Curso: Machine Learning para Economistas
Departamento Académico de Economía
Profesor: Francisco Rosales, PhD.
Fecha: Agosto de 2022

Notas de Clase

Contenidos

1 Machine Learning

Nos referimos con máquina a una función determinista. Decimos que *aprende* porque extrapola un patrón o un conjunto de patrones a partir de datos que se ha observado. Nosotros interactuamos con *algorítmos* de machine learning que son implementaciones particulares de estas funciones.

E1 data de consumidores en un supermercado para segmentación de clientes

E2 data de la economía (variables macro) para predecir la próxima recesión

E3 data de la los compradores de un seguro para predecir si lo seguirán comprando

1.1 Aprendizaje Supervisado

Hay dos tipos de aprendizaje:

- Aprendizaje estadístico

  - Aprendizaje supervisado: usa data (input, output), y le da estructura a la relación entre ambos. Con inputs conocidos puede estimar un output desconocido. (E2, E3).
  - Aprendizaje no-supervisado: usa data input, y le da estructura al input. Con input nuevo puede localizar el input en esa estructura. (E1).
- Aprendizaje Reforzado.

En estas notas veremos únicamente aprendizaje supervisado. El aprendizaje no-supervisado y el aprendizaje por refuerzo serán omitidos.

Anotación 1 (Asumsión de Estructura)**.** *Asumimos que existe una función determinista que vincula inputs con output para un set de datos determinado. Esto significa que existe una estructura/un patrón que descubrir. Este no siempre es el caso. Considere por ejemplo, la predicción de los números de la lotería, o la predicción de un camino aleatorio.*

Ejemplo 1 (Regresión Lineal)**.** *Es un método de aprendizaje supervisado en el que la máquina viene dada por cierta función que mapea desde el espacio de las variables independientes hacia el espacio de la dependiente. Los lenguajes computacionales usan algoritmos (implementaciones) diversos. La implementación afecta la velocidad del cálculo y la estabilidad de la solución.*

1.2 Modelamiento

El problema de aprendizaje supervisado se puede escribir como:

$$Y = f(X) + \epsilon, \quad X = \{X_1, \ldots, X_p\} \tag{1}$$

donde Y es la variable respuesta (dependiente), X_i son variables predictoras (independientes), f es una función determinística, y ϵ es una variable aleatorio, independiente de X, centrada en cero $\mathbb{E}[\epsilon] = 0$.

El problema (1) recibe nombres diferentes dependiendo de la naturaleza de Y.

- Si Y es continuo se llama problema de regresión
- Si Y es discreto[1] se llama problema de clasificación

Anotación 2 (Asumsión de Linealidad)**.** *Es importante anotar que no hemos hecho ninguna asumsión con respecto a f, y que esta función en general no es lineal. En general, en aprendizaje supervisado se usan métodos más flexibles que una regresión lineal. Un método muy popular por su gran flexibilidad es K-nearest[2] neighbors (KNN). En este método la estimación de Y corresponde al valor predominante en el vecindario de la observación x en el espacio de predictores.*

[1] En el curso vamos a ver casos en los que Y es discreto pero no ordinal. Es decir no tocaremos los casos en los que Y tiene etiquetas que tienen algún ordenamiento cualitativo. Considere como ejemplo una variable respuesta con las etiquetas "malo", "regular", bueno.

[2] La noción de cercanía está asociada a cierta norma, típicamente L_2.

1.3 Métricas de desempeño

Las métricas de desempeño son diferentes dependiendo del tipo de problema. Considere $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ la matrix de diseño, $\boldsymbol{y} \in \mathbb{R}^n$ el vector respuesta.

1.3.1 Partición de la data

En ML es conveniente dividir la data de diferentes formas. En principio consideremos la partición de la matriz de diseño en dos conjuntos de observaciones:

- entrenamiento (train): sirve para encontrar los parámetros óptimos de f.
- prueba (test): sirve para evaluar el desemepeno de f usando las métricas antes vistas.

Anotación 3 (Overfitting)**.** *Como se indicó en 4, el MSE en la data de entrenamiento puede ser reducido arbitrariamente. Sin embargo esto no ocurre con el MSE con data de prueba.*

1.3.2 Regresión (Y cont.)

Sea $\boldsymbol{e} = \boldsymbol{y} - \hat{f}(\boldsymbol{X})$, y la norma-p del vector $\boldsymbol{x} \in \mathbb{R}^n$ definida como $\|\boldsymbol{x}\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$, la métrica a considerar es

$$\text{MSE} = n^{-1}\|\boldsymbol{e}\|_2^2 \tag{2}$$

donde MSE es el error cuadrático medio. Otras son

$$\begin{aligned} \text{MAE} &= n^{-1}\|\boldsymbol{e}\|_1^2 && (3)\\ \text{MAPE} &= n^{-1}\|\tilde{\boldsymbol{e}}\|_1^2,\ \tilde{e}_i = e_i/y_i && (4) \end{aligned}$$

donde MAE es el error absoluto medio y MAPE es el error absoluto medio porcentual.

1.3.3 Clasificación (Y disc.)

El ratio de error (ER) es la proporción de errores que realiza el clasificador en un conjunto de datos.

$$\text{ER} = n^{-1}\sum_{i=1}^n \mathbf{1}_{y_i\neq\hat{g}_i} = 1 - n^{-1}\sum_{i=1}^n \mathbf{1}_{y_i=\hat{g}_i}, \tag{5}$$

donde $\mathbf{1}_{\text{cond}(\theta)}$ es una función indicadora que toma el valor de 1 si la condición $\text{cond}(\theta)$ es verdadera, y toma el valor 0 si es falsa.

Ejemplo 2 (Perros, Gatos y Ratones)**.** *Considere que tiene un modelo que clasifica una imagen en Perro, Gato o Ratón de acuerdo a sus características. Dada la siguiente información, ¿cuál es el ratio de error?*

Demostración. [Solución] Es claro que el ratio de error es $1-(0+1+0+0+0+1)/6 = 4/6$. Es decir, el clasificador se equivoca 4 de cada 6 veces y acierta 2 de cada 6 veces. □

Anotación 4 (Overfitting)**.** *Es fácil construir un modelo de muy buen desempeño (a veces perfecto) para un conjunto de datos determinado, pero que no generalice bien a datos no observados. A esto se le llama overfitting. Considere, por ejemplo el caso de un polinomio de grado $n-1$ ajustado a n datos.*

Obs.	Real	Estimado	$\mathbf{1}_{y_i=\hat{g}_i}$	$\mathbf{1}_{y_i\neq\hat{g}_i}$
1	Gato	Ratón	0	1
2	Perro	Perro	1	0
3	Gato	Perro	0	1
4	Ratón	Gato	0	1
5	Gato	Ratón	0	1
6	Ratón	Ratón	1	0

1.4 Tradeoffs

Hay diferentes aspectos qué considerar cuando se selecciona un modelo. En general se deben considerar dos tradeoffs

- Sesgo vs Varianza
- Precisión vs Interpretación

1.4.1 Sesgo vs Varianza

Nuestro interés principal reside en obtener un modelo que minimize el MSE esperado para la data de prueba. Este MSE tiene varios componentes aditivos.

Proposición 1. *El MSE esperado en data de prueba se descompone en tres partes: la varianza de $\hat{f}$, el sesgo al cuadrado de $\hat{f}$ y la varianza del término de error.*

Demostración. Considere el modelo general (1), el MSE esperado en data de prueba viene dado por

$$\begin{aligned}
\mathbb{E}[(Y-\hat{f})^2] &= \mathbb{E}[(f-\hat{f}+\epsilon)^2] && (6)\\
&= \mathbb{E}[(f-\hat{f})^2+2(f-\hat{f})\epsilon+\epsilon^2) && (7)\\
&= \mathbb{E}[(f-\hat{f})^2]+2\mathbb{E}[(f-\hat{f})\epsilon]+\mathbb{E}[\epsilon^2] && (8)\\
&= \mathbb{E}[(f-\hat{f})^2]+2(\mathbb{E}[f\epsilon]-\mathbb{E}[\hat{f}\epsilon])+\mathbb{E}[\epsilon^2] && (9)\\
&= \mathbb{E}[(f-\hat{f})^2]+2(f\mathbb{E}[\epsilon]-\mathbb{E}[\hat{f}]\mathbb{E}[\epsilon])+\mathbb{E}[\epsilon^2] && (10)\\
&= \mathbb{E}[(f-\hat{f})^2]+\mathbb{E}[\epsilon^2] && (11)\\
&= \mathbb{E}[(f-\mathbb{E}\hat{f}+\mathbb{E}\hat{f}-\hat{f})^2]+\mathbb{E}[\epsilon^2] && (12)\\
&= \mathbb{E}[(f-\mathbb{E}\hat{f})-(\hat{f}-\mathbb{E}\hat{f}))^2]+\mathbb{E}[\epsilon^2] && (13)\\
&= \mathbb{E}[(f-\mathbb{E}\hat{f})^2]+2\mathbb{E}[(f-\mathbb{E}\hat{f})(\hat{f}-\mathbb{E}\hat{f})]+\mathbb{E}[(\hat{f}-\mathbb{E}\hat{f})^2]+\mathbb{E}[\epsilon^2] && (14)\\
&= \mathbb{E}[(f-\mathbb{E}\hat{f})^2]+\mathbb{E}[(\hat{f}-\mathbb{E}\hat{f})^2]+\mathbb{E}[\epsilon^2] && (15)\\
&= \text{sesgo}^2(\hat{f})+\mathbb{V}[\hat{f}]+\mathbb{V}[\epsilon]. && (16)
\end{aligned}$$

Luego, minimizar el MSE de prueba involucra reducir el sesgo, la varianza o ambos. Adicionalmente, es obvio que el MSE de prueba nunca puede ser cero, dado que hay un término irreducible que viene dado por la varianza del error. □

Lograr reducir el sesgo y la varianza simultaneamente no siempre es posible. En aprendizaje supervisado se desfavorece el modelo de regresión lineal porque aunque tiene baja varianza, la ruptura del supuesto de linealidad hace que tenga un sesgo grande. En general, buscaremos

opciones con menos sesgo, a costa de una varianza un poco mayor que la que se encuentra en regresión lineal.

1.4.2 Precisión vs Interpretación

En general nos interesan dos cosas:

- una buena precisión
- una buena interpretación

Pero no siempre se pueden lograr ambas cosas. Por ejemplo, la regresión lineal permite muy buena interpretación, pero en general su precisión es baja. Una red neuronal en cambio es muy precisa, pero la relación entre el input y el output es poco clara y algunas personas le llaman método de caja negra. En este curso vamos a explorar métodos que sacrifican interpretabilidad para obtener resultados más precisos.

Anotación 5 (Flexibilidad)**.** *En aprendizaje supervisado es común usar métodos más flexibles que regresión lineal. En general, estos métodos permiten una mayor precisión (a costa de una menor interpretabilidad) y un mayor sesgo (a costa de una menor varianza)*

Ejemplo 3 (KNN para Clasificación)**.** *La tabla muestra data de entrenamiento conteniendo seis observaciones, tres predictores y una respuesta cualitativa.*

Obs.	X_1	X_2	X_3	Y	dist
1	0	3	0	Rojo	3.00
2	2	0	0	Rojo	2.00
3	0	1	3	Rojo	3.16
4	0	1	2	Verde	2.24
5	1	0	1	Verde	1.41
6	1	1	1	Rojo	1.73

Suponga que queremos usar este conjunto de datos para realizar una predicción para Y cuando $X_1 = X_2 = X_3 = 0$ usando KNN para $K = 1$ y para $K = 3$.

Demostración. [Solución] Calcule la distancia euclidiana entre cada observación y el punto de prueba, $X_1 = X_2 = X_3 = 0$. A partir de estos resultados es claro que si $K = 1$, la clasificación sería Verde. Mientras que si $K = 3$, la clasificación sería Rojo. □

2 Modelo Lineal Generalizado

La familia exponencial de distribuciones de probabilidad tiene distribución de probabilidad de la forma:

$$f(y;\theta) = h(y)g(\theta)\exp[\eta(\theta)\cdot T(y)], \tag{17}$$

donde θ representa los parámetros de la distribución, h, g, η, T son funciones en los argumentos indicados.

Anotación 6. *La representación* (17) *no es única. Otras representaciones son:*

$$f(y;\theta) = h(y)\exp[\eta(\theta)\cdot T(y) - A(\theta)] \tag{18}$$
$$f(y;\theta) = \exp[\eta(\theta)\cdot T(y) - A(\theta) + B(y)], \tag{19}$$

donde A y B son funciones adicionales. Todas estas representaciones son equivalentes.

