UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Departamento de Matemática

Tesis de Licenciatura

Análisis Comparativo de Métodos de Clasificación

Noelia Soledad Fernández

Directora: Dra. Ana María Bianco

Febrero de 2025

Para Georgina ,

mi ángel en la tierra...

y para mis ángeles que están en el cielo.

Agradecimientos

Antes que nada, quiero agradecer a mi Directora Ana Bianco, quien además de transmitirme este amor por la estadística, me acompañó en cada etapa del presente trabajo, con consejos, aportes, ideas e infinita, pero infinita paciencia de verdad.

A Die, mi esposo, amigo y compañero de vida. Que siempre cree en mí, aún cuando ni yo misma lo hago, que me banca en todas, y me incita a ser cada día mejor.

A mis papás, Graciela y Jorge, sin ellos ningún logro hubiera sido posible. Les debo todo lo que soy.

A mi pequeño *Cauchy*, por estar horas y horas a mi lado estudiando, haciendo prácticas, preparando finales... Mi compañero incondicional siempre.

A mi familia: mi hermana, ahijado, sobrinos, primos, abuela, suegros, madrina, ... que siempre estuvieron ahí, bancando sin enojarse cuando oían *"voy solo un ratito porque tengo que estudiar"*.

A todos los docentes que me han inspirado con sus conocimientos a lo largo de la carrera: Ana Bianco, Susana Puddu, Juan Sabia, Javier Etcheverry, Ezequiel Rela, Mariela Sued, Gonzalo Chebi, Carolina Mosquera, ...(sería imposible nombrarlos a todos!) de verdad, GRACIAS.

A todos lo que me he cruzado en esta hermosa carrera, que hemos generado lazos más allá de los números: Malena, Antonella, David , Lorena, quienes transitaron junto a mí este camino a la par, y lo hicieron más ameno y agradable, entre mates, charlas, catarsis, risas y estudio.

A los amigos de la vida, Dani, Luz, Romi, Pato, Fede ...

A todos ... GRACIAS, GRACIAS, GRACIAS !!

Resumen

El problema de clasificación o discriminación consiste en predecir una variable categórica a partir de información adicional disponible. Este tipo de problemas es común en diversos ámbitos, tanto en la vida cotidiana como en la industria o las ciencias en general. En esta tesis, abordamos este problema utilizando técnicas de aprendizaje supervisado, que se basan en datos que fueron previamente clasificados con el objetivo de aprender a clasificar nuevas observaciones. Los métodos que analizamos en este estudio se fundamentan en la Regla de Clasificación de Bayes. Nuestro objetivo principal es comparar mediante un estudio de simulación diversos métodos de clasificación, tales como regresión logística, LDA (Análisis Discriminante Lineal), QDA (Análisis Discriminante Cuadrático) y SVM (Máquinas de Vectores de Soporte), en diferentes escenarios.

Índice general

Capítulo 1

Introducción

1.1. Introducción

El *aprendizaje estadístico* puede definirse como una serie de herramientas para modelar y comprender conjuntos complejos de datos (James et al., 2013). Esta disciplina surge de la mixtura entre la estadística y el aprendizaje automático, que desempeña actualmente un papel vital en la organización, segmentación y toma de decisiones en una amplia variedad de contextos. El aprendizaje estadístico es una herramienta de gran importancia en diversas disciplinas, desde la ciencia de datos hasta la ingeniería o la investigación social.

Múltiples problemas de aprendizaje estadístico tienen como objetivo principal predecir valores desconocidos de una determinada variable dependiente, en función de los valores conocidos de una o más variables explicativas. Esto se conoce como *aprendizaje supervisado*. Cuando la variable explicativa es de tipo numérico, las tareas de predicción suelen denominarse *regresión*. Mientras tanto, si la variable explicativa es categórica, el proceso se conoce como *clasificación* (Hastie et al., 2017).

La *clasificación estadística* intenta asignar elementos a clases predefinidas en función de sus características relevantes. La capacidad de predecir la pertenencia o no de una observación a una cierta categoría de forma precisa se traduce en una mejor comprensión de los

datos y, por lo tanto, en una toma de decisiones más informada y eficiente. Esta área está en constante desarrollo, debido a la necesidad de la resolución de diversos problemas de *Big Data*. Dichos problemas, que un tiempo atrás se consideraban irresolubles, hoy hallan solución gracias al aprendizaje estadístico y la evolución de la computación (Jorgensen, 2019).

Los pasos fundamentales en el proceso de clasificación estadística pueden resumirse en las siguientes etapas:

1. Selección de variables: Se toman las variables más importantes que se utilizarán posteriormente como criterios para asignar cada observación del conjunto de datos a una clase determinada. Dichas variables pueden ser tanto cualitativas como cuantitativas.

2. Definición de clases: Se establecen las categorías o clases dentro de las cuales se agruparán los datos. Las mismas deben ser mutuamente excluyentes y en general los datos deberán pertenecer solo a una clase.

3. Asignación de datos: Cada observación es asignada a una clase específica según los valores observados de las variables seleccionadas. Esto puede implicar la comparación directa de los valores numéricos o la consideración de atributos cualitativos.

4. Análisis y resumen: Una vez clasificados los datos, se realiza un análisis de la distribución de los mismos en cada uno de los grupos. Se pueden calcular medidas de resumen tales como promedios, medianas o desvíos estándar para cada clase, lo cual otorga información sobre las propiedades del comportamiento de las observaciones en cada categoría.

5. Interpretación: Los resultados obtenidos luego de la clasificación pueden proporcionar información sobre patrones, tendencias o relaciones entre las variables y las clases. Esto puede ayudar en la toma de decisiones, la formulación de hipótesis y la generación de conclusiones basadas en los datos analizados.

En la presente tesis se pretende comparar cuatro métodos de clasificación muy usados: Regresión Logística, Análisis Discriminante Lineal (LDA), Análisis Discriminante Cuadrático (QDA) y Support Vector Machine (SVM). Las técnicas son contrastadas en primer lugar

desde un punto de vista teórico, analizando las propiedades generales de cada método y sus ventajas. Posteriormente, se utiliza RStudio para realizar simulaciones en laboratorio con datos controlados, con el propósito de observar el comportamiento de las técnicas en diferentes escenarios.

1.2. Clasificación Estadística

El proceso por el cual se busca predecir el valor de una variable respuesta cualitativa Y se conoce como *clasificación*, pues involucra asignar cada observación del conjunto a una determinada clase o categoría (James et al., 2013). Dicho proceso también puede denominarse *aprendizaje supervisado*, *discriminación* o *reconocimiento de patrones* (Wasserman, 2004). El concepto central de la clasificación se basa en el aprendizaje a partir de datos disponibles y la información que se puede extraer a partir de ellos (Jorgensen, 2019).

Matemáticamente, el proceso de clasificación se define de la siguiente manera: sean $(\boldsymbol{X}_1, Y_1)...(\boldsymbol{X}_n, Y_n)$ vectores independientes e idénticamente distribuidos (i.i.d.) donde

$$\boldsymbol{X}_i = (X_{i1}, ..., X_{id})^T \in \mathcal{X} \subset \mathbb{R}^d$$

es un vector d-dimensional e Y_i toma valores en algún conjunto finito $\mathcal{Y}$. Una *regla de clasificación* es una función $g : \mathcal{X} \to \mathcal{Y}$ que asigna a cada valor del espacio $\mathcal{X}$ un valor de $\mathcal{Y}$, o sea una clase posible. Es decir, que al observar un nuevo $\boldsymbol{X}$, se predice que Y será $h(\boldsymbol{X})$ (Wasserman, 2004).

1.3. Clasificación y aprendizaje automático

El *aprendizaje automático supervisado* (también conocido como *supervised machine learning*) se define como una serie de métodos que involucran la construcción de un modelo que permita estimar los valores de una variable dependiente Y (también llamada *output* o variable respuesta) basándose en valores conocidos de una o más variables independientes

(también denominadas *inputs* o variables predictoras) (James et al., 2013). En la práctica para cada observación de la/s variable/s predictora/s $\boldsymbol{x_i}, i = 1, ..., n$ se cuenta con una medida asociada $\boldsymbol{y_i}$ de la variable respuesta, y cuyo interés principal será a partir de una muestra obtener una regla de clasificación que permita predecir la clase a la que pertenece un nuevo individuo, para el que la covariable $\boldsymbol{X}$ toma el valor $\boldsymbol{x}$.

Este enfoque contrasta con el *aprendizaje automático no supervisado* (también conocido como *unsupervised machine learning*). Los métodos de aprendizaje no supervisado son aplicados cuando no se posee una variable que cumpla el rol de predictora, por lo que su objetivo ya no es minimizar el error de predicción, pues no hay valores observados con los cuales comparar los valores predichos. En su lugar, el foco está puesto en observar patrones o asociaciones entre los distintos atributos presentes en el conjunto de datos (Hastie et al., 2017).

Aunque el término *"aprendizaje automático"* sea reciente, muchos de sus conceptos básicos fueron desarrollados hace mucho tiempo. Algunas de las técnicas de aprendizaje supervisado más antiguas que aún se siguen utilizando son el método de los mínimos cuadrados (hoy más conocido como Regresión Lineal), el Análisis Discriminante Lineal y la Regresión Logística. (James et al., 2013).

Gracias al apogeo del aprendizaje automático, se ha revolucionado la forma en que se lleva a cabo la clasificación estadística. Los algoritmos de aprendizaje automático tienen la capacidad de identificar patrones a partir de los datos disponibles, lo que permite modelar relaciones y capturar características de alta dimensionalidad. La disponibilidad de grandes cantidades de datos, junto al aumento en la capacidad de cómputo, han impulsado aún más el desarrollo de algoritmos sofisticados que pueden abordar problemas cada vez más desafiantes y permiten desarrollar modelos de clasificación más precisos y robustos que los métodos estadísticos tradicionales. Algunas de las técnicas que surgieron en décadas más recientes incluyen los métodos basados en árboles (tales como bosques aleatorios o *boosting*), las máquinas de soporte vectorial y los modelos aditivos generalizados.

En las siguientes cuatro secciones (de 2 a 5), se realizará un análisis teórico de cuatro

algoritmos de clasificación supervisada. En la Sección 6, se presentarán métricas para evaluar la calidad de los modelos entrenados en distintos escenarios simulados y los resultados de dichos experimentos numéricos.

A continuación daremos los fundamentos teóricos del problema.

1.3.1. Teorema de Bayes

Dentro del contexto que venimos desarrollando, el *Teorema de Clasificación de Bayes* establece que la regla de clasificación que minimiza el error de clasificación, es aquella que asigna una observación a la clase con mayor probabilidad a posteriori.

Supongamos la situación genérica en la que en un individuo a partir del valor de la variable $X \in \mathcal{X}$ queremos predecir el valor de la clase $Y \in \mathcal{Y}$ a la cual pertenece el individuo y para ello utilizamos una regla $g : \mathcal{X} \to \mathcal{Y}$. Lo que nos interesa es hacer predicciones correctas, en este sentido es que nos ocuparemos del *error de clasificación medio* de una regla g dada, que definiremos como

$$\mathbb{P}(g(X) \neq Y)\,.$$

Una pregunta lógica que surge en este contexto, es pensar cuál es el clasificador óptimo g^{op} con el que debemos separar nuestro conjunto de datos para que minimice el error de clasificación medio. Por simplicidad nos enfocaremos en el caso binario, es decir cuando $\mathcal{Y} = \{0, 1\}$.

Intuitivamente, si $X = x$ clasificaríamos al individuo en la clase con mayor probabilidad condicional, es decir usaríamos la regla dada por

$$g^{op}(x) = \begin{cases} 1 & \text{si} \quad \mathbb{P}(Y = 1|X = x) \geq \mathbb{P}(Y = 0|X = x) \\ 0 & \text{caso contrario} \end{cases}$$

El siguiente resultado establece que g^{op} corresponde al clasificador óptimo de Bayes en el caso binario en el sentido de que minimiza el error de clasificación medio.

Teorema 1. *Dado un clasificador $g : \mathcal{X} \to \mathcal{Y}$ se cumple que*

$$\mathbb{P}(g^{op}(X) \neq Y) \leq \mathbb{P}(g(X) \neq Y)$$

Demostración. Dado un clasificador $g : \mathcal{X} \to \mathcal{Y}$ consideremos su pérdida $L(g) = \mathbb{P}(g(X) \neq Y)$ como su error de clasificación medio. Entonces, tenemos que

$$\begin{aligned} L(g) = \mathbb{P}(g(X) \neq Y) &= E_{XY}\Big[\mathcal{I}_{(g(X)\neq Y)}\Big] \\ &= E_X E_{Y|X}\Big[\mathcal{I}_{(g(X)\neq Y)}\Big] \\ &= E_X\Big[\mathcal{I}_{(g(X)=0)}\mathbb{P}(Y = 1|X) + \mathcal{I}_{(g(X)=1)}\mathbb{P}(Y = 0|X)\Big] \end{aligned}$$

Luego, para minimizar $L(g)$ es suficiente con minimizar para cada x

$$\mathcal{I}_{(g(X)=0)}\mathbb{P}(Y = 1|X = x) + \mathcal{I}_{(g(X)=1)}\mathbb{P}(Y = 0|X = x) \ .$$

Teniendo en cuenta que las dos indicadoras son mutuamente excluyentes, alcanza con tomar

$$\begin{aligned} g^{op}(x) &= \begin{cases} 1 & \text{si} \quad \mathbb{P}(Y = 1|X = x) \geq \mathbb{P}(Y = 0|X = x) \\ 0 & \text{caso contrario} \end{cases} \\ &= \begin{cases} 1 & \text{si} \quad \mathbb{P}(Y = 1|X = x) \geq \mathbb{P}(Y = 0|X = x) \\ 0 & \text{si} \quad \mathbb{P}(Y = 0|X = x) > \mathbb{P}(Y = 1|X = x) \ , \end{cases} \end{aligned}$$

con lo cual queda demostrado el resultado deseado. □

Notemos que

$$\mathbb{P}(Y = 1|X = x) \geq \mathbb{P}(Y = 0|X = x) \Leftrightarrow \mathbb{P}(Y = 1|X = x) \geq \frac{1}{2}$$

de manera que $g^{op}(x)$ puede reescribirse como

$$g^{op}(x) = \begin{cases} 1 & \text{si} \quad \mathbb{P}(Y = 1|X = x) \geq \frac{1}{2} \\ 0 & \text{caso contrario} \end{cases} \tag{1.1}$$

Observemos que, si X es discreta, en un x de masa positiva

$$\mathbb{P}(Y = y|X = x) = \frac{\mathbb{P}(X = x|Y = y)\mathbb{P}(Y = y)}{\mathbb{P}(X = x)}$$

por lo tanto,

$$g^{op}(x) = \begin{cases} 1 & \text{si} \quad \mathbb{P}(X = x|Y = 1)\mathbb{P}(Y = 1) \geq \mathbb{P}(X = x|Y = 0)\mathbb{P}(Y = 0) \\ 0 & \text{si} \quad \mathbb{P}(X = x|Y = 0)\mathbb{P}(Y = 0) > \mathbb{P}(X = x|Y = 1)\mathbb{P}(Y = 1) \end{cases}$$

En el caso en que X es continua, si f_0 denota la densidad de $X|Y = 0$ y f_1 la de $X|Y = 1$, la regla óptima resulta

$$g^{op}(x) = \begin{cases} 1 & \text{si} \quad f_1(x)\mathbb{P}(Y = 1) \geq f_0(x)\mathbb{P}(Y = 0) \\ 0 & \text{caso contrario} \end{cases}$$

Es directa la generalización del resultado del anterior teorema a la situación en que existen más de dos clases, es decir cuando $\mathcal{Y} = \{1, 2, \ldots, K\}$. En este caso la regla óptima es

$$g^{op}(x) = \underset{r}{\text{argmax}}\, \mathbb{P}(Y = r|X = x) = \underset{r}{\text{argmax}}\, \mathbb{P}(Y = r|X = x) = \underset{r}{\text{argmax}}\, \frac{f_r(x)\pi_r}{\sum_{j=1}^{K} f_j(x)\pi_j},$$

donde $\pi_r = \mathbb{P}(Y = r)$ y $f_r(x) = f(x|Y = r)$ denota la función de densidad o de probabilidad puntual condicional de X, según ésta sea continua o discreta.

Capítulo 2

Regresión Logística

La Regresión Logística es un método ampliamente utilizado en problemas de clasificación, particularmente cuando se trabaja con una variable de respuesta binaria. Es un método paramétrico popular que es relativamente fácil de implementar y comprender, además de proveer resultados generalmente muy precisos (Jorgensen, 2019).

Sea Y una variable binaria, es decir que $Y \in \{0, 1\}$. Esta variable es conocida como *variable indicadora* (también llamada variable *dummy*), ya que señala la ocurrencia de un determinado evento (James et al., 2013). Genéricamente, podemos definir a Y de la siguiente forma

$$Y = \begin{cases} 0 & \text{si el evento no ocurre (‘fracaso’)} \\ 1 & \text{si el evento ocurre (‘éxito’)} \quad . \end{cases}$$

que corresponde a una distribución Bernoulli (Wasserman, 2004):

$$Y|X = x \sim Be(\pi) \ ,$$

donde $\pi = P(Y = 1)$, en otras palabras, la probabilidad de que el suceso de interés ocurra. Por ejemplo, un banco podría estar interesado en saber si un determinado cliente realizará adecuadamente un pago, conociendo su fondo bancario disponible. En esta situación, Y indicará si la persona realizó el pago ($Y = 1$) o no ($Y = 0$), mientras que X es el saldo del cliente.

En lugar de modelar directamente a la variable Y, la regresión logística modela la probabilidad (π) de que Y pertenezca a una categoría particular (James et al., 2013). Dicha probabilidad se simboliza de la siguiente manera:

$$\pi(x) = P(Y = 1|X = x).$$

Para el ejemplo mencionado anteriormente con los clientes de un banco, la probabilidad que se desea modelar sería la siguiente:

$$\pi(saldo) = P(pago = si|saldo).$$

Para cualquier valor dado del saldo, se puede realizar una predicción de pago. Por ejemplo, se podría predecir $pago = si$ para cualquier individuo para el cual $\pi(saldo) > \alpha$, con $\alpha \in (0, 1)$, este valor dependerá de cuan conservadora sea la predicción que se desee.

Un primer enfoque para modelizar la relación entre $\pi(x)$ y $\boldsymbol{x}$ podría ser ajustar un modelo de regresión lineal simple, es decir:

$$\pi(x) = \beta_0 + \beta_1 x.$$

Cabe mencionar que, dado que el lado derecho de la expresión anterior depende de $\boldsymbol{\beta} = (\beta_0, \beta_1)$, tenemos que $\pi(x) = \pi(x, \boldsymbol{\beta})$. Sin embargo, el modelo planteado para la probabilidad condicional presenta inconvenientes (James et al., 2013). Para valores muy bajos o muy elevados de $\boldsymbol{x}$, las probabilidades predichas pueden resultar menores a 0 o mayores o iguales a 1 respectivamente. Dado que una probabilidad debe ser siempre superior a 0 e inferior a 1, independientemente de los valores de x, este enfoque no es sensato. En consecuencia, es necesario modelizar la probabilidad condicional utilizando una función que devuelva resultados entre 0 y 1 para cualquier valor de x. Existen múltiples funciones que cumplen esta descripción, entre ellas la función logística que es la opción más frecuente. Esta es la función utilizada en los modelos de regresión logística y la que le otorga su denominación. Por lo

tanto, la relación entre $\pi(x, \boldsymbol{\beta})$ y x se representa de la siguiente manera:

$$\pi(x, \boldsymbol{\beta}) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}.$$

La Figura 2.1, extraída de (James et al., 2013), muestra la diferencia entre ajustar un modelo de regresión lineal simple y un modelo de regresión logística. Como se puede observar, el primer enfoque resulta en algunas probabilidades inferiores a cero. Por el contrario, el modelo de regresión logística asigna una probabilidad de pago cercana (pero nunca inferior) a cero para los saldos con valores bajos, mientras que para saldos altos se asigna una probabilidad de pago cercana (pero nunca superior) a uno. La función logística siempre producirá una curva en forma de "S", en lugar de una línea recta como sucede con regresión lineal.

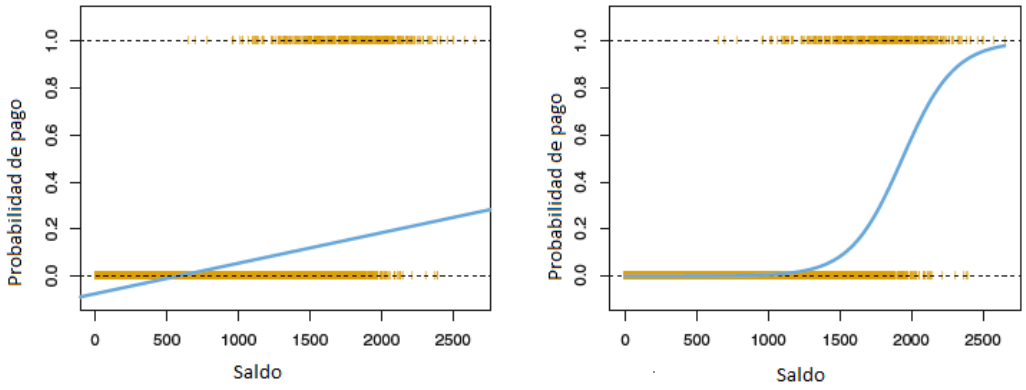

Figura 2.1: Ejemplo de clasificación con una variable respuesta binaria utilizando regresión lineal simple (izquierda) y regresión logística (derecha).

