Clase 1- Módulo 2: Introducción a la analítica

Mauricio Alejandro Mazo Lopera

Universidad Nacional de Colombia
Facultad de Ciencias
Escuela de Estadística
Medellín

Métodos de clasificación

Conectando con el módulo anterior, en este módulo continuaremos con los métodos de clasificación, entre los cuales ya vieron:

- **Regresión logística**.
- **Análisis de discriminante lineal (LDA):** Usa el teorema de Bayes y permite al discriminante considerar un solo predictor lineal o múltiples predictores lineales.

Problema de contextualización:

Deserción Estudiantil

Dado un estudiante que entra a la universidad con un perfil conocido, ¿cuál es la probabilidad de que abandone sus estudios?

Problema de contextualización:

El perfil se puede construir de acuerdo a un amplio conjunto de covariables, como por ejemplo:

$$
\begin{array}{lcl}
X_1 & : & \text{Puntaje en las pruebas Saber 11} \\
X_2 & : & \text{Ingresos mensuales familiares} \\
X_3 & : & \text{Edad al ingresar a la carrera} \\
X_4 & : & \text{Número de créditos en el primer semestre} \\
\vdots & & \vdots
\end{array}
$$

Estas variables se pueden resumir en un vector $\boldsymbol{X}$.

Problema de contextualización:

El perfil se puede construir de acuerdo a un amplio conjunto de covariables, como por ejemplo:

$$
\begin{array}{lcl}
X_1 & : & \text{Puntaje en las pruebas Saber 11} \\
X_2 & : & \text{Ingresos mensuales familiares} \\
X_3 & : & \text{Edad al ingresar a la carrera} \\
X_4 & : & \text{Número de créditos en el primer semestre} \\
\vdots & & \vdots
\end{array}
$$

Estas variables se pueden resumir en un vector $\boldsymbol{X}$.

El objetivo del estudio (deserción) lo podemos relacionar con la variable categórica:

$$
Y = \begin{cases} 0, & \text{No deserta} \\ 1, & \text{Deserta} \end{cases}
$$

Problema de contextualización:

Si se selecciona aleatoriamene un estudiante que acaba de ingresar a la universidad y sabemos que tiene un perfil dado por $\boldsymbol{X} = \boldsymbol{x}$, ¿cuál es la probabilidad de que deserte?

Problema de contextualización:

Si se selecciona aleatoriamene un estudiante que acaba de ingresar a la universidad y sabemos que tiene un perfil dado por $\boldsymbol{X} = \boldsymbol{x}$, ¿cuál es la probabilidad de que deserte?

Esto se traduce en lenguaje estadístico con la siguiente probabilidad condicional

$$P(Y = 1|\boldsymbol{X} = \boldsymbol{x})$$

Problema de contextualización:

Si se selecciona aleatoriamene un estudiante que acaba de ingresar a la universidad y sabemos que tiene un perfil dado por $\boldsymbol{X} = \boldsymbol{x}$, ¿cuál es la probabilidad de que deserte?

Esto se traduce en lenguaje estadístico con la siguiente probabilidad condicional

$$P(Y = 1|\boldsymbol{X} = \boldsymbol{x})$$

$$= \frac{P(\boldsymbol{X} = \boldsymbol{x}|Y = 1)P(Y = 1)}{P(\boldsymbol{X} = \boldsymbol{x}|Y = 0)P(Y = 0) + P(\boldsymbol{X} = \boldsymbol{x}|Y = 1)P(Y = 1)}$$

esto por el Teorema de Bayes.

Problema de contextualización:

Denotando

$$\pi_0 = P(Y = 0),\ \pi_1 = P(Y = 1),$$

$$f_0(\boldsymbol{x}) = P(\boldsymbol{X} = \boldsymbol{x}|Y = 0) \text{ y } f_1(\boldsymbol{x}) = P(\boldsymbol{X} = \boldsymbol{x}|Y = 1)$$