Anotación 7. *La familia exponencial incluye distribuciones de variables aleatorias discretas y continuas. Algunos ejemplos son: normal, exponential, gamma, chi-squared, beta, Dirichlet, Bernoulli, categorical, Poisson, Wishart, inverse Wishart, geometric.*

Ejemplo 4 (Variable Bernoulli)**.** *Considere $y \overset{iid}{\sim}$ Brenoulli(p), tal que $\mathbb{P}[y=1]=p$, $\mathbb{P}[y=0]=1-p$. La distribución Bernoulli pertenece a la familia de distribuciones exponenciales dado que*

$$f(y;p) = p^y(1-p)^{1-y} \tag{20}$$
$$= (1-p)\left[\frac{p}{1-p}\right]^y \tag{21}$$
$$= (1-p)\exp\left[\log\left(\frac{p}{1-p}\right)^y\right] \tag{22}$$
$$= (1-p)\exp\left[y\log\left(\frac{p}{1-p}\right)\right], \tag{23}$$

donde es claro que la función masa de probabilidad tiene la forma (17)*, donde $\theta = p$, $g(p) = (1-p)$, $h(y) = 1$, $T(y) = y$, y η es una función logit de p, i.e., $\eta(p) = \log(p/(1-p))$.*

Ejemplo 5 (Variable Gaussiana)**.** *Considere $y \overset{iid}{\sim} \mathcal{N}(\mu,\sigma)$. La distribución Gaussiana pertenece a la familia de distribuciones exponenciales dado que*

$$f(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}, \tag{24}$$
$$f(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi}}\sigma^{-1}\exp\left\{\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}, \tag{25}$$
$$f(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi}}\exp\{\log(\sigma^{-1})\}\exp\left\{\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}, \tag{26}$$
$$f(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi}}\exp\{-\log(\sigma)\}\exp\left\{\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}, \tag{27}$$
$$f(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi}}\exp\left\{\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2 - \log(\sigma)\right\}, \tag{28}$$
$$= \frac{1}{\sqrt{2\pi}}\exp\left\{\frac{1}{2\sigma^2}y^2 - \frac{\mu}{\sigma^2}y + \frac{\mu^2}{2\sigma^2} - \log\sigma\right\}, \tag{29}$$
$$\tag{30}$$

donde es claro que la función masa de probabilidad tiene la forma (18)*, donde $h(y) = \frac{1}{\sqrt{2\pi}}$, $\eta(\mu,\sigma) = (-\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2})^\top$, $T(y) = (y, y^2)^\top$, y $A(\mu,\sigma) = -\frac{\mu^2}{2\sigma^2} + \log\sigma$.*

Considere el modelo general (1). Un modelo lineal generalizado modela una función $g(\cdot)$ aplicada a la esperanza de la respuesta utilizando una combinación lineal de sus predictores.

$$g(\mu_i) = \beta_0 + \sum_{i=1}^{p} \beta_j x_{ij},\ i = 1, \ldots, n, \tag{31}$$

donde g es una función monótona, creciente y diferenciable llamada función vinculante, $\mu_i = \mathbb{E}[y_i|x_i]$, y las observaciones y_i son iid con distribución de probabilidad perteneciente a la familia exponencial.

Anotación 8. *Note que para data Gaussiana, la función vinculante utilizada en un modelo lineal es la función identidad $g(z) = z$. Para el caso de data Bernoulli, la función vinculante utilizada en un modelo logit es $g(z) = \log(z/(1-z))$.*

2.1 Regresión Lineal Simple

Considere el modelo

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{32}$$

donde $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$, y $\epsilon_i \perp x_i$. Es usual la siguiente formulación

$$\mathbb{E}[y_i|x_i] = \beta_0 + \beta_1 x_i, \tag{33}$$

que indica que la media condicional de y_i dado x_i es una función lineal de x_i.

El modelo de regresión lineal minimiza el MSE en la data disponible en, es decir $(\hat{\beta}_0, \hat{\beta}_1) = \text{argmin}_{\hat{\beta}_0, \hat{\beta}_1} \text{MSE}$, MSE$= n^{-1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$. La solución de este problema es

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{34}$$

$$\hat{\beta}_1 = \hat{\sigma}_{xy} / \hat{\sigma}_x^2, \tag{35}$$

donde $\hat{\sigma}_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})/(n-1)$ y $\hat{\sigma}_x^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)$.

Proposición 2 ($R^2 = \hat{\rho}_{xy}^2$)**.** *Considere la regresión lineal simple (32). El R^2 de la regresión entre las variables y y x es igual a la correlación al cuadrado entre estas mismas variables.*

Demostración. Trivial.

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{f}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{36}$$

$$= \hat{\beta}_1^2 \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{37}$$

$$= \frac{\hat{\sigma}_{xy}^2}{\hat{\sigma}_x^4} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{38}$$

$$= \left(\frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y} \right)^2 \tag{39}$$

$$= \hat{\rho}_{x,y}^2 \tag{40}$$

□

2.2 Regresión Lineal Múltiple

Considere el modelo

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{41}$$

donde $\boldsymbol{y} \in \mathbb{R}^n, \boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I}_n)$, y $\boldsymbol{\epsilon} \perp \boldsymbol{X}$. Es usual la siguiente formulación

$$\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta}, \tag{42}$$

que indica que la media condicional de $\boldsymbol{y}$ dado $\boldsymbol{X}$ es una función lineal de $\boldsymbol{X}$. El modelo de regresión lineal minimiza el MSE en la data disponible en, es decir $\hat{\boldsymbol{\beta}} = \text{argmin}_{\hat{\boldsymbol{\beta}}}\text{MSE}$, MSE$= n^{-1}\|\boldsymbol{e}\|_2^2$. La solución de este problema es $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$, luego $\hat{\boldsymbol{f}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Adicionalmente, es práctico definir el vector de estimadores de $\boldsymbol{y}$ de una manera alternativa

$$\hat{\boldsymbol{f}} \;=\; \boldsymbol{H}\boldsymbol{y}, \quad \boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top, \tag{43}$$

donde $\boldsymbol{H}$ es llamada la matriz de estimación (hat matrix), con algunas propiedades especiales $\boldsymbol{H} = \boldsymbol{H}^\top$, $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$.

El error $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{f}}$ se puede re-escribir como

$$\boldsymbol{e} \;=\; \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}, \tag{44}$$

y su varianza se puede escribir directamente como

$$\begin{aligned}
\mathbb{C}\text{ov}[\boldsymbol{e}] &= \mathbb{E}[(\boldsymbol{e} - \mathbb{E}[\boldsymbol{e}])(\boldsymbol{e} - \mathbb{E}[\boldsymbol{e}])^\top] && (45)\\
&= \mathbb{E}[\boldsymbol{e}\boldsymbol{e}^\top] && (46)\\
&= \mathbb{E}[(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}\boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{H})^\top] && (47)\\
&= (\boldsymbol{I} - \boldsymbol{H})\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top](\boldsymbol{I} - \boldsymbol{H})^\top && (48)\\
&= \sigma_\epsilon^2(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})^\top && (49)\\
&= \sigma_\epsilon^2(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}) && (50)\\
&= \sigma_\epsilon^2(\boldsymbol{I} - \boldsymbol{H}) && (51)
\end{aligned}$$

En consecuencia

$$\hat{\text{std.dev}}[e_i] = \sigma_\epsilon\sqrt{1 - \boldsymbol{H}_{ii}}, \tag{52}$$

donde $\boldsymbol{H}_{ii} \in [0, 1]$, y los errores student-arizados tienen la forma

$$\tilde{e}_i = \frac{e_i}{\hat{\text{std.dev}}[e_i]} = \frac{e_i}{\hat{\sigma}_\epsilon\sqrt{1 - \boldsymbol{H}_{ii}}}, \tag{53}$$

que bajo las asumsiones de normalidad y homocesaticidad, debería tener distribución $\mathcal{N}(0, 1)$.

Proposición 3. $\boldsymbol{H}_{ii} \in [0, 1]$.

Demostración. Sabemos que $\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H}$ y que $\boldsymbol{H} = \boldsymbol{H}^\top$, i.e. $\boldsymbol{H}_{ij} = \boldsymbol{H}_{ji}$. Luego,

$$\begin{aligned} \boldsymbol{H}_{ii} &= \sum_{j=1}^{n} \boldsymbol{H}_{ij}\boldsymbol{H}_{ji} && (54) \\ &= \boldsymbol{H}_{ii}^2 + \sum_{j\neq i}^{n} \boldsymbol{H}_{ij}\boldsymbol{H}_{ji} && (55) \\ &= \boldsymbol{H}_{ii}^2 + \sum_{j\neq i}^{n} \boldsymbol{H}_{ij}^2 && (56) \end{aligned}$$

En consecuencia, $\boldsymbol{H}_{ii} \geq 0$. Adicionalmente, $\boldsymbol{H}_{ii}(1 - \boldsymbol{H}_{ii}) \geq 0$, y es obvio que $\boldsymbol{H}_{ii} \leq 1$. □

2.2.1 Diagnósticos

Diagnósticos de Influencia
Considere la expansión de (43). Las siguientes dos métricas son indicadores de observaciones anómalas

- **Leverage** $\boldsymbol{H}_{ii}$.

$$\begin{aligned} \hat{f}_i &= \sum_{j=1}^{n} \boldsymbol{H}_{ij} y_j && (57) \\ &= \boldsymbol{H}_{ii} y_i + \sum_{j\neq i}^{n} \boldsymbol{H}_{ij} y_j, && (58) \end{aligned}$$

  de modo que $\boldsymbol{H}_{ii}$ mide el efecto potencial de y_i sobre $\hat{f}_i$.

- **Distancia de Cook** D_i.

$$D_i = \frac{\boldsymbol{H}_{ii}}{p\hat{\sigma}_\epsilon^2(1 - \boldsymbol{H}_{ii})^2} e_i^2, \quad (59)$$

  la cual toma valores $D_i > 0$. Se considera que D_i es grande para valores mayores a la unidad.

Diagnóstico de Autocorrelación
La autocorrelación de rezago k viene dada por

$$\hat{\rho}(k) = \frac{\sum_{i=1}^{n} e_i e_{i-k}}{\sum_{i=1}^{n} e_i^2}. \quad (60)$$

Bajo la asumsión de independencia de errores se debe cumplir que $\hat{\rho}(k) \sim \mathcal{N}(0, 1/n)$, de modo que si $|\hat{\rho}(k)| > 2/\sqrt{n}$, se puede decir que se tiene un indicio de que los errores están correlacionados.

Diagnóstico de Dependencia Lineal
Un indicador de colinealidad es el factor de inflación de varianza a nivel de variable:

$$\text{VIF} = \left(1 - R^2_{X_j|X_{-j}}\right)^{-1}, \quad (61)$$

donde se considera que este valor es alto si VIF > 10.

Anotación 9. *Considerando $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ como la matriz de diseño en el problema de regresión (1), el número condicional de $\kappa(\boldsymbol{X}^\top \boldsymbol{X})$ es un indicador su grado de dependencia lineal. Un indicador del nivel de dependencia lineal en una matriz simétrica sem. pos. def. $\boldsymbol{M} \in \mathbb{R}^{p\times p}$ es su número condicional*

$$\kappa(\boldsymbol{M}) = \frac{\sqrt{\lambda_{max}}}{\sqrt{\lambda_{min}}}, \tag{62}$$

donde λ_{min} y λ_{max} corresponden a los valores mínimo y máximo de los valores propios de $\boldsymbol{M}$. Note que $\kappa(\boldsymbol{M}) \in (0, \infty)$. En particular, si $\boldsymbol{M}$ es singular, $\kappa(\boldsymbol{M}) = \infty$. Se considera que un grupo condicional razonable es menor que 10^3.

2.2.2 Inferencia

Considere el modelo (41), bajo las asumsiones antes descritas.

Proposición 4. *El estimador $\hat{\boldsymbol{\beta}}$ resultante bajo la asumsión de normalidad de errores es de la forma*

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma_\epsilon^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right). \tag{63}$$

Demostración. Note que: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{\beta} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\epsilon}$. Luego, es claro que $\hat{\boldsymbol{\beta}}$ hereda la normalidad de $\boldsymbol{\epsilon}$, y que $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$. Adicionalmente

$$\begin{aligned}
\mathbb{C}\text{ov}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top] && (64)\\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top]\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} && (65)\\
&= \sigma_\epsilon^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}. && (66)
\end{aligned}$$

□

Proposición 5 (Intervalo de Confianza)**.** *El intervalo de confianza de $\hat{\beta}_i$ al $(100-\alpha)\%$ viene dado por*

$$\hat{\beta}_i \pm \hat{std.dev}(\hat{\beta}_i) t\left(1 - \frac{\alpha}{2}, n - p\right), \tag{67}$$

donde $t(q, m)$ es el percentil $100 \times q$ de la distribución t con m grados de libertad, y

$$\hat{std.dev}(\hat{\beta}_i) = \hat{\sigma}_\epsilon \sqrt{[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}]_{ii}} \tag{68}$$

Anotación 10. *Estos intervalos de confianza tienen cobertura $100(1-\alpha)\%$ sólo si los errores son independientes y están normalmente distribuídos y la varianza del error es constante.*

Proposición 6 (Bandas de Confianza)**.** *Considere una observación nueva $\boldsymbol{x}_0$. El intervalo de confianza de la estimación de $\boldsymbol{x}_0^\top \boldsymbol{\beta}$ viene dado por*

$$\hat{y}_0 \pm \hat{std.dev.}(\hat{y}_0) t\left(1 - \frac{\alpha}{2}, n - p\right), \tag{69}$$

donde $t(q, m)$ es el percentil $100 \times q$ de la distribución t con m grados de libertad.