Utilizando la función logística, la probabilidad de obtener un fracaso (es decir, $Y = 0$) resulta:

$$P(Y = 0|X = x) = 1 - P(Y = 1|X = x) = 1 - \pi(x, \boldsymbol{\beta}) = 1 - \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{1}{1 + e^{\beta_0 + \beta_1 x}}.$$

El cociente entre una probabilidad (π) y su complemento ($1-\pi$) se denomina *odds*, ocasionalmente traducido como *cuota* o *chance*. Esta medida puede tomar cualquier valor en $[0, +\infty)$, con valores cercanos a 0 indicando probabilidades bajas y valores grandes indicando proba-

bilidades altas (James et al., 2013). En regresión logística, las odds resultan:

$$\frac{\pi(x,\boldsymbol{\beta})}{1-\pi(x,\boldsymbol{\beta})} = \frac{\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{\frac{1}{1+e^{\beta_0+\beta_1 x}}} = e^{\beta_0+\beta_1 x}.$$

Al calcular el logaritmo natural de este cociente, se obtiene el valor conocido como *logit* o *log-odds*:

$$logit[(\pi(x,\boldsymbol{\beta})] = ln\left(\frac{\pi(x,\boldsymbol{\beta})}{1-\pi(x)}\right).$$

El logit incrementa de $-\infty$ a $+\infty$ a medida que la probabilidad incrementa entre 0 y 1 (Efron y Hastie, 2016). En síntesis, el modelo de regresión logística se expresa de la siguiente manera:

$$logit[(\pi(x,\boldsymbol{\beta})] = ln\left(\frac{\pi(x_i)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x,$$

donde:

- β_0 es el *intercept* o constante. Es el valor que tomará el logit cuando $x = 0$.

- β_1 es el coeficiente asociado a x. Indica el cambio en el logit ante incrementos unitarios de X. Esto es equivalente a multiplicar las odds por e^{β_1}. Si $\beta_1 > 0$, un aumento en el valor de x corresponderá a un aumento en $\pi(x,\boldsymbol{\beta})$. Si por el contrario, $\beta_1 < 0$, entonces el aumento de x se asociará con una disminución de $\pi(x,\boldsymbol{\beta})$.

A diferencia de lo que sucede en regresión lineal, un incremento unitario en x no se corresponde con un incremento igual a β_1 en $\pi(x,\boldsymbol{\beta})$, pues la relación entre $\pi(x,\boldsymbol{\beta})$ y x no es lineal (James et al., 2013). Esta última diferencia es también evidente en la Figura 2.1.

La técnica de Regresión Logística puede generalizarse al caso en el cual, en lugar de una única variable explicativa X, se cuenta con un vector d-dimensional de variables predictoras $\boldsymbol{X} = (X_1, X_2, ...X_d)^T$. Dichas variables pueden ser tanto cualitativas como cuantitativas, pero Y sigue siendo una variable binaria (Jorgensen, 2019).

Dado un punto $\boldsymbol{x} = (x_1, x_2, ...x_d)^T$, la probabilidad condicional de la que hablamos es:

$$\pi(\boldsymbol{x}, \boldsymbol{\beta}) = P(Y = 1|X_1 = x_1, X_2 = x_2, ...X_d = x_d)\,,$$

donde ahora $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^T$. El modelo de regresión logística asume que:

$$\pi(\boldsymbol{x}, \boldsymbol{\beta}) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_d x_d}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_d x_d}} = \frac{e^{\beta_0+\sum_{j=1}^{d}\beta_j x_j}}{1 + e^{\beta_0+\sum_{j=1}^{d}\beta_j x_j}}\,. \tag{2.1}$$

Por lo tanto, el modelo de regresión logística múltiple resulta:

$$logit[\pi(\boldsymbol{x}, \boldsymbol{\beta})] = ln\left(\frac{\pi(\boldsymbol{x}, \boldsymbol{\beta})}{1 - \pi(\boldsymbol{x}, \boldsymbol{\beta})}\right) = \beta_0 + \sum_{j=1}^{d}\beta_j x_j,$$

donde:

- β_0 es el valor del logit cuando todas las variables toman el valor 0.
- β_j es el cambio (aumento o decrecimiento) del logit ante incrementos unitarios de x_j, manteniendo constantes los otros términos.

2.1. Estimación de parámetros

Luego de haber planteado el modelo, es necesario hallar estimaciones para sus parámetros según datos observados. Para ello, asumimos como en la sección anterior que disponemos de una muestra de vectores $(\boldsymbol{X}_1, Y_1)...(\boldsymbol{X}_n, Y_n)$ independientes, donde $(\boldsymbol{X}_i, Y_i) \sim (\boldsymbol{X}, Y)$ que cumple

$$Y|\boldsymbol{X} = \boldsymbol{x} \sim Be(\pi(\boldsymbol{x}, \boldsymbol{\beta}))\,,$$

donde $\pi(\boldsymbol{x}, \boldsymbol{\beta})$ satisface (2.1). El método de estimación clásico más utilizado en este contexto es el de *máxima verosmilitud*, que corresponde a elegir los valores de los parámetros que maximizan la probabilidad de observar los resultados obtenidos (James et al., 2013).

Formalmente, el método de máxima verosimilitud busca el vector $\widehat{\boldsymbol{\beta}}_{MV}$ que maximice la *función de verosimilitud*:

$$\mathcal{L}(\boldsymbol{b}) = \mathcal{L}(\boldsymbol{b}, y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \prod_{i=1}^{n} \pi(\boldsymbol{x}_i, \boldsymbol{b})^{y_i} [1 - \pi(\boldsymbol{x}_i, \boldsymbol{b})]^{1-y_i},$$

donde $(\boldsymbol{x}_i, y_i)$ son los pares de datos independientes observados en la muestra y $\boldsymbol{b} = (b_0, b_1, \ldots, b_d)^T \in R^{d+1}$. Matemáticamente, es más sencillo trabajar con el logaritmo de esta función, conocido como *log-verosimilitud*:

$$\ell(\boldsymbol{b})) = \ln[\mathcal{L}(\boldsymbol{b})] = \sum_{i=1}^{n} y_i \ln[\pi(\boldsymbol{x}_i, \boldsymbol{b})] + (1 - y_i) \ln[1 - \pi(\boldsymbol{x}_i, \boldsymbol{b})]$$

Para hallar $\widehat{\boldsymbol{\beta}}_{MV}$, buscamos los puntos críticos derivando $\ell(\boldsymbol{b})$ respecto a $\boldsymbol{b}$ e igualando las derivadas parciales a 0 como mostramos a continuación:

$$\frac{\partial \ell(\mathbf{b})}{\partial b_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{e^{b_0 + \sum_{j=1}^{k} b_j x_{ij}}}{1 + e^{b_0 + \sum_{j=1}^{k} b_j x_{ij}}} = 0$$

$$\frac{\partial \ell(\mathbf{b})}{\partial b_j} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} x_{ij} \frac{e^{b_0 + \sum_{j=1}^{k} b_j x_{ij}}}{1 + e^{b_0 + \sum_{j=1}^{k} b_j x_{ij}}} = 0 \qquad j \neq 0$$

o equivalentemente

$$\frac{\partial \ell(\mathbf{b})}{\partial b_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi(\boldsymbol{x}_i, \mathbf{b}) = 0$$

$$\frac{\partial \ell(\mathbf{b})}{\partial b_j} = \sum_{i=1}^{n} x_{ij} y_i - \sum_{i=1}^{n} x_{ij} \pi(\boldsymbol{x}_i, \mathbf{b}) = 0 \quad j \neq 0\,.$$

Dado que estas ecuaciones son no lineales, deben ser resueltas numéricamente mediante algoritmos iterativos, tales como el método de Newton-Raphson (Jorgensen, 2019). En la práctica, estos procedimientos son realizados casi siempre mediante paquetes de software, como por ejemplo *R*. Las soluciones obtenidas serán los estimadores $\widehat{\boldsymbol{\beta}}_{MV}$ para los parámetros del modelo.

Consideremos por simplicidad nuevamente el caso en que la covariable es univariada.Una vez estimados los parámetros del modelo, se pueden obtener fácilmente predicciones de la probabilidad de $Y = 1$ para cualquier valor dado de X. Por ejemplo, si se desea predecir la probabilidad de ocurrencia del evento dado que $X = 1000$:

$$\hat{\pi}(1000, \widehat{\boldsymbol{\beta}}) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 1000}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 1000}}$$

En el caso de que X sea una variable cualitativa, la estrategia a la que se suele recurrir es a la construcción de una variable indicadora (James et al., 2013). Retomando el ejemplo de las probabilidades de pago, una variable de interés podría ser considerar si el cliente bajo estudio es un estudiante o no. En este caso, la variable indicadora se define de la siguiente manera:

$$X = \begin{cases} 0 & \text{si el cliente no es un estudiante} \\ 1 & \text{si el cliente es un estudiante} \end{cases}$$

De este modo las probabilidades de realizar apropiadamente el pago, considerando el estatus del cliente como estudiante o no se calculan como:

$$\hat{P}(\text{pago=si} \mid \text{estudiante=si}) = \pi(si, \widehat{\boldsymbol{\beta}}) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1}}$$

$$\hat{P}(\text{pago=si} \mid \text{estudiante=no}) = \pi(no, \widehat{\boldsymbol{\beta}}) = \frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}}\,.$$

Finalmente, habiendo estimado las probabilidades $\pi(\boldsymbol{x}, \widehat{\boldsymbol{\beta}})$ para cada observación, estas son asignadas a una de dos clases ($Y = 0$ o $Y = 1$). Para dicho fin, se establece un punto de corte c, siendo $c = 0,5$ que es el valor que resulta de la regla óptima de Bayes tal como indica (1.1). La regla de decisión resultante será asignar la observación a la primera clase ($Y = 0$) si la probabilidad es menor a c, mientras que si la probabilidad estimada es superior a dicho valor, la observación es asignada a la segunda clase ($Y = 1$).

Capítulo 3

Análisis Discriminante Lineal

El Análisis Discriminante Lineal (conocido también por sus siglas en inglés, LDA), es un método originalmente planteado para clasificar observaciones en una de dos categorías, luego expandido al caso de más clases. La idea principal es hallar una combinación lineal de las características de las observaciones que logre una separación óptima entre las diferentes clases (Jorgensen, 2019). En ciertos escenarios, LDA puede llevar a mejores resultados que regresión logística, particularmente cuando las clases están bien separadas entre sí, o cuando el tamaño muestral es pequeño con las variables predictoras siguiendo una distribución normal en cada clase (James et al., 2013).

Sea $\pi_k = P(Y = k)$ la probabilidad de que una observación elegida al azar provenga de la k-ésima clase, con $k = 1, ..., K$. Este valor se denomina *probabilidad a priori*. Sea $f_k(x) = \mathbb{P}(X = x|Y = k)$ la *función de densidad* de X (o eventualmente de probabilidad puntual cuando X es discreta) para una observación de la k-ésima clase. Entonces, $f_k(x)$ tomará valores elevados si, dentro de la clase k, hay altas probabilidades de que $X \approx x$. Análogamente, si hay bajas probabilidades de que $X \approx x$ dentro de la clase k, entonces $f_k(x)$ será menor (James et al., 2013).

A partir del Teorema de Bayes sabemos que:

$$P(Y = k|X = x) = \frac{P(X = x|Y = k)\pi_k}{\sum_{i=1}^{K} P(X = x|Y = i)\pi_i} = \frac{f_k(x)\pi_k}{\sum_{i=1}^{K} f_i(x)\pi_i},$$

y esta resulta ser la *probabilidad a posteriori*, es decir, la probabilidad de que una observación pertenezca a la clase k dado un valor determinado de la variable predictora X. Existen distintas maneras de estimar a partir de los datos esta probabilidad, una de ellas es recurrir a un modelo para obtener una estimación de $f_k(x)$.

En el caso en que $d > 1$, una de las distribuciones que puede escogerse es la Normal (o Gaussiana) multivariada. En esta situación, la densidad de cada clase se modela de la siguiente manera:

$$f_k(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_k|}} e^{\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right]},$$

donde:

- $\boldsymbol{x} \in \mathbb{R}^d$,
- $\boldsymbol{\mu}_k \in \mathbb{R}^d$ es un vector de medias,
- $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times d}$ es una matriz de covarianzas,
- $|\boldsymbol{\Sigma}_k|$ denota al determinante de la matriz.

En el caso de LDA, además, asumimos que todas las clases tienen matrices de covarianzas iguales:

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma},$$

por lo que resulta:

$$f_k(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)} .$$

En resumen, los supuestos de LDA son que cada $X \in \mathbb{R}^d$ proviene de una distribución normal multivariada dentro de cada clase y que todas las clases tienen una matriz de covarianza

común Σ (Jorgensen, 2019). En términos formales, obtenemos:

$$X|Y = k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$$

3.1. LDA para clasificación binaria

Nos enfocaremos en el caso en que Y toma dos valores. Hemos visto que la *Regla de Bayes* asigna a cada observación a la clase para la cual $p_k(X)$ tome el mayor valor (James et al., 2013). Al estimar $f_k(x)$, se puede plantear un clasificador que se aproxime al clasificador bayesiano.

Primero se considera el caso en el cual hay solo dos clases, es decir que $K = 2$. En este contexto, el clasificador bayesiano divide al espacio $\mathbb{R}^d$ en dos regiones, una para los puntos en los cuales $p_1(X) > 0,5$ y otra para los puntos donde $p_0(X) > 0,5$. Al conjunto de puntos en los cuales $p_1(X) = p_0(X) = 0,5$ se lo llama *límite o frontera de decisión*.

Por ejemplo, en el caso univariado, es decir cuando $d = 1$, este límite es una recta en el espacio $\mathbb{R}^2$ (Ghojogh y Crowley, 2019). En general, para obtener la frontera de decisión, primero se plantea:

$$P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = P(Y = 0|\boldsymbol{X} = \boldsymbol{x})\,.$$

Aplicando el Teorema de Bayes a cada lado de la igualdad queda:

$$\frac{f_1(\boldsymbol{x})\pi_1}{\sum_{i=0}^{1} f_i(\boldsymbol{x})\pi_i} = \frac{f_0(\boldsymbol{x})\pi_0}{\sum_{i=0}^{1} f_i(\boldsymbol{x})\pi_i} \Rightarrow f_1(\boldsymbol{x})\pi_1 = f_0(\boldsymbol{x})\pi_0\,. \tag{3.1}$$

Reemplazando $f_i(\boldsymbol{x})$ por la densidad de la distribución normal multivariada y asumiendo:

$$\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}\,,$$

resulta:

$$\pi_0 \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} e^{\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)\right]} = \pi_1 \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} e^{\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right]}.$$

Aplicando el logaritmo natural, simplificando los términos que involucran la raíz cuadrada y reordenando los términos tenemos:

$$-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_0^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0 + \boldsymbol{\mu}_0^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} + \ln(\pi_0) = -\frac{1}{2}\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} + \ln(\pi_1)$$

Multiplicando ambos lados por 2, se obtiene el siguiente hiperplano, correspondiente al límite de decisión entre las dos clases:

$$2(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0))^T \boldsymbol{x} + ((\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)) + 2\ln\left(\frac{\pi_1}{\pi_0}\right) = 0$$

De esta manera, obtenemos que la frontera de decisión es:

$$\eta(\boldsymbol{x}) = 2(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0))^T \boldsymbol{x} + (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1) + 2\ln\left(\frac{\pi_1}{\pi_0}\right),$$

de donde resulta la siguiente regla de decisión:

$$g_{LDA} = \begin{cases} 1 & \text{si } \eta(\boldsymbol{x}) > 0 \\ 0 & \text{si } \eta(\boldsymbol{x}) < 0 \end{cases}$$

La Figura 3.1, tomada de James et al. (2013) muestra un ejemplo para el caso donde $d = 1$ y $K = 2$. El panel izquierdo muestra dos poblaciones, distribuidas normalmente, con la línea punteada indicando el límite de decisión dada por la regla óptima bayesiana. En el panel derecho, se observan dos muestras de tamaño 20 extraídas de dichas poblaciones. En este caso, la línea sólida corresponde al límite de decisión estimado por LDA con datos de las muestras, como se verá en la Sección 3.2.

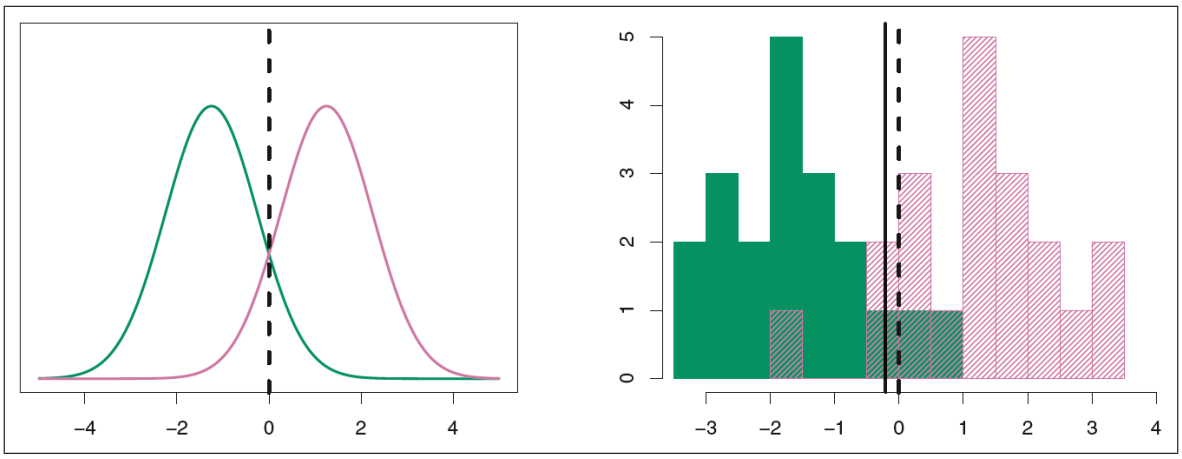

Figura 3.1: Dos poblaciones normales (izquierda) y dos muestras obtenidas de dichas poblaciones (derecha). Las líneas punteadas corresponden al límite de decisión bayesiano y la línea sólida al límite de decisión obtenido con LDA.

Por otro lado, también podemos considerar que clasificaremos un nuevo ítem en la clase 1 si $P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) > P(Y = 0|\boldsymbol{X} = \boldsymbol{x})$ y en la clase 0 en caso contrario. Dicho de otra manera, clasificaremos en la clase $j = 0$ ó 1 con el mayor valor de la expresión

$$\frac{f_j(\boldsymbol{x})\pi_j}{\sum_{i=0}^{1} f_i(\boldsymbol{x})\pi_i},$$

o equivalentemente con mayor

$$\pi_j f_j(\boldsymbol{x}).$$

Reemplazando la densidad condicional f_j por la densidad normal, tenemos que clasificaremos en la clase j con mayor valor :

$$\pi_j \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} e^{\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right]}.$$

Esto equivale, luego de tomar logaritmo, a clasificar en la clase con mayor valor:

$$\delta_j(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j - \frac{1}{2} \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\mu}}_j + \ln(\pi_j), \tag{3.2}$$

que se conoce como *función discriminante*. Notemos que esta función es lineal en $\boldsymbol{x}$, hecho que le da nombre a este método de análisis discriminate. En términos de $\delta_j(\boldsymbol{x})$ la regla de decisión puede reescribirse como:

$$g_{LDA} = \begin{cases} 1 & \text{si } \delta_1(\boldsymbol{x}) > \delta_0(\boldsymbol{x}) \\ 0 & \text{si } \delta_1(\boldsymbol{x}) < \delta_0(\boldsymbol{x}) \end{cases}$$

3.2. Estimación de parámetros

En la práctica, los parámetros de las distribuciones normales son desconocidos, por lo cual es necesario estimarlos usando la información provista por los datos de la muestra (Hastie et al., 2017). Siendo n el tamaño muestral y n_j el número de observaciones pertenecientes a la clase $k = 0, 1$ en la muestra, los parámetros pueden estimarse como:

- $\hat{\pi}_j = \dfrac{n_j}{n}$
- $\hat{\boldsymbol{\mu}}_j = \dfrac{1}{n_j} \sum_{i:y_i=j} \boldsymbol{x}_i$
- $\hat{\boldsymbol{\Sigma}} = \dfrac{1}{n-2} \sum_{k=0}^{1} \sum_{i:y_i=j} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_j)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_j)^T.$

Al reemplazar los parámetros con sus respectivos estimadores, se obtiene la función discriminante lineal estimada (James et al., 2013):

$$\hat{\delta}_j(\boldsymbol{x}) = \boldsymbol{x}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_j - \frac{1}{2} \hat{\boldsymbol{\mu}}_j^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_j + \ln(\hat{\pi}_j)$$

De este modo, también se pueden estimar los límites o fronteras de decisión para cada par de clases (Jorgensen, 2019). Para las clases 1 y 0, el límite se estima como el conjunto de puntos donde $\hat{\delta}_1 = \hat{\delta}_0$, es decir:

$$\boldsymbol{x}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_1 - \frac{1}{2} \hat{\boldsymbol{\mu}}_1^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_1 + \ln(\hat{\pi}_1) = \boldsymbol{x}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_0 - \frac{1}{2} \hat{\boldsymbol{\mu}}_0^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_0 + \ln(\hat{\pi}_0)$$

3.3. LDA y Regresión Logística

A pesar de sus diferentes enfoques, los métodos de LDA y Regresión Logística poseen ciertas semejanzas. Por simplicidad, ilustraremos la relación cuando $d = 1$, en cuyo caso las densidades condicionales $f_j(x)$ corresponden a las de una normal univariada, más precisamente a $N(\mu_j, \sigma^2)$. El logaritmo de los odds de las probabilidades a posteriori puede calcularse como:

$$\ln\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}\right) = \ln\left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right) = \ln\left(P(Y=1|X=x)\right) - \ln\left(P(Y=0|X=x)\right) .$$

Luego, a partir de (3.1) obtenemos:

$$\begin{aligned}
\ln\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}\right) &= x\,\frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \ln(\pi_1) - \left(x\,\frac{\mu_0}{\sigma^2} - \frac{\mu_0^2}{2\sigma^2} + \ln(\pi_0)\right) \\
&= x\left(\frac{\mu_1}{\sigma^2} - \frac{\mu_0}{\sigma^2}\right) + \left(\frac{\mu_0^2}{2\sigma^2} - \frac{\mu_1^2}{2\sigma^2}\right) + \ln\left(\frac{\pi_1(x)}{\pi_0(x)}\right) \\
&= c_0 + c_1 x\,,
\end{aligned}$$

para constantes c_0 y c_1 adecuadamente definidas.

