Regresión logística

Matías Carrasco

5 de octubre de 2023

Índice

1. Un modelo estadístico de clasificación

Veremos como con una modificación del modelo de regresión lineal también podemos aplicarlo al problema de clasificación. El costo será el de no poder calcular los coeficientes de forma explícita. En su lugar, tenemos que recurrir a la optimización numérica para aprender los parámetros del modelo.

Desde una perspectiva estadística, la clasificación equivale a predecir las probabilidades condicionales de cada clase

$$\text{Prob}(y = c \mid \boldsymbol{x})$$

donde y es la etiqueta en $\{1, 2, \ldots, C\}$ y $\boldsymbol{x}$ es el atributo. Es decir, la probabilidad para la clase c dado que conocemos el atributo $\boldsymbol{x}$.

Buscamos construir un clasificador que pueda predecir las probabilidades de clase $\text{Prob}(y \mid \boldsymbol{x})$. Más específicamente, para problemas de clasificación binaria $C = 2$,

donde vamos a suponer que y es 1 o -1, queremos aprender un modelo $g(\boldsymbol{x})$ para el cual

$$\text{Prob}(y = 1 \mid \boldsymbol{x}) \text{ es modelado por } g(\boldsymbol{x}).$$

Tomando complementos el modelo equivale a

$$\text{Prob}(y = -1 \mid \boldsymbol{x}) \text{ es modelado por } 1 - g(\boldsymbol{x}).$$

Dado que $g(\boldsymbol{x})$ es un modelo para una probabilidad, es natural requerir que $0 \leq g(\boldsymbol{x}) \leq 1$ para cualquier $\boldsymbol{x}$.

Por supuesto, aquí tienes la traducción al español:

Para el problema multiclase, hacemos que el clasificador devuelva una función vectorial $\boldsymbol{g}(\mathbf{x})$, donde

$$\begin{bmatrix} \text{Prob}(y = 1 \mid \boldsymbol{x}) \\ \text{Prob}(y = 2 \mid \boldsymbol{x}) \\ \vdots \\ \text{Prob}(y = C \mid \boldsymbol{x}) \end{bmatrix} \text{ es modelado por } \begin{bmatrix} g_1(\boldsymbol{x}) \\ g_2(\boldsymbol{x}) \\ \vdots \\ g_C(\boldsymbol{x}) \end{bmatrix} = \boldsymbol{g}(\boldsymbol{x})$$

Es decir, cada elemento $g_c(\boldsymbol{x})$ de $\boldsymbol{g}(\boldsymbol{x})$ corresponde a la probabilidad condicional de clase $\text{Prob}(y = c \mid \boldsymbol{x})$. Dado que $\boldsymbol{g}(\boldsymbol{x})$ modela un vector de probabilidad, requerimos que cada elemento $g_c(\boldsymbol{x}) \geq 0$ y que $\|\boldsymbol{g}(\boldsymbol{x})\|_1 = \sum_{c=1}^{C} g_c(\boldsymbol{x}) = 1$ para cualquier $\boldsymbol{x}$.

2. Regresión logística para clasificación binaria

La regresión logística puede verse como una modificación del modelo de regresión lineal para que se adapte al problema de clasificación.

Comencemos con la clasificación binaria. Deseamos aprender una función $g(\boldsymbol{x})$ que se aproxime a la probabilidad condicional de la clase positiva. Con este fin, comenzamos con el modelo de regresión lineal que, sin el término de ruido, es:

$$z = b + w_1 x^{(1)} + w_2 x^{(2)} + \cdots + w_D x^{(D)} = \boldsymbol{\theta}^\top \boldsymbol{x}.$$

Esta es una función que toma $\boldsymbol{x}$ y devuelve z, que en este contexto se llama **logit**. Notar que z toma valores en toda la recta real, mientras que necesitamos una función que devuelva un valor en el intervalo $[0, 1]$. La idea principal de la regresión logística es comprimir z al intervalo $[0, 1]$ usando la función logística $h(z) = \frac{e^z}{1+e^z}$. Es decir

$$g(\boldsymbol{x}) = \frac{e^{\boldsymbol{\theta}^\top \boldsymbol{x}}}{1 + e^{\boldsymbol{\theta}^\top \boldsymbol{x}}}$$

que está restringido a $[0, 1]$ y por lo tanto puede interpretarse como una probabilidad.