Demostración. Dada una observación nueva $\boldsymbol{x}_0$, sabemos que $\mathbb{E}[y_0|\boldsymbol{x}_0] = \boldsymbol{x}_0^\top \boldsymbol{\beta}$. Esta cantidad se estima con $\hat{y}_0 = \boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}$, que es un estimador insesgado

$$\mathbb{E}[\hat{y}_0] = \mathbb{E}[\boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}] = \boldsymbol{x}_0^\top \mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{x}_0^\top \boldsymbol{\beta}. \tag{70}$$

Por otro lado

$$\begin{aligned}
\mathbb{V}[\hat{y}_0] &= \mathbb{E}[(\boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}} - \boldsymbol{x}_0^\top \boldsymbol{\beta})(\boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}} - \boldsymbol{x}_0^\top \boldsymbol{\beta})^\top] && (71)\\
&= \mathbb{E}[(\boldsymbol{x}_0^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{x}_0] && (72)\\
&= \boldsymbol{x}_0^\top \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top] \boldsymbol{x}_0 && (73)\\
&= \boldsymbol{x}_0^\top \, \mathbb{C}\mathrm{ov}[\hat{\boldsymbol{\beta}}] \boldsymbol{x}_0 && (74)\\
&= \sigma_\epsilon^2 \boldsymbol{x}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_0, && (75)
\end{aligned}$$

de modo que

$$\hat{\text{std.dev}}(\hat{y}_0) = \hat{\sigma}_\epsilon^2 \sqrt{\boldsymbol{x}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_0} \tag{76}$$

Finalmente, el intervalo de confianza para la estimación de y_0 viene dado por

$$\hat{y}_0 \pm \hat{\text{std.dev.}}(\hat{y}_0) t\left(1 - \frac{\alpha}{2}, n - p\right). \tag{77}$$

□

Proposición 7 (Bandas de Predicción)**.** *Considere una observación nueva $\boldsymbol{x}_0$. El intervalo de predicción de la estimación de $\boldsymbol{x}_0^\top \boldsymbol{\beta} + \epsilon$ viene dado por*

$$\hat{y}_0 \pm \hat{std.dev.}(y - \hat{y}_0) t\left(1 - \frac{\alpha}{2}, n - p\right). \tag{78}$$

Demostración. Dado que se desea un intervalo que incorpore la varianza del error, y éste es independiente de la estimación, es claro que la varianza correspondiente tiene la forma

$$\begin{aligned}
\mathbb{V}[y - \hat{y}_0] &= \mathbb{V}[\hat{y}_0] + \mathbb{V}[\epsilon] && (79)\\
&= \sigma_\epsilon^2 \boldsymbol{x}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_0 + \sigma_\epsilon^2 && (80)\\
&= \sigma_\epsilon^2 (\boldsymbol{x}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_0 + 1) && (81)
\end{aligned}$$

Finalmente, el intervalo de confianza para la estimación de y_0 viene dado por

$$\hat{y}_0 \pm \hat{\text{std.dev.}}(y - \hat{y}_0) t\left(1 - \frac{\alpha}{2}, n - p\right). \tag{82}$$

□

2.3 Clasificación Logit

Un modelo logit es un modelo lineal generalizado (159) con respuesta Bernoulli $\mathbb{E}[y_i|x_{ij}] = \mathbb{P}[y_i = 1|x_{ij}] = p_i$, de la forma

$$g(p_i) = \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j, \tag{83}$$

donde $g(z) = \log\left(\frac{z}{1-z}\right)$, es decir g es una función logit. La siguiente formulación es equivalente

$$p_i = g^{-1}\left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\right), \tag{84}$$

donde $g^{-1}(z) = \frac{e^z}{1+e^z}$.

Anotación 11 (Interpretación de Coeficientes en un Modelo Logit)**.** *Note que β_j no es la contribución marginal de x_j a p, sino la contribución marginal de x_j a los log-odds $g(p)$.*

No se cuenta con una expresión explícita para los estimadores en modelo logit. La estimación se realiza mediante la maximización de una función de verosimilitud.

$$\hat{\beta}_0, \ldots, \hat{\beta}_p = \text{argmax}_{\hat{\beta}_0,\ldots,\hat{\beta}_p} \ell(\hat{\beta}_0, \ldots, \hat{\beta}_p) \tag{85}$$

$$\ell(\hat{\beta}_0, \ldots, \hat{\beta}_p) = \prod_{i:y_i=1} p_i \times \prod_{i:y_i=0} (1-p_i), \tag{86}$$

donde $p_i = \mathbb{P}(y_i = 1|x_i)$, y $\ell(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ es una función convexa. Una vez que los parámetros han sido estimados, se procede con la clasificación

$$\hat{p}_i = \frac{e^{\hat{\beta}_0 + \sum_{i=1}^{p} x_{ij}\hat{\beta}_j}}{1 + e^{\hat{\beta}_0 + \sum_{i=1}^{p} x_{ij}\hat{\beta}_j}}, \quad \hat{y}_i = \begin{cases} 1 & \text{if} \quad \hat{p}_i > H \\ 0 & \text{if} \quad \hat{p}_i \leq H, \end{cases}$$

donde H es una cota, e.g. $H = 0.5$.

El desempeno de un modelo de clasificación logit se puede evaluar con el ratio de error (5). Una inspección más profunda se puede lograr con la matriz de confusión:

	Condición Verdadera	
	P	N
Predicción P	Verdadero Positivo	Falso Positivo (Error Tipo I)
Predicción N	Falso Negativo (Error Tipo II)	Verdadero Negativo

Donde es claro que la matriz de confución es función de H. En base a esta información, podemos calcular el ratio de error como

$$ER = \frac{FP + FN}{VP + VN + FP + FN}, \tag{87}$$

que es equivalente a la ecuación (5). Otras métricas de desempeno importantes para métodos

de clasificación son

$$\text{Accuracy} = \frac{VP + VN}{VP + VN + FP + FN} \tag{88}$$

$$\text{Precision} = \frac{VP}{VP + FN} \tag{89}$$

$$\text{Recall} = \frac{VP}{VP + FP} \tag{90}$$

$$\text{F}_1 = \frac{2VP}{2VP + FN + FP}, \tag{91}$$

las cuales deben ser seleccionadas de acuerdo al problema de clasificación.

Anotación 12. *Note que las métricas (88), (89), (90), (91) son alternativas a maximizar al momento de resolver un problema de clasificación. En un sentido cualitativo, maximizar precision es equivalente a minimizar FN, y maximizar recall es equivalente a minimizar FP.*

Anotación 13. *F_1 es una métrica específica resultante de asignar $\beta = 1$ en la siguiente expresión*

$$F_\beta = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{\beta^2 Precision + Recall}, \quad \beta \geq 0, \tag{92}$$

donde Precision y Recall han sido definidas en (89) y (90) respectivamente. F_2 le da más peso a recall que precision; y $F_{0.5}$ le da más peso a precision que recall.

Anotación 14. *Note que las métricas (88), (89), (90), (91), son todas dependientes de la matriz de confusión, es decir son dependientes de la cota H seleccionada.*

Una métrica de desempeno independiente de la selección de H es el área debajo de la curva ROC. La curva ROC mapea el ratio de verdaderos positivos $RVP(H) : H \to y(x)$ en términos del ratio de falsos positivos $RFP(H) : H \to x$ en el plano

$$(RFP(H), RVP(H)), \quad \text{RVP(H)} = \frac{\text{VP(H)}}{\text{VP(H)} + \text{FN(H)}}, \quad \text{FPR(H)} = \frac{\text{FP(H)}}{\text{FP(H)} + \text{VN(H)}}, \tag{93}$$

con H como variable. La métrica de desempeno derivada de la curva ROC es su íntegral con respecto a H, a veces denominada AUC de la curva ROC.

$$\text{AUC}_{\text{ROC}} = \int_0^1 \text{RVP}(\text{RFP}^{-1}(x))dx, \tag{94}$$

donde es claro que $\text{AUC}_{\text{ROC}} \in [0, 1]$, y que mayores valores indican un mejor desempeno del clasificador.

3 Validación cruzada

Ejemplo 6 (Regresión Polinomial con Grado Desconocido)**.** *Considere el problema de aprendizaje supervisado (1), donde $f \in \mathcal{P}(d)$, es decir*

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots \beta_d x_i^d + \epsilon_i, \tag{95}$$

donde $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$. Note que el problema de estimación involucra no sólo a $\beta_0, \ldots, \beta_d$ sino también a d. El grado del polinomio d, en este caso, es llamado hiper-parámetro.

La validación cruzada es útil para seleccionar hiper-parámetros. La idea es separar algunas observaciones de la data de entrenamiento para validación. La idea es seleccionar el mejor modelo en data de validación como proxy del desempeno del modelo en data de prueba.

Anotación 15. *En el caso de regresión la métrica usual es el MSE. En el caso de clasificación esta métrica puede ser el ER, u alguna de las mencionadas anteriormente,* (88), (89), (90), (91).

3.1 Conjunto de Validación

Descomponer la data de entrenamiento en

$$\text{entrenamiento} = \text{ entrenamiento' (70\%) } + \text{ validación (30\%)},$$

para seleccionar los parámetros del modelo con el menor MSE/ER de "entrenamiento"; y los mejores hiper-parámetros usando el MSE/ER de "validación".

Anotación 16. *Note que este método tiene i) varianza grande, porque las muestras pueden ser muy diferentes; y ii) sesgo grande porque sólo se utiliza una parte de los datos.*

3.2 LOOCV

Descomponer la data de entrenamiento en

$$\text{entrenamiento} = \text{ entrenamiento' } + \text{ validación (1 obs) } (x_i, y_i),$$

para seleccionar los parámetros del modelo con el menor MSE/ER de "entrenamiento" y los mejores hiper-parámetros usando el MSE/ER de "validación". Repetir en un bucle seleccionando como observación de validación a cada uno de los puntos en el conjunto de datos, y calcular la métrica de desempeno

$$\text{CV}_{\text{LOOCV}} = \frac{1}{n}\sum_{i=1}^{n}\text{MSE}_i, \quad \text{METRIC}_i = \left(y_i - \hat{y}_i^{(-i)}\right)^2, \tag{96}$$

en un problema de regresión y

$$\text{CV}_{\text{LOOCV}} = \frac{1}{n}\sum_{i=1}^{n}\text{ER}_i, \quad \text{ER}_i = \mathbf{1}_{\left\{y_i \neq \hat{y}_i^{(-i)}\right\}}, \tag{97}$$

en un problema de clasificación.

Anotación 17. *Note que este método tiene: i) varianza menor que en el enfoque de conjunto de validación; ii) sesgo menor que en el enfoque de validación; iii) pero requiere n ajustes y es computacionalmente costoso.*

3.3 k Dobleces

Idea principal: descomponer la data de entrenamiento en

$$\text{entrenamiento } = \text{ entrenamiento' } + \text{ validación (1 bloque)},$$

para seleccionar los parámetros del modelo con el menor MSE/ER de "entrenamiento" y los mejores hiper-parámetros usando el MSE/ER de "validación". Repetir en un bucle seleccionando como bloque de validación a cada uno de los k bloques disjuntos en la data. Sea $\mathcal{S}_i$ la colección de índices del bloque i, se debe calcular

$$\mathrm{CV}_{\text{K-fold}} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i, \quad \mathrm{MSE}_i = \frac{k}{n}\sum_{j\in\mathcal{S}_i}\left(y_j - \hat{y}_j^{(-\mathcal{S}_i)}\right)^2, \tag{98}$$

para un problema de regresión, y

$$\mathrm{CV}_{\text{K-fold}} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{ER}_i, \quad \mathrm{ER}_i = \frac{k}{n}\sum_{j\in\mathcal{S}_i}\mathbf{1}_{\left\{y_j \neq \hat{y}_j^{(-\mathcal{S}_i)}\right\}}, \tag{99}$$

para un problema de clasificación.

Anotación 18. *Note que este método tiene: i) mayor varianza que LOOCV, ii) mayor sesgo que LOOCV; iii) más velocidad que LOOCV.*

4 Reducción Dimensional

Considere una matriz de diseño $\boldsymbol{X} \in \mathbb{R}^{n\times p}$. Si p es grande, es muy probable que algunas de sus variables sean combinaciones lineales de las otras y agreguen poco a la predicción de la respuesta.

4.1 Métodos de Selección de Variables

Considere el modelo de regresión lineal múltiple (41) en notación escalar

$$y_i = \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij} + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

donde $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$. Típicamente cuando p es grande el estimado $\hat{f}(x_0)$ tiene una varianza grande (por el efecto de la dependencia lineal entre los predictores). Por tanto es conveniente eliminar variables usando algún criterio de relevancia en el modelo.

Algorithm 1 Selección del Mejor Subconjunto (BSS)

Denote al modelo con cero variables como $\mathcal{M}_0$
while $i = 1, 2, \ldots, p$ **do**
  Ajuste los $\binom{p}{i}$ modelos conteniendo i predictores.
  Seleccione el modelo con el menor MSE/ER y denótelo $\mathcal{M}_i$.
Seleccione el modelo óptimo entre $\mathcal{M}_0, \ldots, \mathcal{M}_p$ usando algún criterio.

Proposición 8. *i) El número de modelos evaluados en el algoritmo FWD es igual que el número de modelos evaluados en el algoritmo BKW; ii) El número de modelos evaluados en el algoritmo BSS es mayor o igual que el de los algoritmos FWD (o BKW).*

Algorithm 2 Selección Paso-a-Paso hacia Adelante (FWD)

Denote al modelo con cero variables como $\mathcal{M}_0$
while $i = 0, 1, 2, \ldots, p-1$ **do**
  Ajuste de los $p - i$ modelos que resultan de agregar un predictor a $\mathcal{M}_i$.
  Seleccione el modelo con el menor MSE/ER y denótelo $\mathcal{M}_{i+1}$.
Seleccione el modelo óptimo entre $\mathcal{M}_0, \ldots, \mathcal{M}_p$ usando algún criterio.

Algorithm 3 Selección Paso-a-Paso hacia Atrás (BKW)

Denote el modelo que utiliza todas las variables como $\mathcal{M}_p$.
while $i = p - 1 \ldots, 1$ **do**
  Ajuste de los $i - 1$ modelos resultantes de sustraer un predictor de $\mathcal{M}_i$.
  Seleccione el modelo con el menor MSE/ER y denótelo $\mathcal{M}_{i-1}$.
Seleccione el modelo óptimo entre $\mathcal{M}_0, \ldots, \mathcal{M}_p$ usando algún criterio.

Demostración. El número de posibles modelos evaluados en BSS es:

$$\sum_{i=0}^{p} \binom{p}{i} = \binom{p}{0} + \binom{p}{1} + \cdots + \binom{p}{p} = 2^p,$$

mientras que el número de modelos evaluados en FWD es

$$1 + p + (p-1) + \cdots + 1 = 1 + \frac{p(p+1)}{2}.$$

Es fácil verificar que número de modelos evaluados en BKW es exáctamente igual que el correspondiente a FWD. Finalmente, es inmediato que $1 + p(p+1)/2 \leq 2^p$, lo cual prueba la proposición. □

Selección del Modelo Óptimo

En los algoritmos BSS, FWD y BKW, no se ha especificado cómo seleccionar el mejor modelo en el último paso. Para seleccionar el mejor modelo se puede usar el error de predicción de validación cruzada.