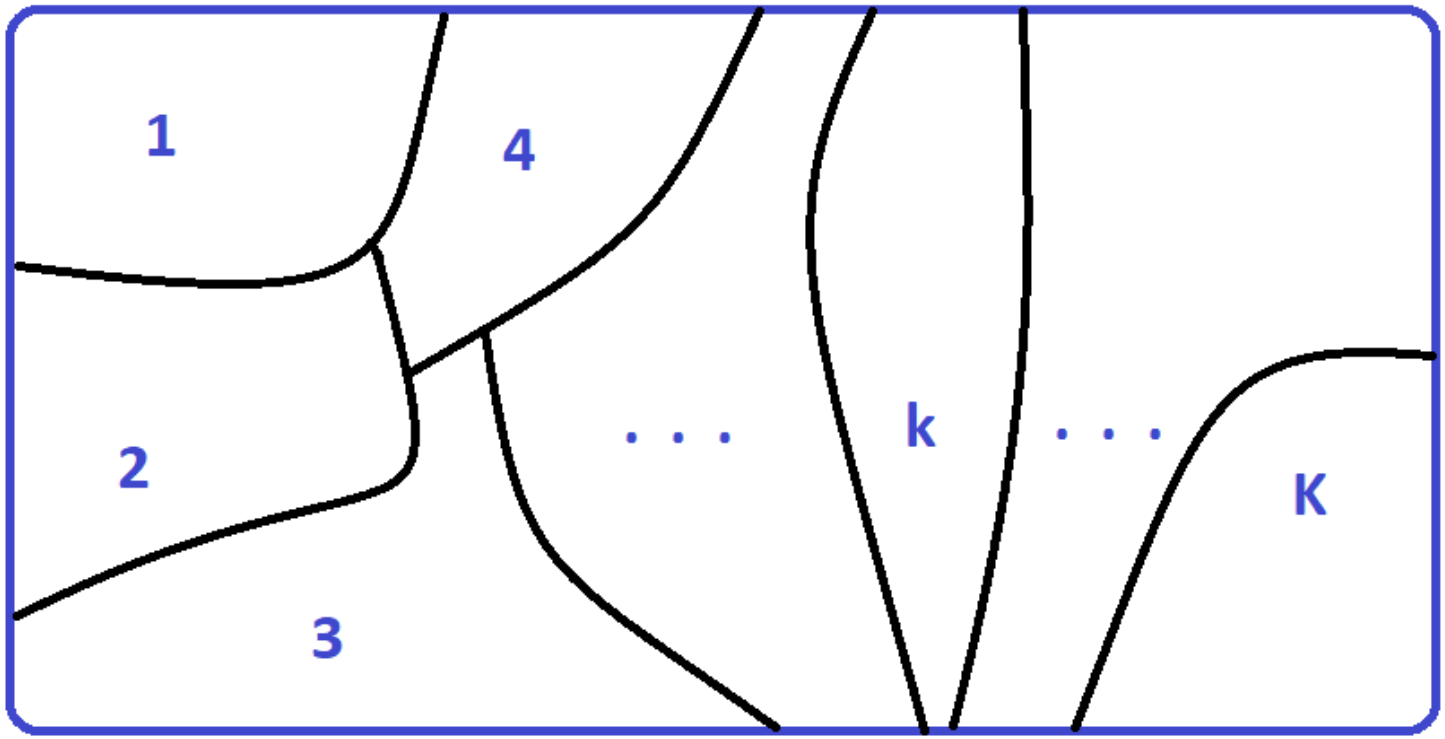
se tiene que

$$P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \frac{\pi_1 f_1(\boldsymbol{x})}{\pi_0 f_0(\boldsymbol{x}) + \pi_1 f_1(\boldsymbol{x})}$$

Surge entonces la pregunta: ¿cómo definir $\pi_0, \pi_1, f_0(\boldsymbol{x})$ y $f_1(\boldsymbol{x})$?

Teorema de Bayes

Si tienen en total K categorías, la pregunta que surge es: ¿cuál es la probabilidad de pertenecer a la k-ésima categoría, dado que se tiene un perfil caracterizado por $\boldsymbol{X} = \boldsymbol{x}$?



Teorema de Bayes

Usando el teorema de Bayes, dicha probabilidad se calcula como:

$$P(Y = k|\boldsymbol{X} = \boldsymbol{x}) = \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{j=1}^{K} \pi_j f_j(\boldsymbol{x})}$$

Teorema de Bayes

Usando el teorema de Bayes, dicha probabilidad se calcula como:

$$P(Y = k|\boldsymbol{X} = \boldsymbol{x}) = \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{j=1}^{K} \pi_j f_j(\boldsymbol{x})}$$

donde:

- $\pi_1, \pi_2, \ldots, \pi_k$ son las probabilidades apriori de pertenecer a cada una de las categorías.

Teorema de Bayes

Usando el teorema de Bayes, dicha probabilidad se calcula como:

$$P(Y = k|\boldsymbol{X} = \boldsymbol{x}) = \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{j=1}^{K} \pi_j f_j(\boldsymbol{x})}$$

donde:

- $\pi_1, \pi_2, \ldots, \pi_k$ son las probabilidades apriori de pertenecer a cada una de las categorías.
- $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^T$ es un vector de valores que toma el vector de covariables $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T$.

Teorema de Bayes

Usando el teorema de Bayes, dicha probabilidad se calcula como:

$$P(Y = k|\boldsymbol{X} = \boldsymbol{x}) = \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{j=1}^{K} \pi_j f_j(\boldsymbol{x})}$$

donde:

- $\pi_1, \pi_2, \ldots, \pi_k$ son las probabilidades apriori de pertenecer a cada una de las categorías.
- $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^T$ es un vector de valores que toma el vector de covariables $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T$.
- $f_j(\boldsymbol{x})$ es la función densidad conjunta del vector $\boldsymbol{X}$ dentro de la categoría j, para cada $j = 1, 2, \ldots, K$.

Análisis de discriminante lineal (LDA)

En análisis de discriminante lineal asumieron que

$\boldsymbol{X} \sim N(\boldsymbol{\mu}_k, \Sigma)$, para todo $k = 1, 2, \ldots, K$, es decir,

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right)$$

Es decir, cada categoría tiene su vector de medias para las covariables, $\boldsymbol{X}$, pero la matriz de covarianzas, Σ, es la misma para todas.

Análisis de discriminante lineal (LDA)

En análisis de discriminante lineal asumieron que

$\boldsymbol{X} \sim N(\boldsymbol{\mu}_k, \Sigma)$, para todo $k = 1, 2, \ldots, K$, es decir,

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right)$$

Es decir, cada categoría tiene su vector de medias para las covariables, $\boldsymbol{X}$, pero la matriz de covarianzas, Σ, es la misma para todas.

¿Estadísticamente qué significa lo anterior?

Análisis de discriminante lineal (LDA)

Volviendo a la fórmula del teorema de Bayes:

$$P(Y = k|\boldsymbol{X} = \boldsymbol{x}) \quad = \quad \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{j=1}^{K} \pi_j f_j(\boldsymbol{x})}$$

Análisis de discriminante lineal (LDA)

Volviendo a la fórmula del teorema de Bayes:

$$
\begin{aligned}
P(Y = k|\boldsymbol{X} = \boldsymbol{x}) &= \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{j=1}^{K} \pi_j f_j(\boldsymbol{x})} \\
&= \frac{\pi_k \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)}{\sum_{j=1}^{K} \pi_j \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right)}
\end{aligned}
$$

Análisis de discriminante lineal (LDA)

Volviendo a la fórmula del teorema de Bayes:

$$
\begin{aligned}
P(Y=k|\boldsymbol{X}=\boldsymbol{x}) &= \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{j=1}^{K}\pi_j f_j(\boldsymbol{x})} \\
&= \frac{\pi_k \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)}{\sum_{j=1}^{K}\pi_j \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right)} \\
&= \frac{\pi_k \cancel{\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)}{\cancel{\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}} \sum_{j=1}^{K}\pi_j \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right)}
\end{aligned}
$$