El modelo de regresión logística es para $\text{Prob}(y = 1 \mid \boldsymbol{x})$. Notar que también da implícitamente un modelo para $\text{Prob}(y = -1 \mid \boldsymbol{x})$:

$$1 - g(\boldsymbol{x}) = 1 - \frac{e^{\boldsymbol{\theta}^\top \boldsymbol{x}}}{1 + e^{\boldsymbol{\theta}^\top \boldsymbol{x}}} = \frac{1}{1 + e^{\boldsymbol{\theta}^\top \boldsymbol{x}}} = \frac{e^{-\boldsymbol{\theta}^\top \boldsymbol{x}}}{1 + e^{-\boldsymbol{\theta}^\top \boldsymbol{x}}}.$$

En resumen, el modelo de regresión logística es la regresión lineal complementada con la función logística. Esta es la razón del confuso nombre. La razón por la que no hay un término de ruido ϵ, como teníamos en el modelo de regresión lineal, es que la aleatoriedad en la clasificación está modelada estadísticamente por la probabilidad de clase.

2.1. Máxima verosimilitud

Para aprender $\boldsymbol{\theta}$ a partir de los datos de entrenamiento $S = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$, comenzamos con el enfoque de máxima verosimilitud. Esto implica resolver:

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \text{Prob}(\boldsymbol{y} \mid \boldsymbol{X}; \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ln \text{Prob}\left(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}\right),$$

donde asumimos que los datos de entrenamiento son independientes.

$$\ln \text{Prob}\left(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}\right) = \begin{cases} \ln g\left(\boldsymbol{x}_i; \boldsymbol{\theta}\right) & \text{si } y_i = 1 \\ \ln\left(1 - g\left(\boldsymbol{x}_i; \boldsymbol{\theta}\right)\right) & \text{si } y_i = -1 \end{cases}$$

Utilizando el negativa del logaritmo de la verosimilitud como función de pérdida, obtenemos

$$L(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \underbrace{\begin{cases} -\ln g\left(\boldsymbol{x}_i; \boldsymbol{\theta}\right) & \text{si } y_i = 1, \\ -\ln\left(1 - g\left(\boldsymbol{x}_i; \boldsymbol{\theta}\right)\right) & \text{si } y_i = -1 \end{cases}}_{\text{Binary Cross Entropy (BCE)}}$$

Esta función de pérdida se llama **entropía cruzada**. Puede ser usada para cualquier clasificador binario que prediga probabilidades de clase $g(\boldsymbol{x}; \boldsymbol{\theta})$.

Sin embargo, ahora consideraremos específicamente el modelo de regresión logística, para el cual podemos detallar más la función de coste (3.32). Al hacerlo, la elección particular de etiquetado $\{-1, 1\}$ resulta ser conveniente. Para $y_i = 1$ podemos escribir:

$$g\left(\mathbf{x}_i; \boldsymbol{\theta}\right) = \frac{e^{\boldsymbol{\theta}^\top \mathbf{x}_i}}{1 + e^{\boldsymbol{\theta}^\top \mathbf{x}_i}} = \frac{e^{y_i \boldsymbol{\theta}^\top \mathbf{x}_i}}{1 + e^{y_i \boldsymbol{\theta}^\top \mathbf{x}_i}},$$

y para $y_i = -1$:

$$1 - g\left(\mathbf{x}_i; \boldsymbol{\theta}\right) = \frac{e^{-\boldsymbol{\theta}^\top \mathbf{x}_i}}{1 + e^{-\boldsymbol{\theta}^\top \mathbf{x}_i}} = \frac{e^{y_i \boldsymbol{\theta}^\top \mathbf{x}_i}}{1 + e^{y_i \boldsymbol{\theta}^\top \mathbf{x}_i}}.$$

