Tarea 2: Clasificación supervisada

Dr. Marco Antonio Aquino López
Maestría en Probabilidad y Estadística,
CIMAT

Agosto–Diciembre 2025

Objetivo:
Aplicar los conceptos de clasificación supervisada en un caso real, utilizando el *Bank Marketing Dataset*, disponible en el repositorio UCI Machine Learning:

https://archive.ics.uci.edu/dataset/222/bank+marketing

El trabajo debe mostrar no solo la implementación técnica de los clasificadores vistos en clase, sino también la justificación teórica de las decisiones tomadas. Además, se espera que los alumnos investiguen y documenten:

- El origen de los datos: £qué institución los generó y con qué propósito?
- La naturaleza de la variable respuesta y de los predictores.
- Posibles problemas en la base: desbalance de clases, variables categóricas de difícil codificación, redundancia o correlaciones.
- El tipo de conclusiones que pueden extraerse después del análisis.

Instrucciones generales:

- El proyecto se realizará en equipos de 3 estudiantes, asignados aleatoriamente.
- El código deberá entregarse en Python, con comentarios claros.
- Se debe incluir un informe escrito (máximo 10 páginas) con explicaciones formales, gráficos y conclusiones.
- Toda afirmación debe estar sustentada en resultados numéricos, propiedades estadísticas o literatura de referencia.

Lineamientos del proyecto:

1. **Exploración inicial:** describir la base de datos, número de variables, número de observaciones y tipos de datos. Identificar el tipo de variable respuesta y su distribución.
2. **Preprocesamiento:** aplicar codificación y escalamiento cuando sea necesario. Justificar cada decisión en términos prácticos y teóricos.
3. **Modelado:** entrenar al menos los clasificadores discutidos en clase (Naive Bayes, LDA, QDA, Fisher y k-NN), tomando decisiones fundamentadas sobre su entrenamiento y validación.
4. **Evaluación:** calcular estimadores de desempeño (exactitud, sensibilidad, especificidad, F1, AUC) mediante técnicas de validación apropiadas.

5. **Análisis crítico:** comparar y discutir los resultados de los clasificadores, reflexionando sobre cómo la naturaleza de los datos influye en el desempeño y qué aportes ofrece cada técnica.

Entrega:

- Código reproducible en un archivo `.py`.
- Informe en PDF: máximo 10 páginas, incluyendo gráficas, tablas y referencias.
- Un repositorio GitHub por equipo, con un `README.md` describiendo el proyecto, su uso y la participación del equipo.
- Fecha de entrega: *(1/10/25) antes de las 11:59 Pm.*

Parte II. Análisis computacional mediante simulaciones

El propósito de esta segunda parte es estudiar, en un entorno controlado donde la distribución verdadera es conocida, el comportamiento de los clasificadores vistos en el curso frente al clasificador óptimo (de Bayes). Para ello, deberán diseñar y ejecutar simulaciones con datos sintéticos generados a partir de distribuciones normales multivariadas, y comparar el riesgo de clasificación de cada método tanto contra el riesgo óptimo como contra estimadores obtenidos por métodos de validación.

Planteamiento general (abierto)

- Considere dos clases $Y \in \{0, 1\}$ con priors π_0, π_1 (pueden tomar $\pi_0 = \pi_1 = 0{,}5$ y/o explorar desbalance). Genere muestras i.i.d. de

$$X \mid Y = 0 \sim \mathcal{N}_p(\mu_0, \Sigma_0), \qquad X \mid Y = 1 \sim \mathcal{N}_p(\mu_1, \Sigma_1).$$

- Estudie al menos dos escenarios: (i) *covarianzas iguales* $\Sigma_0 = \Sigma_1$ (LDA coincide con Bayes) y (ii) *covarianzas distintas* $\Sigma_0 \neq \Sigma_1$ (QDA coincide con Bayes). Proponga parámetros (μ_k, Σ_k) de manera que las distribuciones de las clases presenten distintos grados de separación (casos fáciles, intermedios y difíciles de clasificar).
- Varíe **tamaño muestral** n (por clase), y el hiperparámetro de **k-NN** (número de vecinos k), documentando sus decisiones.

Métodos a comparar

- Clasificador óptimo (Bayes): impleméntelo usando las densidades gaussianas verdaderas y los priors fijados. Úselo como *línea base*.
- Naive Bayes (gaussiano), LDA, QDA, criterio de Fisher (proyección 1D con umbral), k-NN (distintas k; opcional: ponderado por distancia).

Riesgo verdadero vs. validación

- **Riesgo verdadero $L(g)$:** dado que las distribuciones de las clases son conocidas, el riesgo de Bayes y el de cada clasificador pueden calcularse en forma exacta (o al menos en forma cerrada en escenarios gaussianos simples).
- **Riesgo estimado por validación:** para cada método, estime el desempeño a partir de los datos generados empleando validación cruzada estratificada (y opcionalmente bootstrap .632/.632+). Compare estos estimadores frente al riesgo verdadero, discutiendo su sesgo y variabilidad.

Diseño experimental sugerido (flexible)

- **Barridos de parámetros:**

$$n \in \{50, 100, 200, 500\} \text{ (por clase)}, \quad k \in \{1, 3, 5, 11, 21\}.$$

- **Replicación:** para cada combinación, realice R réplicas independientes (p. ej. $R = 20$) para promediar y reportar media $\pm$ desviación estándar de L.
- **Escenarios:** (i) $\Sigma_0 = \Sigma_1$ (LDA óptimo), (ii) $\Sigma_0 \neq \Sigma_1$ (QDA óptimo), (iii) *desbalance* de clases (p. ej. $\pi_1 = 0{,}2$), (iv) *correlaciones fuertes* y/o *mal condicionamiento* en Σ_k.

Productos esperados (mínimos)

- **Gráficas de evolución:**
  - $L(g)$ vs. n (por método).
  - L(k-NN) vs. k (curvas por n).
  - Brechas $L(g) - L$(Bayes) vs. n y/o k (líneas o *heatmaps*).
  - Comparación *validación* vs. *Monte Carlo*: $L_{\mathrm{CV}}(g)$ frente a $L(g)$
- **Tablas resumidas:** medias y desviaciones estándar de L por método y condición.
- **Discusión crítica:** cuándo y por qué cada método se acerca (o no) a Bayes; efectos de n, p y k; sensibilidad a supuestos (igualdad de covarianzas, correlaciones, desbalance).

Notas finales

- Esta sección **no cuenta en el límite de páginas del reporte de la Parte I**. Puede extenderse tanto como el equipo considere necesario.
- Los puntos anteriores constituyen *sugerencias*. Se espera que los estudiantes exploren estas propiedades de manera amplia, con un trabajo bien documentado, resultados numéricos y referencias bibliográficas pertinentes.