Como se puede apreciar, el logit calculado tanto con Regresión Logística como con LDA son funciones lineales de x (James et al., 2013). Dicha relación puede extenderse en forma directa también al caso multidimensional, es decir $d > 1$. Por ende, ambas técnicas producen límites de decisión lineales similares y suelen derivar en resultados muy parecidos. La diferencia radica principalmente en la forma en la que los parámetros son estimados, pues Regresión Logística utiliza el método de máxima verosimilitud (como se explica en la sección 2.1) y LDA utiliza datos de la muestra (que se asume proveniente de una población con distribución normal) (Jorgensen, 2019). Dicho de otra forma: Regresión Logística realiza supuestos sobre las probabilidades a posteriori mientras que LDA trabaja con supuestos sobre las probabilidades a priori y la función de densidad (Ghojogh y Crowley, 2019).

La decisión entre aplicar una técnica o la otra dependerá principalmente del cumplimiento de los supuestos de LDA. Se ha comprobado que LDA produce resultados superiores en

escenarios en los que se cumplen los supuestos mencionados en la introducción al Capítulo (Hastie et al., 2017). Esta técnica también se recomienda en los casos en los que el tamaño muestral y el número de parámetros a estimar es reducido (Jorgensen, 2019).

Por su parte, Regresión Logística es un método mucho más flexible, ya que no realiza supuestos sobre las distribuciones de las variables predictoras. Mientras que LDA asume que las densidades condicionales $f_k(\boldsymbol{x})$ siguen una normal multivariada, Regresión Logística no realiza ni supuestos ni estimaciones sobre ella (Wasserman, 2004). Esta característica lleva a un mejor desempeño en la clasificación cuando se incumplen los supuestos de LDA (James et al., 2013). Otra ventaja es que presenta cierta robustez frente a los valores extremos, en tanto que LDA es muy sensible a ellos (Jorgensen, 2019).

Una desventaja que comparten tanto Regresión Logística como LDA es que sus límites o fronteras de decisión pueden llegar a ser demasiado simples; sin embargo, en tal caso, Regresión Logística permitiría complejizar el modelo en términos de las covariables. A menudo, resulta de interés separar las clases con límites de decisión más irregulares y no solo lineales (Hastie et al., 2017). Si el tamaño muestral es lo suficientemente grande, se puede recurrir a una generalización de LDA conocida como Análisis Discriminante Cuadrático. Este tema se desarrolla en la siguiente sección.

Capítulo 4

Análisis Discriminante Cuadrático

Como hemos visto Regresión Logística y LDA son métodos de clasificación lineales. El Análisis Discriminante Cuadrático (por sus siglas en inglés, QDA) es una extensión no lineal de LDA que resulta más flexible y que se prefiere cuando los límites de decisión bayesianos son cuadráticos (Jorgensen, 2019).

A diferencia de LDA, QDA relaja el supuesto de igualdad de las matrices de covarianzas. Ahora, se permite que cada clase tenga su propia matriz de covarianza Σ_k, es decir, no se asume que sean iguales Σ_1 y Σ_0. Aunque las covarianzas puedan ser distintas, esto no significa que deban serlo necesariamente. En el caso particular cuando todas las matrices son iguales, los límites de decisión serán lineales y QDA se reduce a LDA (Ghojogh y Crowley, 2019).

Asumiremos, entonces, que:

$$X|Y = k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma_k}) \quad k = 0, 1\,.$$

Para obtener la función discriminante en este caso, se utiliza el mismo procedimiento que en la sección anterior, excepto que ahora se utiliza una matriz distinta por cada clase. Debido a esto, algunas cancelaciones no podrán realizarse y, en particular, el término cuadrático de $\boldsymbol{x}$ permanece, dándole su nombre a la técnica (Hastie et al., 2017). Nuevamente, siguiendo la

regla óptima, clasificaremos una observación en la categoría 1 si

$$P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) > P(Y = 0|\boldsymbol{X} = \boldsymbol{x})\,,$$

que como hemos visto, es equivalente a:

$$f_1(\boldsymbol{x})\pi_1 > f_0(\boldsymbol{x})\pi_0\,, \tag{4.1}$$

donde f_k corresponde a la densidad normal en $\mathbb{R}^d$. En consecuencia, clasificaremos en la clase 1 si

$$\frac{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)}}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}}\pi_1 > \frac{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)}}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_0|^{\frac{1}{2}}}\pi_0\,.$$

Después de tomar logaritmo natural y simplificar, vemos que se clasifica en la clase 1 si

$$-\left(\mathbf{x}-\boldsymbol{\mu}_1\right)^T\boldsymbol{\Sigma}_1^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_1\right)+\left(\mathbf{x}-\boldsymbol{\mu}_0\right)^T\boldsymbol{\Sigma}_0^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_0\right)+\ln\left(\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|}\right)+2\ln\left(\frac{\pi_1}{\pi_0}\right)>0\,,$$

o equivalentemente, si

$$\mathbf{x}^T\left(\boldsymbol{\Sigma}_0^{-1}-\boldsymbol{\Sigma}_1^{-1}\right)\mathbf{x}+2\left(\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\right)\mathbf{x}+\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0-\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1+\ln\left(\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|}\right)+2\ln\left(\frac{\pi_1}{\pi_0}\right)>0\,.$$

Notemos la frontera de decisión entre las dos clases ya no queda lineal sino cuadrática ya que la expresión de arriba es de la forma:

$$\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}+\mathbf{b}^{\mathrm{T}}\mathbf{x}+c$$

para una matriz $\mathbf{A}$, un vector $\mathbf{b}$ y una constante c adecuados. Luego, si definimos

$$\eta^*(\mathbf{x})=\mathbf{x}^T\left(\boldsymbol{\Sigma}_0^{-1}-\boldsymbol{\Sigma}_1^{-1}\right)\mathbf{x}+2\left(\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\right)\mathbf{x}+\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0-\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1+\ln\left(\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|}\right)+2\ln\left(\frac{\pi_1}{\pi_0}\right)\,,$$

la regla de decisión puede escribirse como

$$g_{QDA} = \begin{cases} 1 & \text{si } \eta^*(\boldsymbol{x}) > 0 \\ 0 & \text{si } \eta^*(\boldsymbol{x}) < 0 \, . \end{cases}$$

Como en el capítulo anterior, podemos enfocarnos en la función discriminante cuadrática. Para ello, volvamos a notar que clasificaremos en la clase k con mayor valor:

$$\pi_k \, f_k(\boldsymbol{x}) \, ,$$

es decir con mayor

$$\pi_k \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_k|}} e^{\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right]} \, .$$

Dado que la función logaritmo es estrictamente creciente podemos tomar logaritmo de esta expresión y clasificar en la clase k con mayor:

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

En resumen, una vez más, la regla de decisión es asignar a cada unidad a la clase k que maximice su valor de $\delta_k^*(\boldsymbol{x})$ y podemos reescribir la regla como:

$$g_{QDA} = \begin{cases} 1 & \text{si } \delta_1^*(\boldsymbol{x}) > \delta_0^*(\boldsymbol{x}) \\ 0 & \text{si } \delta_1^*(\boldsymbol{x}) < \delta_0^*(\boldsymbol{x}) \, . \end{cases}$$

4.1. Estimación de parámetros

Las estimaciones para QDA son similares a las estimaciones para LDA mencionadas en la Sección 3.2, excepto que ahora se deben estimar matrices de covarianza separadas para cada clase (Hastie et al., 2017). El estimador de máxima verosimilitud para la k-ésima clase

es:

$$\hat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k} \sum_{i:y_i=k} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)^T$$

Alternativamente, se puede utilizar el estimador insesgado:

$$\hat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k - 1} \sum_{i:y_i=k} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)^T \, .$$

4.2. QDA y LDA

Al aplicar un método de aprendizaje estadístico, se desea mantener simultáneamente un bajo sesgo y una baja varianza. Un método con sesgo alto y baja varianza lleva a un infrajuste de los datos, mientras que un método con bajo sesgo pero alta varianza producirá sobreajuste. Se denomina compromiso sesgo-varianza (en inglés *Bias-Variance Trade-off*) al intento de evitar tanto el infrajuste como el sobreajuste en el modelo (Jorgensen, 2019). En general, los métodos más complejos (es decir, con más parámetros), tendrán menor sesgo pero mayor varianza.

La decisión de utilizar LDA o QDA depende del compromiso sesgo-varianza. Para d variables, aplicar LDA requiere estimar una única matriz de covarianzas con $\frac{d(d+1)}{2}$ parámetros. Por otro lado, QDA trabaja con matrices separadas para cada clase, así que se deberá estimar $K\frac{d(d+1)}{2}$ parámetros (en nuestro caso $K = 2$), un número muy superior. Consecuentemente, LDA es una técnica que puede tener un muy buen desempeño en la clasificación que requiere estimar menos parámetros, pero si no se cumple el supuesto de igualdad de covarianzas, el sesgo puede llegar a ser elevado (James et al., 2013). El método de QDA, que requiere la estimación de un número mayor de parámetros, suele recomendarse cuando el tamaño de muestra es suficientemente grande.

La Figura 4.1, tomada de James et al. (2013), muestra dos escenarios, ambos para el caso donde $d = 2$ y $K = 2$. En el panel izquierdo se refleja un ejemplo en el cual $\Sigma_1 = \Sigma_0$ con una correlación igual a 0,7. En el panel de la derecha se ilustra con la situación en que $\Sigma_1 \neq \Sigma_0$, con correlación 0,7 y $-0{,}7$ respectivamente. En ambos casos, la línea punteada violeta es el límite

bayesiano, la línea punteada negra es el límite de decisión estimado por LDA y la curva verde es el límite estimado por QDA. En el panel de la izquierda, el límite de decisión de Bayes es lineal y observamos que LDA da una buena aproximación a la frontera de decisión. Como notan James et al. (2013), el límite de decisión de QDA tiene un desempeño inferior, porque sufre de una mayor varianza sin una disminución correspondiente del sesgo. Se observa que, al cumplirse el supuesto de igualdad de varianzas, el límite de clasificación bayesiano es lineal y por lo tanto, LDA produce un mejor ajuste. Por el contrario, en el lado derecho vemos que al violarse dicho supuesto, QDA resulta en una mejor aproximación. al límite de decisión que ya no es lineal.

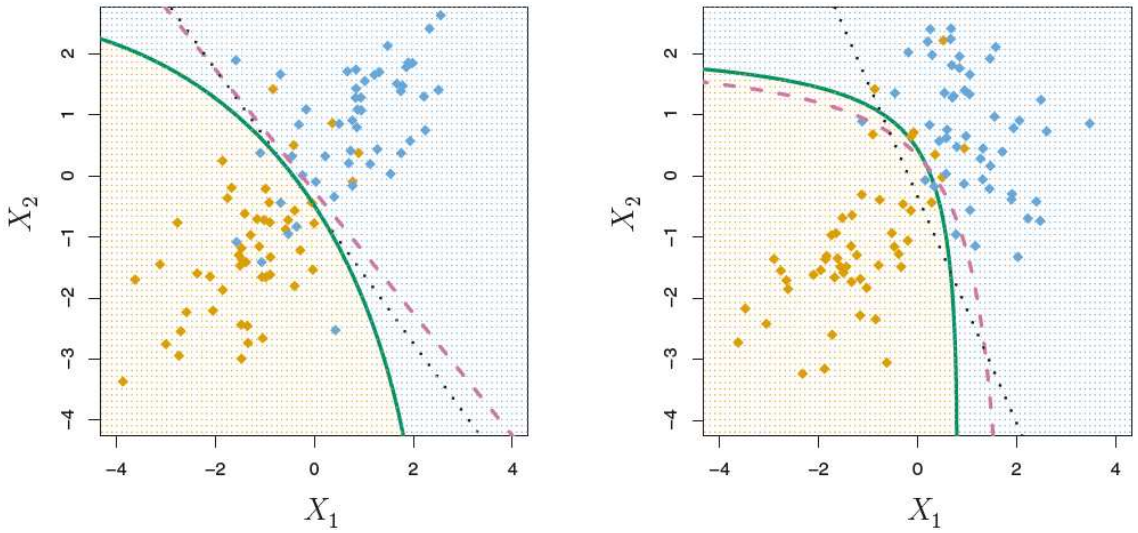

Figura 4.1: Ejemplo de un problema de clasificación para los casos $\Sigma_1 = \Sigma_2$ (izquierda) y $\Sigma_1 \neq \Sigma_2$ (derecha). Las lineas violeta, negra y verde corresponden a los límites de decisión bayesiano, estimado por LDA y estimado por QDA respectivamente.

4.3. QDA y Regresión Logística

Como se mencionó al principio de la Sección 4, tanto LDA como Regresión Logística, tal como lo hemos planteado, son métodos de clasificación lineales. Esto quiere decir que, cuando el verdadero límite entre dos poblaciones es no lineal, ambas técnicas tenderán al infrajuste (Jorgensen, 2019). Por lo tanto, es más recomendable aplicar metodologías que

permitan ajustar límites de decisión no lineales, tales como QDA. Cabe mencionar que, al tratarse de un método de regresión, Regresión Logística permitiría la posibilidad de plantear un modelo más complejo en términos de las covariables, haciéndolo más flexible.

En contraste, cuando los verdaderos límites son lineales, QDA ajustará límites de decisión más flexibles de lo necesario. De esta manera, el método presentará mayor varianza al mismo tiempo que no logrará reducir el sesgo de forma notable (James et al., 2013).

De manera similar a lo ocurrido entre LDA y Regresión Logística (Sección 3.3), QDA también realiza supuestos sobre la función de densidad y las probabilidades a priori, mientras que Regresión Logística hace supuestos sobre las probabilidades a posteriori. Por ende, QDA también tendrá usualmente un desempeño peor cuando no se cumplan los supuestos correspondientes, en otras palabras, cuando las variables no provengan de una distribución normal multivariada dentro de cada clase (Ghojogh y Crowley, 2019). James et al., (2013) destacan que se ha comprobado que QDA es incluso más sensible que LDA ante la falta de normalidad, por lo que no resulta provechoso aplicarlo si se viola este supuesto.

Finalmente, cuando los datos provienen de poblaciones que siguen funciones no lineales más complejas, Regresión Logística, LDA y QDA presentarán todos una baja precisión en la clasificación, aunque se espera que QDA sea levemente superior a los otros dos métodos (James et al., 2013). En estos casos, será necesario aplicar metodologías no paramétricas más flexibles, como la que se verá a continuación.

Capítulo 5

Support Vector Machine

Las máquinas de vectores de soporte, más conocidas como *support vector machine* o SVM, son una técnica cuyo desarrollo comenzó en la década de 1990 en la comunidad de Ciencias de la Computación y ha ganado popularidad debido a su buen desempeño en tareas de clasificación (James et al., 2013) tanto en escenarios lineales como no lineales. A diferencia con Regresión Logística, no se ajusta un modelo para $P(Y = 1|X = x)$, sino que se construye el clasificador directamente. Efron y Hastie (2016) mencionan que en general, SVM suele ser preferible cuando el énfasis está puesto en maximizar la precisión en la clasificación y no en la inferencia estadística.

SVM es una generalización del método de clasificador de vectores de soporte (o SVC por sus siglas en inglés), que es a su vez una generalización del método de *clasificadores de máximo margen* . Comparado con estas dos técnicas, SVM es más flexible y puede ser aplicado en un rango de situaciones más amplio.

Para un espacio de dimensión d, se define *hiperplano* como el subespacio de dimensión $d - 1$ conformado por puntos con coordenadas $(X_1, X_2, ...X_d)$ que verifican la siguiente ecuación:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_d X_d = \beta_0 + \sum_{i=1}^{d} \beta_i X_i = 0\,.$$

Para el caso en el cual $d = 2$, este hiperplano es una recta, mientras que para $d = 3$, la ecuación

define a un plano. Para todo punto que no satisface la ecuación, hay dos casos posibles: o bien $\beta_0 + \sum_{i=1}^{d} \beta_i X_i > 0$, o bien $\beta_0 + \sum_{i=1}^{d} \beta_i X_i < 0$. Esto quiere decir que el hiperplano dividirá al espacio en dos regiones separadas. Suponiendo que Y es una variable binaria que indica dos clases distintas, se dice que los datos son *linealmente separables* si existe un hiperplano que divide perfectamente a ambas clases.

La técnica de *clasificadores de máximo margen* (MMC, por sus siglas en inglés), consiste en encontrar el hiperplano que separe ambas clases y maximice la distancia euclídea entre las observaciones y dicho separador. A la mínima distancia entre el hiperplano separador elegido y alguno de los puntos se la llama *margen*, dándole su nombre al clasificador. Todo punto que esté sobre los márgenes se denomina *vector de soporte*, pues la ubicación del hiperplano óptimo depende de ellos (Wasserman, 2004).

Cabe mencionar que, por razones prácticas, en el contexto de Support Vector Machine la representación de la variable binaria se cambia a los valores 1 y -1, es decir, que ahora $Y \in \mathcal{Y} = \{-1, 1\}$ en lugar de $\{0, 1\}$ como veníamos trabajando.

Podemos pensar que tenemos una matriz de datos $\mathbb{X}$, de dimensión $n \times d$, con filas $\boldsymbol{X}_i^T$, que consiste en nuestros n datos de entrenamiento que caen en dos clases: $\{-1, 1\}$ que se registran en las respuestas $Y_1, \ldots, Y_n$. Supongamos que es posible construir un hiperplano que separe las observaciones de entrenamiento perfectamente a un lado y otro del hiperplano de acuerdo con sus etiquetas de clase. Es decir, etiquetamos las observaciones de la clase azul como $y_i = 1$ y por otro lado, las de la clase magenta como $Y_i = -1$. Bajo separación perfecta, si denotamos $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$, un hiperplano separador cumple:

$$\boldsymbol{\beta}^T \boldsymbol{X}_i + \beta_0 > 0 \text{ si } Y_i = 1$$

$$\boldsymbol{\beta}^T \boldsymbol{X}_i + \beta_0 < 0 \text{ si } Y_i = -1$$

o equivalentemente

$$Y_i(\boldsymbol{\beta}^T \boldsymbol{X}_i + \beta_0) > 0, i = 1, \ldots, n\,.$$

Esto nos llevaría a un clasificador del tipo: $sgn(\boldsymbol{\beta}^T \boldsymbol{X}_i + \beta_0)$. Lo que se conoce como el *clasi-*

ficador de máximo margen es el separador que está más lejos de las observaciones de entrenamiento. En efecto, podemos calcular la distancia perpendicular de cada observación a un hiperplano de separación dado; la distancia más pequeña de este tipo se conoce como *margin* (o margen). El *hiperplano de máximo margen* es el hiperplano de separación para el cual el margen es mayor, o sea es el hiperplano que tiene la mayor distancia mínima a las observaciones de entrenamiento. MMC es un clasificador fácil de implementar e interpretar, ya que la regla de decisión simplemente consiste en asignarle a cada observación la clase indicada por su ubicación respecto al hiperplano. Sin embargo, no puede aplicarse en los casos no linealmente separables (los cuales, en la práctica, serán los más usuales) y además es una técnica poco robusta, es decir, que es muy afectada por la presencia de valores extremos en los datos.

Formalmente, decimos que las clases son linealmente separables si existen $\boldsymbol{\beta}$ y β_0 tales que $Y_i z(\mathbf{X}_i) > 0$, para todo punto en la muestra de entrenamiento, siendo $z(\boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{x} + \beta_0$.