La elección particular de etiquetado $\{-1, 1\}$ resulta ser conveniente. Para $y_i = 1$ podemos escribir:

$$g\left(\boldsymbol{x}_i; \boldsymbol{\theta}\right) = \frac{e^{\boldsymbol{\theta}^\top \boldsymbol{x}_i}}{1 + e^{\boldsymbol{\theta}^\top \boldsymbol{x}_i}} = \frac{e^{y_i \boldsymbol{\theta}^\top \boldsymbol{x}_i}}{1 + e^{y_i \boldsymbol{\theta}^\top \boldsymbol{x}_i}}$$

y para $y_i = -1$:

$$1 - g\left(\boldsymbol{x}_i; \boldsymbol{\theta}\right) = \frac{e^{-\boldsymbol{\theta}^\top \boldsymbol{x}_i}}{1 + e^{-\boldsymbol{\theta}^\top \boldsymbol{x}_i}} = \frac{e^{y_i \boldsymbol{\theta}^\top \boldsymbol{x}_i}}{1 + e^{y_i \boldsymbol{\theta}^\top \boldsymbol{x}_i}}.$$

Por lo tanto, obtenemos la misma expresión en ambos casos y podemos escribir la función de costo de manera compacta como:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \frac{1}{N} \sum_{i=1}^{N} - \ln \frac{e^{y_i \boldsymbol{\theta}^\top \boldsymbol{x}_i}}{1 + e^{y_i \boldsymbol{\theta}^\top \boldsymbol{x}_i}} \\ &= \frac{1}{N} \sum_{i=1}^{N} - \ln \frac{1}{1 + e^{-y_i \boldsymbol{\theta}^\top \boldsymbol{x}_i}} \\ &= \frac{1}{N} \sum_{i=1}^{N} \underbrace{\ln\left(1 + e^{-y_i \boldsymbol{\theta}^\top \boldsymbol{x}_i}\right)}_{\text{Pérdida logística } L(\boldsymbol{x}_i, y_i, \boldsymbol{\theta})}. \end{aligned}$$

La función de pérdida $L\left(\boldsymbol{x}_i, y_i, \boldsymbol{\theta}\right)$, que es un caso especial de la pérdida de entropía cruzada, se llama pérdida logística. Aprender un modelo de regresión logística equivale a resolver:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \ln\left(1 + e^{-y_i \boldsymbol{\theta}^\top \boldsymbol{x}_i}\right).$$

Este problema no tiene una solución en forma cerrada, por lo que debemos usar optimización numérica en su lugar.

2.2. Relación con la 0-1 loss

Generalmente se define a partir del modelo de regresión logística siguiente el clasificador:

$$\hat{y} = \begin{cases} 1 & \text{si } \operatorname{Prob}(y = 1 \mid \boldsymbol{x}; \boldsymbol{\theta}) \geq \beta \\ -1 & \text{si } \operatorname{Prob}(y = 1 \mid \boldsymbol{x}; \boldsymbol{\theta}) < \beta \end{cases} = \begin{cases} 1 & \text{si } g(\boldsymbol{x}; \boldsymbol{\theta}) \geq \beta \\ -1 & \text{si } g(\boldsymbol{x}; \boldsymbol{\theta}) < \beta \end{cases}$$

