Índice

1. Modelos Supervisados

1.1. Modelos de Regresión

Introducción

Los problemas de clasificación se predice un 'sí' o un 'no' y después nos fijamos en la probabilidad de equivocarnos, o el número de veces qué nos equivocamos. Sin embargo, las predicciones son de naturaleza probabilística y, en cambio, queremos predecir la probabilidad de un resultado (por ejemplo, la probabilidad de lluvia).

¿Cómo podríamos analizar las predicciones? Por ejemplo, si tenemos dos pronósticos, uno que predice con un 70 % que llueva y el otro predice con una probabilidad del 80 % , ¿cómo podríamos compararlos si llueve? Podemos definir a x como las condiciones climáticas, y a

$$y = \begin{cases} 1 & \text{si llueve} \\ 0 & \text{si no llueve} \end{cases}$$

Deseamos estimar la probabilidad de que llueva dadas las condiciones iniciales, la cual denotaremos por p(x). Observemos que

$$\begin{aligned} p(x) &= P[y = 1|x] \\ &= 1 \cdot P[y = 1|x] + 0 \cdot P[y = 0|x] \\ &= E[y|x]. \end{aligned}$$

Esta última formulación es más general y nos permite predecir, por ejemplo, el número de pulgadas de lluvia que esperamos (una pulgada de lluvia que cae sobre 1 acre de tierra equivale a unos 27 154 galones y pesa unas 113 toneladas).

Este tipo de problemas son llamados de regresión. Estamos tratando de estimar un número de valor real.

En nuestro ejemplo, supongamos que un instituto usa $h_1(x)$ para estimar $p(x)$, mientras que el Servicio Meteorológico utiliza $h_2(x)$. Queremos ver si $h_1(x)$ o $h_2(x)$ predice mejor $p_(x)$, sabiendo que tampoco conocemos $p(x)$.

Por ello, tenemos que confiar solo en los resultados observados y encontrar una manera de "puntuar"las predicciones. Consideraremos una función de "pérdida", que proporciona una medida de cómo puntuar un error. En nuestro primer vistazo, usaremos una función de pérdida cuadrática o cuadrada, que es simplemente el cuadrado de la diferencia entre una probabilidad o expectativa pronosticada y su valor real :

$$L = (h(x) - y)^2$$

Queremos minimizar el valor esperado de nuestra pérdida:

$$E[(h(x) - y)^2]$$

Recordemos que si g es una función y X una v.a discreta, entonces

$$E[g(X)] = \sum_{x \in Dom(X)} P[X = x] g(X)$$

De esta forma, notemos que al minimizar el riesgo, en realidad estamos obligando a h(x) a estar lo más cerca posible de $p(x) = P[y = 1|x]$, pues

$$\begin{aligned} E[(h(x) - y)^2] &= P[y = 1|x](h(x) - y)^2 + P[y = 0|x](h(x) - y)^2 \\ &= p(x)(h(x) - 1)^2 + (1 - p(x))(h(x) - 0)^2 \\ &= p(x)(h(x) - 1)^2 + (1 - p(x))(h(x))^2 \end{aligned}$$

Ahora, para minimizarla, derivamos respecto a h, pues son los valores que deseamos que se aproximen a y e igualamos a cero,

$$\begin{aligned} 0 &= \frac{d}{dh} E[(h(x) - y)^2] \\ &= p(x) \cdot 2(h(x) - 1)(1) + (1 - p(x)) \cdot 2(h(x))(1) \\ &= 2p(x)(h(x) - 1) + 2(1 - p(x))(h(x)) \\ &= 2p(x)h(x) - 2p(x) + 2h(x) - 2p(x)h(x) \\ &= -2p(x) + 2h(x) \\ &= 2[h(x) - p(x)] \end{aligned}$$

La cual se resuelve cuando $h(x) = p(x)$.

Así, el enfoque de esta sección es el aprendizaje por predicción lineal con el enfoque ERM (Empirical risk minimization). Aunque no podemos saber exactamente qué tan bien funcionará un algoritmo en la práctica -el verdadero riesgo- porque no sabemos la verdadera distribución de los datos en los que funcionará el algoritmo, pero podemos medir su desempeño en un conjunto conocido de datos de entrenamiento -el riesgo empírico-.

Se define la clase de funciones afines L_d como

$$L_d = \{h_{w,b} : w \in \mathbb{R}^r, b \in \mathbb{R}\}$$

donde $h_{w,b} : \mathbb{R}^d \to \mathbb{R}$ es tal que :

$$\begin{aligned} h_{w,b}(x) &= \langle w,x \rangle + b \\ &= \sum_{i=1}^{d} w_i x_i + b \end{aligned}$$

De esta manera, L_d es el conjunto de funciones, donde cada una de ellas es parametrizada por $w \in \mathbb{R}^d$ y $b \in \mathbb{R}$, y recibe como input un vector x y regresa un escalar $\langle w,x \rangle + b$. A b se le conoce como el *bias*, el cual se podría interpretar como el error de la aproximación.

