Aprendizaje supervisado y no supervisado

2.1. Aprendizaje supervisado

En regresión, $\mathcal{Y} = \mathbb{R}$ y típicamente $L(y, \hat{y}) = (y - \hat{y})^2$.
En clasificación, $\mathcal{Y} = 1, \ldots, K$ y puede tomarse

$$L(y, \hat{y}) = \mathbf{1}(y \neq \hat{y}).$$

2.2. Aprendizaje no supervisado

En ausencia de variable respuesta, el aprendizaje se formula como un problema de optimización geométrica.

PCA. Sea $X \in \mathbb{R}^{n \times p}$ la matriz de datos centrados. PCA busca

$$\max_{|w|=1} w^\top S_T w, \qquad S_T = \frac{1}{n} X^\top X.$$

k-means.

$$\min_{c, \mu} \sum_{i=1}^n |X_i - \mu_{c(i)}|^2.$$

3. Sesgo, varianza y complejidad

La descomposición sesgo-varianza en regresión cuadrática es

$$\mathbb{E}[(Y - \hat{f}(X))^2] = \text{sesgo}^2 + \text{varianza} + \sigma^2.$$

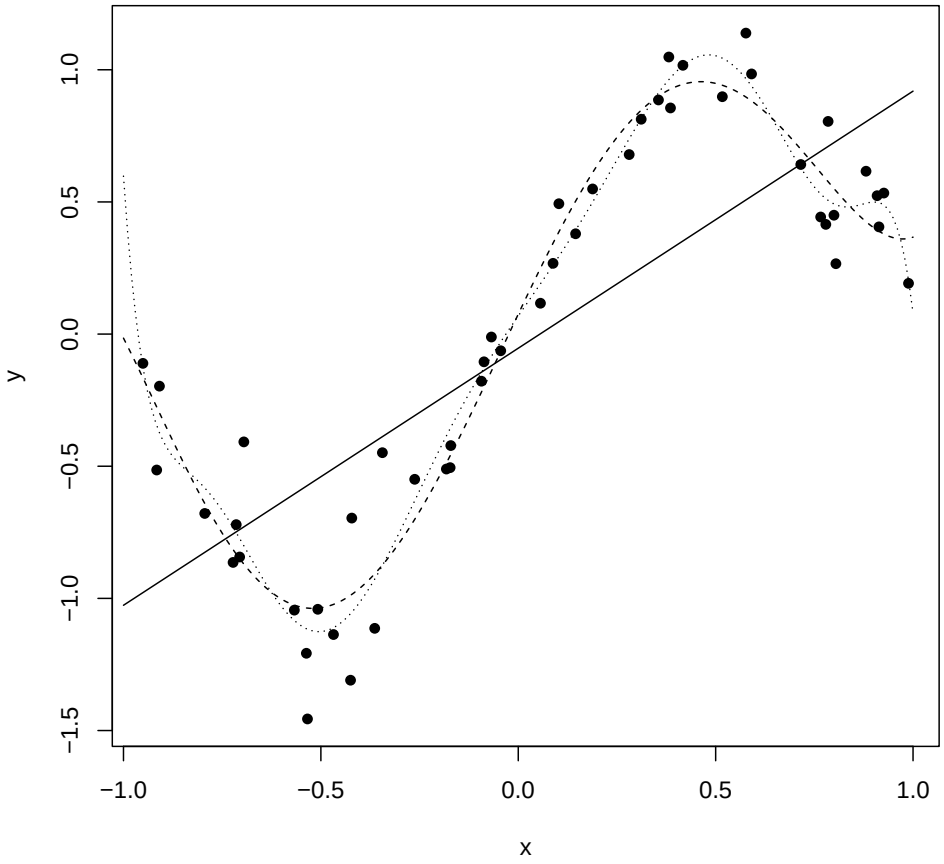

Figura 1: Ejemplo de infraajuste y sobreajuste.

4. Validación de modelos

4.1. Hold-out

Se divide la muestra en conjuntos de entrenamiento y prueba:

$$\mathcal{D}_n = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{test}}.$$

4.2. Validación cruzada

En la validación cruzada k-fold, la muestra se divide en k subconjuntos y se define

$$\hat{R}_{CV}(f) = \frac{1}{k} \sum_{j=1}^{k} \hat{R}_{n_j}^{(-j)}(f).$$

El caso límite $k = n$ corresponde a Leave-One-Out (DUF).