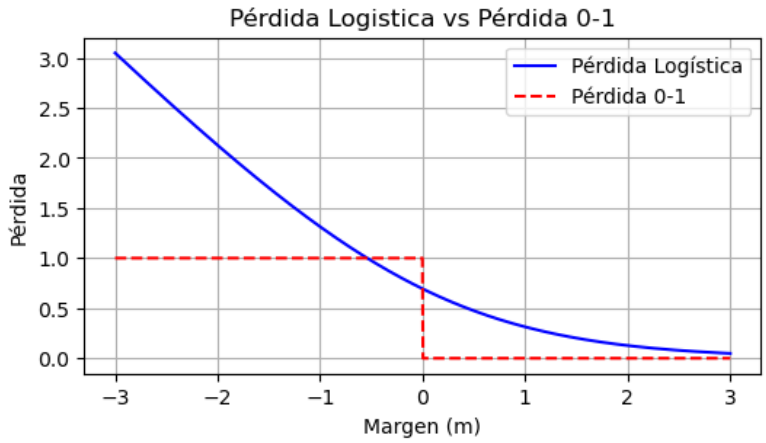


Figura 1: Gráfico que compara la pérdida logística con la pérdida 0-1 en función del margen m. La pérdida logística es una función suave, mientras que la pérdida 0-1 es una función escalonada que toma el valor de 1 cuando $m < 0$ y 0 cuando $m > 0$.

en donde $\beta \in [0, 1]$ es un umbral a elegir. Es usual elegir $\beta = 1/2$, y en este caso el clasificador se puede escribir de forma más compacta como

$$\hat{y} = \text{signo}\left[\boldsymbol{\theta}^\top \boldsymbol{x}\right].$$

Idealmente queremos obtener un clasificador que minimice el error de clasificación con la 0-1 loss $L(y, \boldsymbol{x}, \boldsymbol{\theta}) = \mathbb{1}_{y \neq \hat{y}}$ en donde $\hat{y}$ está dado por la ecuación de arriba. Sin embargo, esto generalmente no es posible debido a la naturaleza discreta de la pérdida cero-uno: la pérdida cero-uno no tiene gradiente y, por lo tanto, la minimización bajo la pérdida cero-uno es esencialmente una optimización combinatoria computacionalmente intratable en la práctica.

Para enfrentar este problema, usamos una pérdida **sustituta** $L(\boldsymbol{x}, y, \boldsymbol{\theta})$ que tiende a tomar un valor pequeño para un **margen** grande $m = y\boldsymbol{\theta}^\top \boldsymbol{x}$. Por lo tanto, con la formulación basada en la pérdida sustituta, se incentiva un margen más grande.

Para un clasificador lineal (como la regresión logística), el modelo intenta encontrar un hiperplano que separe las clases en el espacio de atributos. En 2D, este hiperplano es simplemente una línea; en 3D, es un plano; y en dimensiones más altas, es un hiperplano. La ecuación de dicho hiperplano, para el caso de umbral $\beta = 1/2$ es $\boldsymbol{\theta}^\top \boldsymbol{x} = 0$.

El significado del margen es:

- Si el margen es **positivo** para un punto dado, eso indica que el punto está correctamente clasificado y del lado correcto del hiperplano.

- Si el margen es **negativo**, el punto está mal clasificado y está del lado incorrecto del hiperplano.
- Si el margen es **cero**, el punto está justo en el hiperplano de decisión.

De hecho, la pérdida logística es una función del margen:

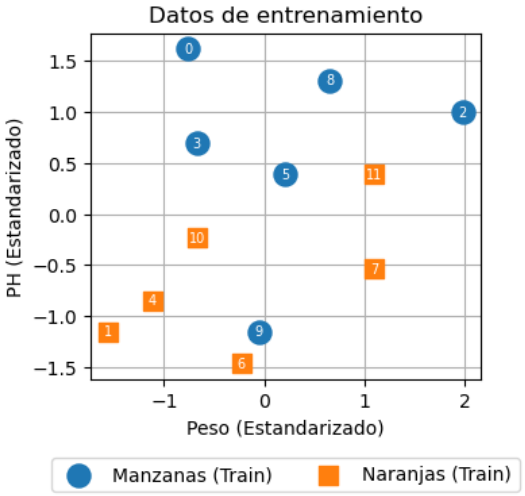
$$L(y, \boldsymbol{x}, \boldsymbol{\theta}) = \ln\left(1 + e^{-y\boldsymbol{\theta}^\top \boldsymbol{x}}\right) = \ln\left(1 + e^{-m}\right) = L(m).$$

La Fig. 1 muestra una comparación entre la pérdida logística y la pérdida 0-1.