Otra forma de denotar el espacio L_d es

$$L_d = \{x \to \langle w,x \rangle + b, w \in \mathbb{R}^d, b \in \mathbb{R}\}$$

A veces es conveniente incorporar a b como la primera coordenada de w , mientras que se añade un 1 en la primera de x. Es decir, se definen w' y x' como

$$w' = (b, w_1, w_2, ..., w_d) \in \mathbb{R}^d \quad \text{y} \quad x' = (1, x_1, x_2, ..., x_d) \in \mathbb{R}^{d+1}$$

tales que

$$h_{w,b}(x) = \langle w,x \rangle + b = \langle w', x' \rangle$$

Así, cada función afín en $\mathbb{R}^d$ puede ser reescrita como una función lineal homogénea en $\mathbb{R}^{d+1}$ bajo la transformación que agrega la constante 1 a la primera entrada de cada vector. De esta manera, para simplificar notación a veces se omitirá el ***bias*** cuando se hable del conjunto L_d.

La familia de hipótesis de clases de la predicción lineal incluye espacios binarios, predicción por regresión lineal y predicción por regresión logística.

Espacios Binarios (Halfspace)

Consideremos un problema de clasificación en el espacio $\gamma = \{-1, +1\}$ y recordemos a la función signo, la cual denotaremos por $\phi_{sign} : \mathbb{R} \to \gamma$, donde para $a \in \mathbb{R}$,

$$\phi_{sign}(a) = \begin{cases} +1 & \text{si } a \geq 0 \\ -1 & \text{si } a < 0 \end{cases}$$

La clase de espacios binarios o de semiespacios HS_d se define como

$$\begin{aligned} HS_d &= \phi_{sign} \circ L_d \\ &= \{x \to \phi_{sign}(h_{w,b}(x)) : h_{w,b} \in L_d\} \end{aligned}$$

Ejemplo. Para ilustrarlo geométricamente,consideremos el caso $d = 2$ con $w = (w_1, w_2) \in$

$\mathbb{R}^2$, $b \in \mathbb{R}$. En este caso, las hipótesis de clase de espacios binarios HS_d son H_1 y H_2 cuya frontera está delimitada por el hiperplano L_H

$$H_1 := w_1x_1 + w_2x_2 + b \geq 0$$
$$H_2 := w_1x_1 + w_2x_2 + b < 0$$
$$L_H := w_1x_1 + w_2x_2 + b = 0$$

Dada L_H, ésta es perpendicular a w , pues para cualesquiera $a, b \in L_H$, $w\dot{(}a-b) = 0$.Las regiones H_1, H_2 intersectan en el punto $\left(0, -\frac{b}{w_2}\right)$ (ordenada al origen de L_H) con el eje x_2. Aquellas $x \in H_1$ serán etiquetadas con +1 y si $x \in H_2$ entonces con -1. Figura 1.

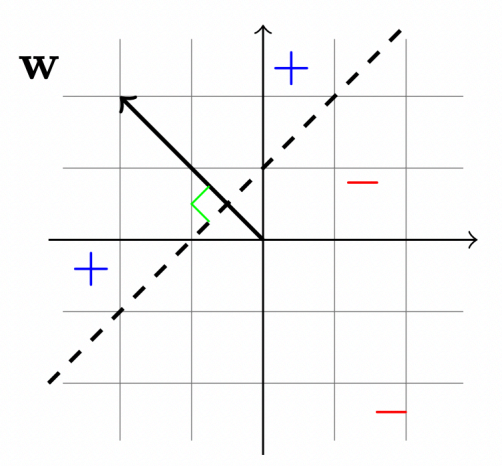

Figura 1: Ejemplo para $\mathbb{R}^2$

En el contexto de Espacios binarios, un caso realizable o separable es aquel donde es posible separar con un hiperplano todos los ejemplares positivos de los negativos y esto se puede encontrar con el enfoque ERM. Por otro lado, para los casos no separables o agnósticos, es computacionalmente difícil con este enfoque.

Observación 3.1. Las diferentes clases de hipótesis para la predicción lineal son composiciones de funciones $\phi : \mathbb{R} \to \gamma$ en el espacio L_d. Por ejemplo, para los espacios binarios ϕ es la función signo y $\gamma = \{+1, -1\}$, mientras que para los problemas de regresión ϕ es simplemente la función identidad y $\gamma = \mathbb{R}$.