Volviendo a la fórmula del teorema de Bayes:

$$
\begin{aligned}
P(Y=k|\boldsymbol{X}=\boldsymbol{x}) &= \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{j=1}^{K}\pi_j f_j(\boldsymbol{x})} \\
&= \frac{\pi_k \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)}{\sum_{j=1}^{K}\pi_j \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right)} \\
&= \frac{\pi_k \cancel{\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)}{\cancel{\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}} \sum_{j=1}^{K}\pi_j \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right)} \\
&= \frac{\pi_k \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)}{\sum_{j=1}^{K}\pi_j \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right)}
\end{aligned}
$$

Análisis de discriminante lineal (LDA)

Volviendo a la fórmula del teorema de Bayes:

$$
\begin{aligned}
P(Y=k|\boldsymbol{X}=\boldsymbol{x}) &= \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{j=1}^{K}\pi_j f_j(\boldsymbol{x})} \\
&= \frac{\pi_k \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)}{\sum_{j=1}^{K}\pi_j \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right)} \\
&= \frac{\pi_k \cancel{\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)}{\cancel{\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}} \sum_{j=1}^{K}\pi_j \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right)} \\
&= \frac{\pi_k \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)}{\sum_{j=1}^{K}\pi_j \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right)}
\end{aligned}
$$

Así, la categoría donde esta expresión es mayor para el perfil $\boldsymbol{X}=\boldsymbol{x}$, se alcanza cuando se obtiene el mayor valor en el numerador, es decir, cuando

$$\pi_k \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right) \text{ tiene el mayor valor.}$$

Análisis de discriminante lineal (LDA)

Para simplificar, tomamos logaritmo natural en la expresión anterior, lo cual no afecta el hecho de que buscamos el mayor valor originalmente para $P(Y = k|\boldsymbol{X} = \boldsymbol{x})$.

Análisis de discriminante lineal (LDA)

Para simplificar, tomamos logaritmo natural en la expresión anterior, lo cual no afecta el hecho de que buscamos el mayor valor originalmente para $P(Y = k|\boldsymbol{X} = \boldsymbol{x})$.

Así, nuestro objetivo es clasificar el perfil $\boldsymbol{X} = \boldsymbol{x}$ en la clase k donde

$$\delta_k(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) + \log(\pi_k)$$

tiene el mayor valor.

Análisis de discriminante lineal (LDA)

Para simplificar, tomamos logaritmo natural en la expresión anterior, lo cual no afecta el hecho de que buscamos el mayor valor originalmente para $P(Y = k|\boldsymbol{X} = \boldsymbol{x})$.

Así, nuestro objetivo es clasificar el perfil $\boldsymbol{X} = \boldsymbol{x}$ en la clase k donde

$$\delta_k(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) + \log(\pi_k)$$

tiene el mayor valor.

Simplificando (**TAREA**), podemos escribir

$$\delta_k(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log(\pi_k)$$

De esta expresión surge el nombre "Discriminante lineal", ya que $\delta_k(\boldsymbol{x})$ es una función lineal de $\boldsymbol{x}$.

Análisis de discriminante lineal (LDA)

¿Cómo se estiman $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}_k$ y π_k para $k = 1, 2, \ldots, K$?

Análisis de discriminante lineal (LDA)

¿Cómo se estiman $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}_k$ y π_k para $k = 1, 2, \ldots, K$?

En este caso las probabilidades apriori se calculan con los datos de `entrenamiento`, contando cuántos individuos pertenecen a la k-ésima clase, es decir, para una base de datos de entrenamiento con n_e individuos,

$$\widehat{\pi}_k = \frac{\#\text{ de individuos que pertenecen a la clase } k}{n_e}$$

Análisis de discriminante lineal (LDA)

¿Cómo se estiman $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}_k$ y π_k para $k = 1, 2, \ldots, K$?

En este caso las probabilidades apriori se calculan con los datos de `entrenamiento`, contando cuántos individuos pertenecen a la k-ésima clase, es decir, para una base de datos de entrenamiento con n_e individuos,

$$\widehat{\pi}_k = \frac{\#\text{ de individuos que pertenecen a la clase } k}{n_e}$$

Así mismo, si denotamos el número de indiviuos en la k-ésima categoría como n_k, la estimación del vector $\boldsymbol{\mu}_k$ y de la matriz $\boldsymbol{\Sigma}$ están dadas, respectivamente, por:

$$\widehat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^{n_k} \boldsymbol{x}_{i,k}}{n_k}$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} (\boldsymbol{x}_{i,k} - \widehat{\boldsymbol{\mu}_k})(\boldsymbol{x}_{i,k} - \widehat{\boldsymbol{\mu}_k})^\top}{n - K}$$

Análisis de discriminante cuadrático (QDA)

En análisis de discriminante cuadrático se asume que $\boldsymbol{X} \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, para todo $k = 1, 2, \ldots, K$, es decir,

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right)$$

Análisis de discriminante cuadrático (QDA)

Por el teorema de Bayes se tiene que:

$$P(Y = k|\boldsymbol{X} = \boldsymbol{x}) \quad = \quad \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{j=1}^{K} \pi_j f_j(\boldsymbol{x})}$$

Análisis de discriminante cuadrático (QDA)

Por el teorema de Bayes se tiene que:

$$
\begin{aligned}
P(Y = k|\boldsymbol{X} = \boldsymbol{x}) &= \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{j=1}^{K} \pi_j f_j(\boldsymbol{x})} \\
&= \frac{\pi_k \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)}{\sum_{j=1}^{K} \pi_j \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_j|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right)}
\end{aligned}
$$

Análisis de discriminante cuadrático (QDA)

Por el teorema de Bayes se tiene que:

$$\begin{aligned} P(Y=k|\boldsymbol{X}=\boldsymbol{x}) &= \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{j=1}^{K}\pi_j f_j(\boldsymbol{x})} \\ &= \frac{\pi_k \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)}{\sum_{j=1}^{K}\pi_j \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_j|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right)} \end{aligned}$$

Así, la categoría donde esta expresión es mayor para el perfil $\boldsymbol{X}=\boldsymbol{x}$, se alcanza cuando se obtiene el mayor valor en el numerador, es decir, cuando

$$\pi_k \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)$$

tiene el mayor valor.