5. Análisis discriminante lineal (LDA)

Sea K el número de clases. Definimos:

$$\bar{x}_k = \frac{1}{n_k} \sum_{i:Y_i=k} X_i, \qquad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

5.1. Matriz within-class

$$S_W = \sum_{k=1}^{K} \sum_{i:Y_i=k} (X_i - \bar{x}_k)(X_i - \bar{x}_k)^\top.$$

5.2. Matriz between-class

$$S_B = \sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^\top.$$

5.3. Criterio de Fisher

El análisis discriminante lineal busca direcciones $w \in \mathbb{R}^p$ que maximicen el cociente

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

Esto conduce al problema de autovalores generalizado

$$S_B w = \lambda S_W w.$$

6. Comparación PCA vs LDA

PCA maximiza la varianza total:

$$\max_{|w|=1} w^\top S_T w, \qquad S_T = S_W + S_B.$$

LDA maximiza la separabilidad entre clases relativa a la dispersión interna.

7. Unificación mediante optimización regularizada

Muchos métodos de aprendizaje pueden escribirse como

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \left\{ \hat{R}_n(f) + \Omega(f) \right\},$$

donde $\Omega(f)$ es un término de regularización.
Ejemplos:

- Ridge: $\Omega(f) = \lambda |\beta|^2$.
- Lasso: $\Omega(f) = \lambda |\beta|_1$.
- PCA: restricción sobre el rango.
- LDA: cociente de Rayleigh.

8. Conclusión

El aprendizaje automático puede entenderse como una extensión predictiva del análisis multivariante clásico, formulada como un problema de optimización estadística bajo incertidumbre.

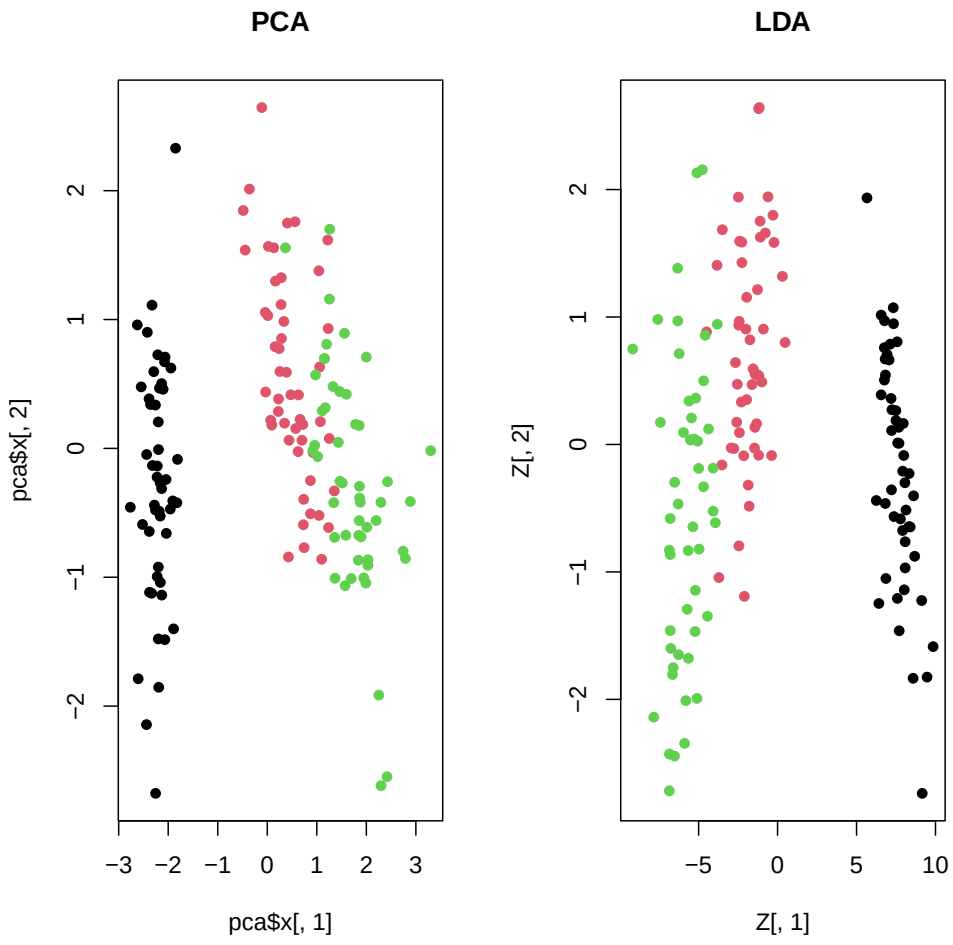

Figura 2: Proyección PCA vs LDA en el conjunto iris.