Dado un punto $\boldsymbol{x}$, sea $\boldsymbol{x}^*$ su proyección ortogonal a la frontera definida por un hiperplano separador $\boldsymbol{\beta}^T \boldsymbol{x} + \beta_0 = 0$. Observemos que:

- $\boldsymbol{x}^*$ cumple: $\boldsymbol{\beta}^T \boldsymbol{x}^* + \beta_0 = 0$
- $(\boldsymbol{x} - \boldsymbol{x}^*)$ es ortogonal a hiperplano, cuyo vector normal es $\boldsymbol{\beta}$
- $|\boldsymbol{\beta}^T (\boldsymbol{x} - \boldsymbol{x}^*)| = ||\boldsymbol{\beta}||\, ||\boldsymbol{x} - \boldsymbol{x}^*||$
- $\boldsymbol{\beta}^T (\boldsymbol{x} - \boldsymbol{x}^*) = \boldsymbol{\beta}^T (\boldsymbol{x} - \boldsymbol{x}^* \pm \beta_0) = z(\boldsymbol{x})$
- Por lo tanto: $||\boldsymbol{x} - \boldsymbol{x}^*|| = \dfrac{|z(\boldsymbol{x})|}{||\boldsymbol{\beta}^T||}$
- Luego para cada punto de la muestra tenemos que la distancia es:

$$\frac{|z(\boldsymbol{X}_i)|}{||\boldsymbol{\beta}||} = \frac{Y_i\, z(\boldsymbol{X}_i)}{||\boldsymbol{\beta}||}$$

Por lo tanto, si definimos

$$M = M(\boldsymbol{\beta}, \beta_0) = \min_i \frac{Y_i\, z(\boldsymbol{X}_i)}{||\boldsymbol{\beta}||} = \frac{1}{||\boldsymbol{\beta}||} \min_i Y_i\, (\boldsymbol{\beta}^T \boldsymbol{X}_i + \beta_0) = \frac{Y_\ell\, (\boldsymbol{\beta}^T \boldsymbol{X}_\ell + \beta_0)}{||\boldsymbol{\beta}||}$$

queremos maximizar el margen $M(\boldsymbol{\beta}, \beta_0)$ sujeto a que $Y_i\, z(\mathbf{X}_i) > 0, i = 1, \ldots, n$. El problema de maximizar el margen M puede escribirse como:

$$\underset{\boldsymbol{\beta},\beta_0}{\operatorname{argmax}} \left\{ \frac{1}{||\boldsymbol{\beta}||} \min_i \{Y_i\, (\boldsymbol{\beta}^T \boldsymbol{X}_i + \beta_0)\} \right\}$$

o como suele escribirse en general como $\underset{\boldsymbol{\beta},\beta_0,||\boldsymbol{\beta}||=1}{\max} M$ sujeto a $Y_i\, (\boldsymbol{\beta}^T \boldsymbol{X}_i + \beta_0) \geq M,\ i = 1, \ldots, n$.

5.1. Support Vector Classifier

El clasificador de vectores de soporte (SVC), también llamado clasificador de margen blando (*Soft-Margin Classifier*), es una generalización del MMC que puede aplicarse en un escenario donde los datos no son linealmente separables. James et al. (2013) observan que, comparado con MMC, SVC suele ser menos sensible a la influencia de observaciones individuales y producir mejores resultados en la clasificación .

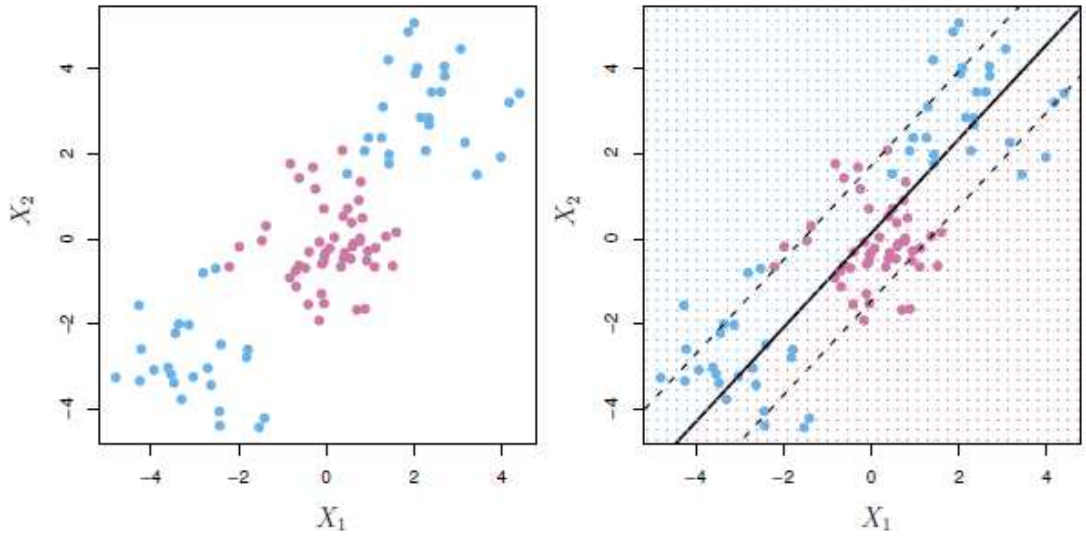

Figura 5.1: Izquierda: Las observaciones se dividen en dos clases, con un límite no lineal entre ellas. Derecha: el SVC busca un límite lineal, y en consecuencia tiene un rendimiento muy pobre. Fuente: James et al. (2013).

La generalización ocurre al permitir que algunos puntos violen el margen, en otras palabras, puede haber observaciones en el lado incorrecto del margen o del hiperplano, como se puede observar en la Figura 5.1. La forma de lograr esto es introduciendo un nuevo parámetro, indicado como C, que indica la cantidad de observaciones que se está dispuesto a clasificar incorrectamente en el conjunto de entrenamiento. De esta manera, el conjunto de hiperplanos que el algoritmo puede considerar aumenta, pues ahora no se limita a los casos donde este hiperplano logra una separación óptima de las clases (Efron y Hastie, 2016).

Cuánto mayor sea C, mayor será el nuevo conjunto de soporte (ya que se permite un mayor número de observaciones mal clasificadas), disminuyendo la variabilidad pero aumentando el sesgo. Mientras tanto, valores pequeños de C producirán reglas con bajo sesgo pero gran variabilidad (James et al., 2013). Para el caso particular cuando $C = 0$, el método resulta idéntico a MMC. En la práctica, el valor de C será tuneado mediante validación cruzada.

La Figura 5.2, tomada de Hastie et al. (2017), ilustra el caso separable en el panel de la izquierda y el no separable en el de la derecha. La frontera de decisión es la línea sólida, mientras que las líneas punteadas delimitan el margen máximo sombreado, con un valor de ancho $2C = 1/||\boldsymbol{\beta}||$. El panel de la derecha muestra el caso no separable. Los puntos indicados con ξ^* yacen del lado incorrecto del margen por una cantidad $\xi_i^* = C\xi_i$, mientras que los puntos ubicados del lado correcto corresponden a $\xi_i^* = 0$. El margen es maximizado sujeto a que $\sum \xi_i < C$, por lo que $\sum \xi_i^* < cte$ es la distancia total de los puntos que están del lado incorrecto de sus márgenes.

En este sentido, el problema de maximizar ahora es:

$$
\begin{aligned}
&\underset{\beta,\boldsymbol{\beta}_0,\xi_i}{\text{máx}}\ M \\
&\text{sujeto a } ||\boldsymbol{\beta}|| = 1 \\
&Y_i\,(\boldsymbol{\beta}^T\mathbf{x}_i + \beta_0) \geq M(1-\xi_i),\ i = 1,\ldots,n \\
&\xi_i \geq 0,\ \sum_{i=1}^{n} \xi_i \leq C, C > 0
\end{aligned}
$$

donde C representa el costo y es un hiperparámetro que es *data–driven*.

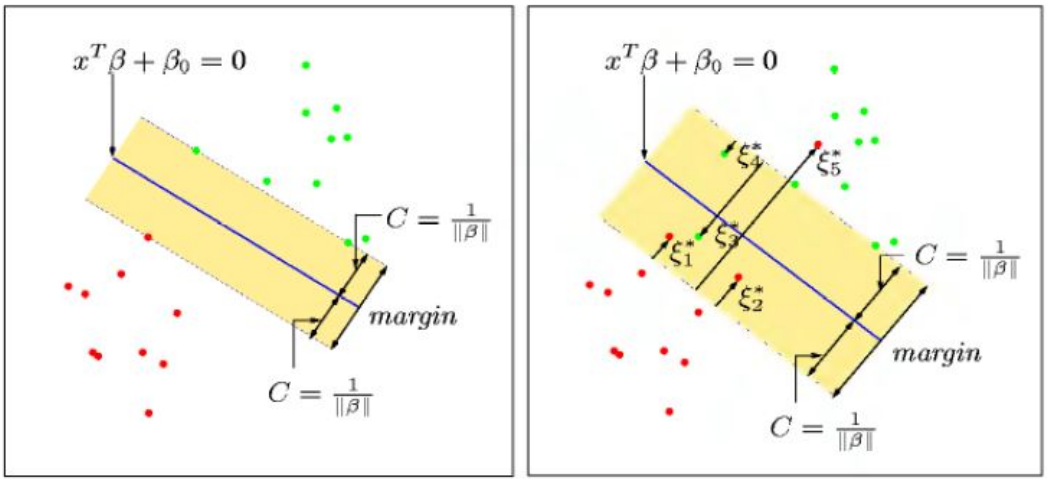

Figura 5.2: El panel de la izquierda representa el caso separable, mientras que el de la derecha muestra el caso no separable. Fuente: Hastie et al. (2017).

El clasificador se representa con la siguiente fórmula:

$$f(\boldsymbol{X}) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle \boldsymbol{X}, \boldsymbol{X}_i \rangle \,,$$

donde:

- $\boldsymbol{X} = (X_1, X_2, ..., X_d)$ es un vector de coordenadas para un individuo que se desea clasificar.
- $\langle \boldsymbol{X}, \boldsymbol{X}_i \rangle$ es el producto interno entre vectores, es decir que $\langle \boldsymbol{X}_i, \boldsymbol{X}_k \rangle = \sum_{j=1}^{d} X_{ij} X_{kj}$.
- $\boldsymbol{\alpha_i}$ son coeficientes para cada una de las n observaciones.

Los parámetros $\boldsymbol{\alpha_i}$ que definen al hiperplano son estimados de la misma manera que ocurría antes con los parámetros β_i, pero esta vez sujeto a restricciones impuestas por el parámetro C (Wasserman, 2004). Los puntos para los cuáles $\boldsymbol{\alpha_i} \neq 0$ conforman los vectores de soporte. Dado que para cualquier observación que no sea un vector de soporte los coeficientes serán iguales a cero, la regla de decisión implementada por SVC depende únicamente de los

productos internos entre las unidades a clasificar y los vectores de soporte. Finalmente, β_0 es estimado utilizando las estimaciones $\hat{\alpha}$. La regla de clasificación, una vez más, es dada por la región en la cual queda ubicada cada observación, indicada según el signo de $\hat{f}(\boldsymbol{X})$.

5.2. Kernelización

Si bien SVC presenta mejoras notables respecto a MMC y puede aplicarse en una mayor variedad de situaciones, aún posee una importante restricción: únicamente es capaz de hallar límites lineales (Hastie et al., 2017). Para generalizar esta situación y permitir límites no lineales, se recurre a un proceso conocido como kernelización. La idea básica es proyectar la covariable $\boldsymbol{X}$ a un espacio de mayor dimensión y aplicar el clasificador en este nuevo espacio, logrando así mayor flexibilidad sin perder demasiada simplicidad computacional (Wasserman, 2004). Por ejemplo: si los grupos son separables por una elipse en el espacio $\mathbb{R}^2$, al proyectar las variables $\boldsymbol{X}_i$ al espacio $\mathbb{R}^3$, es posible calcular un límite de decisión lineal.

Así, es posible generalizar la fórmula para $f(\boldsymbol{X})$ vista anteriormente utilizando una mayor variedad de funciones en lugar de recurrir únicamente al producto interno. Esto produce un clasificador más flexible sin perder las ventajas de utilizar un límite de decisión lineal. Dicho clasificador es conocido como Support Vector Machine. Su fórmula resulta:

$$f(\boldsymbol{X}) = \beta_0 + \sum_{i=1}^{n} \alpha_i K(\boldsymbol{X}, \boldsymbol{X}_i)\,,$$

donde $K(\boldsymbol{X}_i, \boldsymbol{X}_k)$ es una función llamada *kernel*, que cuantifica la semejanza entre dos observaciones. Si $K(\boldsymbol{X}_i, \boldsymbol{X}_k) = \sum_{j=1}^{d} X_{ij}X_{kj}$, esto se conoce como kernel lineal. Para aplicar SVM, se suele elegir una función no lineal. Algunas alternativas utilizadas comúnmente incluyen:

- Kernel polinómico (de orden d): $K(\boldsymbol{X}_i, \boldsymbol{X}_k) = \left(1 + \sum_{j=1}^{d} X_{ij}X_{kj}\right)^d$, siendo $d \in \mathbb{Z} > 1$.
- Kernel radial: $K(\boldsymbol{X}_i, \boldsymbol{X}_k) = e^{[-\gamma \times \sum_{j=1}^{d}(X_{ij}-X_{kj})^2]}$, siendo $\gamma > 0$.
- Kernel Gaussiano: $K(\boldsymbol{X}_i, \boldsymbol{X}_k) = e^{\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{d}(X_{ik}-X_{jk})^2\right)}$, siendo $\sigma > 0$.

- Kernel sigmoide: $K(\boldsymbol{X}_i, \boldsymbol{X}_k) = \tanh\Big(a \sum_{j=1}^{d} X_{ij} X_{kj} + b\Big)$.

La Figura 5.3 ejemplifica la sensibilidad que posee la construcción de los hiperplanos ante la variación o incorporación de un dato. La adición de una sola observación en el panel derecho conduce a un cambio dramático en el hiperplano óptimo.

Como observan James et al. (2013), esto ocurre debido a que el kernel radial presenta un comportamiento más "local" que el kernel polinómico: si una determinada observación del conjunto de prueba $\boldsymbol{X}^*$ se encuentra muy alejada de una observación del conjunto de entrenamiento x_i, entonces $\sum_{j=1}^{d}(X_j^* - X_{kj})^2$ será muy elevado y por lo tanto $K(\boldsymbol{X}^*, \boldsymbol{X}_i) = e^{[-\gamma \times \sum_{j=1}^{d}(X_j^* - X_{kj})^2]}$ presentará un valor muy bajo. En consecuencia, únicamente las observaciones cercanas a $\boldsymbol{X}^*$ tendrán un rol importante en su clasificación.

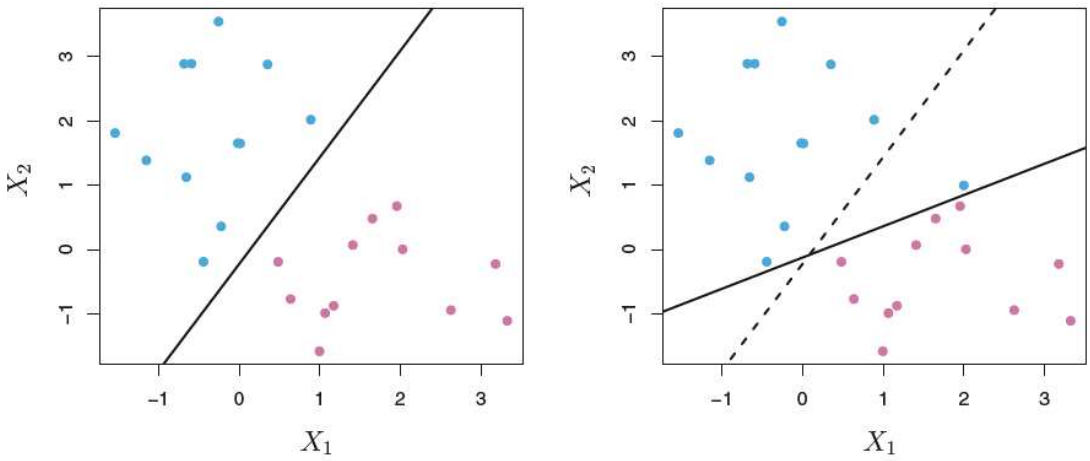

Figura 5.3: Ejemplo de ajuste de hiperplano . Izquierda: Se observan dos tipos de datos,que se muestran en azul y en violeta, junto con el hiperplano de óptimo. Derecha: se añade un dato adicional azul, lo que lleva a un cambio dramático en el hiperplano óptimo, representado como una línea continua. La línea discontinua indica el antiguo hiperplano óptimo,que se obtuvo en ausencia de este punto adicional. Fuente: James et al). (2013.

Una alternativa a la kernelización consiste en trabajar con funciones polinómicas de las variables predictoras originales. Por ejemplo, en lugar de trabajar con las d variables $X_1, X_2, ..., X_d$, se puede operar con las $2d$ variables $X_1, X_1^2, X_2, X_2^2, ..., X_d, X_d^2$. La desventaja de

esta estrategia es que, si d es demasiado grande, los costos computacionales pueden aumentar drásticamente. Utilizar cualquiera de los kernels mencionados previamente solo requiere computar los $\binom{n}{2}$ pares $K(\boldsymbol{X}_i, \boldsymbol{X}_k)$, sin necesidad de trabajar explícitamente sobre el nuevo espacio de mayor dimensión.

En resumen, el proceso de kernelización radica en encontrar una aplicación

$$\phi : \mathcal{X} \to \mathcal{Z}$$

(donde $\mathcal{Z}$ posee una dimensión superior a $\mathcal{X}$) y una función K tal que

$$K(\boldsymbol{X}_i, \boldsymbol{X}_k) = \langle \phi(\boldsymbol{X}_i), \phi(\boldsymbol{X}_k) \rangle$$

En la práctica, solamente será necesario especificar un kernel y no se construirá directamente ϕ.

5.3. Estimación del hiperplano

Habiendo seleccionado un kernel, el hiperplano para SVM es estimado siguiendo un procedimiento muy similar al utilizado para SVC (ver Sección 5.1). La principal diferencia es que, en lugar de $\langle \boldsymbol{X}_i, \boldsymbol{X}_k \rangle$, se emplea $K(\boldsymbol{X}_i, \boldsymbol{X}_k)$. Los parámetros α_i se estiman maximizando la siguiente expresión:

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k K(\boldsymbol{X}_i, \boldsymbol{X}_k)$$

El hiperplano estimado resulta:

$$\hat{f}(\boldsymbol{x}) = \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\alpha}_i K(\boldsymbol{X}, \boldsymbol{X}_i) \, .$$

Una vez más, el valor asignado al parámetro C influye en las estimaciones de α_i y en la forma del límite de decisión resultante. Elegir valores grandes para el parámetro provocará

que los límites en el espacio original sean más ondulados y con tendencia al sobreajuste, mientras que valores pequeños originarán límites más suaves (Hastie et al., 2017). El hecho de que sólo los vectores de soporte afecten al clasificador está en línea con la afirmación anterior de que C controla la compensación sesgo-varianza. Cuando el parámetro de sintonización C es grande, entonces el margen es ancho, muchas observaciones violan el margen, por lo que hay muchos vectores de apoyo . En este caso, muchas observaciones están involucradas en la determinación del hiperplano. El panel superior izquierdo de la Figura 5.4 ilustra esta configuración: el clasificador tiene una varianza baja (ya que muchas observaciones son vectores de soporte pero tienen un sesgo potencialmente alto). Por el contrario, si C es pequeño, habrá menos vectores de soporte y, por lo tanto, el clasificador resultante tendrá un sesgo bajo, pero alta varianza. El panel inferior derecho de la Figura 5.4 ilustra esta configuración: con sólo ocho vectores de soporte. El hecho de que la regla de decisión del clasificador de vectores de soporte se base únicamente en un subconjunto potencialmente pequeño de las observaciones de entrenamiento (los vectores de soporte) significa que es bastante robusto al comportamiento de las observaciones que están muy lejos del hiperplano.

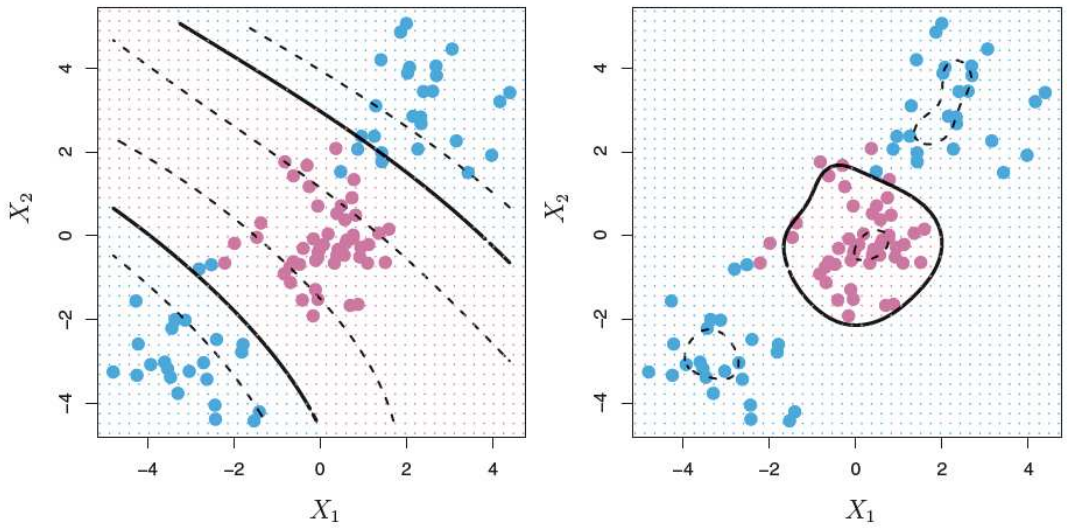

Figura 5.4: Ejemplo donde se ajustó un clasificador utilizando dos valores diferentes del parámetro de C. El mayor valor de C se utilizó en la parte izquierda del panel, y se utilizó un valor más pequeño en la parte derecha. Cuando C es grande, entonces hay una alta tolerancia para que las observaciones se encuentren en el lado equivocado del hiperplano, por lo que el margen será grande. A medida que C disminuye,la tolerancia para que las observaciones estén en el lado equivocado del hiperplano disminuyen,y el margen se estrecha. Fuente: James et al. (2013)

Capítulo 6

Experimentos Numéricos

6.1. Métricas de Evaluación

Para poder comparar el desempeño de distintos modelos, es necesario contar con alguna métrica que permita evaluar la calidad de las predicciones de cada uno de ellos (Jorgensen, 2019). Mientras que en problemas de regresión suelen calcularse valores tales como el error cuadrático medio, que proveen una medida numérica de la diferencia entre el valor estimado $\hat{y}_i$ y el verdadero valor y_i, en los problemas de clasificación los cálculos deben realizarse teniendo en cuenta que Y es una variable categórica y no cuantitativa.