Análisis de discriminante cuadrático (QDA)

El logarítmo natural de esta expresión quedaría:

$$\log(\pi_k) - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_k|) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)$$

descartando la constante y reescribiendo la expresión, nuestro objetivo es clasificar el perfil $\boldsymbol{X} = \boldsymbol{x}$, en la clase k donde

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

tiene el mayor valor.

Análisis de discriminante cuadrático (QDA)

¿Cómo se estiman $\boldsymbol{\Sigma}_k$, $\boldsymbol{\mu}_k$ y π_k para $k = 1, 2, \ldots, K$?

Análisis de discriminante cuadrático (QDA)

¿Cómo se estiman $\boldsymbol{\Sigma}_k$, $\boldsymbol{\mu}_k$ y π_k para $k = 1, 2, \ldots, K$?

Igual que para LDA:

$$\widehat{\pi}_k = \frac{\#\text{ de individuos que pertenecen a la clase } k}{n_e}$$

Análisis de discriminante cuadrático (QDA)

¿Cómo se estiman $\boldsymbol{\Sigma}_k$, $\boldsymbol{\mu}_k$ y π_k para $k = 1, 2, \ldots, K$?

Igual que para LDA:

$$\widehat{\pi}_k = \frac{\#\text{ de individuos que pertenecen a la clase } k}{n_e}$$

Así mismo, si denotamos el número de indiviuos en la k-ésima categoría como n_k, la estimación del vector $\boldsymbol{\mu}_k$ y de la matriz $\boldsymbol{\Sigma}_k$ están dadas, respectivamente, por:

$$\widehat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^{n_k} \boldsymbol{x}_{i,k}}{n_k}$$

$$\widehat{\boldsymbol{\Sigma}}_k = \frac{\sum_{i=1}^{n_k} (\boldsymbol{x}_{i,k} - \widehat{\boldsymbol{\mu}_k})(\boldsymbol{x}_{i,k} - \widehat{\boldsymbol{\mu}_k})^\top}{n_k - 1}$$

Comparando los métodos de clasificación

Los métodos de clasificación expuestos fueron:

- KNN.
- Regresión logística.
- LDA.
- QDA.

Escenarios simulados para la comparación:

Se generaron 6 escenarios de simulación de tal manera que:

1. La variable respuesta Y consiste de dos categorías.

Escenarios simulados para la comparación:

Se generaron 6 escenarios de simulación de tal manera que:

1. La variable respuesta Y consiste de dos categorías.
2. Los tres primeros (1, 2, 3) fueron creados considerando una frontera de Bayes lineal.

Escenarios simulados para la comparación:

Se generaron 6 escenarios de simulación de tal manera que:

1. La variable respuesta Y consiste de dos categorías.
2. Los tres primeros (1, 2, 3) fueron creados considerando una frontera de Bayes lineal.
3. Los tres siguientes (4, 5, 6) considerando una frontera no lineal.

Escenarios simulados para la comparación:

Se generaron 6 escenarios de simulación de tal manera que:

1. La variable respuesta Y consiste de dos categorías.
2. Los tres primeros (1, 2, 3) fueron creados considerando una frontera de Bayes lineal.
3. Los tres siguientes (4, 5, 6) considerando una frontera no lineal.
4. En cada escenario se consideraron 2 covariables X_1 y X_2.

Escenarios simulados para la comparación:

Se generaron 6 escenarios de simulación de tal manera que:

1. La variable respuesta Y consiste de dos categorías.
2. Los tres primeros (1, 2, 3) fueron creados considerando una frontera de Bayes lineal.
3. Los tres siguientes (4, 5, 6) considerando una frontera no lineal.
4. En cada escenario se consideraron 2 covariables X_1 y X_2.
5. Se produjeron 100 repeticiones de cada escenario y se ajustaron los 5 métodos de clasificación en cada una: KNN-1 (KNN con un vecino), KNN-CV (KNN calculando el número de vecinos con validación cruzada, método que veremos más adelante), LDA, QDA y logístico.

Escenarios simulados para la comparación:

Se generaron 6 escenarios de simulación de tal manera que:

1. La variable respuesta Y consiste de dos categorías.
2. Los tres primeros (1, 2, 3) fueron creados considerando una frontera de Bayes lineal.
3. Los tres siguientes (4, 5, 6) considerando una frontera no lineal.
4. En cada escenario se consideraron 2 covariables X_1 y X_2.
5. Se produjeron 100 repeticiones de cada escenario y se ajustaron los 5 métodos de clasificación en cada una: KNN-1 (KNN con un vecino), KNN-CV (KNN calculando el número de vecinos con validación cruzada, método que veremos más adelante), LDA, QDA y logístico.
6. Para cada uno de los cinco métodos de clasificación se realizaron los boxplot de las tasas de error considerando, en cada caso, datos de test simulados.

Escenarios simulados para la comparación:

- **Escenario 1:** Dos clases cada una con 20 observaciones (de entrenamiento). Se asume que

$$X_{1,i} \sim N(\mu_1, \sigma^2) \quad \text{y} \quad X_{2,j} \sim N(\mu_2, \sigma^2), \quad i, j = 1, \ldots, 20$$

donde $Cor(X_{1,r}, X_{1,s}) = 0$, para $r \neq s$ y $Cor(X_{2,t}, X_{2,u}) = 0$, para $t \neq u$.