4.2 Regresión con Componentes Principales

Considere una matriz estandarizada[3] $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, con filas dadas por las observaciones $\boldsymbol{x}_j \in \mathbb{R}^p$, y un vector unitario $\boldsymbol{v} \in \mathbb{R}^p$.

4.2.1 Paso 1: Reducción Dimensional

Considere la proyección escalar de $\boldsymbol{x}$ sobre $\boldsymbol{v}$ dada por

$$s_j = \cos(\theta) \times \|\boldsymbol{x}\|_2 = \left(\frac{\boldsymbol{x}^\top \boldsymbol{v}}{\|\boldsymbol{x}\|_2 \|\boldsymbol{v}\|_2} \right) \times \|\boldsymbol{x}\|_2 = \frac{\boldsymbol{x}^\top \boldsymbol{v}}{\|\boldsymbol{v}\|_2} = \frac{\boldsymbol{x}^\top \boldsymbol{v}}{\boldsymbol{v}^\top \boldsymbol{v}} = \boldsymbol{x}^\top \boldsymbol{v}. \tag{100}$$

[3] Una matriz de diseno estandarizada es aquella en la que cada columna de $\boldsymbol{X}$ tiene media cero y desviación estandar unitaria.

Adicionalmente, si h_j representa la distancia entre $\boldsymbol{x}_j$ y $\boldsymbol{v}$, es obvio que $h = \sqrt{\|\boldsymbol{x}_j\|_2^2 - s^2}$, es decir que el valor que minimiza h_j es el valor que maximiza s_j y viceversa. La maximización la magnitud de s_j corresponde a maximizar s_j^2, y el problema de maximizar la suma de las magnitudes de cada s_j corresponde a maximizar

$$\sum_{j=1}^{p} (\boldsymbol{x}_j^\top \boldsymbol{v})^2 = \sum_{j=1}^{p} \boldsymbol{v}^\top \boldsymbol{x}_j \boldsymbol{x}_j^\top \boldsymbol{v} = \boldsymbol{v}^\top \sum_{j=1}^{p} (\boldsymbol{x}_j \boldsymbol{x}_j^\top) \boldsymbol{v} = \boldsymbol{v}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{v}. \tag{101}$$

Considerando la restricción sobre el vector unitario, el problema de maximización es

$$\boldsymbol{v}^* = \text{argmax}_{\boldsymbol{v}} \left\{ \boldsymbol{v}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{v} : \|\boldsymbol{v}\|_2^2 = 1 \right\} = \text{argmax}_{\boldsymbol{v}}\, \ell(\boldsymbol{v}), \tag{102}$$

donde $\ell(\cdot)$ es la función lagrangiana

$$\ell(\boldsymbol{v}) = \boldsymbol{v}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{v} - \lambda(\boldsymbol{v}^\top \boldsymbol{v} - 1) \tag{103}$$

La solución de (103) viene dada por el vector $\boldsymbol{v}$ que resuelve las condiciones de primer orden

$$\frac{\partial \ell(\boldsymbol{v})}{\partial \boldsymbol{v}} = 2\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{v} - 2\lambda \boldsymbol{v} = 0 \tag{104}$$

$$\frac{\partial \ell(\boldsymbol{v})}{\partial \lambda} = \boldsymbol{v}^\top \boldsymbol{v} - 1 = 0. \tag{105}$$

De estas ecuaciones es claro que

$$\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{v} = \lambda \boldsymbol{v}, \quad \boldsymbol{v}^\top \boldsymbol{v} = 1, \tag{106}$$

y por tanto $\boldsymbol{v}$ es un vector propio de $\boldsymbol{X}^\top \boldsymbol{X}$ y λ es el valor propio asociado a él. De manera más general, sabemos que (106) se satisface para los p pares de vectores propios $\boldsymbol{v}_j$ y correspondientes valores propios λ_j, es decir

$$\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{v}_j = \lambda_j \boldsymbol{v}_j, \quad \boldsymbol{v}_j^\top \boldsymbol{v}_j = 1. \tag{107}$$

Adicionalmente, dado que $\boldsymbol{X}^\top \boldsymbol{X}$ es real simétrica, se sabe que sus vectores propios son ortogonales, i.e. $\boldsymbol{v}_j^\top \boldsymbol{v}_i = \mathbf{1}_{i=j}$, y que sus valores propios son todos innegativos. Sin pérdida de generalidad, podemos ordenar los valores propios de manera descendente

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0, \tag{108}$$

y escribir la matriz de vectores propios asociada a este orden $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p]$. En consecuencia

$$\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{V} \boldsymbol{\Lambda} \boldsymbol{V}^\top = \sum_{i=1}^{p} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top. \tag{109}$$

El truncamiento de la suma en (109) hasta $r < p$, da lugar a la aproximación

$$\boldsymbol{X}^\top \boldsymbol{X} \approx \sum_{i=1}^{r} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top, \tag{110}$$

la cual utiliza únicamente los r primeros vectores propios. Note que la expresión (110) representa una descomposición aditiva de la matriz de covarianzas de $\boldsymbol{X}$. En este sentido se dice que el truncamiento en el componente r permite explicar el

$$\left(\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \right) \times 100\% \tag{111}$$

de la varianza total.

4.2.2 Paso 2: Construcción de una Regresión de Baja Dimensión

Dados los r primeros vectores unitarios $\boldsymbol{v}_j$, podemos calcular la proyección escalar de la j-ésima observación $\boldsymbol{x}_j$ sobre cada uno de ellos

$$\boldsymbol{z}_j = [\boldsymbol{x}_j^\top \boldsymbol{v}_1 \quad \boldsymbol{x}_j^\top \boldsymbol{v}_2 \quad \dots \quad \boldsymbol{x}_j^\top \boldsymbol{v}_r] = \boldsymbol{x}_j^\top \tilde{\boldsymbol{V}}, \tag{112}$$

y denotarlo j-ésimo componente principal. De manera más general se dice que

$$\boldsymbol{Z} = [\boldsymbol{X}\boldsymbol{v}_1, \boldsymbol{X}\boldsymbol{v}_2, \dots, \boldsymbol{X}\boldsymbol{v}_r] = \boldsymbol{X}\tilde{\boldsymbol{V}}, \tag{113}$$

contiene a los $r < p$ componentes principales de la data $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ en sus r columnas. La regresión por componentes principales de una variable respuesta Y explicada por las variables latentes $Z_1, Z_2, \dots, Z_p$ corresponde a la regresión lineal

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{114}$$

bajo los supuestos habituales.

Anotación 19 (Limitaciones de PCR)**.** *La selección de variables en PCR sólo considera la estructura de la matriz de diseno, pero no considera la relación entre los componentes extraídos y la respuesta.*

Ejemplo 7 (Correlación)**.** *Considere la matriz de correlación*

$$\boldsymbol{X}^\top \boldsymbol{X} = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}, \tag{115}$$

calcule los vectores unitarios $\boldsymbol{v}_1$ y $\boldsymbol{v}_2$ y el porcentaje de la varianza que explica el primer componente principal.

Solución. La descomposición de valores propios da como resultado:

$$\boldsymbol{X}^\top \boldsymbol{X} = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 3/2 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}, \tag{116}$$

donde es claro que el primer componente principal cubre el 75% de la varianza. □

4.3 Mínimos Cuadrados Parciales

Considere una matriz estandarizada $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, con filas dadas por las observaciones $\boldsymbol{x}_j \in \mathbb{R}^p$, y un vector unitario $\boldsymbol{v} \in \mathbb{R}^p$. Usaremos las siguientes dos representaciones

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{bmatrix}, \quad \boldsymbol{X} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_p], \tag{117}$$

cuando sea conveniente.

4.3.1 Paso 1: Encontrar la variable más Correlacionada

Considere el primer componente de la descomposición dado por,

$$\boldsymbol{z}_1 = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^\top \boldsymbol{v} \\ \boldsymbol{x}_2^\top \boldsymbol{v} \\ \vdots \\ \boldsymbol{x}_n^\top \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{bmatrix} \boldsymbol{v} = \boldsymbol{X}\boldsymbol{v}, \tag{118}$$

es decir con una expresión similar a (113), que representa los componentes principales en PCR.

Nuestro objetivo será maximizar el cuadrado de la covarianza entre $\boldsymbol{z}$ y la data $\boldsymbol{y}$,

$$\mathbb{C}\text{ov}^2(\boldsymbol{z}_1, \boldsymbol{y}) = \mathbb{C}\text{ov}^2(\boldsymbol{X}\boldsymbol{v}, \boldsymbol{y}) = (\boldsymbol{v}^\top \boldsymbol{X}^\top \boldsymbol{y})(\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{v}) = \boldsymbol{v}^\top (\boldsymbol{X}^\top \boldsymbol{y}\boldsymbol{y}^\top \boldsymbol{X})\boldsymbol{v}, \tag{119}$$

con la restricción de que $\boldsymbol{v}$ es un vector unitario. Es decir, el problema se puede escribir como

$$\boldsymbol{v}^* = \text{argmax}_{\boldsymbol{v}} \left\{ \boldsymbol{v}^\top (\boldsymbol{X}^\top \boldsymbol{y}\boldsymbol{y}^\top \boldsymbol{X})\boldsymbol{v} : \|\boldsymbol{v}\|_2 = 1 \right\} = \text{argmax}_{\boldsymbol{v}} \ell(\boldsymbol{v}, \lambda), \tag{120}$$

donde $\ell(\boldsymbol{v}, \lambda)$ es la función lagrangiana

$$\ell(\boldsymbol{v}, \lambda) = \boldsymbol{v}^\top (\boldsymbol{X}^\top \boldsymbol{y}\boldsymbol{y}^\top \boldsymbol{X})\boldsymbol{v} - \lambda(1 - \boldsymbol{v}^\top \boldsymbol{v}) \tag{121}$$

La solución de (121) viene dada por el vector $\boldsymbol{v}$ que resuelve las condiciones de primer orden

$$\frac{\partial \ell(\boldsymbol{v})}{\partial \boldsymbol{v}} = 2(\boldsymbol{X}^\top \boldsymbol{y}\boldsymbol{y}^\top \boldsymbol{X})\boldsymbol{v} - 2\lambda \boldsymbol{v} = 0 \tag{122}$$

$$\frac{\partial \ell(\boldsymbol{v})}{\partial \lambda} = \boldsymbol{v}^\top \boldsymbol{v} - 1 = 0. \tag{123}$$

De estas ecuaciones es claro que

$$(\boldsymbol{X}^\top \boldsymbol{y}\boldsymbol{y}^\top \boldsymbol{X})\boldsymbol{v} = \lambda \boldsymbol{v}, \quad \boldsymbol{v}^\top \boldsymbol{v} = 1, \tag{124}$$

que se satisface para $\boldsymbol{v}^* = \boldsymbol{X}^\top \boldsymbol{y} / \|\boldsymbol{X}^\top \boldsymbol{y}\|_2$,

$$(\boldsymbol{X}^\top \boldsymbol{y}\boldsymbol{y}^\top \boldsymbol{X}) \frac{\boldsymbol{X}^\top \boldsymbol{y}}{\|\boldsymbol{X}^\top \boldsymbol{y}\|_2} = (\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{y}) \frac{\boldsymbol{X}^\top \boldsymbol{y}}{\|\boldsymbol{X}^\top \boldsymbol{y}\|_2} = \lambda \frac{\boldsymbol{X}^\top \boldsymbol{y}}{\|\boldsymbol{X}^\top \boldsymbol{y}\|_2}, \tag{125}$$

en cuyo caso $\lambda^* = \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{y}$.

4.3.2 Paso 2: Predecir $\boldsymbol{y}$ con la Variable Score $\boldsymbol{z}_1$

Dado que $\boldsymbol{z}_1$ maximiza la covarianza con $\boldsymbol{y}$, el mejor predictor de $\boldsymbol{y}$ viene dado por

$$\hat{\boldsymbol{y}} = \hat{\theta} \boldsymbol{z}_1, \quad \hat{\theta} = \frac{\mathbb{C}\text{ov}(\boldsymbol{z}_1, \boldsymbol{y})}{\text{Var}(\boldsymbol{z}_1)} = \frac{\boldsymbol{v}^{*\top} \boldsymbol{X}^\top \boldsymbol{y}}{\boldsymbol{v}^{*\top} \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{v}^*} \tag{126}$$

4.3.3 Paso 3: Repetir el Proceso para el Residuo

Construimos una nueva respuesta $\tilde{\boldsymbol{y}}$, que sea ortogonal al estimador de la respuesta calculado en el paso anterior $\hat{\boldsymbol{y}}$. Note que

$$\begin{aligned} \hat{\boldsymbol{y}} &= \boldsymbol{H}\boldsymbol{y} && (127)\\ \tilde{\boldsymbol{y}} &= (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{y}, && (128) \end{aligned}$$

y es claro que $\hat{\boldsymbol{y}}$ es ortogonal a $\tilde{\boldsymbol{y}}$:

$$\begin{aligned} \hat{\boldsymbol{y}}^\top \tilde{\boldsymbol{y}} &= \boldsymbol{y}^\top \boldsymbol{H}^\top (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{y} && (129)\\ &= \boldsymbol{y}^\top \boldsymbol{H}(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{y} && (130)\\ &= \boldsymbol{y}^\top (\boldsymbol{H} - \boldsymbol{H}^2)\boldsymbol{y} && (131)\\ &= \boldsymbol{y}^\top (\boldsymbol{H} - \boldsymbol{H})\boldsymbol{y} && (132)\\ &= 0, && (133) \end{aligned}$$

es decir, $\tilde{\boldsymbol{y}}$ es ortogonal a $\hat{\boldsymbol{y}}$.