La medida más intuitiva que puede utilizarse es la *tasa de error de clasificación*, que computa la proporción de observaciones incorrectamente clasificadas (James et al., 2013). Siendo Y_i los valores observados en el conjunto de datos y, por otro lado, $\hat{Y}_i$ los valores predichos por el modelo, se define *tasa de error en el conjunto de entrenamiento* como:

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{I}(Y_i \neq \hat{Y}_i),$$

donde $\mathcal{I}(Y_i \neq \hat{Y}_i)$ es una variable indicadora que tomará el valor 0 si $Y_i = \hat{Y}_i$ (la unidad fue correctamente clasificada) o bien el valor 1 si $Y_i \neq \hat{Y}_i$ (la unidad fue incorrectamente clasificada) y n corresponde al tamaño del conjunto de entrenamiento.

Sin embargo, a modo de evitar sobreajuste, en general, se desea evaluar el funcionamiento de la regla de clasificación en observaciones que no participaron en el proceso de ajuste de la misma, por lo que esta medida suele ser aplicada en un nuevo conjunto de datos denominado conjunto de prueba. Para ello, el universo de observaciones se divide en dos grandes conjuntos disjuntos: $\mathcal{T}$ y $\mathcal{V}$, conjuntos de entrenamiento (o training, de tamaño $n = 200$ en nuestro caso), y de validación (o prueba, de tamaño $m = 100$ en nuestro experimento), respectivamente.

Luego, la *tasa de error en el conjunto de prueba* la calcularemos como:

$$\frac{1}{\#\{\mathcal{V}\}} \sum_{i:i\in\mathcal{V}} \mathcal{I}(Y_i \neq \hat{Y}_i),$$

siendo $\#\{\mathcal{V}\}$ el cardinal del conjunto de validación $\mathcal{V}$. Consecuentemente, se considerará como el mejor modelo a aquel que minimice la tasa de error en el conjunto de prueba, que como dijimos en nuestro caso tiene tamaño 100.

Tras haber examinado los aspectos teóricos de los algoritmos bajo análisis, se intenta comprobar empíricamente el rendimiento de los mismos mediante un estudio con datos simulados. Se desea generar distintos conjuntos de datos para evaluar el desempeño de las cuatro técnicas bajo diversas condiciones experimentales. Para tal fin, se recurre al software R, utilizando el entorno de desarrollo RStudio. Se emplean las librerías *MASS*, *mvtnorm*, *class*, *naivebayes*, *e1071*, *Matrix* y *matrixStats*. Para cada una de las cuatro técnicas, se utilizaron los parámetros establecidos por defecto dentro de sus respectivas funciones en R. La única excepción fue SVM, donde se aplicó el kernel lineal en lugar del kernel radial indicado por defecto.

Comenzaremos por analizar el caso gaussiano y seguidamente, a modo de enriquecer el análisis del comportamiento de los algoritmos analizados, se evaluarán otras distribuciones, como la distribución *Cauchy*, *lognormal*(μ, σ^2) y datos definidos mediante una estructura de Corona.

El estudio numérico se basa en todos los casos en $Nrep = 1000$ replicaciones.

6.2. Escenario Normal o Gaussiano

Se define una nueva función que crea vectores $(X_1, X_2) \in \mathbb{R}^2$, generados aleatoriamente a partir de una distribución $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. La función posee los siguientes parámetros:

- **n**: Indica el tamaño de la muestra que se extraerá del total de datos generados.
- $\boldsymbol{\Delta} = (\Delta, \Delta)^T$: Define la separación entre las medias de las poblaciones.
- ρ: la correlación entre las variables X_1 y X_2.

En cada uno de los tres escenarios siguientes, se genera el conjunto de datos a partir de una distribución normal multivariada, cuya estructura depende de una constante Δ asociada a la distancia entre las medias de las distribuciones y una correlación ρ. Seguidamente, se extrae una muestra de tamaño n y se entrenan los cuatro métodos de clasificación (Regresión Logística, Análisis Discriminante Lineal, Análisis Discriminante Cuadrático y Support Vector Machine) utilizando los datos de la muestra como conjunto de entrenamiento. Estos modelos son posteriormente evaluados con datos fuera de la muestra, que se consideran los datos de validación o testeo y luego, se computan los errores de clasificación. Todo este proceso se repite un total de $Nrep = 1000$ veces y se calculan medidas resumen para las tasas de error de clasificación obtenidas en cada repetición. Se estableció una semilla para garantizar la reproducibilidad de los resultados obtenidos en cada simulación.

A continuación, se describirá en más detalle cada uno de los escenarios, los parámetros utilizados y las propiedades de los algoritmos que se busca evaluar en cada uno de ellos. Todos ellos fueron evaluados para valores de $n = \{50, 100, 150, 200\}$, $\rho = \{\ 0, 0.5, 0.7, 0.9\ \}$ y $\Delta = \{0.5, 1, 1.5, 2, 3\ \}$, pero, para hacer más amena la lectura de tablas, solo se incorporan al análisis los valores más significativos de cada algoritmo.

6.2.1. Escenario I

En el primer escenario, se busca evaluar el efecto que produce la variación del tamaño muestral sobre el desempeño de los modelos de clasificación cuando las poblaciones tienen

igual matriz de covarianza.

En este caso f_1 corresponde a la densidad de una $N_2(\mathbf{0}, \mathbf{\Sigma})$ y f_0 a la densidad de una $N_2(\mathbf{\Delta}, \mathbf{\Sigma})$, para $\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

Para la mayoría de los métodos estadísticos, se considera positivo trabajar con tamaños de muestra más elevados, pues de esta manera disminuye la variabilidad de los estimadores. Los resultados obtenidos a partir de muestras muy pequeñas son, a menudo, poco confiables. En este escenario, se prueban cuatro valores diferentes para el tamaño muestral: $n = \{50, 100, 150, 200\}$. Se evalúan también distintas magnitudes de los parámetros Δ y ρ para estudiar su interacción con n. Los valores evaluados son Δ = { 0.5, 1, 2 , 3 } y ρ = {0; 0.2, 0.5, 0.7 , 0.9 } Como se tienen clases con medias diferentes pero estructura común de covarianzas, se cumplen los supuestos para aplicar LDA.

Para ilustrar los datasets creados en la Figura 6.1, se presentan gráficos de dispersión de X_1 vs. X_2 para dos elecciones distintas de Δ donde en rojo se representan los puntos con $Y_i = 0$ y en azul aquellos con $Y_i = 1$, siendo $\rho = 0{,}5$. El panel de la izquierda corresponde a Δ =0.5 mientras que el de la derecha a $\Delta = 3$. Luego de calcular las tasas de error para cada escenario, estos resultados se resumen en diagramas de caja (*boxplot*).

Como se mencionó en la Sección 3.3, en general se considera que LDA produce mejores resultados que Regresión Logística cuando el tamaño muestral es reducido, ya que hay pocos parámetros que deberán ser estimados. Sin embargo, algunos estudios contradicen este concepto (Jorgensen, 2019). Tanto LDA como Regresión Logística suelen tener resultados bastante pobres si el número de variables predictoras es muy elevado al mismo tiempo que el tamaño muestral es pequeño, ya que esto lleva al sobreajuste del modelo.

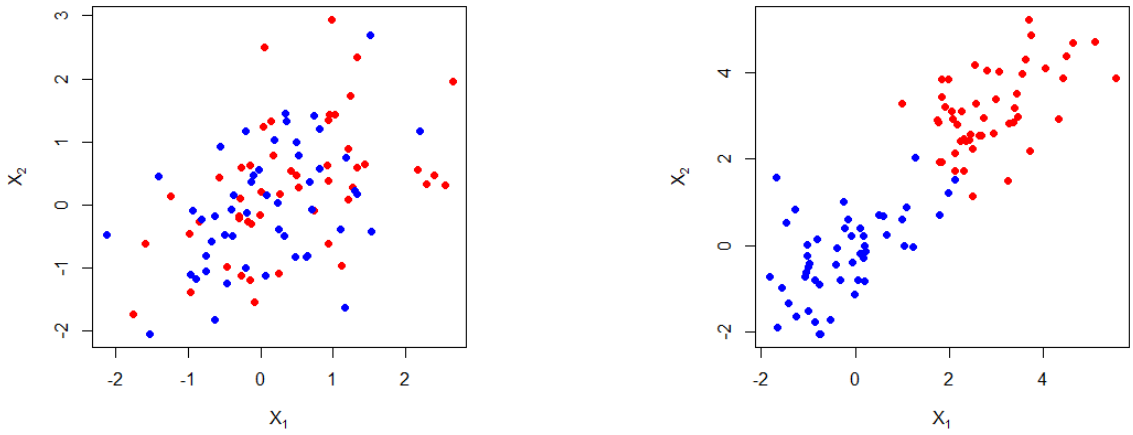

Figura 6.1: Escenario I: Diagramas de dispersión con $n = 50$ y correlación ρ =0.5. El panel de la izquierda corresponde a Δ =0.5, mientras que el de la derecha a Δ =3.

Como en James et al. (2013) observan, LDA también podría resultar una mejor alternativa a QDA ante tamaños muestrales pequeños. Esto se debe al compromiso sesgo-varianza, explicado en la Sección 4.2. Cuando se trabaja con pocas observaciones en el conjunto de entrenamiento, reducir la varianza resulta crucial; por ende, se recomienda aplicar LDA. Por el contrario, si el número de observaciones en el conjunto es grande, QDA puede ser una mejor alternativa.

Para ejemplificar el comportamiento, mostramos la Figura 6.2 que contiene los boxplots de los errores de clasificación en la muestra de validación obtenidos en las 1000 replicaciones para cada uno de los métodos para $n = 200$ en el panel de la derecha y $n = 50$ en el panel de la izquierda cuando $\Delta = 3$ y $\rho = 0$. A partir de esta figura, es evidente que el rendimiento de los métodos es muy similar para n pequeño, pero mejora notablemente el de LDA al considerar un tamaño $n = 200$. Un análisis más extenso se hará a partir de las tablas de resumen más adelante. Se puede observar que el entrenar con un valor de n menor, produce mayor cantidad de valores *atípicos* que al hacerlo con un n mayor.

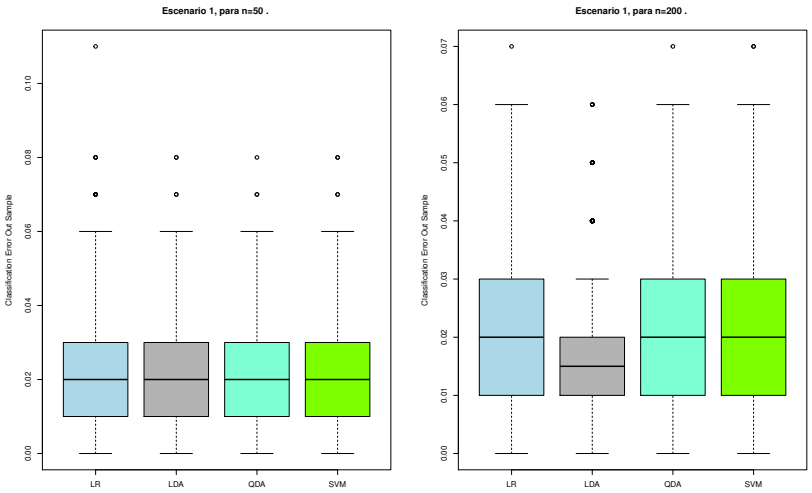

Figura 6.2: Escenario I: Boxplots de los errores de clasificación en la muestra de validación para $\Delta = 3$ y $\rho = 0$ del Escenario 1. Izquierda: los procedimientos se entrenan con $n = 50$. Derecha: los procedimientos se entrenan con n=200.

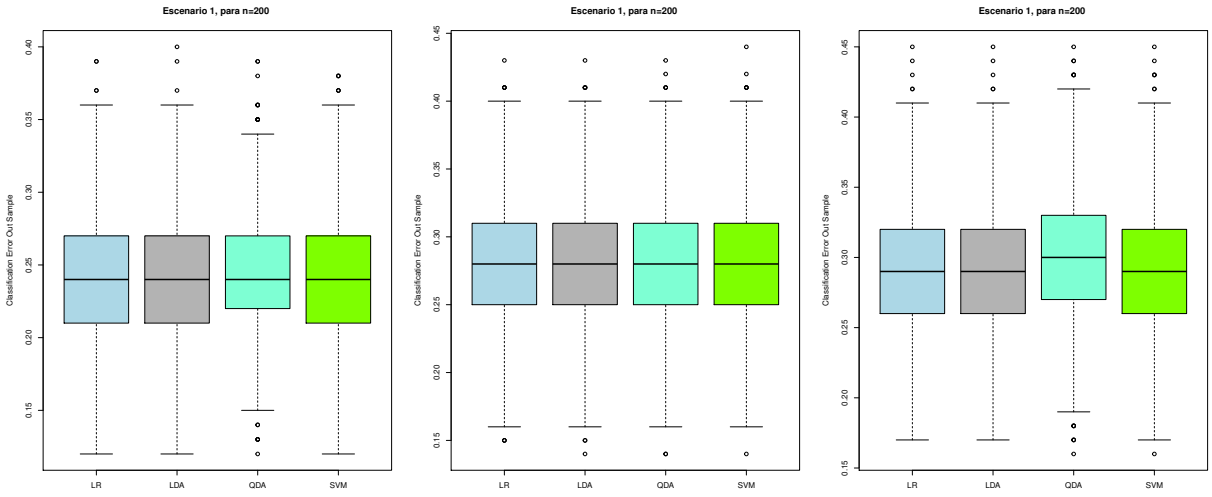

Figura 6.3: Escenario I: Boxplots de los errores de clasificación en la muestra de validación para $n = 200$, $\Delta = 1$, y variando la correlación $\rho = 0$, 0.5 y 0.7 respectivamente.

La Figura 6.3 ilustra con boxplots de los errores de clasificación en la muestra de valida-

ción para $n = 200$, cuando $\Delta = 1$, el comportamiento de los distintos métodos según varía la correlación $\rho = 0$, 0.5 y 0.7, mostrando un peor comportamiento relativo de QDA cuando la correlación es más alta.

6.2.2. Escenario II

En el segundo escenario, se consideran nuevamente dos poblaciones normales, pero con distinta matriz de covarianza. Por simplicidad, el tamaño de muestra queda fijo en el valor $n = 200$ (el cual se espera que producirá los mejores resultados, en base a lo mencionado para el escenario anterior) y se va variando el parámetro Δ, que indica la separación entre las medias de las poblaciones. Un mayor valor de Δ corresponde a una mayor separación entre las clases. Los valores que se evaluarán son $\Delta = \{0.5, 1, 1.5, 2, 3\}$. Cuanto más cerca estén las clases entre sí, más difícil suele ser la tarea de clasificación.

En este caso f_1 corresponde a la densidad de una $N_2(\mathbf{0}, \mathbf{\Sigma_1})$ y f_0 a la densidad de una $N_2(\mathbf{\Delta}, \mathbf{\Sigma}_0)$, donde $\mathbf{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, mientras que $\mathbf{\Sigma}_0 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

Una desventaja de la técnica de Regresión Logística es que usualmente no puede aplicarse cuando las clases son perfectamente separables en el conjunto de entrenamiento. La razón de esto es que el algoritmo de máxima verosimilitud se vuelve muy inestable, con las estimaciones de los parámetros tendiendo a infinito, e incluso puede no converger (Efron y Hastie, 2016). Por su parte, LDA no sufre de esta limitación, por lo que puede resultar una alternativa más apropiada cuando las clases están muy separadas (James et al., 2013). Aún así, Jorgensen (2019) comenta que hay evidencia de que la separación de las categorías no afecta significativamente a la diferencia en el desempeño para la clasificación entre LDA y Regresión Logística .

SVM es otra metodología que puede brindar un mejor desempeño cuando las clases están muy separadas, tal como ya se ha mencionado anteriormente. En general, se espera que Regresión Logística sea un método superior en los casos cuando Δ sea pequeño, mientras que SVM o LDA deberían superarlo en casos cuando Δ sea alto, es decir, cuando las clases están

muy separadas, como se puede observar en la Figura 6.4.

A modo de ilustración la Figura 6.4 presenta los boxplots de los errores de clasificación en la muestra de validación para $n = 200$, $\rho = 0.9$ y $\Delta = 0.5$,2 y 3, respectivamente. A medida que Δ decrece, es decir de izquierda a derecha, se puede observar lo distintivo del comportamiento de QDA respecto a los otros algoritmos, tendiendo a dar errores de clasificación en la muestra de validación más pequeños que los demás para Δ más chicos. En otras palabras, cuando la diferencia entre las medias es grande el comportamiento de los cuatro procedimientos es parecido y la ganancia de QDA es notable para valores pequeños de la diferencia.

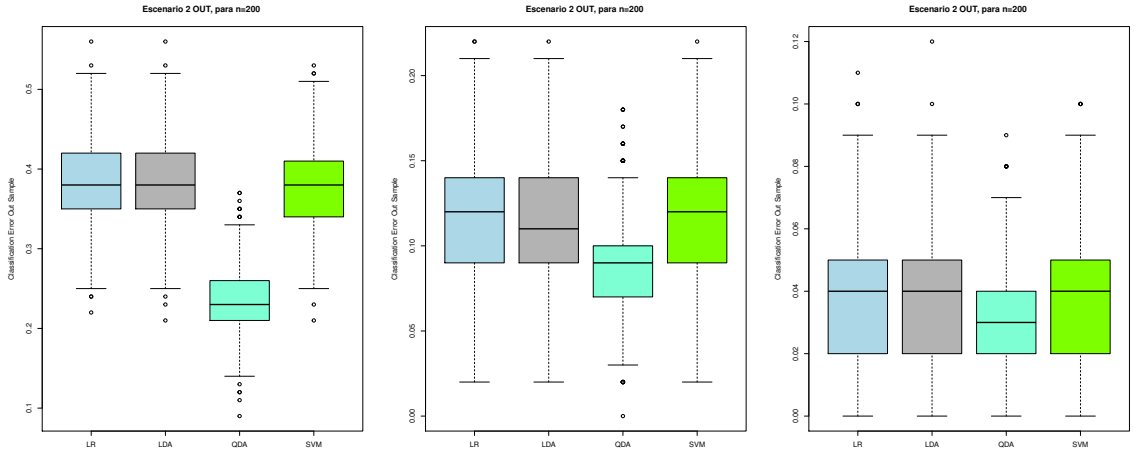

Figura 6.4: Escenario II: Boxplots de los errores de clasificación en la muestra de validación para $n = 200$, $\rho = 0{,}9$ y $\Delta = 0{,}5$, 2 y 3, respectivamente.

Adicionalmente, se experimenta con el uso de distintas matrices de covarianzas para ambas poblaciones. Para una de las poblaciones, la correlación permanece fija (el valor ρ es constante) mientras que para la segunda población este parámetro varía entre $\rho = \{0, 0{,}5, 0{,}7, 0{,}9\}$. De esta manera, no se cumple uno de los supuestos de LDA, pues las matrices resultarán desiguales entre las dos clases.

Como se mencionó en secciones anteriores, LDA tiene un buen desempeño en tareas de clasificación siempre que se cumplan ciertos supuestos, es decir cuando la matriz de covarianza sea común. Es de esperar que de no cumplirse los supuestos, la técnica presentará

peores resultados (James et al., 2013). En particular, QDA podría tener un desempeño destacable si la separación entre las clases tiende a ser cuadrática en lugar de lineal al aumentar ρ (Jorgensen, 2019).

6.2.3. Escenario III

Al desarrollar este escenario, definimos en este caso f_1 que corresponde a la densidad de una $N_2(\mathbf{0}, \mathbf{\Sigma_1})$ y f_0 a la densidad de una $N_2(\mathbf{\Delta}, \mathbf{\Sigma}_0)$, donde $\mathbf{\Sigma}_1 = \begin{pmatrix} 0{,}5 & \rho.(0{,}5) \\ \rho(0{,}5) & 0{,}5 \end{pmatrix}$, mientras que $\mathbf{\Sigma}_0 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

En el tercer escenario, el interés está puesto sobre el parámetro ρ, que es la correlación entre las variables X_1 y X_2. Cuando ρ toma valores bajos, la relación entre ambas variables es débil. Gráficamente, se observa un patrón de "nube de puntos" al plasmar los datos en un gráfico de dispersión (donde el eje de las abscisas corresponde a X_1 y el eje de las ordenadas a X_2). Si ρ es alto, esto indica una relación fuerte entre X_1 y X_2. Esta relación se visualiza en los gráficos de dispersión, donde las observaciones formarán un patrón lineal, como se puede visualizar en la Figura 6.5.

La correlación puede ser positiva (si a mayores valores de una variable corresponden también mayores valores de la segunda variable) o negativa (si al aumentar una variable, la otra disminuye). Sin embargo, estudios por simulación realizados previamente muestran que el signo de la correlación no afecta notablemente a la clasificación (James et al., 2013).

Cuando dos variables predictoras presentan una correlación excesivamente elevada, se da el fenómeno conocido como *multicolinealidad*. Esto puede perjudicar a los modelos, pues el error estándar de los estimadores aumenta, volviéndolos más imprecisos y reduciendo la potencia de las pruebas de significación. En particular, las técnicas de regresión, tanto lineal como logística, son vulnerables a la multicolinealidad (James et al., 2013).

Los parámetros que se prueban son $\rho = \{0, 0{,}5, 0{,}7, 0{,}9\}$, $n = 200$ y $\Delta = \{1, 2, 3\}$. A diferencia del escenario anterior, esta vez las matrices varían en forma proporcional para

ambas poblaciones. Por lo tanto, así como en el escenario I, se cumplen los supuestos de QDA.