Programación Lineal para Espacios Binarios

Un problema de programación lineal (LP) pueden ser expresados como la maximización de una función lineal sujeta a inecuaciones lineales, es decir, deseamos encontrar el vector de variables $w \in \mathbb{R}^d$ tal que

$$max_{w\in\mathbb{R}^d} \langle u, v\rangle \tag{1}$$
$$Aw \geq v \tag{2}$$

Donde $A \in M_{m\times d}(\mathbb{R})$ y $v \in \mathbb{R}^m$, $u \in \mathbb{R}^d$ vectores.

Los problema ERM para semiespacios pueden ser expresado como un problema LP. Por simplicidad, consideremos el caso homogéneo. Sea $S = \{(x_i, y_i) : i \in \{1,...,m\}\}$ el conjunto de entrenamiento de tamaño m. Así el conjunto $\{x_i\}$ es el de vectores de las variables a ingresar y $\{y_i\}$ es el de los escalares a regresar, como estamos en HS_d, entonces los escalares y_i toman valores en el conjunto $\{+1, -1\}$. Considerando un caso realizable, el predictor *ERM* debe de tener cero errores en el conjunto S. De esta manera, buscamos $w \in \mathbb{R}^d$ tal que para toda $i \in \{1,...,m\}$,

$$\phi(\langle w, x_i\rangle) = y_i \tag{3}$$

donde ϕ es la función signo.Observemos que $y_i = +1$ si y sólo si $\langle w, x_i\rangle \geq 0$. Del mismo modo, $y_i = -1$ si y sólo si $\langle w, x_i\rangle < 0$. Así, la 3 es equivalente a

$$y_i \langle w, x_i\rangle > 0 \tag{4}$$

Sea w^* el vector que satisface 4 y $\overline{w} = \frac{w^*}{\frac{1}{\gamma}}$, donde

$$\gamma = \min_{i\in\{1,...,m\}} y_i \langle w^*, x_i\rangle > 0$$

Entonces para toda $i \in \{1,...,m\}$,

$$y_i \langle w^*, x_i\rangle \geq \gamma$$
$$\Rightarrow \frac{1}{\gamma} y_i \langle w^*, x_i\rangle \geq 1$$

Lo que implica que

$$y_i \langle \overline{w}, x_i\rangle \geq 1 \tag{5}$$

Hemos probado que existe un vector que satisface $y_i \langle w, x_i\rangle \geq$ para toda $i \in \{1,...,m\}$. Dicho vector es un predictor *ERM*. Para encontrarlo, usaremos LP. Sea $A \in M_{m\times d}(\mathbb{R})$, tal que para $i \in \{1,...,m\}$ y $j \in \{1,...,d\}$, $A_{i,j} = y_i x_{i,j}$, donde $x_{i,j}$ es el j-ésimo elemento del vector $x_i = (x_{i,1}, x_{i,2}, ..., x_{i,d})$.Sea $v = (1,...,1) \in \mathbb{R}^m$. Entonces la 5 puede ser reescrita como

$$Aw \geq v \tag{6}$$

Se requiere maximizar una ecuación, como todas las w que cumplen con la restricción 6 son candidatas iguales como hipótesis resultante, se propone la ecuación 1 con $u = (0,...,0) \in \mathbb{R}^d$.

Perceptrón para Espacios Binarios

Una implementación con enfoque ERM distinta está dada por el algoritmo llamado Perceptrón (Rosenblatt 1958). Este algoritmo iterativo construye una secuencia de vectores $w^{(1)}, w^{(2)}, ...$ Se inicializa con $w^{(1)} = (0,...,) \in \mathbb{R}^d$ y empieza a iterar. Para la t-ésima iteración, verifica si existe $i \in \{1,...,m\}$ tal que cumpla la condición

$$y_i \langle w^t, x_i \rangle \leq 0$$

(observar que es la negación de 5) , si es así, entonces se define

$$w^{(t+1)} = w^{(t)} + y_i x_i$$

y procede a realizar la iteración $t+1$, sino se satisface la condición, entonces el vector $w^{(t)}$ es el que buscamos y termina el algoritmo.

El *Teorema 3.1* nos afirma que si tenemos un caso realizable, entonces el algoritmo Perceptrón clasifica correctamente los puntos y existe un número máximo de iteraciones en el mismo.

Teorema 3.1.