Recuerde que la proyección del vector $\boldsymbol{e}_j$ sobre $\boldsymbol{z}_1$ viene dada por

$$\text{Proj}_{\boldsymbol{z}_1}(\boldsymbol{e}_j) = \left(\frac{\boldsymbol{e}_j^\top \boldsymbol{z}_1}{\boldsymbol{z}_1^\top \boldsymbol{z}_1}\right)\boldsymbol{z}_1, \tag{134}$$

que se puede re-escribir como

$$\text{Proj}_{\boldsymbol{z}_1}(\boldsymbol{e}_j) = \left(\frac{\boldsymbol{e}_j^\top \boldsymbol{z}_1}{\boldsymbol{z}_1^\top \boldsymbol{z}_1}\right)\boldsymbol{z}_1 = \left(\frac{\boldsymbol{z}_1 \boldsymbol{z}_1^\top}{\|\boldsymbol{z}_1\|_2^2}\right)\boldsymbol{e}_j = \boldsymbol{Q}\boldsymbol{e}_j, \quad \boldsymbol{Q} = \frac{\boldsymbol{z}_1 \boldsymbol{z}_1^\top}{\|\boldsymbol{z}_1\|_2^2}. \tag{135}$$

Proposición 9 (Matriz de Proyección)**.** $\boldsymbol{Q}$ *es una matriz de proyección ortogonal.*

Demostración. $\boldsymbol{Q}$ es una matriz de proyección si $\boldsymbol{Q} = \boldsymbol{Q}^2$. Note que

$$\boldsymbol{Q}^2 = \frac{\boldsymbol{z}_1 \boldsymbol{z}_1^\top}{\|\boldsymbol{z}_1\|_2^2}\frac{\boldsymbol{z}_1 \boldsymbol{z}_1^\top}{\|\boldsymbol{z}_1\|_2^2} = \frac{\boldsymbol{z}_1^\top \boldsymbol{z}_1}{\|\boldsymbol{z}_1\|_2^2}\frac{\boldsymbol{z}_1 \boldsymbol{z}_1^\top}{\|\boldsymbol{z}_1\|_2^2} = \frac{\|\boldsymbol{z}_1\|_2^2}{\|\boldsymbol{z}_1\|_2^2}\frac{\boldsymbol{z}_1 \boldsymbol{z}_1^\top}{\|\boldsymbol{z}_1\|_2^2} = \frac{\boldsymbol{z}_1 \boldsymbol{z}_1^\top}{\|\boldsymbol{z}_1\|_2^2} = \boldsymbol{Q}, \tag{136}$$

Adicionalmente, $\boldsymbol{Q}$ es una matriz de proyección ortogonal si $\boldsymbol{Q} = \boldsymbol{Q}^\top$. En este caso

$$\boldsymbol{Q} = \frac{\boldsymbol{z}_1 \boldsymbol{z}_1^\top}{\|\boldsymbol{z}_1\|_2^2} = \boldsymbol{Q}^\top, \tag{137}$$

por lo tanto $\boldsymbol{Q}$ es una matriz de proyección ortogonal. □

Adicionalmente note que

$$[\text{Proj}_{\boldsymbol{z}_1}(\boldsymbol{e}_1) \quad \ldots \quad \text{Proj}_{\boldsymbol{z}_1}(\boldsymbol{e}_p)] = [\boldsymbol{Q}\boldsymbol{e}_1 \quad \ldots \quad \boldsymbol{Q}\boldsymbol{e}_p] = \boldsymbol{Q}\boldsymbol{X}. \tag{138}$$

De modo que la información ortogonal a $\boldsymbol{Q}\boldsymbol{X}$ es

$$\tilde{\boldsymbol{X}} = (\boldsymbol{I}_n - \boldsymbol{Q})\boldsymbol{X}. \tag{139}$$

El segundo vector score $\boldsymbol{z}_2$ se puede obtener repitiendo el paso 1 con $\tilde{\boldsymbol{X}}$ y $\tilde{\boldsymbol{y}}$ reemplazando a $\boldsymbol{X}$ y $\boldsymbol{y}$ respectivamente. Usando los vectores $\boldsymbol{z}_j$, $j = 1, \ldots, m$, $r < p$, se obtiene el predictor

$$\hat{\boldsymbol{y}} = \boldsymbol{Z}\hat{\boldsymbol{\theta}} \tag{140}$$

5 Modelos Lineales Regularizados

Anotación 20. *En esta sección, igual que en la sección anterior, consideraremos estandarizar la data (la matriz de diseno $\boldsymbol{X}$ y la respuesta yvec) antes de la aplicación del método.*

Anotación 21. *Recuerde que la norma ℓ de un vector $\boldsymbol{x} \in \mathbb{R}^n$ es definida como*

$$\|\boldsymbol{x}\|_\ell = \left(\sum_{i=1}^{n} |x_i|^\ell \right)^{\frac{1}{\ell}}. \tag{141}$$

En esta sección haremos uso de los casos en los que $\ell = 1$ y $\ell = 2$.

Teorema 1 (SVD Reducido)**.** *Toda matriz $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, tiene una descomposición de valores singulares reducida*

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top, \tag{142}$$

donde $\boldsymbol{U} \in \mathbb{R}^{n\times p}$ y $\boldsymbol{V} \in \mathbb{R}^{p\times p}$ son matrices ortogonales. Es decir: $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_p$, $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}_p$, y $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ es una matriz diagonal. Adicionalmente, las columnas de $\boldsymbol{U}$ son denominadas vectores singulares por izquierda y las de $\boldsymbol{V}$ vectores singulares por derecha.

Proposición 10. *Note que si $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, luego*

$$\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^\top \tag{143}$$

$$\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^\top \tag{144}$$

Demostración. Trivial. □

Ejemplo 8 (Matriz de Proyección en Regresión Lineal)**.** *Considere*

$$\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}, \quad \boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top. \tag{145}$$

Para $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, la expresión anterior se simplifica a

$$\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}, \quad \boldsymbol{H} = \boldsymbol{U}\boldsymbol{U}^\top. \tag{146}$$

Note que $\boldsymbol{U}^\top\boldsymbol{y}$ es un elemento en el span generado por las columnas $\boldsymbol{u}_j$, considerando como pesos los elementos de la respuesta y_j.

En la sección 4 se vieron modelos de reducción dimensional para reducir la varianza del estimador de $\hat{\boldsymbol{y}}$ (incrementando el sesgo). Una alternativa a esto es "encoger" los parámetros del modelo. Es decir, transformar el problema de regresión lineal en

$$\hat{\boldsymbol{\beta}}^* = \text{argmin}_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 : \|\boldsymbol{\beta}\|_\ell^\ell \leq s \right\}, \tag{147}$$

que es un modelo lineal "regularizado", o "penalizado". Adicionalmente, si $\ell = 2$ nombramos al modelo regresión Ridge, mientras que si $\ell = 1$, lo nombramos regresión LASSO (Least Absolute Shrinkage and Selection Operator). Es fácil ver que el problema (148) es equivalente a

$$\hat{\boldsymbol{\beta}}^* = \text{argmin}_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_\ell^\ell \right\}, \tag{148}$$

es decir, que para cada λ hay un valor de s que permite que ambas formulaciones den los mismos resultados.

5.1 Regresión Ridge

El problema (148) para $\ell = 2$ se puede re-escribir como

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2 \right\}, \quad \lambda > 0, \tag{149}$$

que tiene una solución dada por:

$$\hat{\boldsymbol{\beta}} \;=\; \left[(\boldsymbol{X}^\top \boldsymbol{X}) + \lambda \boldsymbol{I}_p \right]^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \tag{150}$$

el cual puede calcularse incluso si $\boldsymbol{X}^\top \boldsymbol{X}$ es singular para algunos valores de $\lambda > 0$.

Anotación 22. *Note que si $\lambda = 0$, el problema se reduce al problema de regresión lineal simple. Si $\lambda > 0$ se genera un costo adicional en (148) que incrementa el sesgo (reduce la varianza). En particular, si $\lambda \to \infty$, todo el control viene dado por la restricción.*

Anotación 23. *Note que*

$$\hat{\boldsymbol{y}} = \boldsymbol{H}_\lambda \boldsymbol{y}, \quad \boldsymbol{H}_\lambda = \boldsymbol{X} \left[(\boldsymbol{X}^\top \boldsymbol{X}) + \lambda \boldsymbol{I}_p \right]^{-1} \boldsymbol{X}^\top, \tag{151}$$

donde $\boldsymbol{H}_\lambda$ no es una matriz de proyección.

Proposición 11. *Los grados de libertad del modelo se pueden calcular como*

$$df(f_\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}, \tag{152}$$

donde d_j son los elementos de la diagonal principal de $\boldsymbol{D}$ cuando $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$.

Demostración. Reemplace $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ en

$$\boldsymbol{H}_\lambda = \boldsymbol{X} \left[(\boldsymbol{X}^\top \boldsymbol{X}) + \lambda \boldsymbol{I}_p \right]^{-1} \boldsymbol{X}^\top = \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{D}\boldsymbol{U}^\top = \boldsymbol{U}\text{diag}\left\{ \frac{d_j^2}{d_j^2 + \lambda} \right\} \boldsymbol{U}^\top, \tag{153}$$

de modo que

$$\text{df}(f_\lambda) = \text{tr}(\boldsymbol{H}_\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}. \tag{154}$$

□

Anotación 24. *Note que*

$$\hat{\boldsymbol{y}} = \sum_{j=1}^{p} \left(\frac{d_j^2}{d_j^2 + \lambda} \boldsymbol{u}_j \boldsymbol{u}_j^\top \right) \boldsymbol{y}, \tag{155}$$

donde es claro que se agrega una mayor penalización (relativamente) a los casos en los que d_j es más pequeno. Se puede interpretar d_j^2 como los valores propios en $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^\top$. De manera que se agrega más penalización a los vectores del span de $\boldsymbol{U}$ que son "menos importantes" en la matriz de diseno.

5.2 Regresión Lasso

El problema (148) para $\ell = 1$ se puede re-escribir como

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 \right\}, \quad \lambda > 0, \tag{156}$$

que tiene una solución dada por el $\hat{\boldsymbol{\beta}}$ que satisface:

$$(\boldsymbol{X}^\top \boldsymbol{X})\boldsymbol{\beta} + \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \quad = \quad \boldsymbol{X}^\top \boldsymbol{y}, \tag{157}$$

que es un sistema no-lineal de ecuaciones.

Anotación 25. *Note que LASSO penaliza una norma diferente de $\boldsymbol{\beta}$, y que con esta norma se seleccionan algunos β's exactamente iguales a cero.*

Anotación 26 (Selección de Hiper-parámetros)**.** *En RR y LASSO, los hiper-parámetros λ se han considerado como dados. En la realidad estos parámetros deben ser estimados utilizando la data mediante validación cruzada, por ejemplo usando k-dobleces.*

5.3 Regresión Elastic-Net

El problema es una combinación directa de RR y LASSO. Es decir, el problema se puede escribir como

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\alpha \|\boldsymbol{\beta}\|_2^2 + \lambda(1-\alpha)\|\boldsymbol{\beta}\|_1 \right\}, \quad \lambda > 0, \tag{158}$$

donde $\alpha \in [0, 1]$ es, por lo general, seleccionado por el modelador.

6 Modelos Aditivos Generalizados

Un modelo aditivo generalizado retira la asumsión de linealidad del modelo lineal generalizado (159), y lo generaliza de la siguiente manera:

$$g(\mu_i) = \beta_0 + \sum_{i=1}^{p} f_j(x_{ij}), \ i = 1, \ldots, n, \tag{159}$$

donde g es una función vinculante, $u_i = \mathbb{E}[y_i|x_i]$, $f_1, \ldots, f_p$ son funciones "suaves", i.e. primera y segunda derivada continua. Las funciones f_j son modeladas utilizando diferentes métodos.

6.1 Modelos Univariados

Considere (159) para g la función identidad, es decir $y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n$, donde $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$, y $\epsilon_i \perp x_i$. Es usual la siguiente formulación $\mathbb{E}[y_i|x_i] = f(x_i)$, que indica que la media condicional de y_i dado x_i es una función $f(x_i)$. En general, la asumsión central de los modelos de regresión es que

$$f \in \mathcal{S}, \quad \mathcal{S} \in \text{span}\{b_1, b_2, \ldots, b_p\}, \tag{160}$$

donde $\mathcal{S}$ es cierto espacio funcional generado por una base $\{b_1, \ldots, b_p\}$.

Anotación 27. *En el caso especial de los modelos de regresión polinomial se cumple que $\mathcal{S}$ es un espacio de funciones polinomiales y que la base que genera el espacio viene dada por*

$$\{1, x, x^2, \ldots, x^p\}. \tag{161}$$

El caso de regresión lineal es el caso especial en el que $p = 1$.

6.1.1 Serie de Fourier

Considere (160), para el caso en el que $\mathcal{S}$ es un espacio de funciones continuas y periódicas, generado por la base

$$\{\sin(\pi x_i), \sin(2\pi x_i), \ldots, \sin(p\pi x_i), \cos(\pi x_i), \cos(2\pi x_i), \ldots, \cos(p\pi x_i)\}, \tag{162}$$

usualmente llamada "base de Fourier".

Teorema 2 (Teorema de Fourier)**.** *Considere una función periódica $f(x_i)$, $x_i \in [0, 1]$. Bajo ciertas condiciones de regularidad f admite la siguiente representación:*

$$f(x_i) = const + \sum_{j=1}^{\infty} \alpha_j \sin(j\pi x_i) + \beta_j \cos(j\pi x_i), \tag{163}$$

donde $\alpha_j, \beta_j \in \mathbb{R}$, $j \in \mathbb{N}$.

Note que los parámetros α_j, β_j en el teorema 2 controlan la amplitud de las funciones trigonométricas, mientras j controla su periodo/fase (mientras j es más grande la función es más oscilatoria). Esto sugiere el modelo:

$$\hat{f}(x_i) = \text{const} + \sum_{j=1}^{p} \hat{\alpha}_j \sin(j\pi x_i) + \hat{\beta}_j \cos(j\pi x_i), \tag{164}$$

otros ejemplos son las bases polinomiales, las bases b-spline, o las bases wavelet.