Debido a la existencia de matrices comunes, es posible que LDA tenga un desempeño superior a QDA en este escenario, como se observa en la Figura 6.6. Como se explicó en la Sección 4.2 y en la 4.3, esta última técnica puede llegar a presentar una varianza muy elevada sin compensarlo con una reducción relevante en el sesgo.

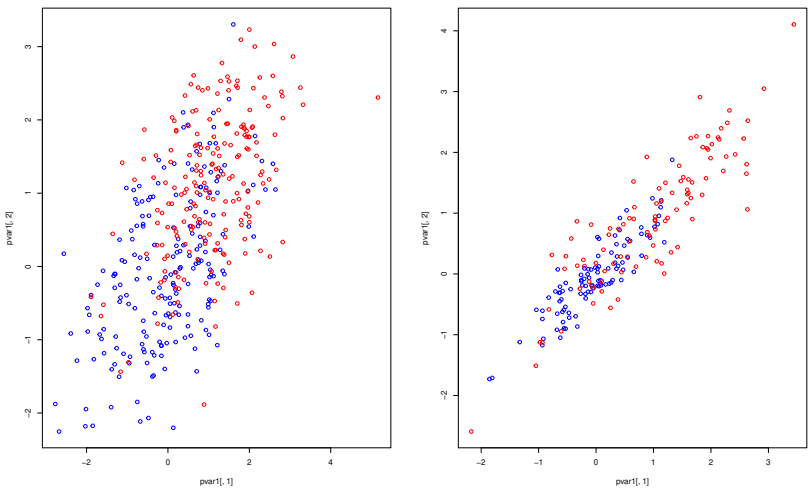

Figura 6.5: Escenario III: Gráfico de Dispersión para $n = 200$, $\Delta = 1$ y $\rho = 0.5$ y 0.9.

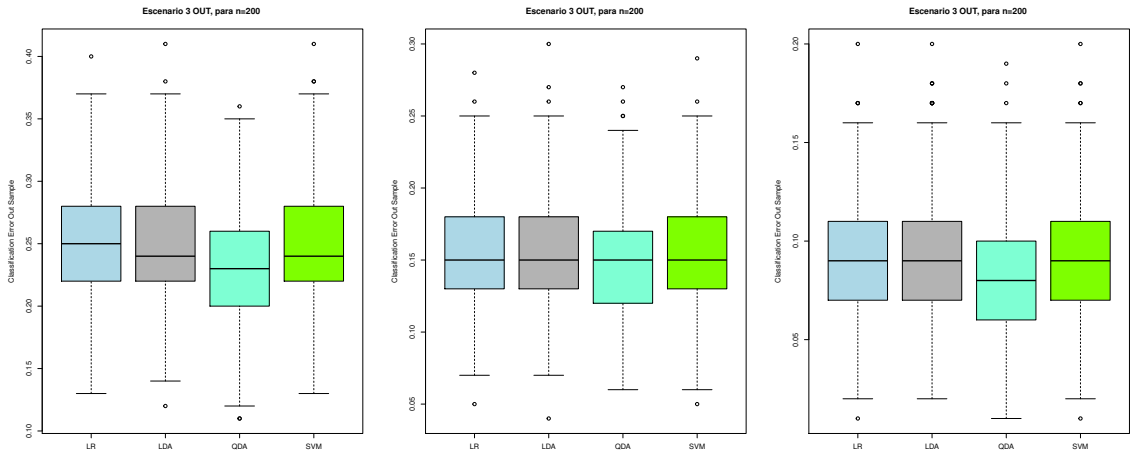

Figura 6.6: Muestras realizadas en laboratorio para los valores de $\Delta = 1,\ 1{,}5\ y\ 2$ respectivamente y $\rho = 0{,}5$ del escenario 3, donde se observa que LDA posee un comportamiento superior a QDA con valores de Δ mayores.

6.3. Resultados

En esta sección se analizan los resultados obtenidos en cada uno de los escenarios. Se presentan medidas resumen correspondientes a las tasas de error, tanto en el conjunto de entrenamiento como en el conjunto de prueba, para cada combinación de técnicas empleadas y parámetros seleccionados.

6.3.1. Escenario I

El Cuadro 6.1 indica las tasas de error promedio y sus desvíos estándar para cada técnica según valores de los parámetros n, Δ y ρ, para los datos pertenecientes al conjunto de entrenamiento. Mientras que el Cuadro 6.2 indica las tasas de error promedio y sus desvíos estándar para cada técnica según valores de los parámetros n, Δ y ρ, para los datos pertenecientes al conjunto de prueba.

En general, en todos los escenarios que mostraremos, observaremos que los errores repor-

tados tienden a ser menores en el cuadro correspondiente a los conjuntos de entrenamiento que los informados para los conjuntos de prueba, lo que resulta natural ya que los ajustes fueron realizados a partir de las muestras de entrenamiento. Por este motivo, en general, analizaremos las tablas para los conjuntos de prueba que reflejan los errores en futuras predicciones.

Conjunto de Entrenamiento										
n	Δ	ρ	RL		LDA		QDA		SVM	
			Media	D.E.	Media	D.E.	Media	D.E.	Media	D.E.
50	0,5	0	0,356	0,047	0,356	0,047	0,346	0,046	0,354	0,047
50	0,5	0,7	0,384	0,048	0,384	0,048	0,373	0,045	0,382	0,047
50	2	0	0,073	0,027	0,074	0,026	0,073	0,026	0,073	0,026
50	2	0,7	0,136	0,035	0,137	0,034	0,136	0,035	0,136	0,034
100	0,5	0	0,358	0,035	0,358	0,035	0,354	0,035	0,358	0,035
100	0,5	0,7	0,390	0,034	0,390	0,034	0,384	0,033	0,389	0,033
100	2	0	0,076	0,019	0,077	0,019	0,077	0,019	0,076	0,019
100	2	0,7	0,137	0,023	0,138	0,023	0,137	0,023	0,138	0,023
150	0,5	0	0,361	0,027	0,361	0,027	0,358	0,027	0,360	0,028
150	0,5	0,7	0,390	0,028	0,390	0,028	0,386	0,028	0,389	0,028
150	2	0	0,076	0,015	0,077	0,015	0,077	0,015	0,076	0,015
150	2	0,7	0,138	0,019	0,138	0,019	0,138	0,020	0,138	0,019
200	0,5	0	0,359	0,023	0,359	0,023	0,357	0,023	0,359	0,023
200	0,5	0,7	0,391	0,024	0,391	0,024	0,388	0,024	0,390	0,024
200	2	0	0,076	0,013	0,077	0,013	0,077	0,013	0,076	0,013
200	2	0,7	0,137	0,017	0,137	0,017	0,137	0,017	0,137	0,017

Cuadro 6.1: Medias y desvíos estándar de la tasa de error en el conjunto de entrenamiento del Escenario I según valores de *n*.

Conjunto de Prueba										
n	Δ	ρ	RL		LDA		QDA		SVM	
			Media	D.E.	Media	D.E.	Media	D.E.	Media	D.E.
50	0,5	0	0,372	0,048	0,372	0,048	0,382	0,050	0,373	0,049
50	0,5	0,7	0,402	0,051	0,402	0,051	0,415	0,053	0,404	0,050
50	2	0	0,084	0,028	0,082	0,027	0,082	0,027	0,084	0,027
50	2	0,7	0,145	0,035	0,144	0,035	0,145	0,035	0,144	0,035
100	0,5	0	0,363	0,048	0,362	0,048	0,366	0,048	0,363	0,048
100	0,5	0,7	0,397	0,050	0,398	0,049	0,405	0,049	0,398	0,049
100	2	0	0,077	0,027	0,076	0,027	0,077	0,027	0,077	0,027
100	2	0,7	0,140	0,033	0,140	0,034	0,140	0,034	0,140	0,033
150	0,5	0	0,364	0,049	0,364	0,049	0,365	0,049	0,364	0,049
150	0,5	0,7	0,397	0,048	0,396	0,048	0,402	0,048	0,397	0,049
150	2	0	0,080	0,025	0,079	0,025	0,079	0,025	0,080	0,025
150	2	0,7	0,140	0,034	0,140	0,034	0,140	0,034	0,140	0,034
200	0,5	0	0,395	0,049	0,395	0,049	0,398	0,049	0,396	0,049
200	0,5	0,7	0,364	0,048	0,364	0,048	0,366	0,048	0,364	0,048
200	2	0	0,081	0,027	0,080	0,027	0,080	0,027	0,081	0,027
200	2	0.7	0,140	0,035	0,140	0,035	0,140	0,035	0,140	0,035

Cuadro 6.2: Medias y desvíos estándar de la tasa de error en el conjunto de prueba para el Escenario I según valores de n.

Nos concentraremos en el Cuadro 6.2, ya que, como explicamos más arriba, es el de mayor interés. Al aumentar el tamaño muestral n, fijando la separación entre poblaciones Δ y la correlación ρ, se aprecia que el rango de las tasas de error tiende a disminuir notablemente, así como también la variabilidad.

La mejora en la precisión de la clasificación parece ser más importante cuando la se-

paración entre las clases es baja. Por ejemplo, la disminución en la variabilidad parece ser mayor cuando se pasa del caso $n = 50, \Delta = 0{,}5, \rho = 0{,}7$ al caso $n = 200, \Delta = 0{,}5, \rho = 0{,}7$, que cuando se pasa de $n = 50, \Delta = 2, \rho = 0{,}7$ a $n = 200, \Delta = 2, \rho = 0{,}7$. Este mismo fenómeno se puede observar para la correlación entre las variables. Pasar de la situación $n = 50, \Delta = 0,5, \rho = 0{,}7$ a la situación $n = 200$, Δ =0.5, $\rho = 0.7$ parece producir una mejoría más marcada que pasar de la situación $n = 50$, Δ =0.5, $\rho = 0$ a la situación $n = 200$, Δ =0.5, ρ =0.

Como es de esperar, aumentar la separación de clases produce una disminución destacable en las tasas de error. Para cada caso evaluado, el escenario donde $\Delta = 2$ produjo mejores resultados que su contraparte con $\Delta = 0,5$. Los efectos de este parámetro serán analizados más profundamente en el siguiente escenario.

Las condiciones bajo las cuales se observó la menor tasa de error en promedio, de los cuatro métodos para el conjunto de prueba fueron cundo $n = 100$, $\Delta = 2$ y $\rho = 0$, siendo este promedio igual a 0.077. Por el contrario, los parámetros bajo los cuales el error fue mayor en promedio fueron $n = 50$, $\Delta = 0{,}5$ y $\rho = 0{,}7$ siendo 0.402 el valor obtenido.

En la mayoría de los casos, se aprecia que las cuatro técnicas presentaron un desempeño muy similar. En algunas simulaciones, LDA presentó resultados ligeramente superiores a las demás metodologías, seguido de regresión logística. En contraste, QDA resultó más deficiente que las otras técnicas en diversos casos, particularmente cuando Δ =0.5.

No se apreció que alguna de las técnicas presente consistentemente una mayor variabilidad en las tasas de error comparada a las demás.

6.3.2. Escenario II

El Cuadro 6.3 indica las tasas de error promedio y sus desvíos estándar para cada técnica según valores de los parámetros Δ y ρ, para el conjunto de datos de entrenamiento, mientras que el Cuadro 6.4 lo hace para el conjunto de datos de prueba, cabe recordar, que el tamaño de $n = 200$ está fijo.

Nos enfocaremos en el Cuadro 6.4, ya que es el de mayor interés dado que reporta el error fuera de la muestra de entrenamiento. Así como se había observado en el Escenario I, se puede ver que, naturalmente, al aumentar la separación entre las clases, el error de clasificación disminuye. En las simulaciones donde Δ =0.5 la mayoría de las repeticiones presentaron un porcentaje de error de clasificación superior al 20 % con cualquier técnica. Para $\Delta = 3$, este error rara vez superó el 10 %.

En los escenarios donde $\Delta = 0.5$ o $\Delta = 1$, se aprecia que las tasas de error muestran una variabilidad muy elevada. A medida que las poblaciones se van separando, la variabilidad disminuye.

Para valores fijos del parámetro Δ, los resultados parecen en general empeorar a medida que aumenta el parámetro ρ. Por ejemplo, para $\Delta = 3$ y $\rho = 0$, las medianas de los errores de clasificación se ubican en torno al valor 0.02. Cuando ρ aumenta a 0.9, estas medianas ahora se encuentran alrededor del valor 0.04. La variabilidad no parece ser afectada por los valores de este parámetro.

Las cuatro metodologías bajo análisis presentaron desempeños muy parecidos, especialmente Regresión Logística y SVM. LDA presentó los mejores resultados en la simulación donde $\Delta = 3$ y $\rho = 0$, es decir, el escenario que mejor se apega a los supuestos de esta técnica. Aún en los casos donde los supuestos son incumplidos de forma notoria, el desempeño de LDA no pareció empeorar tan drásticamente y fue muy similar al obtenido con regresión logística o SVM. Finalmente, QDA presentó resultados ligeramente inferiores cuando ρ tomó valores pequeños, pero al aumentar este parámetro fue la técnica que menor tasa de error presentó. De hecho, cuando ρ =0,9, QDA superó de manera muy marcada a las otras tres metodologías y esto corresponde al caso en que la matriz $\mathbf{\Sigma}_0$ está más lejos de la matriz identidad.

Al contrastar los resultados obtenidos en el conjunto de entrenamiento y en el conjunto de prueba, se puede notar que en varios casos QDA presenta tasa de error menores al resto de técnicas al momento de ajustar el modelo, pero luego esta diferencia disminuye o desaparece al aplicarlo a nuevos datos. Esto es un indicador de que QDA parece ser muy vulnerable al

sobreajuste, a comparación de los demás métodos. Dicho efecto es menos notorio a medida que aumenta ρ.

Conjunto de Entrenamiento									
Δ	ρ	RL		LDA		QDA		SVM	
		Media	D.E.	Media	D.E.	Media	D.E.	Media	D.E.
0,5	0	0,359	0,023	0,359	0,023	0,357	0,023	0,359	0,023
0,5	0,7	0,376	0,024	0,376	0,024	0,319	0,022	0,372	0,023
1	0,5	0,261	0,021	0,260	0,021	0,248	0,021	0,260	0,021
1	0,7	0,265	0,022	0,265	0,022	0,235	0,020	0,263	0,021
1,5	0	0,142	0,017	0,142	0,017	0,142	0,017	0,142	0,017
1,5	0,9	0,182	0,019	0,181	0,018	0,134	0,016	0,181	0,018
2	0	0,076	0,013	0,077	0,013	0,077	0,013	0,076	0,013
3	0	0,015	0,006	0,016	0,006	0,016	0,006	0,015	0,006
3	0,5	0,026	0,008	0,028	0,008	0,002	0,008	0,026	0,008
3	0,7	0,030	0,008	0,033	0,008	0,028	0,008	0,030	0,008
3	0,9	0,034	0,009	0,038	0,009	0,026	0,008	0,034	0,009

Cuadro 6.3: Medias y desvíos estándar de la tasa de error en el conjunto de entrenamiento el Escenario II según valores de Δ y ρ.

Conjunto de Prueba									
Δ	ρ	RL		LDA		QDA		SVM	
		Media	D.E.	Media	D.E.	Media	D.E.	Media	D.E.
0,5	0	0,364	0,048	0,364	0,048	0,366	0,048	0,364	0,048
0,5	0,7	0,380	0,048	0,380	0,048	0,328	0,047	0,377	0,048
1	0,5	0,261	0,045	0,261	0,045	0,248	0,044	0,260	0,045
1	0,7	0,267	0,044	0,267	0,044	0,238	0,043	0,266	0,044
1,5	0	0,145	0,035	0,145	0,035	0,145	0,035	0,145	0,035
1,5	0,9	0,185	0,040	0,183	0,040	0,136	0,034	0,184	0,040
2	0	0,081	0,027	0,080	0,027	0,080	0,027	0,081	0,027
3	0	0,017	0,013	0,016	0,012	0,017	0,013	0,017	0,013
3	0,5	0,029	0,016	0,029	0,016	0,027	0,016	0,029	0,016
3	0,7	0,033	0,017	0,034	0,018	0,030	0,017	0,033	0,018
3	0,9	0,037	0,018	0,039	0,018	0,028	0,016	0,037	0,018

Cuadro 6.4: Medias y desvíos estándar de la tasa de error en el conjunto de prueba para el Escenario II según valores de Δ y ρ.

6.3.3. Escenario III

El Cuadro 6.5 indica las tasas de error promedio y sus desvíos estándar para cada técnica según valores de los parámetros ρ y Δ, para el conjunto T, mientras que el Cuadro 6.6 lo hace para el conjunto V. Para un mismo Δ, la tasa de error media parece aumentar a medida que la correlación aumenta. Sin embargo, el efecto no parece tan importante como el de la separación entre clases examinado en el escenario anterior. Este mismo fenómeno se había observado en el escenario I. El caso con peores resultados en general fue el que presentó $\Delta = 1$ y $\rho = 0{,}7$.

Conjunto de Entrenamiento									
Δ	ρ	RL		LDA		QDA		SVM	
		Media	D.E.	Media	D.E.	Media	D.E.	Media	D.E.
1	0	0,200	0,029	0,199	0,029	0,188	0,028	0,198	0,029
1	0,7	0,258	0,031	0,257	0,031	0,239	0,030	0,255	0,031
2	0	0,047	0,015	0,050	0,015	0,046	0,015	0,047	0,015
2	0,7	0,099	0,021	0,101	0,021	0,095	0,021	0,099	0,021
3	0	0,003	0,005	0,008	0,006	0,005	0,005	0,005	0,005
3	0,5	0,018	0,010	0,023	0,010	0,019	0,009	0,019	0,009
3	0,7	0,026	0,012	0,030	0,011	0,026	0,011	0,026	0,011
3	0,9	0,033	0,013	0,037	0,012	0,033	0,012	0,034	0,012

Cuadro 6.5: Medias y desvíos estándar de la tasa de error en el conjunto de entrenamiento para el Escenario III según valores de ρ y Δ.

Conjunto de Prueba									
Δ	ρ	RL		LDA		QDA		SVM	
		Media	D.E.	Media	D.E.	Media	D.E.	Media	D.E.
1	0	0,202	0,040	0,201	0,040	0,191	0,038	0,201	0,040
1	0,7	0,261	0,045	0,260	0,045	0,246	0,043	0,260	0,045
2	0	0,049	0,022	0,051	0,022	0,047	0,021	0,049	0,022
2	0,7	0,101	0,030	0,101	0,029	0,097	0,028	0,101	0,030
3	0	0,009	0,010	0,009	0,009	0,006	0,008	0,007	0,008
3	0,5	0,023	0,015	0,025	0,016	0,020	0,014	0,021	0,015
3	0,7	0,030	0,017	0,031	0,018	0,027	0,016	0,028	0,017
3	0,9	0,037	0,019	0,038	0,020	0,034	0,018	0,035	0,018

Cuadro 6.6: Medias y desvíos estándar de la tasa de error en el conjunto de prueba para el Escenario III según valores de ρ y Δ.

La variabilidad también parece aumentar a mayores valores de ρ, así como se había podido observar en el escenario I para $\rho = 0$ y $\rho = 0{,}7$. La situación más variable fue una vez más el caso donde $\Delta = 1$ y $\rho = 0{,}7$.

Como ya se vio en el escenario II, una mayor separación entre poblaciones resulta en un mejor desempeño para todos los métodos. La simulación que produjo los mejores resultados fue aquella en la que $\Delta = 3$ y $\rho = 0$

Las cuatro técnicas presentan nuevamente un desempeño muy similar en las distintas simulaciones. QDA pareció lograr los mejores resultados en general, seguido de Regresión Logística y SVM. LDA fue el método cuyo desempeño varió más para distintos valores de los parámetros. Cuando ρ es bajo, LDA presenta en general baja variabilidad y tasa de error, pero al aumentar estos valores la técnica parece volverse más inestable y producir más errores de clasificación.

Tanto SVM como regresión logística produjeron resultados muy similares, aunque SVM

parece tener una variabilidad ligeramente inferior. Esta misma situación ocurrió en los escenarios anteriores.

A diferencia de lo ocurrido en el escenario II, no se aprecia que el sobreajuste afecte más a alguna técnica en particular. Tampoco hay una relación clara entre los valores de ρ y las diferencias entre el error en el conjunto de entrenamiento y el conjunto de prueba.

6.3.4. Distribuciones No Normales

Luego del análisis realizado para los algoritmos Regresión Lineal, LDA, QDA y SVM, bajo condiciones normales de los valores del conjunto de entrenamiento y testeo, es lógico preguntarse qué sucedería si estas condiciones de normalidad no se cumplen.

Se abre un abanico infinito en materia de estudio para abordar este cuestionamiento tan interesante. En esta vasta diversidad, analizaremos tres tipos de distribuciones diferentes, dentro del universo de las distribuciones *No Normales,* para ver el comportamiento sobre estos algoritmos.

A modo de poder conocer el comportamiento en este contexto no normal, realizaremos unas pruebas de laboratorio, donde utilizaremos los algoritmos para un valor de $n = 200$, donde se harán $Nrep = 1000$ para cada uno.

Las variables elegidas para investigar son: ***Cauchy***, ***LogNormal***, donde se evaluarán Δ y ρ, dentro de los valores ya utilizados en el estudio de los escenarios del I al III y datos aleatorios provenientes de anillos concéntricos, también llamados ***Coronas*** donde se evaluará sobre el valor ρ, que en este caso será el radio de las mencionadas coronas.