3. Ejemplo - Clasificando manzanas y naranjas

Consideremos los siguientes datos de entrenamiento:

id	Peso (g)	PH	Fruta
0	139	3.9	manzana
1	130	3.0	naranja
2	170	3.7	manzana
3	140	3.6	manzana
4	135	3.1	naranja
5	150	3.5	manzana
6	145	2.9	naranja
7	160	3.2	naranja
8	155	3.8	manzana
9	147	3.0	manzana
10	140	3.3	naranja
11	160	3.5	naranja



Nuestro objetivo es clasificar las manzanas y las naranjas.

3.1. Caso univariado

Consideremos primero un caso univariando utilizando solamente el atributo Peso (g). Veamos cómo afectan los parámetros la forma del modelo. La Fig.2 muestra qué ocurre con $g(x; b, w)$ cuando variamos b y w.

3.2. Caso bivariado

Consideremos ahora ambos atributos, Peso y PH. La Fig. 3 muestra las curvas de nivel de la función de pérdida logística en función de los coeficientes w_1 (Peso) y w_2 (PH) correspondientes a los atributos estandarizados. Notar que si bien dichas

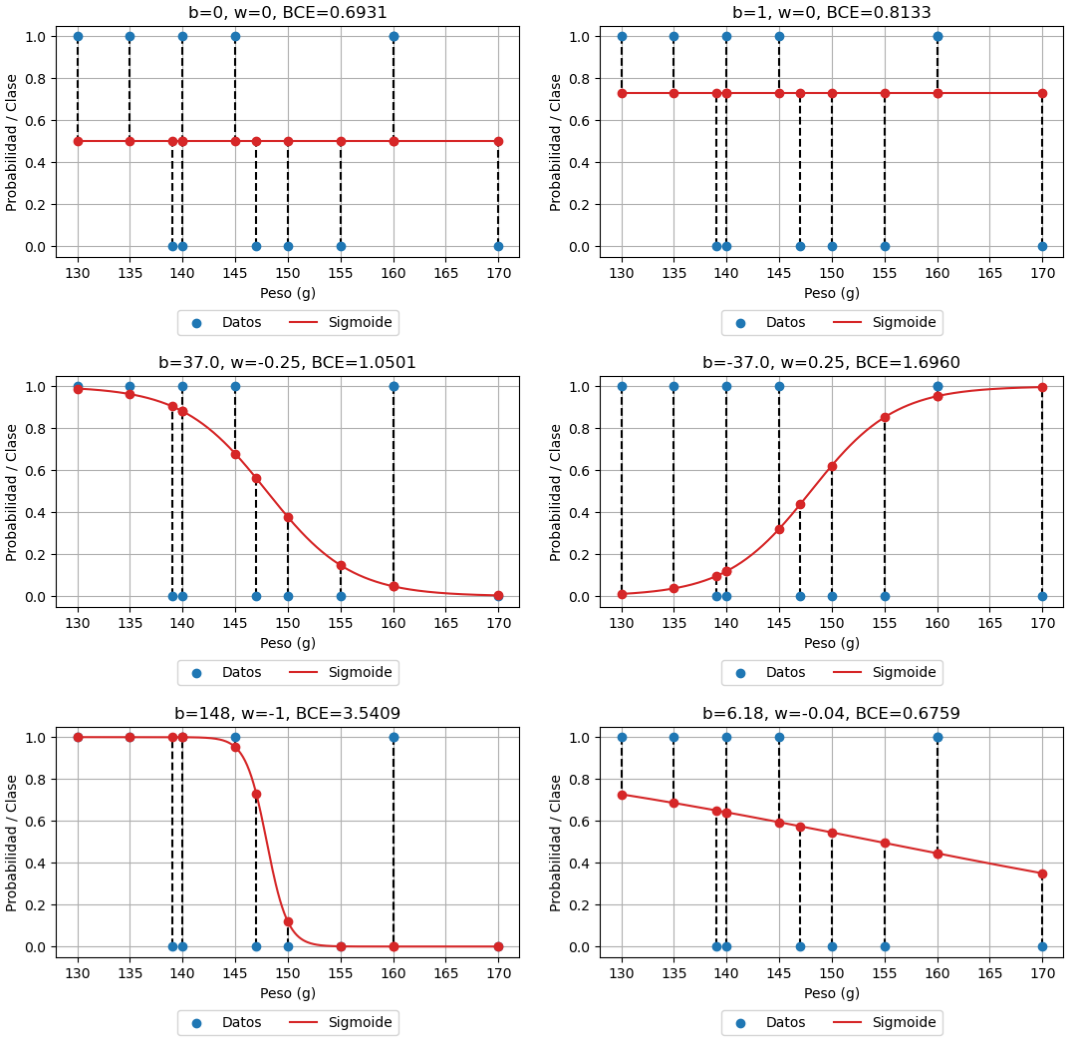