6.1.2 Funciones Kernel

Considere (160), para el caso en el que $\mathcal{S}$ es un espacio de funciones de polinomios locales, centrados en x_i, generado por la base

$$\{1, (x_j - x_i), (x_j - x_i)^2, \ldots, (x_j - x_i)^p\}. \tag{165}$$

En este sentido, la función $f(x_i)$ toma la forma

$$f(x_i) = \beta_0 + \beta_1 z + \beta_2 z^2 + \cdots + \beta_p z^p, \quad z = x_j - x_i, \tag{166}$$

es decir la función f evaluada en el punto x_i es un polinomio en $x_j - x_i$.

Definición 1 (Función Kernel)**.** *Una función kernel es una función $K_h(x; x_0)$ que cumple con la condición $\int_{-\infty}^{\infty} K_h(x; x_0)dx = 1$. Esta función asigna pesos a los puntos x al rededor de x_0.*

Ejemplo 9 (Funciones Kernel)**.** *Algunos ejemplos de funciones Kernel se presentan a continuación:*

Nombre	Función	Soporte
Uniforme	$K(x;x_0) = \frac{1}{2}$	$\lvert x - x_0\rvert \leq 1$
Triangular	$K(x;x_0) = (1 - \lvert x - x_0\rvert)$	$\lvert x - x_0\rvert \leq 1$
Epanechnikov	$K(x;x_0) = \frac{3}{4}[1 - (x - x_0)^2]$	$\lvert x - x_0\rvert \leq 1$
Gaussiano	$K_h(x;x_0) = \frac{1}{\sqrt{2\pi}h} \exp\{-\frac{1}{2}\left(\frac{x-x_0}{h}\right)^2\}$	$\lvert x - x_0\rvert \in \mathbb{R}$

Anotación 28. *Note que la función kernel Gaussiana presenta un parámetro adicional h. Este parámetro es denominada "bandwith", y controla la velocidad de decaimiento de la función con respecto a* $\min x \leq x_0 \leq \max x$.

Los coeficientes óptimos de la regresión utilizando funciones kernel son la solución de

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{W}^{1/2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\|_2^2 \right\}, \tag{167}$$

que es un problema de mínimos cuadrados ponderados, donde

$$\boldsymbol{X} = \begin{bmatrix} 1 & (x_1 - x_i) & \dots & (x_1 - x_i)^p \\ 1 & (x_2 - x_i) & \dots & (x_2 - x_i)^p \\ \vdots & \vdots & \dots & \vdots \\ 1 & (x_n - x_i) & \dots & (x_n - x_i)^p \end{bmatrix}; \quad \boldsymbol{W} = \begin{bmatrix} K_h(x_1;x_i) & 0 & \dots & 0 \\ 0 & K_h(x_2;x_i) & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & K_h(x_n;x_i) \end{bmatrix} \tag{168}$$

donde $K_h(\cdot)$ es una función de pesos kernel, e.g. kernel Gaussiano. Es fácil ver que la solución del problema (167) viene dada por

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{y}. \tag{169}$$

Ejemplo 10 (Estimador Nadaraya-Watson)**.** *El estimador Nadaraya-Watson se obtiene de reemplazar $p = 0$ en la expresión de $\boldsymbol{X}$ en (168). En este caso particular, el estimador tiene la forma*

$$\hat{f}(x_i; 0, h) = \frac{\sum_{j=1}^n K_h(x_j - x_i) y_j}{\sum_{j=1}^n K_h(x_j - x_i)}. \tag{170}$$

6.1.3 Splines Penalizados

Considere (160), para el caso en el que $\mathcal{S}$ es un espacio de funciones de "suaves" $f \in \mathbb{C}^2$, generadas por la base

$$\{b_1, b_2, \dots, b_p\}. \tag{171}$$

En este sentido, la función $f(x_i)$ toma la forma $f(x_i) = \sum_{j=1}^p b_j(x_i)\theta_j$. Los coeficientes $\boldsymbol{\theta}$ óptimos son la solución del siguiente problema de optimización

$$\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \left\{ \|\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\theta}\|_2^2 + \lambda \boldsymbol{\theta}^\top \boldsymbol{D} \boldsymbol{\theta} \right\}, \quad \lambda \geq 0, \tag{172}$$

donde $\boldsymbol{B} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\theta} \in \mathbb{R}^p$, y típicamente $p \approx n$. Adicionalmente, el término penalizado tiene una forma especial:

$$\boldsymbol{\theta}^\top \boldsymbol{D} \boldsymbol{\theta} = \int_{-\infty}^{\infty} \left\{ f^{(2)}(x) \right\}^2 dx. \tag{173}$$

La solución de (172), viene dada por:

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{D}\right)^{-1} \boldsymbol{B}^\top \boldsymbol{y}, \tag{174}$$

$$\hat{\boldsymbol{f}} = \boldsymbol{S}_\lambda \boldsymbol{y}, \quad \boldsymbol{S}_\lambda = \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{D}\right)^{-1} \boldsymbol{B}^\top. \tag{175}$$

Ejemplo 11 (Polinomios Truncados)**.** *Considere el siguiente caso especial del modelo* (172) *dado por*

$$f(x_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_i^j + \sum_{\ell=1}^{r} \alpha_\ell (x_i - \kappa_\ell)_+^q, \tag{176}$$

donde $x_+ = \max\{x, 0\}$ *para cualquier* $x \in \mathbb{R}$*, y los* κ_ℓ*,* $1 \leq \ell \leq r$*, son valores predeterminados, usualmente equidistantes en el rango de* x*, i.e.:* $\min x \leq \kappa_1 \leq \kappa_2 \leq \cdots \leq \kappa_r \leq \max x$*. Note que* (176) *es una función lineal por intervalos, y que estos intervalos están unidos en los puntos* κ_ℓ*, denominados nudos (knots).*

En particular, la solución viene dada por las ecuaciones (174) *y* (175)*, con las siguientes selecciones de las matrices:*

$$\boldsymbol{\theta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \\ \alpha_1 \\ \vdots \\ \alpha_q \end{bmatrix}; \quad \boldsymbol{B} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^p & (x_1 - \kappa_1)_+^q & (x_1 - \kappa_2)_+^q & (x_1 - \kappa_r)_+^q \\ 1 & x_2 & x_2^2 & \dots & x_2^p & (x_2 - \kappa_1)_+^q & (x_2 - \kappa_2)_+^q & (x_1 - \kappa_r)_+^q \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^p & (x_n - \kappa_1)_+^q & (x_n - \kappa_2)_+^q & (x_1 - \kappa_r)_+^q \end{bmatrix}. \tag{177}$$

Adicionalmente, únicamente se incorporan penalizaciones para los parámetros asociados a los polinomios truncados. De manera que:

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{0}_{(1+p)\times(1+p)} & \boldsymbol{0}_{(1+p)\times(r)} \\ \boldsymbol{0}_{r\times(1+p)} & \boldsymbol{I}_{r\times r} \end{bmatrix}. \tag{178}$$

Anotación 29. *Note que la selección de* λ *se realiza utilizando validación cruzada. Además* $r \approx n$*, y usualmente* $q = 2$*.*

Ejemplo 12 (Polinomios Truncados con $r = 1$)**.** *Las variables necesarias para resolver* (172) *son:*

$$\boldsymbol{\theta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \alpha_1 \\ \vdots \\ \alpha_L \end{bmatrix}; \quad \boldsymbol{B} = \begin{bmatrix} 1 & x_1 & (x_1 - \kappa_1)_+ & (x_1 - \kappa_2)_+ & (x_1 - \kappa_r)_+ \\ 1 & x_2 & (x_2 - \kappa_1)_+ & (x_2 - \kappa_2)_+ & (x_1 - \kappa_r)_+ \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & (x_n - \kappa_1)_+ & (x_n - \kappa_2)_+ & (x_1 - \kappa_r)_+ \end{bmatrix}; \tag{179}$$

Adicionalmente, la matriz de penalización es de la forma

$$\boldsymbol{D} = \begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times r} \\ \mathbf{0}_{r\times 2} & \boldsymbol{I}_{r\times r} \end{bmatrix}. \tag{180}$$

Este planteamiento genera una función $\hat{f}(x_i)$ suave, pero su estimación numérica es inestable por la colinealidad en la matriz de diseno.

Anotación 30 (Cálculo Numérico)**.** *Para acelerar el cálculo numérico de $\hat{f}(x_i)$ es conveniente utilizar la ortogonalización Demmler-Reinsch. Considere la siguiente descomposición de Cholesky para $\boldsymbol{B}^\top\boldsymbol{B}$ (simétrica definida positiva):*

$$\boldsymbol{B}^\top\boldsymbol{B} = \boldsymbol{R}^\top\boldsymbol{R}, \tag{181}$$

y la descomposición de valores propios

$$\boldsymbol{R}^{-\top}\boldsymbol{D}\boldsymbol{R}^{-1} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top, \tag{182}$$

luego, reemplazando (181) y (182) en (175), note que

$$\begin{aligned}
\hat{\boldsymbol{f}} &= \boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B} + \lambda\boldsymbol{D}\right)^{-1}\boldsymbol{B}^\top\boldsymbol{y} && (183)\\
&= \boldsymbol{B}\left(\boldsymbol{R}^\top\boldsymbol{R} + \lambda\boldsymbol{R}^\top\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top\boldsymbol{R}\right)^{-1}\boldsymbol{B}^\top\boldsymbol{y} && (184)\\
&= \boldsymbol{B}\boldsymbol{R}^{-1}\left(\boldsymbol{I} + \lambda\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top\right)^{-1}\boldsymbol{R}^{-\top}\boldsymbol{B}^\top\boldsymbol{y} && (185)\\
&= \boldsymbol{B}\boldsymbol{R}^{-1}\left(\boldsymbol{V}\boldsymbol{V}^\top + \lambda\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top\right)^{-1}\boldsymbol{R}^{-\top}\boldsymbol{B}^\top\boldsymbol{y} && (186)\\
&= \boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{V}\left(\boldsymbol{I} + \lambda\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{V}^\top\boldsymbol{R}^{-\top}\boldsymbol{B}^\top\boldsymbol{y} && (187)\\
&= \boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{V}\,diag\left\{\frac{1}{1+\lambda\rho_j}\right\}\boldsymbol{V}^\top\boldsymbol{R}^{-\top}\boldsymbol{B}^\top\boldsymbol{y} && (188)\\
&= \boldsymbol{U}\,diag\left\{\frac{1}{1+\lambda\rho_j}\right\}\boldsymbol{U}^\top\boldsymbol{y} && (189)
\end{aligned}$$

donde $\boldsymbol{U} = \boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{V}$.

Anotación 31 (Base B-Splines)**.** *En la práctica es inusual utilizar bases de polinomios truncados. La base de B-splines tiene mejores propiedades numéricas y es la comúnmente utilizada. Algunas modificaciones adicionales se deben realizar en estas bases para reducir los efectos de borde, pero serán omitidas en estas notas.*

En adelante, consideraremos el problema de minimización (172), utilizando b-splines cúbicos.

6.2 Modelos Multivariados

La estimación de la función de regresión para un modelo aditivo generalizado se obtiene mediante la solución del siguiente problema de optimización

$$\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}}\left\{\|\boldsymbol{y} - \sum_{i=1}^{p}\boldsymbol{B}\boldsymbol{\theta}_i\|_2^2 + \sum_{i=1}^{p}\lambda_i\boldsymbol{\theta}_i^\top\boldsymbol{D}\boldsymbol{\theta}_i\right\}, \quad \lambda_i \geq 0, \tag{190}$$

donde se ha asumido que todos los predictores emplean las mismas bases $\boldsymbol{B}$ y las mismas matrices de penalización $\boldsymbol{D}$.

De manera más general, el problema de optimización en (190) se puede escribir como

$$\hat{\boldsymbol{\Theta}} = \text{argmin}_{\boldsymbol{\Theta}} \left\{ \|\boldsymbol{y} - \mathcal{B}\boldsymbol{\Theta}\|_2^2 + \boldsymbol{\Theta}^\top \mathcal{D} \boldsymbol{\Theta} \right\}, \quad \lambda_i \geq 0, \tag{191}$$

donde $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p]^\top$, $\mathcal{B} = [\boldsymbol{B}_1, \ldots, \boldsymbol{B}_p]$, y $\mathcal{D} = \text{blockdiag}\{\lambda_i \boldsymbol{D}\}$, de manera que se pueden utilizar métodos muy similares a los de splines univariados.