Escenarios simulados para la comparación:

- **Escenario 1:** Dos clases cada una con 20 observaciones (de entrenamiento). Se asume que

$$X_{1,i} \sim N(\mu_1, \sigma^2) \quad \text{y} \quad X_{2,j} \sim N(\mu_2, \sigma^2), \quad i, j = 1, \ldots, 20$$

  donde $Cor(X_{1,r}, X_{1,s}) = 0$, para $r \neq s$ y $Cor(X_{2,t}, X_{2,u}) = 0$, para $t \neq u$.

- **Escenario 2:** Dos clases cada una con 20 observaciones (de entrenamiento). Se asume que

$$X_{1,i} \sim N(\mu_1, \sigma^2) \quad \text{y} \quad X_{2,j} \sim N(\mu_2, \sigma^2), \quad i, j = 1, \ldots, 20$$

  donde $Cor(X_{1,r}, X_{1,s}) = -0.5$, para $r \neq s$ y $Cor(X_{2,t}, X_{2,u}) = -0.5$, para $t \neq u$.

Escenarios simulados para la comparación:

- **Escenario 3:** Dos clases cada una con 50 observaciones (de entrenamiento). Se asume que

$$X_{1,i} \sim t(n_1) \quad \text{y} \quad X_{2,j} \sim t(n_2), \quad i, j = 1, \ldots, 50$$

Escenarios simulados para la comparación:

- **Escenario 3:** Dos clases cada una con 50 observaciones (de entrenamiento). Se asume que

$$X_{1,i} \sim t(n_1) \quad \text{y} \quad X_{2,j} \sim t(n_2), \quad i,j = 1,\ldots,50$$

- **Escenario 4:** Dos clases cada una con 20 observaciones (de entrenamiento). Se asume que

$$X_{1,i} \sim N(\mu_1, \sigma_1^2) \quad \text{y} \quad X_{2,j} \sim N(\mu_2, \sigma_2^2), \quad i,j = 1,\ldots,20$$

donde $Cor(X_{1,r}, X_{1,s}) = 0.5$, para $r \neq s$ y
$Cor(X_{2,t}, X_{2,u}) = -0.5$, para $t \neq u$.

Escenarios simulados para la comparación:

- **Escenario 5:** Dos clases cada una con 20 observaciones (de entrenamiento). Se asume que

$$X_{1,i} \sim N(\mu_1, \sigma_1^2) \quad \text{y} \quad X_{2,j} \sim N(\mu_2, \sigma_2^2), \quad i, j = 1, \ldots, 20$$

donde $Cor(X_{1,r}, X_{1,s}) = 0$, para $r \neq s$ y $Cor(X_{2,t}, X_{2,u}) = 0$, para $t \neq u$. Las respuestas Y fueron obtenidas por medio de una función logística con X_1^2, X_2^2 y $X_1 \times X_2$ como covariables.

Escenarios simulados para la comparación:

- **Escenario 5:** Dos clases cada una con 20 observaciones (de entrenamiento). Se asume que

$$X_{1,i} \sim N(\mu_1, \sigma_1^2) \quad \text{y} \quad X_{2,j} \sim N(\mu_2, \sigma_2^2), \quad i, j = 1, \ldots, 20$$

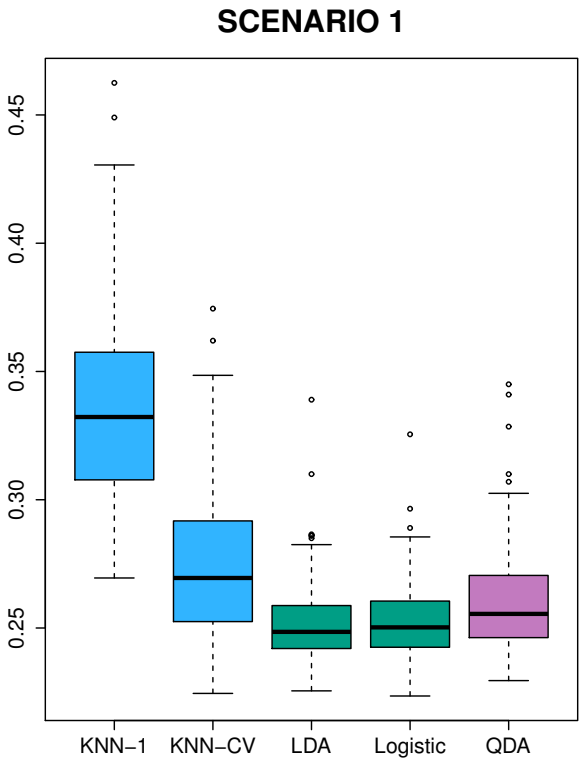
donde $Cor(X_{1,r}, X_{1,s}) = 0$, para $r \neq s$ y $Cor(X_{2,t}, X_{2,u}) = 0$, para $t \neq u$. Las respuestas Y fueron obtenidas por medio de una función logística con X_1^2, X_2^2 y $X_1 \times X_2$ como covariables.