Escenario IV: Distribución Cauchy

Para realizar las pruebas, elegimos la distribución de Cauchy que tiene algunas características distintivas:

- La distribución de Cauchy tiene colas más pesadas en comparación con la distribución

Normal. Esto significa que hay una mayor probabilidad de observar *outliers*.

- La distribución de Cauchy no tiene ni primer ni segundo momento. Esta característica es importante porque muchos métodos estadísticos, a menudo suponen que los datos tienen una media y varianza finitas.

Definimos las densidades f_1 y f_0 para el desarrollo de este escenario, que corresponden a la distribución conjunta de dos variables independientes, cada una con distribución univariada Cauchy, f_1 está centrada en $(0, 0)$, mientras que f_0 está centrada en $\mathbf{\Delta} = (\Delta, \Delta)$.

Como anteriormente se mencionó, tanto para la distribución Cauchy como para otras dos pertenecientes a las *"No Normales"*, se realizaron pruebas variando Δ y ρ .
En el Cuadro 6.7, se ilustran algunos de los valores obtenidos en las simulaciones de laboratorio más relevantes. No colocamos todos, con el fin de hacer más ameno el análisis.
Aquí podemos observar que los valores no se ven alterados por la modificación del ρ, sino por el valor que varíe de Δ, es decir que la separabilidad de las clases no sería un factor relevante a la hora de mejorar el comportamiento de los algoritmos bajo esta variable.

Conjunto de Entrenamiento									
Δ	ρ	RL		LDA		QDA		SVM	
		Media	D.E.	Media	D.E.	Media	D.E.	Media	D.E.
1	0	0.429	0.008	0.441	0.007	0.493	0.0002	0.473	0.002
1	0,7	0.429	0.008	0.441	0.007	0.493	0.0002	0.473	0.002
2	0	0.314	0.013	0.346	0.012	0.489	0.001	0.389	0.011
2	0,7	0.314	0.013	0.346	0.012	0.489	0.001	0.389	0.011
3	0	0.227	0.012	0.262	0.015	0.482	0.002	0.280	0.017
3	0,7	0.227	0.012	0.262	0.015	0.482	0.002	0.280	0.017

Cuadro 6.7: Medias y desvíos estándar de la tasa de error en el conjunto de Entrenamiento para el Escenario No Normal-Cauchy según valores de ρ y Δ.

Al analizar los datos en conjunto, observamos que el comportamiento de los cuatro al-

goritmos poseen un punto en común, a medida que aumenta el valor de Δ, el error medio disminuye, pero aumenta la variabilidad. Como ya hemos visto en la Sección 2, el algoritmo de Regresión Logística modela la probabilidad condicional de que una observación pertenezca a una clase en particular.

Cuando se aplica el algoritmo de Regresión Logística a datos que originados por una variable aleatoria Cauchy, se pueden observar varios comportamientos:

- Como la distribución de Cauchy tiene colas pesadas, la Regresión Logística puede volverse muy sensible a los valores extremos, esto puede ocasionar la inestabilidad de los coeficientes del modelo, donde, con pequeños cambios en las observaciones, pueden resultar en cambios muy significativos en las predicciones del mismo.

- Debido a la falta de media y varianza definidas para la variable Cauchy, en la Regresión Logística observamos que hay dificultades para encontrar una separación correcta entre las clases, lo que deriva en un ajuste deficiente. Como la Regresión Logística asume que la relación entre los predictores y la log-odds de la variable dependiente es lineal, si esto no sucede y los datos poseen una estructura más compleja, como la que podría presentarse con una distribución Cauchy, la frontera de decisión resultante puede ser inapropiada y no expresar bien la complejidad de los datos.

Podemos observar en el Cuadro 6.8, como los datos no son sensibles al cambio de valor del parámetro ρ, pero si lo son al valor de Δ, como se visualiza también en el conjunto de Entrenamiento, descripto en el Cuadro 6.7. A causa de la inestabilidad del modelo, las probabilidades predichas pueden no ser confiables. Esto conlleva a errores de clasificación más altos, especialmente en las clases que están más cerca de los valores extremos.

Conjunto de Prueba									
Δ	ρ	RL		LDA		QDA		SVM	
		Media	D.E.	Media	D.E.	Media	D.E.	Media	D.E.
1	0	0.437	0.009	0.448	0.008	0.500	0.001	0.482	0.002
1	0,7	0.437	0.009	0.448	0.008	0.500	0.001	0.482	0.002
2	0	0.317	0.014	0.350	0.014	0.495	0.001	0.398	0.011
2	0,7	0.317	0.014	0.350	0.014	0.495	0.001	0.398	0.011
3	0	0.232	0.014	0.266	0.016	0.488	0.002	0.286	0.018
3	0,7	0.232	0.014	0.266	0.016	0.488	0.002	0.286	0.018

Cuadro 6.8: Medias y desvíos estándar de la tasa de error en el conjunto de Prueba para el Escenario No Normal-Cauchy según valores de ρ y Δ.

Esto se ve reflejado en la tasa de error, que es más alta en comparación con modelos aplicados a datos que siguen una distribución normal, ya que la Regresión Logística no puede manejar adecuadamente la heterogeneidad de las observaciones, como muestra la Figura 6.7

Como ya se ha mencionado, LDA se basa en varias suposiciones:

- Asumimos que las características dentro de cada clase siguen una distribución normal multivariada.
- Las diferentes clases comparten la misma matriz de covarianza.

Cuando los datos observados siguen una distribución Cauchy, los supuestos asumidos por LDA se violan severamente, ya que la distribución Cauchy no es una distribución Normal (sus colas pesadas y la falta de momentos definidos rompen la premisa de normalidad), así como también la presencia de valores extremos puede alterar las estimaciones de la matriz de covarianza, afectando la igualdad asumida entre clases.

La capacidad que tiene LDA para clasificar correctamente nuevas observaciones depende

de la precisión de las estimaciones de los parámetros del modelo. Con datos generados por una distribución de Cauchy observamos:

- La presencia de valores extremos incrementa la tasa de error de clasificación, ya que el modelo puede ajustar sus parámetros basándose en datos atípicos que no representan la verdadera distribución subyacente (sobreajuste).
- El modelo no puede generalizar bien a nuevos datos, especialmente si estos también contienen valores extremos.

El Análisis Discriminante Lineal, al estar basado en supuestos de normalidad y covarianzas homogéneas, se ve seriamente afectado cuando los datos son generados por una distribución de Cauchy, como podemos observar en la Figura 6.7 en la que se muestran resultados correspondientes a una selección de valores de Δ y $\rho = 1$.

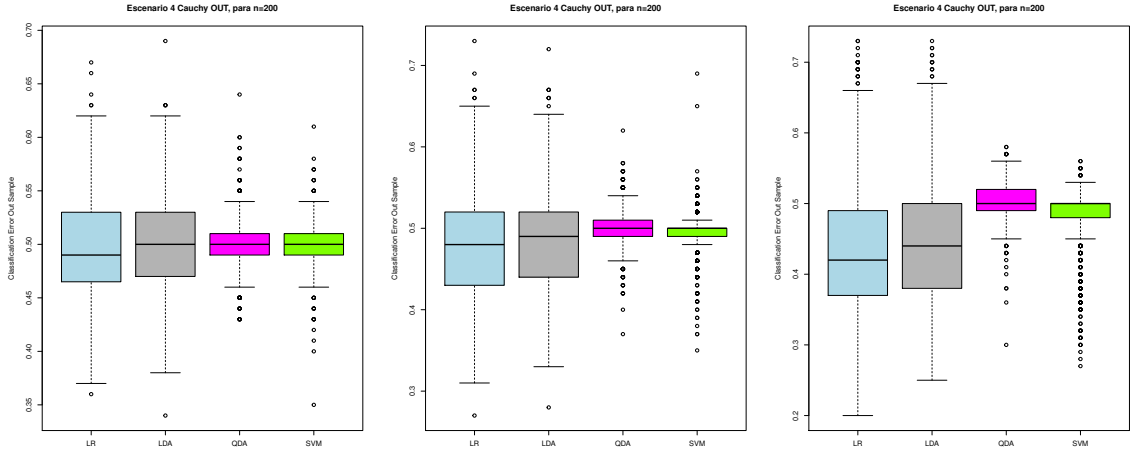

Figura 6.7: Muestras realizadas en laboratorio para los valores de $\Delta = 0{,}1,\ 0{,}5\ y\ 3$ respectivamente y $\rho = 1$ del Escenario IV con v.a. Cauchy. Se puede observar que en todos los casos, existe una gran presencia de outliers, y la variabilidad de los modelos.

Como ya hemos desarrollado en la Sección 4, QDA se basa en varias suposiciones clave para funcionar correctamente:

- Cada clase se asume que sigue una distribución normal multivariada.

- Cada clase puede tener su propia matriz de covarianza, permitiendo fronteras de decisión cuadráticas.

Aunque QDA también asume normalidad dentro de cada clase, al permitir matrices de covarianza diferentes por clase, puede ofrecer una mayor flexibilidad con respecto a LDA, sin embargo, no resuelve completamente el problema de los outliers.

QDA estima las medias y las matrices de covarianza para cada clase. La presencia de valores extremos puede distorsionar significativamente estas estimaciones, ya que QDA utiliza la media muestral y la matriz de covarianza muestral, que son sensibles a outliers.

Rendimiento de clasificación reducido, debido a las estimaciones inestables de los parámetros y las fronteras de decisión distorsionadas, el rendimiento de QDA en términos de precisión de clasificación puede verse significativamente afectado:

- La presencia de outliers lleva a una mayor tasa de error, ya que el modelo se ajusta a valores atípicos que no representan adecuadamente la distribución de las clases.

- El modelo puede no generalizar bien a nuevos datos, especialmente si estos también contienen valores extremos, lo que reduce la utilidad práctica de QDA en este contexto.

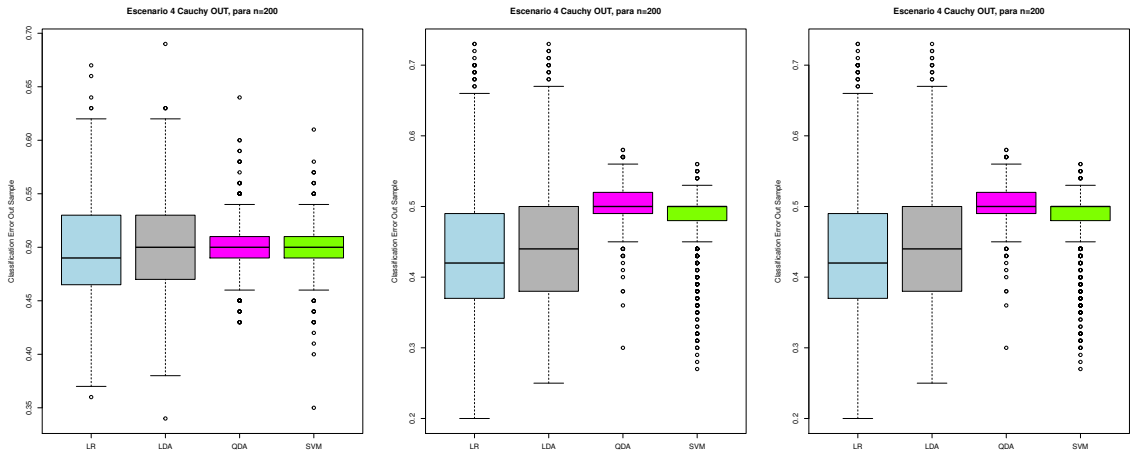

Figura 6.8: Muestras realizadas en laboratorio para los valores de $\Delta = 0{,}1,\ 1$ *y* 1 respectivamente y $\rho = 1 \quad \rho = 2 \quad \rho = 3$ repectivamente del Escenario IV con v.a. Cauchy. Se puede observar que en todos los casos, existe una gran presencia de outliers, y la variabilidad de los modelos. sin importar el aumento o disminución de Δ o ρ .

Si observamos bien, a diferencia de LDA y QDA, SVM no hace suposiciones fuertes sobre la distribución de los datos, que asumen normalidad. Sin embargo, algunos aspectos clave afectan su rendimiento. Como se desarrolló en la Sección 5.1, este algoritmo busca maximizar el margen entre las clases, utilizando puntos de soporte clave, y es relativamente sensible a los outliers, especialmente si se usa un margen rígido, ya que los vectores de soporte determinan la frontera de decisión. La característica más notoria de la distribución de Cauchy es la presencia de outliers frecuentes y extremos, lo que tiene un impacto directo en el comportamiento de SVM, lo que se ve reflejado en:

- Los vectores de soporte,son las observaciones que están más cerca del hiperplano de separación. Si los datos contienen muchos outliers debido a la distribución de Cauchy, estos valores extremos pueden actuar como vectores de soporte, distorsionando el hiperplano de decisión y volviendolo inestable.

- En lugar de obtener un hiperplano que generalice bien, los outliers pueden forzar a SVM a ajustar la frontera de decisión para acomodar estos valores extremos, lo que

puede llevar a un rendimiento deficiente en la clasificación de nuevas muestras, y aún más deficiente la generalización.

- El parámetro de penalización C controla la tolerancia de SVM a los errores de clasificación. Si C es pequeño, SVM permitirá que algunos puntos (potencialmente outliers) se clasifiquen incorrectamente, lo que mejora la generalización. Sin embargo, si C es demasiado grande, SVM intentará clasificar incluso a los outliers correctamente, lo que puede llevar a un sobreajuste.

Un kernel lineal(el que utilizamos en las pruebas de laboratorio), puede ser demasiado rígido si los datos no son linealmente separables debido a la dispersión de los valores extremos, por lo que los outliers generados por una distribución de Cauchy pueden distorsionar severamente el hiperplano lineal, como se puede observar en la Figura 6.8.

Escenario IV - LogNormal

La distribución lognormal tiene las siguientes características clave:

- La distribución lognormal es asimétrica, con una cola más larga hacia la derecha. Esto significa que la mayoría de los datos están concentrados en valores más bajos, pero pueden existir valores extremos altos. En una distribución lognormal, la varianza aumenta con la media, lo que puede llevar a heterocedasticidad (varianza no constante) en los datos.

- Si una variable aleatoria Y sigue una distribución lognormal, entonces su logaritmo natural $X = ln(Y)$ sigue una distribución normal. Esto implica que las transformaciones y relaciones en el espacio logarítmico son más lineales.

Conjunto de Entrenamiento									
Δ	ρ	RL		LDA		QDA		SVM	
		Media	D.E.	Media	D.E.	Media	D.E.	Media	D.E.
0.1	0.5	0.455	0.002	0.455	0.002	0.464	0.001	0.459	0.001
0.1	0.8	0.456	0.002	0.456	0.002	0.464	0.001	0.462	0.001
0.2	0.5	0.447	0.002	0.447	0.002	0.457	0.001	0.451	0.001
0.2	0.8	0.451	0.002	0.452	0.002	0.460	0.001	0.456	0.001
0.6	0.5	0.376	0.002	0.381	0.002	0.406	0.002	0.384	0.002
0.6	0.8	0.387	0.003	0.392	0.002	0.415	0.002	0.397	0.002

Cuadro 6.9: Medias y desvíos estándar de la tasa de error en el conjunto de Entrenamiento para el Escenario IV LogNormal según valores de ρ y Δ.

La Regresión Logística es adecuada para modelar relaciones *"no lineales"* entre las variables independientes y la probabilidad de pertenencia a una clase, como se vió en la Sección 2. La naturaleza asimétrica de la distribución lognormal puede ser manejada razonablemente bien por la Regresión Logística, especialmente si las características están bien seleccionadas. En general, puede ajustar adecuadamente las probabilidades de clasificación en datos lognormales, ya que puede manejar bien las no linealidades en las relaciones.

Conjunto de Prueba									
Δ	ρ	RL		LDA		QDA		SVM	
		Media	D.E.	Media	D.E.	Media	D.E.	Media	D.E.
0.1	0.5	0.494	0.002	0.494	0.002	0.495	0.001	0.495	0.001
0.1	0,8	0.492	0.002	0.493	0.002	0.497	0.001	0.496	0.001
0.2	0.5	0.477	0.003	0.476	0.003	0.487	0.002	0.483	0.002
0.2	0.8	0.480	0.003	0.481	0.003	0.488	0.002	0.487	0.001

0.6	0.5	0.384	0.002	0.389	0.002	0.419	0.002	0.401	0.002
0.6	0.8	0.398	0.003	0.401	0.002	0.428	0.002	0.415	0.002

Cuadro 6.10: Medias y desvíos estándar de la tasa de error en el conjunto de Prueba para el Escenario IV -LogNormal según valores de ρ y Δ.

Si los datos originales son lognormales, podría ser beneficioso aplicar una transformación logarítmica a las variables independientes antes de aplicar la Regresión Logística. Esto puede ayudar a *"normalizar"* la distribución y reducir la asimetría, facilitando un mejor ajuste del modelo, aunque es más robusta que otros modelos en términos de sensibilidad a outliers, la presencia de valores extremos derivados de la distribución lognormal puede influir en la estimación de los coeficientes, aunque no de manera tan drástica como vimos en el caso de la distribución Cauchy.

En el Cuadro 6.10 podemos apreciar como los 4 algoritmos se comportan de manera similar, teniendo un sesgo relativamente alto, comparado a los obtenidos con la distribución $Normal(\mu, \sigma^2)$, pero con una baja varianza.

Al analizar los métodos, Análisis Discriminante Lineal (LDA) y el Análisis Discriminante Cuadrático (QDA), como se vio en las Secciones 3 y 4 respectivamente, que son dos procedimientos supervisados ampliamente utilizados para clasificar observaciones en diferentes clases basándose en características continuas, observamos que estos métodos asumen ciertas propiedades sobre la distribución de los datos dentro de cada clase. Cuando los datos que alimentan a estos algoritmos son generados por una distribución lognormal, surgen varias consideraciones que afectan su desempeño y efectividad; como por ejemplo: LDA y QDA asumen que las características dentro de cada clase siguen una distribución normal multivariada e igualdad de matrices de covarianza, y la distribución lognormal no respeta. La asimetría positiva de la distribución lognormal viola la suposición de simetría de la distribución normal. Esto puede llevar a estimaciones sesgadas de las medias y matrices de covarianza. Aunque la distribución lognormal no tiene colas tan pesadas como la distribución de Cauchy,

aún así puede presentar valores extremos que afectan las estimaciones de los parámetros, especialmente en QDA que maneja matrices de covarianza individuales.

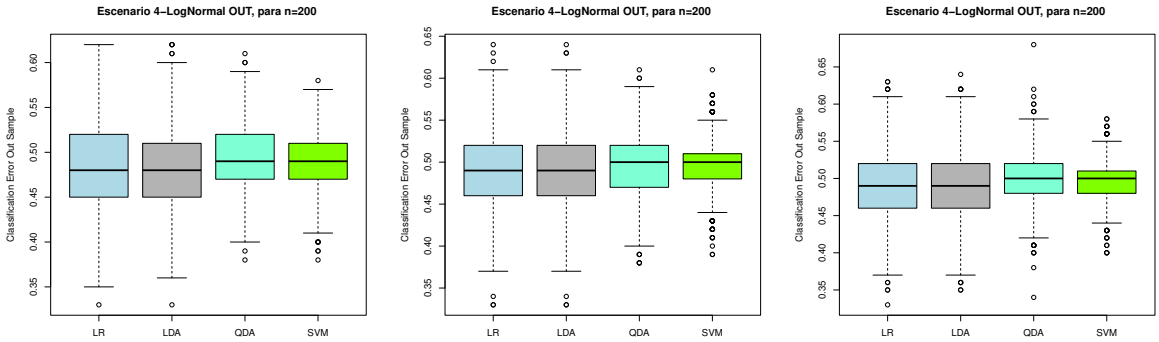

Figura 6.9: Muestras realizadas en laboratorio para los valores de $\Delta = 0{,}01,\ 0{,}1$ *y* $0{,}1$ respectivamente y $\rho = 0$, $\rho = 0{,}5$ y $\rho = 0{,}99$ del Escenario IV Lognormal. Se puede observar que en todos los casos,existe una gran presencia de outliers, y la variabilidad de los modelos, sin importar el aumento o disminución de Δ o ρ

.

Aunque QDA permite matrices de covarianza distintas, si las distribuciones lognormales generan matrices de covarianza altamente variables o mal condicionadas, la estabilidad y precisión de QDA pueden verse comprometidas. Los modelos pueden no generalizar bien a nuevos datos si las estimaciones de los parámetros están sesgadas o son inestables debido a la distribución lognormal.

Como ya vimos en la Sección 5, las SVM operan bajo ciertas premisas que afectan cómo manejan las diferentes distribuciones de datos.

Se busca el hiperplano que maximiza el margen entre las clases, es decir, la distancia mínima entre el hiperplano y los vectores de soporte de cada clase, se utilizan kernels,que permiten transformar los datos a un espacio de mayor dimensión donde las clases puedan ser separadas linealmente, facilitando la clasificación de relaciones no lineales.

La asimetría positiva de la distribución lognormal tiene como consecuencia que hay una concentración de datos cerca de un valor mínimo y una larga cola hacia valores mayores. Esto

puede afectar la forma en que la SVM identifica el margen óptimo, y así, la SVM podría inclinarse hacia la derecha para acomodar la cola larga, lo que puede distorsionar el hiperplano de separación y reducir la capacidad de generalización del modelo.

En una distribución lognormal, la varianza de los datos aumenta con la media. Esto significa que las observaciones con valores más altos tendrán una mayor dispersión. La SVM podría tener dificultades para determinar un margen óptimo consistente, ya que la dispersión variable afecta la identificación de vectores de soporte relevantes, y a pesar de no tener colas pesadas como la variable Cauchy, aún asi se pueden generar valores extremos que se comportan como outliers. Estos outliers pueden convertirse en vectores de soporte, influenciando desproporcionadamente la posición y orientación del hiperplano, lo que conlleva a una menor precisión en la clasificación de nuevas observaciones.