Figura 2: Gráficos del modelo $g(x; b, w)$ variando los coeficientes b y w. El gráfico abajo a la derecha muestra el modelo óptimo.

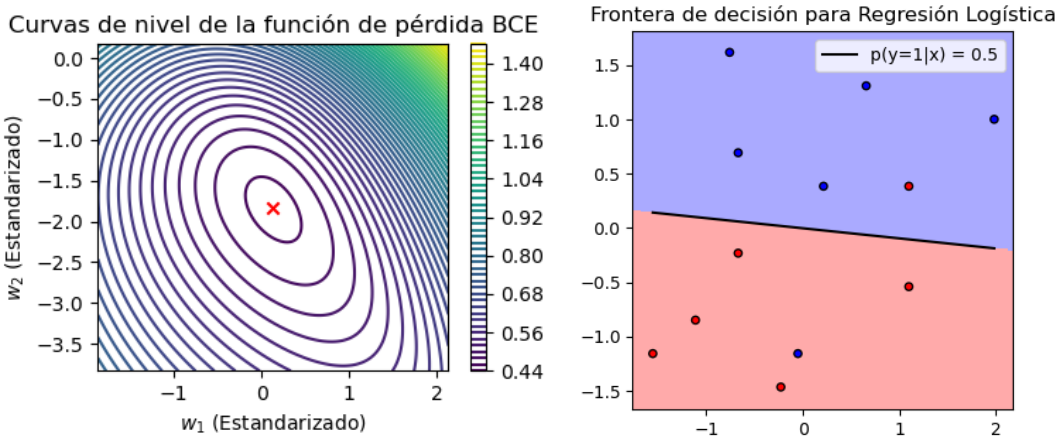


Figura 3: A la izquierda se muestran las curvas de nivel de la función de pérdida en función de w_1 y w_2. A la derecha se muestra la frontera de decisión.

curvas no son exactamente elipses, son curvas cerradas que encierran al punto óptimo.

La **frontera de decisión** es la recta $\boldsymbol{x}^\top \boldsymbol{\theta} = 0$.

También se puede hacer regresión logística con funciones base:

$$g(\boldsymbol{x}; \boldsymbol{\theta}) = \sigma\left(b + w_1 h_1(\boldsymbol{x}) + w_2 h_2(\boldsymbol{x}) + \cdots + w_K h_K(\boldsymbol{x})\right)$$

Casos particulares son:

- **Regresión polinomial**: cuando $h_j(\boldsymbol{x})$ son productos de potencias $x^{(j)}$.
- **Regresión trigonométrica**: cuando $h_j(\boldsymbol{x})$ son funciones trigonométricas.