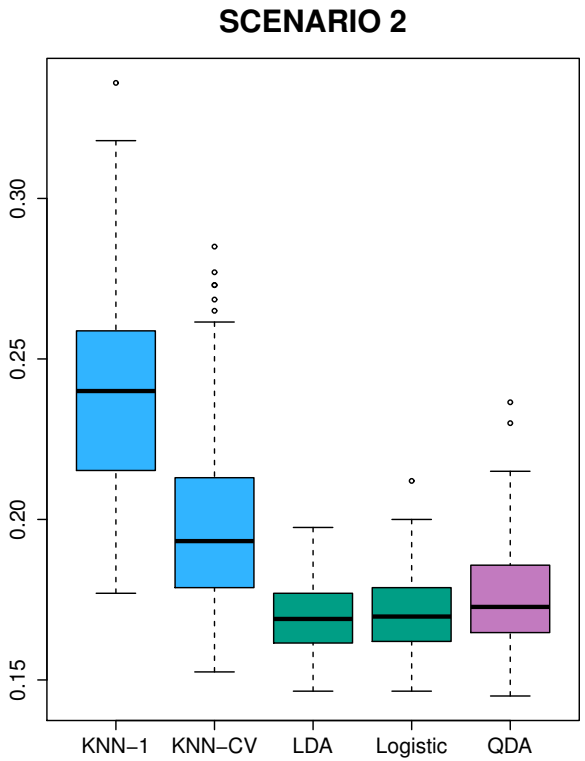
- **Escenario 6:** Dos clases cada una con 20 observaciones (de entrenamiento). Se asume que

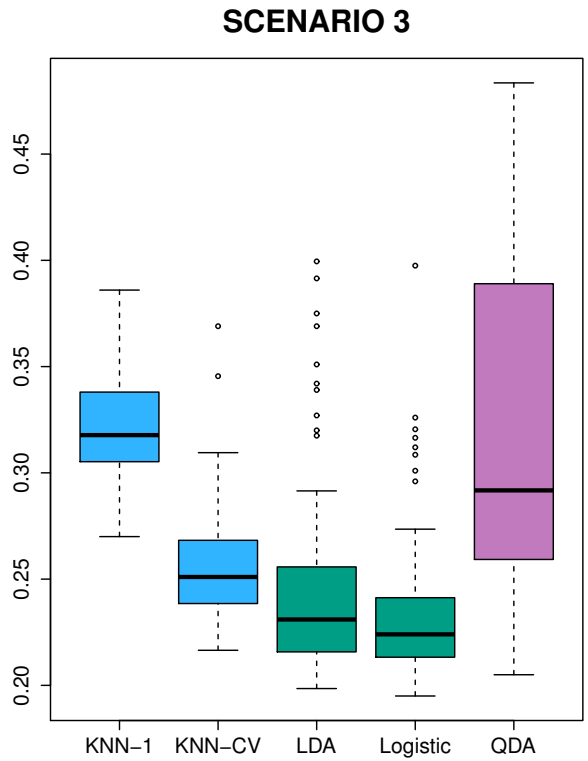
$$X_{1,i} \sim N(\mu_1, \sigma_1^2) \quad \text{y} \quad X_{2,j} \sim N(\mu_2, \sigma_2^2), \quad i, j = 1, \ldots, 20$$

donde $Cor(X_{1,r}, X_{1,s}) = 0$, para $r \neq s$ y $Cor(X_{2,t}, X_{2,u}) = 0$, para $t \neq u$. Las respuestas Y fueron obtenidas con una función no lineal más complicada que una cuadrática.

Resultados de los escenarios de simulación 1, 2 y 3:







Interpretación de los escenarios 1, 2 y 3:

- **Escenario 1:** LDA fue el método donde se obtuvieron mejores resultados, ya que este escenario se construyó con base en dicho método. El método logístico también presentó buenos resultados, debido a que éste asume también una estructura lineal. El método KNN no funcionó bien ya que éste no tiene en cuenta la estructura de variación. El QDA sufrió de un "sobreajuste" que se reflejó en un mal rendimiento.

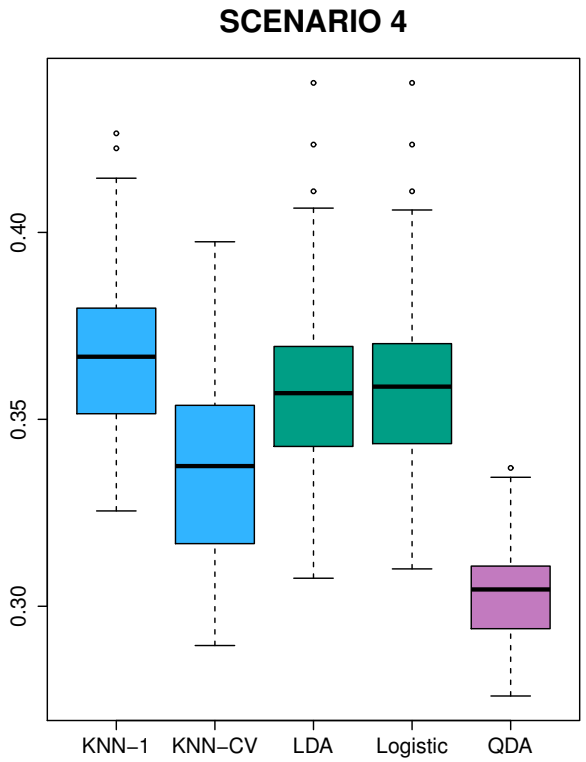
Interpretación de los escenarios 1, 2 y 3:

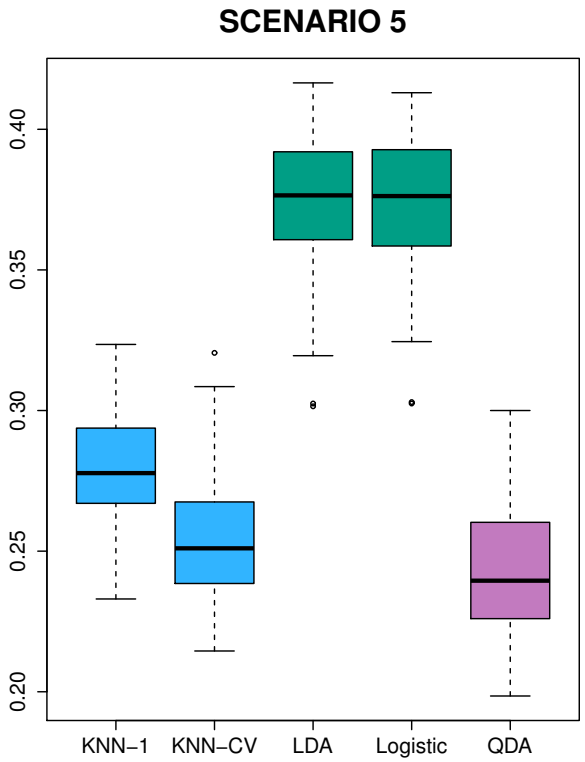
- **Escenario 1:** LDA fue el método donde se obtuvieron mejores resultados, ya que este escenario se construyó con base en dicho método. El método logístico también presentó buenos resultados, debido a que éste asume también una estructura lineal. El método KNN no funcionó bien ya que éste no tiene en cuenta la estructura de variación. El QDA sufrió de un "sobreajuste" que se reflejó en un mal rendimiento.
- **Escenario 2:** Era de esperarse que los resultados fueran muy parecidos a los del escenario 1, ya que el único cambio fue tener en cuenta una correlación intra-clase -**0.5** entre X_1 y X_2, la cual se tiene en cuenta cuando se ajusta LDA.