La distribución lognormal puede no ser linealmente separable en el espacio original, lo que requiere el uso de kernels no lineales para capturar la complejidad de la separación, y un mejor funcionamiento del método, pero no lo garantizan, pues la presencia de asimetría y heterocedasticidad puede aún afectar negativamente el desempeño del modelo.

Escalar las observaciones *(Normalizarlas)* para que estén dentro de un rango específico, como [0, 1], puede mejorar la convergencia y el desempeño de la SVM.

Escenario IV - Coronas

Los datos en forma de *Coronas* presentan características distintivas muy específicas, como por ejemplo, poseer una estructura radial, donde los datos se agrupan en anillos alrededor de un centro, donde cada anillo puede representar diferentes clases o categorías. Debido a esto, la separación entre las clases no es lineal; las clases están organizadas en estructuras anulares, lo que complica la clasificación,como podemos ver en la Figura 6.10.

Gráfico de Dispersión del Escenario IV Coronas

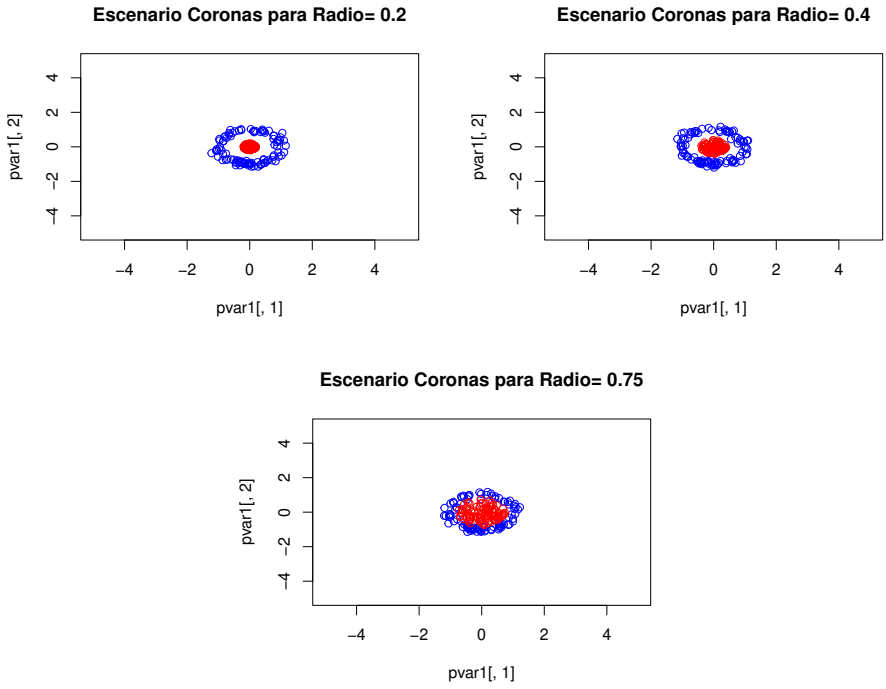

Figura 6.10: Muestras realizadas en laboratorio para los valores de $n = 200$ y $\rho = 0{,}2,\ 0{,}4$ y $0{,}75$, respectivamente donde se observa el formato de la distribución de los datos en forma de coronas concéntricas, dejando fijo el anillo azul y variando el rojo.

Para poder generar los datos de la variable *Coronas*, se definieron las variables:

- $\theta \sim 2\pi U(m_1)$, donde m_1 la definimos en 50, valor fijo.
- ρ , que lo haremos variar entre (0,1] obteniendo $r = \rho * \sqrt{(U(m0))}$
- y la construcción de los puntos mediante : $(x, y) = (r * \cos(\theta), r * \sin(\theta))$,

donde la corona externa (azul) está fija y los puntos rojos irán variando su dimensión como se observa en la Figura 6.10

Conjunto de Entrenamiento								
ρ	RL		LDA		QDA		SVM	
	Media	D.E.	Media	D.E.	Media	D.E.	Media	D.E.
0.1	0.349	0.005	0.349	0.005	0.000	0	0.302	0.0005
0.2	0.409	0.003	0.409	0.003	0.000	0	0.295	0.0005
0.5	0.449	0.001	0.449	0.001	0.003	2.22E-05	0.336	0.001
0,75	0.456	0.001	0.456	0.001	0.057	0.0002	0.411	0.001
0.9	0.458	0.001	0.458	0.001	0.134	0.0007	0.434	0.001
1	0.460	0.001	0.460	0.001	0.206	0.0012	0.445	0.001

Cuadro 6.11: Medias y desvíos estándar de la tasa de error en el conjunto de Entrenamiento para el Escenario IV-Coronas según valores de ρ.

La Regresión Logística que hemos ajustado asume que hay una relación lineal entre las variables independientes y la *log-odds* de la variable dependiente. Debido a la naturaleza anular de los datos, la frontera de decisión que intenta crear la Regresión Logística será una línea recta, lo que no se ajusta de la mejor manera a la separación radial que poseen las clases. En consecuencia, el modelo de Regresión Logística puede no ser capaz de clasificar correctamente las observaciones, ya que no puede capturar la complejidad de la estructura en anillos que poseen los datos. Esta incapacidad para representar la separación de clases de forma adecuada en datos de coronas puede resultar en una alta tasa de error en la clasificación de nuevas observaciones. Las clases que están más cerca de los bordes de los anillos pueden ser particularmente difíciles de clasificar correctamente.

Conjunto de Prueba								
ρ	RL		LDA		QDA		SVM	
	Media	D.E.	Media	D.E.	Media	D.E.	Media	D.E.
0.1	0.379	0.006	0.379	0.006	0	0	0.329	0.001
0.2	0.442	0.004	0.442	0.004	0.0000	0.0000	0.322	0.001
0.5	0.484	0.003	0.484	0.003	0.006	0.000	0.366	0.002
0,75	0.494	0.003	0.494	0.003	0.074	0.001	0.445	0.002
0.9	0.495	0.003	0.495	0.003	0.158	0.002	0.471	0.003
0,1	0.499	0.003	0.499	0.003	0.240	0.003	0.483	0.003

Cuadro 6.12: Medias y desvíos estándar de la tasa de error en el conjunto de prueba para el Escenario No Normal-oronas según valores de ρ.

Tanto LDA como QDA asumen *normalidad*, que es violada en el caso de los datos que hemos generado en forma de Corona.

Dado que LDA busca una frontera de decisión lineal, no puede capturar eficazmente la separación circular entre clases. Esto resulta en una frontera que no refleja la verdadera estructura de los datos.

Aunque QDA es más flexible que LDA al permitir matrices de covarianza distintas, la estructura anular aún puede no alinearse bien con las suposiciones de normalidad, afectando la precisión.

La presencia de puntos cercanos al centro o en los límites de los anillos puede influir significativamente en las estimaciones de las matrices de covarianza y las medias, llevando a fronteras de decisión distorsionadas como observamos en la Figura 6.11 .

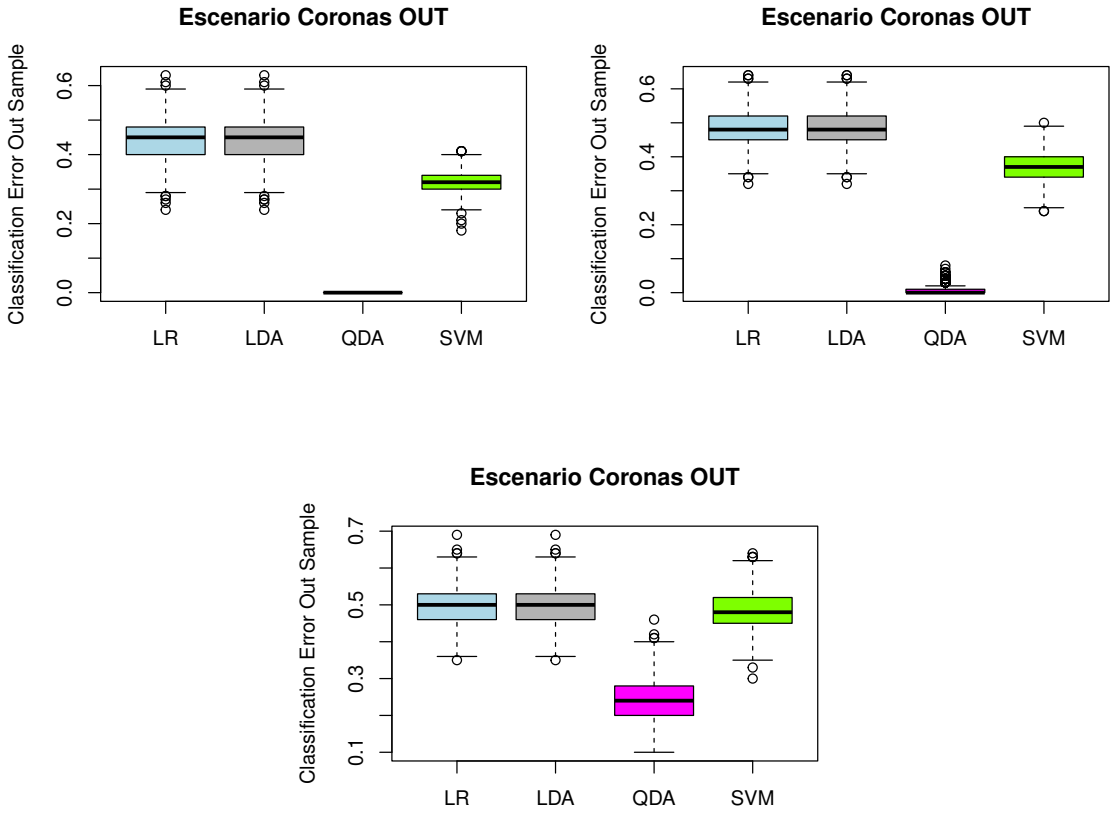

Figura 6.11: Muestras realizadas en laboratorio para los valores de $\rho = 0{,}2,\ 0{,}5\ y\ 1$ respectivamente del Escenario IV Coronas. Se puede observar que en todos los casos, existe una gran presencia de outliers, y la variabilidad de los modelos, sin importar el aumento o disminución de ρ

.

QDA puede manejar mejor la separación no lineal que LDA, reduciendo la tasa de error de clasificación, pero aún así es deficiente para separaciones complejas, pues en estructuras anulares con múltiples anillos o más complejas, QDA puede no ser suficiente para capturar todas las relaciones no lineales, resultando en errores de clasificación, como se puede observar en la Figura 6.11.

Una SVM lineal intenta encontrar un hiperplano (en 2D, una línea) que separa las clases. En una distribución en forma de corona, donde las clases están organizadas en anillos

concéntricos, no existe una línea recta que pueda separar eficazmente las clases, luego SVM lineal fallará en clasificar correctamente los datos, ya que no puede capturar la relación no lineal entre las clases. La tasa de error de clasificación será alta, y el modelo no generalizará bien a nuevos datos.

La capacidad de la SVM para manejar datos en forma de corona depende en gran medida de la elección del kernel. Un kernel adecuado es esencial para capturar la estructura no lineal de los datos. En nuestro estudio de laboratorio, sólo hemos utilizado el kernel lineal en todos los algoritmos.

Cuando se utiliza un kernel no lineal adecuado, como por ejemplos RBF, y se ajustan correctamente los parámetros, la SVM puede lograr una alta precisión en la clasificación de datos distribuidos en forma de corona.

Sin una elección y ajuste adecuados del kernel y sus parámetros, la SVM puede fallar en capturar la estructura anular, resultando en una baja precisión y mala generalización.

Capítulo 7

Conclusiones

En este trabajo se propuso como objetivo contrastar cuatro algoritmos de clasificación estadística: Regresión Logística, Análisis Discriminante Lineal, Análisis Discriminante Cuadrático y Support Vector Machine. Dichas metodologías fueron primero explicadas individualmente, analizando sus supuestos, reglas de decisión y la estimación de sus parámetros. Seguidamente, se compararon las técnicas teniendo en cuenta sus ventajas y desventajas y las situaciones en las cuales resulta conveniente aplicar cada una de ellas.

Regresión Logística es un método que consiste en modelar la relación entre la variable respuesta Y y las variable explicativas $X_1, X_2, ...X_p$ utilizando la transformación logit (logaritmo de las odds). La función resultante tiene forma de letra "S", con sus parámetros β siendo fácilmente interpretables. Las estimaciones para dichos parámetros se obtienen mediante el método de máxima verosimilitud, que posee buenas propiedades pero ocasiona problemas si existe una separación perfecta entre clases. Esta técnica no hace supuestos fuertes sobre las distribuciones de Y y X, haciéndola una técnica flexible y robusta a los valores extremos.

El Análisis Discriminante Lineal consiste en encontrar límites de decisión que logre una separación óptima de las observaciones en diferentes clases. Los parámetros de los límites se estiman mediante información provista por los datos de la muestra. Esta técnica hace supuestos fuertes sobre la distribución de las variables explicativas: se debe asumir que las mismas provienen de una distribución normal multivariada (dentro de cada clase) con una matriz de

covarianzas común. De cumplirse los supuestos, el desempeño del Análisis Discriminante puede llegar a ser igual o superior al de Regresión Logística. Sin embargo, cuando no se cumplen los supuestos, los resultados obtenidos suelen ser bastante pobres.

El Análisis Discriminante Cuadrático es una generalización del Análisis Discriminante Lineal que relaja el supuesto de igualdad de matrices de covarianzas. Esto permite que los límites de decisión sean cuadráticos y no solamente lineales. Las estimaciones se realizan de manera análoga a la metodología anterior, excepto que ahora es necesario estimar una matriz distinta por clase. El Análisis Discriminante Cuadrático resulta una alternativa viable en los casos donde el tamaño muestral es elevado y no se puede garantizar el cumplimiento de los supuestos para el análisis lineal.

Support Vector Machine es un algoritmo que construye un clasificador mediante el ajuste de un hiperplano. Gracias a la kernelización, la técnica resulta muy flexible sin perder simplicidad, pudiéndose aplicar distintas funciones según la situación bajo estudio. El hiperplano estimado dependerá de un parámetro C que se selecciona mediante validación cruzada. Las máquinas de soporte vectorial suelen aplicarse cuando se prioriza la capacidad predictiva por sobre la inferencia estadística, particularmente en los casos donde hay una separación muy marcada entre las clases.

Luego de esta revisión teórica, se recurrió a simulaciones en laboratorio con datos controlados para evaluar de forma empírica el desempeño de los cuatro métodos bajo diversas condiciones. Aplicando librerías del software R, se generaron múltiples conjuntos de datos con distintas características que se deseaban estudiar. Las simulaciones se agruparon en tres escenarios:

En el primer escenario, se estudió el efecto del tamaño muestral sobre la precisión de las clasificaciones. Se probaron cuatro valores distintos, evaluando también dos valores distintos para la separación entre las variables y la correlación entre ambas. En este escenario, se cumplen los supuestos de LDA.

En el segundo escenario, el foco se puso en analizar el comportamiento de los modelos ante distintas magnitudes de separación entre las poblaciones. Se generaron conjuntos de

datos con cinco valores diferentes del parámetro Δ, con mayores valores correspondiendo a mayor separación. Además, se trabajó con matrices de covarianza separadas para cada población, violando así uno de los supuestos de LDA.

En el tercer y último escenario, se evaluó el desempeño de los modelos con diferentes niveles de correlación entre las variables. Se realizaron ensayos con cuatro valores distintos para el parámetro ρ, donde valores más elevados indican una relación más fuerte entre X_1 y X_2. En esta ocasión, las matrices de covarianza variaron de forma proporcional, por lo que se cumple el supuesto de matrices comunes de LDA.

Las simulaciones se replicaron múltiples veces (= 1,000) , registrando en cada repetición la tasa de error en el conjunto de entrenamiento y en el conjunto de prueba. Una vez efectuadas todas las simulaciones, se resumieron los resultados obtenidos para cada técnica dentro de cada escenario y combinación de parámetros seleccionada.

En el Escenario I, se pudo observar un desempeño similar entre las cuatro técnicas analizadas. LDA presentó resultados ligeramente mejores que las demás, así como también menor variabilidad. Esta técnica funcionó mejor que las demás cuando el tamaño muestral fue reducido. Por el contrario, QDA resultó el método menos preciso, particularmente cuando las clases estuvieron muy superpuestas. Asimismo, presentó el mayor error promedio entre los cuatro métodos.

En el Escenario II, se apreció que las cuatro técnicas presentaron nuevamente desempeños muy similares. LDA mostró resultados superiores cuando se cumplieron sus supuestos, mientras que QDA fue la mejor técnica para valores altos de ρ. Una mayor separación entre las observaciones de las poblaciones resultó generalmente en menores tasas de error, así como en menor variabilidad. Aumentar la correlación entre las dos variables predictoras en una de las poblaciones pareció perjudicar la clasificación para cualquier técnica.

Finalmente, en el Escenario III se pudo apreciar que aumentar ρ para ambas poblaciones empeoró el desempeño de los métodos de clasificación, observándose mayores tasas de error y más variabilidad. En particular, LDA parece haber sido más afectado por el aumento de la correlación.

Como era de suponer, aumentar el tamaño muestral resultó en una reducción en la variabilidad del error de clasificación para las cuatro técnicas. Esto también permitió disminuir el promedio de las tasas de error, aunque en menor medida de lo que se redujo la variabilidad. Aunque se esperaba que aumentar n afectara positivamente a QDA en mayor proporción, no pareció observarse una mejora más notable que para las otras tres técnicas.

Del mismo modo, aunque se creía que LDA tendría un desempeño mejor en muestras pequeñas que Regresión Logística, en la práctica el desempeño de estas dos técnicas fue muy similar. Aún así, LDA si pareció funcionar mejor ante tamaños de muestra reducidos. Debido a que se trabajó siempre con el mismo número de variables explicativas (2), no pudo evaluarse el caso donde se tienen muchos predictores para una muestra pequeña, situación que probablemente hubiera benefciado a SVM por encima de los demás métodos

Así como se esperaba, cuando Δ tomó valores elevados, LDA obtuvo resultados favorables, pero solamente en los casos donde las correlaciones eran iguales o similares para ambas poblaciones. Aunque se esperaba que SVM produjera una mejor clasificación ante poblaciones muy separadas y Regresión Logística hiciera lo propio para poblaciones superpuestas, ambas metodologías presentaron resultados muy similares en casi todas las situaciones planteadas. Esto podría deberse, en parte, al kernel elegido, por lo que en futuras investigaciones podría estudiarse el efecto de utilizar otras funciones como por ejemplo un kernel radial.

A pesar de que se creía que Regresión Logística podría ser afectada negativamente por la multicolinealidad, no se apreció que este método resultara notablemente peor que el resto al aumentar la correlación entre X_1 y X_2. Por otra parte, si se cumplió que QDA resultó en un mejor desempeño cuando los límites de decisión dejan de ser lineales al aumentar ρ en el Escenario III.

Regresión Logística y SVM presentaron resultados muy similares en las tres situaciones, aunque SVM tuvo menor variabilidad en el Escenario I y Regresión Logística tuvo un desempeño ligeramente superior en los Escenarios I y II. Por su lado, la precisión de las clasificaciones obtenidas con LDA y QDA dependió en gran manera del cumplimiento de los supuestos de ambas, mostrando que suelen ser menos flexibles que las otras dos técnicas.

Generamos de manera adicional un Escenario IV, compuesto por las distribuciones *Cauchy*, *Lognormal* y datos en forma de *Corona*, para poder analizar de forma epírica que sucede con Regresión Logística, LDA, QDA y SVM ante la posiblidad de trabajar con datos *no Normales*, violando incluso los supuestos de algunos modelos. El rendimiento no fue adecuado, propio a la distribución de los datos. Se observó un elevado sesgo en todos los escenarios, y baja varianza, por lo que concluimos que antes distribuciones más atípicas, se debe analizar bien con qué método clasificarlas. Si bien existe cierta flexibilidad en los 4 métodos estudiados, esa flexibilidad no es infinita.

En síntesis, no existe una única metodología que sea universalmente superior en todos los casos. Cada técnica posee ventajas y desventajas, obteniendo un desempeño mejor o peor al resto según el escenario con el que se trabaje. No obstante, se pudo comprobar que los cuatro métodos producen buenos resultados en diferentes situaciones, aún cuando no se cumplen las condiciones ideales para su aplicación. De esta manera, se puede confiar en que seleccionar cualquiera de estas técnicas resultará en una clasificación apropiada en la mayoría de los casos, por lo que dicha elección puede depender de criterios adicionales tales como su facilidad de interpretación o la rapidez de su implementación.

Bibliografía

[Efron Hastie(2016)] Efron, B., y Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press.

[Ghojogh(2019)] Ghojogh, B., y Crowley, M. (2019). Linear and Quadratic Discriminant Analysis: Tutorial. *ArXiv*. https://arxiv.org/abs/1906.02590

[Hastie(2017)] Hastie, T., Friedman, J., y Tisbshirani, R. (2017). *The elements of Statistical Learning: Data Mining, Inference, and prediction*. Springer.

[James(2013)] James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer.

[Jorgensen(2019)] Jörgensen, S. (2019). *A Comparative Simulation Study of Logistic Regression and Linear Discriminant Analysis for Classification* [Tésis de grado]. Universidad de Estocolmo. https://kurser.math.su.se/pluginfile.php/20130/mod_folder/content/0/Kandidat/2019/2019_10_report.pdf

[Wasserman(2004)] Wasserman, L. (2004). *All of statistics: A concise course in statistical inference*. Springer Texts in Statistics.