Por ejemplo, la Fig. 4 muestra las fronteras de decisión en el caso de regresión logística con polinomios para diversos grados. Notar que la frontera deja de ser lineal y pasa a ser una curva, que incluso puede ser disconexa.

4. Regresión logística multiclase

La regresión logística también se puede utilizar para el problema multiclase cuando hay más de dos clases, $C > 2$. Existen varias maneras de generalizar la regresión logística a este escenario. Utilizaremos la llamada función **softmax**, la cual es también muy utilizada en redes neuronales.

Para el problema binario, usamos la función logística para diseñar un modelo $g(\boldsymbol{x})$, una función escalar que representa $\text{Prob}(y = 1 \mid \boldsymbol{x})$. Para el problema multicla-

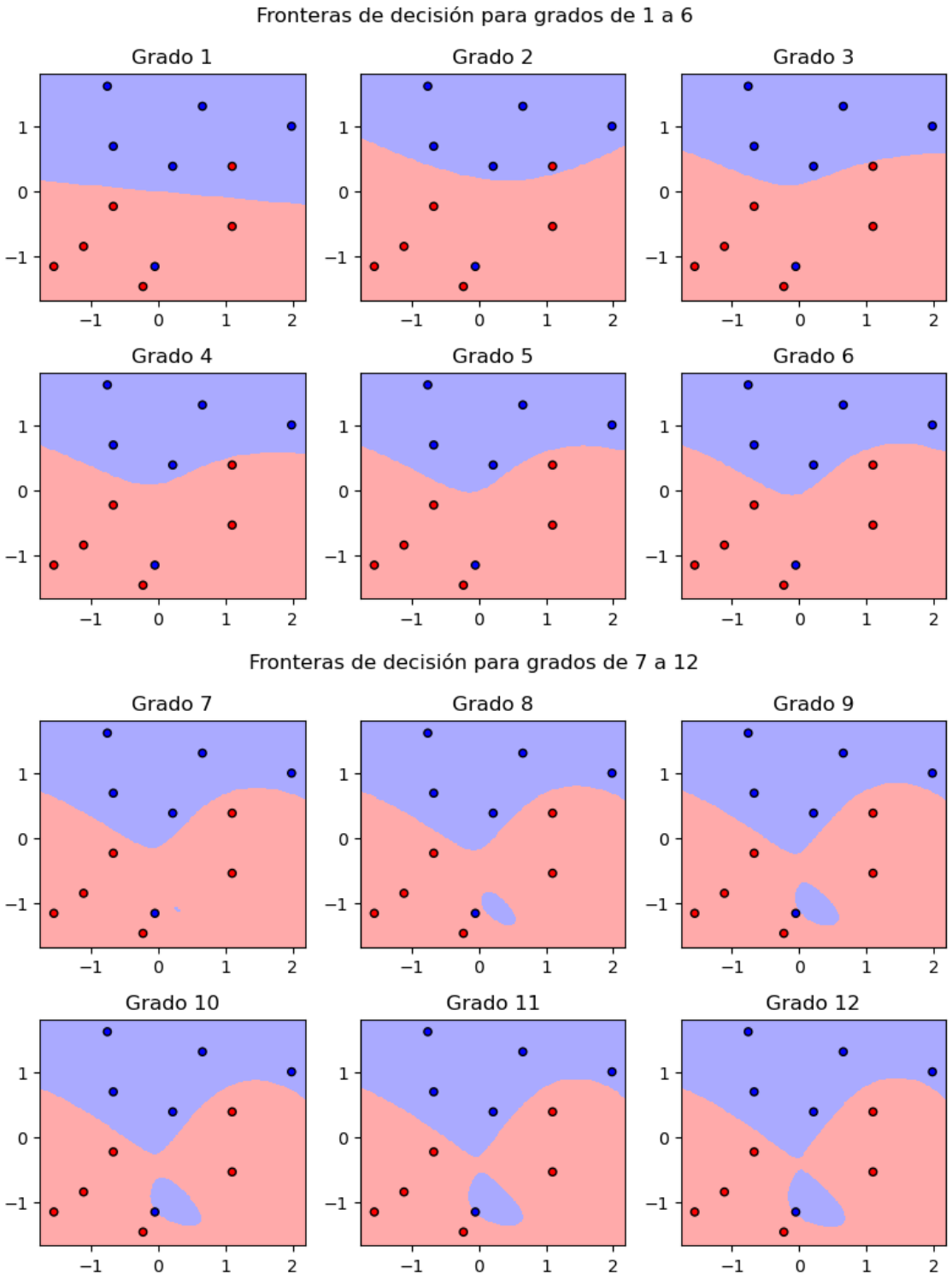


Figura 4: Fronteras de decisión para regresión logística polinomial variando el grado de 1 a 12.

se, debemos diseñar una función vectorial $\boldsymbol{g}(\boldsymbol{x})$ cuyos elementos deberían ser no negativos y sumar uno.

Con este propósito, primero usamos C instancias lineales, cada una denotada como z_c y cada una con un conjunto diferente de parámetros $\boldsymbol{\theta}_c$, $z_c = \boldsymbol{\theta}_c^\top \boldsymbol{x}$. Apilamos todos los z_c en un vector de logits

$$\boldsymbol{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_C \end{bmatrix}$$

y utilizamos la función softmax como una generalización vectorial de la función logística:

$$\text{softmax}(\boldsymbol{z}) = \frac{1}{\sum_{c=1}^{C} e^{z_c}} \begin{bmatrix} e^{z_1} \\ e^{z_2} \\ \vdots \\ e^{z_M} \end{bmatrix}$$