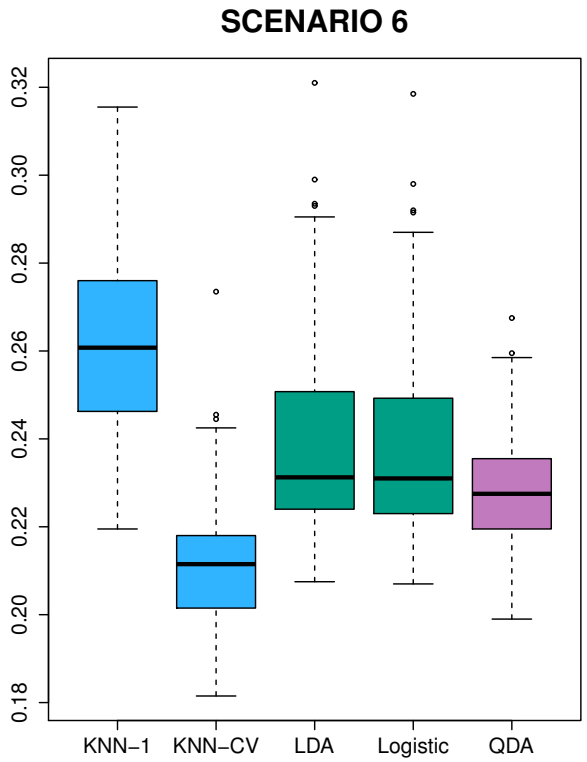
Interpretación de los escenarios 1, 2 y 3:

- **Escenario 1:** LDA fue el método donde se obtuvieron mejores resultados, ya que este escenario se construyó con base en dicho método. El método logístico también presentó buenos resultados, debido a que éste asume también una estructura lineal. El método KNN no funcionó bien ya que éste no tiene en cuenta la estructura de variación. El QDA sufrió de un "sobreajuste" que se reflejó en un mal rendimiento.
- **Escenario 2:** Era de esperarse que los resultados fueran muy parecidos a los del escenario 1, ya que el único cambio fue tener en cuenta una correlación intra-clase -**0.5** entre X_1 y X_2, la cual se tiene en cuenta cuando se ajusta LDA.
- **Escenario 3:** La violación del supuesto de normalidad afectó a LDA y QDA, lo cual era de esperarse, ya que ambos métodos se basan en dicho supuesto. El logístico no se vio muy afectado, ya que no se basa en supuestos de normalidad. Finalmente, KNN tuvo un mejor comportamiento que en los escenarios 1 y 2.

Resultados de los escenarios de simulación 4, 5 y 6:







Interpretación de los escenarios 4, 5 y 6:

- **Escenario 4:** QDA fue el método donde se obtuvieron mejores resultados, ya que este escenario se construyó con base en dicho método. El método KNN no funcionó bien ya que éste no tiene en cuenta la estructura de variación. El LDA y logístico se vieron seriamente afectados por la no linealidad.

Interpretación de los escenarios 4, 5 y 6:

- **Escenario 4:** QDA fue el método donde se obtuvieron mejores resultados, ya que este escenario se construyó con base en dicho método. El método KNN no funcionó bien ya que éste no tiene en cuenta la estructura de variación. El LDA y logístico se vieron seriamente afectados por la no linealidad.

- **Escenario 5:** Era de esperarse que los resultados fueran muy parecidos a los del escenario 5, ya que este escenario consideró una estructura cuadrática en la frontera de desición.

Interpretación de los escenarios 4, 5 y 6:

- **Escenario 4:** QDA fue el método donde se obtuvieron mejores resultados, ya que este escenario se construyó con base en dicho método. El método KNN no funcionó bien ya que éste no tiene en cuenta la estructura de variación. El LDA y logístico se vieron seriamente afectados por la no linealidad.

- **Escenario 5:** Era de esperarse que los resultados fueran muy parecidos a los del escenario 5, ya que este escenario consideró una estructura cuadrática en la frontera de desición.

- **Escenario 6:** Este escenario se construyó con una estructura más compleja que la lineal o que la cuadrática. Esto conlleva a que un método no-paramétrico como KNN-CV tenga mejor rendimiento que los demás métodos, sin embargo, cuando no se consideró un nivel de suavizamiento apropiado como en KNN-1, los resultados no son buenos.