Anotación 32. *Note que la especificación (190) abarca los casos de estimación semi-paramétrica, es decir, aquellos casos en los que variables categóricas y términos "no suavizados" son admisibles.*

6.3 Inferencia en GAM

Para esta sección considere el caso univariado y errores $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$. La estimación de la función de regresión tiene la siguiente forma

$$\hat{\boldsymbol{f}} = \boldsymbol{S}_\lambda \boldsymbol{y}; \quad \boldsymbol{S}_\lambda = \boldsymbol{B} \left(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{D} \right)^{-1} \boldsymbol{B}^\top, \tag{192}$$

en particular, si $\boldsymbol{x}_0$ es una observación fuera de la muestra, definamos el vector

$$\boldsymbol{b}_0 = [b_1(\boldsymbol{x}_0), \ldots, b_p(\boldsymbol{x}_0)]^\top, \tag{193}$$

luego, la estimación

$$\hat{f}(\boldsymbol{x}_0) = \boldsymbol{S}_\lambda(\boldsymbol{x}_0)\boldsymbol{y}; \quad \boldsymbol{S}_\lambda(\boldsymbol{x}_0) = \boldsymbol{b}_0^\top \left(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{D} \right)^{-1} \boldsymbol{B}^\top. \tag{194}$$

En este sentido,

$$\begin{aligned} \text{Var}\left(\hat{f}(\boldsymbol{x}_0)\right) &= \boldsymbol{S}_\lambda(\boldsymbol{x}_0) \mathbb{E}(\boldsymbol{y}\boldsymbol{y}^\top) \boldsymbol{S}_\lambda(\boldsymbol{x}_0)^\top & (195) \\ &= \sigma_\epsilon^2 \boldsymbol{S}_\lambda(\boldsymbol{x}_0) \boldsymbol{S}_\lambda(\boldsymbol{x}_0)^\top. & (196) \end{aligned}$$

Considerando la normalidad del error, es claro que

$$\hat{f}(\boldsymbol{x}_0) \sim \mathcal{N}(\mathbb{E}[\hat{f}(\boldsymbol{x}_0)], \sigma_\epsilon^2 \boldsymbol{S}_\lambda(\boldsymbol{x}_0) \boldsymbol{S}_\lambda(\boldsymbol{x}_0)^\top), \tag{197}$$

o, equivalentemente

$$\frac{\hat{f}(\boldsymbol{x}_0) - \mathbb{E}[\hat{f}(\boldsymbol{x}_0)]}{\sigma_\epsilon \sqrt{\boldsymbol{S}_\lambda(\boldsymbol{x}_0) \boldsymbol{S}_\lambda(\boldsymbol{x}_0)^\top}} \sim \mathcal{N}(0, 1). \tag{198}$$

Sin embargo, cuando, σ_ϵ^2 es estimado mediante $\hat{\sigma}_\epsilon^2$ la distribución anterior toma la forma

$$\frac{\hat{f}(\boldsymbol{x}_0) - \mathbb{E}[\hat{f}(\boldsymbol{x}_0)]}{\hat{\sigma}_\epsilon \sqrt{\boldsymbol{S}_\lambda(\boldsymbol{x}_0) \boldsymbol{S}_\lambda(\boldsymbol{x}_0)^\top}} \sim t_{[n - df(f_\lambda)]}, \tag{199}$$

donde $df(f_\lambda) = \text{tr}(\boldsymbol{S}_\lambda)$, y $[z]$ denota el entero más cercano a $z \in \mathbb{R}$. Luego, el intervalo de confianza en el punto $\boldsymbol{x}_0$ viene dado por

$$\hat{f}(\boldsymbol{x}_0) \pm t\left(1 - \frac{\alpha}{2}; [n - df(f_\lambda)]\right) \times \hat{\sigma}_\epsilon \sqrt{\boldsymbol{S}_\lambda(\boldsymbol{x}_0)\boldsymbol{S}_\lambda(\boldsymbol{x}_0)^\top}, \tag{200}$$

donde $t(q, m)$ es el percentil $100 \times q$ de la distribución t con m grados de libertad.

Adicionalmente, es fácil mostrar que

$$\hat{f}(\boldsymbol{x}_0) \pm t\left(1 - \frac{\alpha}{2}; [n - df(f_\lambda)]\right) \times \hat{\sigma}_\epsilon \sqrt{1 + \boldsymbol{S}_\lambda(\boldsymbol{x}_0)\boldsymbol{S}_\lambda(\boldsymbol{x}_0)^\top}, \tag{201}$$

donde el incremento de la varianza es producto de la inclusión de la varianza del error.

7 Métodos Basados en Árboles

La idea general es segmentar el espacio de predictores en regiones rectangulares disjuntas y predecir la respuesta usando la media/moda de la respuesta en cada región. Son muy útiles para interpretación, pero no muy buen desempeno en predicción. Esto se resuelve con otros algoritmos basados en árboles, e.g. bagging, random forest y boosting.

Anotación 33. *Se llaman "árboles" porque el conjunto de reglas de segmentación se puede representar con una estructura de árbol.*

7.1 Generalidades

Un modelo CART tiene la siguiente forma

$$y_i = f(x_i) + \epsilon_i, \quad f(x_i) = \sum_{j=1}^{m} \beta_j \mathbf{1}_{\{x_i \in R_j\}}, \tag{202}$$

donde R_j representa una región en la partición $R = \sqcup_{j=1}^{m} R_j$, y R representa el espacio de predictores. Este modelo puede ser estimado mediante

$$\hat{y}_i = \hat{f}(x_i), \quad \hat{f}(x_i) = \sum_{j=1}^{m} \hat{\beta}_j \mathbf{1}_{\{x_i \in R_j\}}, \tag{203}$$

donde $\hat{\beta}_j$ se puede estimar como el promedio de y_i en R_j (para el problema de regresión), o como la moda de y_i (para el problema de clasificación). Note que esta estimación está condicionada al conocimiento de la partición de R.

Anotación 34. *En estas notas asumiremos que las regiones se construyen por segmentación binaria. Esto involucra seleccionar una variable j y un punto de corte s secuencialmente, siguiendo algún criterio.*

Ejemplo 13 (Segmentación Binaria)**.** *Considere el problema de segmentar la región R en dos sub-regiones R_1 y R_2, de la siguiente manera:*

$$R_1(j, s) = \{\boldsymbol{X} | X_j \leq s\} \quad y \quad R_2(j, s) = \{\boldsymbol{X} | X_j > s\}, \tag{204}$$

donde $R = R_1 \cup R_2$, y, en este sentido, se dice que la segmentación binaria forma una partición.

El criterio para seleccionar el mejor corte depende del tipo de problema que se enfrenta (regresión, clasificación). A continuación se explora cada caso.

7.2 Problemas de Regresión

Considere el problema de regresión, i.e. donde y_i es una respuesta continua. En este caso la selección de j y s en (204) corresponde a la regla

$$\hat{j}, \hat{s} = \text{argmin}_{j,s} \left\{ \sum_{i:x_i \in R_1(j,s)} \left(y_i - \hat{y}_{R_1(j,s)}\right)^2 + \sum_{i:x_i \in R_2(j,s)} \left(y_i - \hat{y}_{R_2(j,s)}\right)^2 \right\}, \tag{205}$$

la cual se puede aplicar iterativamente sobre las nuevas regiones R_1 y R_2 hasta que se cumpla algún criterio, e.g. # obs. por región < 5.

7.3 Problemas de Clasificación

Considere la respuesta discreta $y_i \in \{1, 2, \ldots, K\}$ y denote

$$\hat{p}_{m,k} = \frac{1}{N_m} \sum_{i:x_i \in R_m} \mathbf{1}_{\{y_i = k\}}$$

la proporción de k observaciones en el nodo m. La clasificación resultante para R_m es $k(R_m) = \text{argmax}_k \, \hat{p}_{m,k}$. Algunas medidas de impureza para problemas de clasificación son:

Nombre	Función	Descripción
Ratio de Error	$E = 1 - \max_k(\hat{p}_{m,k})$	Ratio de error
Índice Gini	$G = \sum_{k=1}^{K} \hat{p}_{m,k}(1 - \hat{p}_{m,k})$	Variabilidad
Entropía cruzada	$D = -\sum_{k=1}^{K} \hat{p}_{m,k} \log(\hat{p}_{m,k})$	Entropía

Cuando y_i es una respuesta discreta, la selección de j y s en (204) corresponde a la regla

$$\hat{j}, \hat{s} = \text{argmin}_{j,s} \left\{ N_1 E_{R_1(j,s)} + N_2 E_{R_2(j,s)} \right\}, \tag{206}$$

donde E representa el ratio de error pero las métricas D o G también son válidad. Esta regla se puede aplicar iterativamente igual que con los árboles de regresión.

7.4 Podado de Árboles

Para evitar el sobreajuste a los datos, se recomienda primero ajustar un árbol grande T_0 y luego "podarlo". Podar un árbol significa encontrar el árbol $T \subset T_0$ que minimiza la siguiente función de costo:

$$C_\alpha(T) = \underbrace{\sum_{m=1}^{|T|} N_m Q_m(T)}_{\text{Costo por Impureza}} + \underbrace{\alpha |T|}_{\text{Costo por complejidad}}, \quad \alpha > 0, \tag{207}$$

donde $N_m = \#\{x_i \in R_m\}$, α es un parámetro de penalización, $|T|$ el número de regiones en el árbol T.

Para el caso de regresión, debemos considerar

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2, \quad \hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i.$$

Mientras que para el caso de clasificación, podemos considerar

$$Q_m(T) = E_{R_m},$$

o alguna otra métrica de impureza.

7.5 Algoritmo de Estimación

Algorithm 4 Árbol de Regresión

Tome su data de entrenamiento y realice segmentación binaria para ajustar un árbol T_0 tal que cada nodo terminal tenga como máximo 5 obs.
for $\alpha_k = \alpha_1, \alpha_2, \ldots, \alpha_K$, **do**
  Resuelva $T_{\alpha_k} = \text{argmin}_{T \subset T_0}\{C_{\alpha_k}(T)\}$.
  Guarde T_{α_k}.
Seleccione $\alpha^* = \text{argmin}_\alpha$ MSE/ER mediante validación cruzada, y seleccione el árbol T_{α^*}.

Proposición 12 (Árboles de Regresión vs. Regresión Lineal)**.** *Una regresión lineal es un modelo de la forma*

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i.$$

Un árbol de regresión es un modelo de la forma

$$y_i = \beta_0 + \sum_{j=1}^{p} c_j \cdot 1_{\{x_i \in R_j\}} + \epsilon_i.$$

En este sentido existe una diferencia fundamental en el proceso generador de datos, dado que en el primer caso la función de regresión es lineal en los parámetros y en el segundo caso no. Además en el primero se hacen algunas asumsiones de independencia con respecto al error, pero en el segundo no.

Proposición 13 (Ventajas y Desventajas de usar Árboles)**.** *Algunas ventajas son las siguientes: i) son fáciles de graficar, explicar e interpretar; ii) algunas personas piensan que simulan bien el proceso de decisión humano; iii) pueden manejar predictores cualitativos sin usar variables dummy. Algunas desventajas son las siguientes: i) presentan alta varianza, y alto sesgo; ii) no capturan la estructura aditiva entre las variables.*

7.6 Modelos Ensamblados

La idea es combinar modelos (usar el promedio de los estimadores en regresión, o la moda del predictor en clasificación). Existen diferentes maneras de lograr este objetivo. Un ejemplo es bagging, otro es boosting. Bagging es el principio fundamental usado en Random Forest, y Boosting es el principio fundamental usado en Gradient Boosting Trees.

7.6.1 Bagging

Definición 2 (Bagging)**.** *Bagging es un procedimiento general (no de uso exclusivo para árboles) que se utiliza para reducir la varianza de un método de aprendizaje estadístico.*

Ejemplo 14 (Promedio de Variables Aleatorias)**.** *Recuerde que promediar un conjunto de variables aleatorias independientes reduce la varianza. De hecho, dadas variables aleatorias independientes $z_1, \ldots, z_n$, cada una con varianza σ^2, es claro que la varianza de $\bar{z}$ es σ^2/n.*

El uso de bagging para métodos basados en árboles consiste en usar bootstrap para generar n conjuntos de entrenamiento, construir diferentes árboles para cada conjunto, y promediar las predicciones

$$\hat{f}_{\text{bagging}}(x) = \frac{1}{n}\sum_{\ell=1}^{n} \hat{f}^{(\ell)}(x),$$

donde $\hat{f}^{(\ell)}(x)$ es la predicción con el conjunto de entrenamiento ℓ, y n es el número de conjuntos de entrenamiento utilizados. La idea detrás de este ensamble es que al generar muestras usando bootstrap, estas muestras deben ser independientes y que con ello se logra reducir la varianza del estimador final.

Anotación 35 (Random Forest)**.** *Random Forest es un caso especial de bagging en el que cada vez que se considera un par (j, s), se realiza la selección de j entre $m \ll p$ predictores ($m \approx \sqrt{p}$) aleatoriamente. Esto contribuye a generar estimadores $\hat{f}^b$ independientes entre sí, y contribuye a reducir la varianza de $\hat{f}_{bagging}(x)$.*

7.6.2 Boosting

Igual que bagging, boosting utiliza bootstrap para generar n conjuntos de entrenamiento, ajusta estos modelos "de alguna manera" y luego los combina

$$\hat{f}_{\text{boosting}}(x) = \frac{1}{n}\sum_{\ell=1}^{n} \omega_\ell \hat{f}^{(\ell)}(x),$$

donde ω_ℓ es un peso directamente proporcional al desempeno de $\hat{f}^{(\ell)}(x)$. La principal diferencia es que el ajuste de cada modelo $\hat{f}^{(\ell)}(x)$ no se realiza de manera independiente, sino de manera secuencial.

Dicho de manera general, la función estimadora $\hat{f}^{(\ell)}(x)$ es construída en base a los resultados de la función estimadora $\hat{f}^{(\ell-1)}(x)$ al momento de construír la muestra usando bootstrap. En este paso se da un peso adicional a aquellas observaciones que hayan sido mal clasificadas/predichas por $\hat{f}^{(\ell)}(x)$. En este sentido, boosting es un método que sobre todo reduce el sesgo del estimador.

Anotación 36. *Los hiper-parámetros en bagging y en boosting son i) el número de árboles, ii) el número de cortes, y otros. Algunos se pueden escoger mediante validación cruzada. Otros se pueden seleccionar utilizando algunos resultados heurísticos.*

8 Máquinas de Vectores de Soporte

8.1 Clasificadores Lineales

En esta sección veremos al modelo Logit como un modelo que contiene un separador lineal subyacente, e introduciremos un modelo naive que será de utilidad para entender los conceptos básicos de SVM.

Ejemplo 15 (Modelo Logit)**.** *El modelo logit construye un separador lineal de la data. Considere datos con respuesta binaria $y_i \in \{0,1\}$, y una cota de $\bar{p} = 0.5$, es decir,*

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i, \tag{208}$$

$$p_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}, \tag{209}$$

tal que la predicción viene dada por

$$\hat{y}_i = \begin{cases} 1, & si \quad p_i \geq 0.5 \\ 0, & si \quad p_i < 0.5. \end{cases} \tag{210}$$

Note que dado $p_i \geq 0.5$, se cumple que $\beta_0 + \beta_1 x_i \geq 0$, y que dado $p_i < 0.5$, se cumple que $\beta_0 + \beta_1 x_i < 0$. En este sentido, $\beta_0 + \beta_1 x$ es un hiperplano que separa las dos clases.