El argumento $\boldsymbol{z}$ para la función softmax es un vector de dimensión C, y devuelve un vector de la misma dimensión. Por construcción, el vector de salida de la función softmax siempre suma 1, y cada elemento es siempre ≥ 0.

En resumen:

$$\boldsymbol{g}(\boldsymbol{x}) = \text{softmax}(\boldsymbol{z}), \text{ donde } \boldsymbol{z} = \begin{bmatrix} \boldsymbol{\theta}_1^\top \boldsymbol{x} \\ \boldsymbol{\theta}_2^\top \boldsymbol{x} \\ \vdots \\ \boldsymbol{\theta}_C^\top \boldsymbol{x} \end{bmatrix}$$

Las probabilidades de clase individuales, es decir, los elementos del vector $\boldsymbol{g}(\boldsymbol{x})$ son

$$g_c(\boldsymbol{x}) = \frac{e^{\boldsymbol{\theta}_c^\top \boldsymbol{x}}}{\sum_{j=1}^{C} e^{\boldsymbol{\theta}_j^\top \boldsymbol{x}}}, \quad c = 1, \ldots, C.$$

Este es el modelo de regresión logística multiclase. Utiliza C vectores de parámetros $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_C$ (uno para cada clase), lo que significa que el número de parámetros a aprender crece con C.

Al igual que con la regresión logística binaria, podemos aprender esos parámetros usando el método de máxima verosimilitud. Utilizamos $\boldsymbol{\theta}$ para denotar todos los parámetros del modelo

$$\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_C\}.$$

Dado que $g_c(\boldsymbol{x}_i; \boldsymbol{\theta})$ es nuestro modelo para Prob $(y_i = c \mid \boldsymbol{x}_i)$, la función de pérdida de entropía cruzada (log-verosimilitud negativa) para el problema multiclase es

$$L(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \underbrace{-\ln g_{y_i}(\boldsymbol{x}_i; \boldsymbol{\theta})}_{L(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}.$$