Trabajando en R-Studio:

Para ajustar los métodos de clasificación LDA y QDA se pueden utilizar las funciones **lda** y **qda** de la librería **MASS**. Para obtener información acerca de dichas funciones podemos escribir en R el siguiente código:

```
require(MASS)
?lda
?qda
```

Para ver cómo funcionan dichas funciones consideremos la base de datos **Smarket** de la librería **ISLR**:

Trabajando en R-Studio:

```
require(ISLR)
names(Smarket)
```

```
## [1] "Year"      "Lag1"      "Lag2"      "Lag3"      "Lag4"      "Lag5"
## [7] "Volume"    "Today"     "Direction"
```

```
dim(Smarket)
```

```
## [1] 1250    9
```

```
head(Smarket)
```

```
##   Year   Lag1   Lag2   Lag3   Lag4   Lag5 Volume  Today Direction
## 1 2001  0.381 -0.192 -2.624 -1.055  5.010 1.1913  0.959        Up
## 2 2001  0.959  0.381 -0.192 -2.624 -1.055 1.2965  1.032        Up
## 3 2001  1.032  0.959  0.381 -0.192 -2.624 1.4112 -0.623      Down
## 4 2001 -0.623  1.032  0.959  0.381 -0.192 1.2760  0.614        Up
## 5 2001  0.614 -0.623  1.032  0.959  0.381 1.2057  0.213        Up
## 6 2001  0.213  0.614 -0.623  1.032  0.959 1.3491  1.392        Up
```

Trabajando en R-Studio:

Para entender esta base de datos consideremos a P_t como el precio de una acción el i-ésimo día.

Para entender esta base de datos consideremos a P_t como el precio de una acción el i-ésimo día.

Los retornos en el día t se pueden calcular como:

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right) \times 100$$

y se pueden interpretar como el porcentaje de ganancia (si es positivo) o de pérdida (si es negativo) en un día t con respecto al día anterior $(t-1)$.

Para entender esta base de datos consideremos a P_t como el precio de una acción el i-ésimo día.

Los retornos en el día t se pueden calcular como:

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right) \times 100$$

y se pueden interpretar como el porcentaje de ganancia (si es positivo) o de pérdida (si es negativo) en un día t con respecto al día anterior $(t-1)$.

El volúmen de un activo en el día t, denotado por V_t se define como el número total de transacciones (compras y ventas de una acción) el día t. Entre mayor sea, se dice que mayor es la liquidez de un activo, es decir, mayor facilidad de venderlo.

Trabajando en R-Studio:

¿Será que podemos predecir si sube o baja el precio del índice bursátil S&P500 de una acción en con los porcentajes de retorno de los últimos 5 días y del volumen de la acción hoy?

¿Será que podemos predecir si sube o baja el precio del índice bursátil S&P500 de una acción en con los porcentajes de retorno de los últimos 5 días y del volumen de la acción hoy?

Las variables son:

- `Year`: El año al que pertenece cada registro. Los datos van del 2001 al 2005.
- `Lag1`: Retorno del día anterior.
- `Lag2`: Retorno dos días antes.
- `Lag3`: Retorno tres días antes.
- `Lag4`: Retorno cuatro días antes.
- `Lag5`: Retorno cinco días antes.
- `Volume`: Volúmen de la acción.
- `Today`: Retorno del día de hoy.
- `Direction`: Hoy subió o bajo la acción con respecto al día de ayer.

Trabajando en R-Studio:

Utilicemos como datos de entrenamiento los años 2001 a 2004 y como datos de test los del 2005.

```
train <-(Smarket$Year<2005)
length(train)
```

```
## [1] 1250
```

```
head(train)
```

```
## [1] TRUE TRUE TRUE TRUE TRUE TRUE
```

```
tail(train)
```

```
## [1] FALSE FALSE FALSE FALSE FALSE FALSE
```

Trabajando en R-Studio - Ajuste LDA:

```
library (MASS)
Ajuste.lda<-lda(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+
                    Volume,data=Smarket,subset =train)
Ajuste.lda
```

```
## Call:
## lda(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume, data = Smarket,
##     subset = train)
##
## Prior probabilities of groups:
##     Down       Up
## 0.491984 0.508016
##
## Group means:
##             Lag1        Lag2         Lag3         Lag4          Lag5   Volume
## Down  0.04279022  0.03389409 -0.009806517 -0.010598778  0.0043665988 1.371843
## Up   -0.03954635 -0.03132544  0.005834320  0.003110454 -0.0006508876 1.363210
##
## Coefficients of linear discriminants:
##                LD1
## Lag1   -0.58081056
## Lag2   -0.49111007
## Lag3    0.07707664
## Lag4    0.06904095
## Lag5   -0.04549853
## Volume -1.24678716
```

Trabajando en R-Studio - Ajuste QDA:

```
library (MASS)
Ajuste.qda<-qda(Direction~Lag1+Lag2+Lag3+Lag4+Lag5
                +Volume,data=Smarket,subset =train)
Ajuste.qda
```

```
## Call:
## qda(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume, data = Smarket,
##     subset = train)
##
## Prior probabilities of groups:
##     Down       Up
## 0.491984 0.508016
##
## Group means:
##             Lag1        Lag2         Lag3         Lag4          Lag5   Volume
## Down  0.04279022  0.03389409 -0.009806517 -0.010598778  0.0043665988 1.371843
## Up   -0.03954635 -0.03132544  0.005834320  0.003110454 -0.0006508876 1.363210
```

Seleccionamos los datos de test:

```
Smarket.2005<-Smarket[!train,] # Datos de test del 2005
dim(Smarket.2005)
```

```
## [1] 252   9
```

```
Direction.2005<-Smarket$Direction[!train]
```

Trabajando en R-Studio - Matrices de confusión:

```
Pred.lda<-predict(Ajuste.lda , Smarket.2005)
Clase.lda =Pred.lda$class
table(Clase.lda,Direction.2005)
```

```
##           Direction.2005
## Clase.lda Down Up
##      Down   77 97
##      Up     34 44
```

```
mean(Clase.lda != Direction.2005)
```

```
## [1] 0.5198413
```

```
Pred.qda<-predict(Ajuste.qda , Smarket.2005)
Clase.qda =Pred.qda$class
table(Clase.qda, Direction.2005)
```

```
##           Direction.2005
## Clase.qda Down  Up
##      Down   82 111
##      Up     29  30
```

```
mean(Clase.qda != Direction.2005)
```