Ejemplo 16 (Modelo Naive)**.** *Considere datos con respuesta binaria $y_i \in \{-1,1\}$, donde existen m_+ observaciones de la clase $+1$ y m_- observaciones de la clase -1. Note que el centro para cada conjunto de observaciones se puede calcular como*

$$\boldsymbol{c}_p = \frac{1}{m_+} \sum_{i:y_i=+1} \boldsymbol{x}_i \tag{211}$$

$$\boldsymbol{c}_n = \frac{1}{m_-} \sum_{i:y_i=-1} \boldsymbol{x}_i. \tag{212}$$

Otras dos cantidades se pueden calcular con facilidad:

$$\boldsymbol{c} = (\boldsymbol{c}_p + \boldsymbol{c}_m)/2, \tag{213}$$

$$\boldsymbol{w} = \boldsymbol{c}_p - \boldsymbol{c}_m, \tag{214}$$

donde $\boldsymbol{c}$ es el vector que ubicado en la posición media entre $\boldsymbol{c}_p$ y $\boldsymbol{c}_m$, y $\boldsymbol{w}$ es el vector que une los centros $\boldsymbol{c}_p$ y $\boldsymbol{c}_m$. Usando esta información es fácil ver que la clasificación de una observación nueva $\boldsymbol{x}$ viene dada por

$$\boldsymbol{y} = sgn\left(\langle \boldsymbol{x} - \boldsymbol{c}, \boldsymbol{w} \rangle\right) \tag{215}$$

Proposición 14. *Es fácil ver que*

$$\begin{aligned} \langle \boldsymbol{x} - \boldsymbol{c}, \boldsymbol{w} \rangle = & \frac{1}{m_+} \sum_{i:y_i=+1} \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle - \frac{1}{m_-} \sum_{i:y_i=-1} \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle \\ & + \frac{1}{2} \left(\frac{1}{m_-^2} \sum_{(i,j):y_i=-1} \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle - \frac{1}{m_+^2} \sum_{(i,j):y_i=+1} \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \right). \end{aligned} \tag{216}$$

Demostración. Trivial

$$\begin{aligned}
\langle \boldsymbol{x} - \boldsymbol{c}, \boldsymbol{w} \rangle &= \left\langle \boldsymbol{x} - \frac{\boldsymbol{c}_p + \boldsymbol{c}_m}{2}, \boldsymbol{c}_p - \boldsymbol{c}_m \right\rangle && (217)\\
&= \langle \boldsymbol{x}, \boldsymbol{c}_p \rangle - \langle \boldsymbol{x}, \boldsymbol{c}_m \rangle - \frac{1}{2} \langle \boldsymbol{c}_p + \boldsymbol{c}_m, \boldsymbol{c}_p - \boldsymbol{c}_m \rangle && (218)\\
&= \langle \boldsymbol{x}, \boldsymbol{c}_p \rangle - \langle \boldsymbol{x}, \boldsymbol{c}_m \rangle + \frac{1}{2} \left(\|\boldsymbol{c}_m\|_2^2 - \|\boldsymbol{c}_p\|_2^2 \right) && (219)\\
&= \left\langle \boldsymbol{x}, \frac{1}{m_+} \sum_{i:y_i=+1} \boldsymbol{x}_i \right\rangle - \left\langle \boldsymbol{x}, \frac{1}{m_-} \sum_{i:y_i=-1} \boldsymbol{x}_i \right\rangle + \\
&\quad \frac{1}{2} \left(\left\langle \frac{1}{m_-} \sum_{i:y_i=-1} \boldsymbol{x}_i, \frac{1}{m_-} \sum_{i:y_i=-1} \boldsymbol{x}_i \right\rangle - \left\langle \frac{1}{m_+} \sum_{i:y_i=+1} \boldsymbol{x}_i, \frac{1}{m_+} \sum_{i:y_i=+1} \boldsymbol{x}_i \right\rangle \right) \\
&= \frac{1}{m_+} \sum_{i:y_i=+1} \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle - \frac{1}{m_-} \sum_{i:y_i=-1} \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle \\
&\quad + \frac{1}{2} \left(\frac{1}{m_-^2} \sum_{(i,j):y_i=-1} \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle - \frac{1}{m_+^2} \sum_{(i,j):y_i=+1} \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \right). && (220)
\end{aligned}$$

□

Note que reemplazando (216) en (215), se obtiene

$$\begin{aligned}
\boldsymbol{y} &= \operatorname{sgn} \left(\langle \boldsymbol{x} - \boldsymbol{c}, \boldsymbol{w} \rangle \right) && (221)\\
&= \operatorname{sgn} \left(\frac{1}{m_+} \sum_{i:y_i=+1} \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle - \frac{1}{m_-} \sum_{i:y_i=-1} \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle \right. && (222)\\
&\quad \left. + \frac{1}{2} \left(\frac{1}{m_-^2} \sum_{(i,j):y_i=-1} \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle - \frac{1}{m_+^2} \sum_{(i,j):y_i=+1} \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \right) \right). && (223)
\end{aligned}$$

Considere $k(\boldsymbol{x}, \boldsymbol{x}_i) = \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}_i) \rangle$, como una métrica de similaridad. En este caso la expresión anterior se puede re-escribir como

$$\begin{aligned}
\boldsymbol{y} &= \operatorname{sgn} \left(\frac{1}{m_+} \sum_{i:y_i=+1} k(\boldsymbol{x}, \boldsymbol{x}_i) - \frac{1}{m_-} \sum_{i:y_i=-1} k(\boldsymbol{x}, \boldsymbol{x}_i) \right. && (224)\\
&\quad \left. + \frac{1}{2} \left(\frac{1}{m_-^2} \sum_{(i,j):y_i=-1} k(\boldsymbol{x}_i, \boldsymbol{x}_j) - \frac{1}{m_+^2} \sum_{(i,j):y_i=+1} k(\boldsymbol{x}_i, \boldsymbol{x}_j) \right) \right). && (225)
\end{aligned}$$

8.2 SVM de Clasificación

Definición 3 (Hiperplano)**.** *Sea $\mathcal{S}$ un espacio p-dimensional. $\mathcal{H}$ es un hiperplano en $\mathcal{S}$ si es un subespacio afín de $\mathcal{S}$, i.e. es un sub-espacio de $\mathcal{S}$ que no contiene el elemento nulo. En general, para valores $\beta_0, \beta_1, \ldots, \beta_p \in \mathbb{R}^p$,*

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = 0, \tag{226}$$

define un hiperplano en $\mathbb{R}^p$ en el sentido de que cualquier $\boldsymbol{x} \in \mathbb{R}^p$ que cumpla con (226) es un punto en el hiperplano.

Anotación 37. *Note que i) cualquier línea en $\mathbb{R}^2$ es un hiperplano en $\mathbb{R}^2$, pero sólo aquellas que cruzan el origen son subespacios de $\mathbb{R}^2$; ii) cualquier plano en $\mathbb{R}^3$ es un hiperplano en $\mathbb{R}^3$, pero sólo aquellas que cruzan el origen son subespacios de $\mathbb{R}^3$.*

Definición 4 (Hiperplano Separador)**.** *Considere las observaciones $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})^\top \in \mathbb{R}^p$, y una respuesta binaria $y_i \in \{-1, 1\}$. Un hiperplano separador es un hiperplano para el cual se cumple que*

$$\begin{aligned} \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} > 0 \quad & \text{if } y_i = +1 \\ \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} < 0 \quad & \text{if } y_i = -1, \end{aligned}$$

o equivalentemente

$$y_i(\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}) > 0 \quad \Leftrightarrow \quad y_i(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta}) > 0$$

donde $\boldsymbol{x}_i$ representa a observación i.

Anotación 38 (Clasificación usando un Hiperplano Separador)**.** *Note que: i) un hiperplano separador no es único; ii) todos los puntos utilizados para realizar la clasificación no son igualmente importantes. Dado $\boldsymbol{x}_i^*$, el clasificador es más sensible mientras la cantidad $\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta}$ se acerca a cero.*

Definición 5 (Margen)**.** *El margen es la distancia mínima desde cualquier observación del conjunto de entrenamiento a un hiperplano separador determinado.*

Definición 6 (Hiperplano de Margen Máximo)**.** *El hiperplano de margen máximo es el hiperplano separador que tiene el margen más grande.*

Definición 7 (Vectores de Soporte)**.** *Todas las observaciones equidistantes al hiperplano de margen máximo son llamados vectores de soporte.*

Anotación 39. *Note que i) el hiperplano de margen máximo depende directamente de los vectores de soporte, pero no de las demás observaciones; ii) la técnica de clasificación que utiliza el hiperplano de margen máximo se llama clasificador de margen máximo.*

Los estimadores del clasificador de margen máximo vienen dados por los valores $\hat{\boldsymbol{\beta}}$ que resuelven

$$\text{argmax}_{\beta_0, \boldsymbol{\beta}} \left\{ M : y_i(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta}) \geq M, \text{ para } i = 1, \ldots, n, \ y \ \sum_{j=0}^{p} \beta_j^2 = 1 \right\}.$$

Sin embargo, un hiperplano separador puede no existir. Esto se puede resolver usando un "margen laxo". La generalización del clasificador de margen máximo al caso no separable se conoce como el clasificador de vectores de soporte. Los estimadores del hiperplano de vectores de soporte vienen dados por los valores $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\epsilon}})$ que resuelven

$$\begin{aligned} \text{argmax}_{\beta_0, \boldsymbol{\beta}, \boldsymbol{\epsilon}} \Bigg\{ & M : y_i(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta}) \geq M(1 - \epsilon_i), \sum_{j=0}^{p} \beta_j^2 = 1, \\ & \epsilon_i \geq 0, \quad \boldsymbol{\epsilon}^\top \mathbf{1} \leq C, \quad i = 1, \ldots, n \Bigg\}, \end{aligned}$$

donde los ϵ_i, son llamadas variables de holgura. Note que: i) si $\epsilon_i > 0$, la observación i está en el lado incorrecto del margen; ii) si $\epsilon_i > 1$, la observación i está en el lado incorrecto del hiperplano; iii) todos los vectores en los que $\epsilon_i > 0$ son vectores de soporte. iv) el parámetro C controla la severidad de las violaciones; v) si $C \downarrow$, $\downarrow$ sesgo, $\uparrow$ var. Si $C \uparrow$, $\uparrow$ sesgo, $\downarrow$ var.

Proposición 15 (Hiperplano de Vectores de Soporte)**.** *El hiperplano de vectores de soporte puede ser representado como*

$$f(\boldsymbol{x}) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle, \quad \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle = \boldsymbol{x}^\top \boldsymbol{x}_i \tag{227}$$

donde $\mathcal{S}$ es el conjunto de índices correspondientes a los vectores de soporte, y $\alpha_i > 0$.

Demostración. En base a la FOC del problema SVC se sabe que: $\boldsymbol{\beta} = \sum_{i=1}^n (\delta_i y_i) \boldsymbol{x}_i$. Luego

$$\begin{aligned} f(\boldsymbol{x}) &= \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta} \\ &= \beta_0 + \boldsymbol{x}^\top \left(\sum_{i=1}^n \delta_i y_i \boldsymbol{x}_i \right) \\ &= \beta_0 + \sum_{i \in \mathcal{S}} \delta_i y_i \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle + \sum_{i \notin \mathcal{S}}^n \delta_i y_i \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle \\ &= \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle, \end{aligned}$$

donde la última expresión sigue de $\alpha_i = \delta_i y_i$, y el hecho de que $\alpha_i \neq 0$ sólo para los vectores de soporte. □

Definición 8 (Máquinas de Vectores de Soporte)**.** *SVM's son una extensión de SVC que resulta de agrandar el espacio de predictores usando kernels,*

$$f(\boldsymbol{x}) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle \quad \rightarrow \quad f(\boldsymbol{x}) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i),$$

donde $K(\boldsymbol{x}, \boldsymbol{x}_i) = \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle$ es llamado kernel lineal.

Otros ejemplos de kernels son el Kernel polinomial $K(\boldsymbol{x}, \boldsymbol{x}_i) = \left(1 + \boldsymbol{x}^\top \boldsymbol{x}_i\right)^d, \quad d > 0$; y el Kernel radial $K(\boldsymbol{x}, \boldsymbol{x}_i) = \exp\left(-\gamma \|\boldsymbol{x} - \boldsymbol{x}_i\|_2^2\right), \quad \gamma > 0$.

Anotación 40 (SVMs: Más de dos clases)**.** *Considere k posibles clases para la respuesta. Se pueden usar dos enfoques:*

1. *Uno vs. uno. Hacer SVM por pares para los $\binom{k}{2}$ casos posibles. Hacer una votación por observación y clasificar por votación.*
2. *Uno vs. todos.*
   - *Para todas las observaciones en la clase 1, escribir $+1$ como respuesta, y -1 para todas las demás clases. Hacer SVM y calcular $\beta_{0,1}, \beta_{1,1}, \ldots, \beta_{p,1}$.*
   - *Repetir el paso anterior para las clases $2, 3, \ldots, k$ y recolectar $\beta_{0,i}, \beta_{1,i}, \ldots, \beta_{p,i}$, $i = 1, \ldots, k$ en cada caso.*

- *Para una observación fuera de la muestra* $\boldsymbol{x}^*$*, calcular*

$$f\left(\boldsymbol{x}^*; \boldsymbol{\beta}_i\right) = \boldsymbol{\beta}_0 + \boldsymbol{x}^{*\top} \boldsymbol{\beta}_i, \quad i = 1, \ldots, k,$$

  y seleccionar el caso i *para el cual* $f\left(\boldsymbol{x}^*; \boldsymbol{\beta}_i\right)$ *es mayor.*

8.3 SVM de Regresión