Supongamos que $S = \{(x_i, y_i) : i\{1,...,m\}\}$ es un conjunto separable (caso realizable). Sea

$$B = min\{||w|| : \forall i \in \{1,...,m\}, y_i \langle w, x_i \rangle \geq 1\}$$

y

$$R = \text{máx}\{||x_i|| : i \in \{1,...,m\}\}$$

El algorimto del Perceptrón se detiene a lo más en $(RB)^2$ iteraciones y si se detiene en la t-ésima iteración, entonces para toda $i \in \{1,...,m\}$,

$$y_i \left\langle w^{(t)}, x_i \right\rangle > 0$$

Demostración.

Por la definición de la condición de paro del algoritmo del Perceptrón, pues si se detiene en la t-ésima iteración, implica que $\forall i \in \{1,...,m\}$, $y_i \left\langle w^{(t)}, x_i \right\rangle > 0$, es decir, cumple con la ecuación 5, por lo tanto se ha clasificado correctamente.

Probemos que si el algoritmo ha iterado T veces, entonces $T \leq (RB)^2$, es decir, que el algoritmo se ejecuta a lo más $(RB)^2$ veces.

Sea $w^* \in B$, sabemos que el conjunto B es no vacío por 1.

Dado que $w^{(1)} = (0,...,0)$, entonces $\left\langle w^*, w^{(1)} \right\rangle = 0$. Para la t-ésima iteración con $t < T$, supongamos que se actualizó $w^{(t+1)}$ con la i-ésima muestra (x_i, y_i) donde $i \in \{1,...,m\}$. De esta forma, $y_i \left\langle w^{(t)}, x_i \right\rangle \leq 0$ y se actualizó $w^{(t+1)} = w^{(t)} + y_i x_i$. Por lo tanto

$$\begin{aligned}
\left\langle w^*, w^{(t+1)} \right\rangle - \left\langle w^*, w^{(t+1)} \right\rangle &= \left\langle w^*, w^{(t+1)} - w^{(t)} \right\rangle \\
&= \left\langle w^*, w^{(t)} + y_i x_i - w^{(t)} \right\rangle \\
&= \langle w^*, y_i x_i \rangle \\
&= y_i \langle w^*, x_i \rangle \\
&\geq 1 \text{ por } 1
\end{aligned}$$

Lo que implica que

$$\left\langle w^*, w^{(T+1)} \right\rangle = \sum_{t=1}^{T} \left(\left\langle w^*, w^{(t+1)} \right\rangle - \left\langle w^*, w^{(t)} \right\rangle \right) \geq \sum_{t=1}^{T} 1 = T \tag{7}$$

Además, para cada $t < T$, se tiene

$$\begin{aligned}
||w^{(t+1)}||^2 &= ||w^{(t)} + y_i x_i||^2 \\
&= ||w^{(t)}||^2 + 2y_i \left\langle w^{(t)}, x_i \right\rangle + y_i^2 ||x_i||^2
\end{aligned}$$

Debido a que $||x_i|| \leq R$ por la definición de R, implica que $||x_i||^2 \leq R^2$. Por otro lado, sin importar que $y_i = 1$ o $y_i = -1$, ocurre que $y_i^2 = 1$. Además se sabe que $y_i \left\langle w^{(t)}, x_i \right\rangle \leq 0$. De esta forma,

$$||w^{(t+1)}||^2 \leq ||w^{(t)}|| + 2 \cdot 0 + 1 \cdot R^2 = ||w^{(t)}|| + R^2$$

Ahora, probemos que para cualquier iteración t, ocurre que $||w^{(t+1)}||^2 \leq tR^2$.

i) Para $t = 1$. Como $||w(1)|| = 0$ y $||w^{(2)}||^2 \leq ||w^{(1)}||^2 + R^2$, entonces

$$||w^{(2)}||^2 \leq 0^2 + R^2 = R^2$$

ii) Supongamos que para $t \in \mathbb{N}$ se cumple que

$$||w^{(t+1)}||^2 \leq tR^2$$

iii) Probemos para $t+1$. Sabemos que $||w^{((t+1)+1)}||^2 \leq ||w^{(t+1)}|| + R^2$. Por hipótesis,

$$||w^{((t+1)+1)}||^2 \leq tR^2 + R^2 = (t+1)R^2$$

En particular, se cumple para T, por lo tanto

$$||w^{(T+1)}||^2 \leq TR^2 \quad \Rightarrow \quad ||w^{(T+1)}|| \leq \sqrt{T}R \tag{8}$$

Por definición de B, $||w^*|| = B \geq 0$. Usando 8,

$$||w^*||||w^{(T+1)}|| \leq B\sqrt{T}R \quad \Rightarrow \quad \frac{1}{||w^*||||w^{(T+1)}||} \geq \frac{1}{B\sqrt{T}R}$$

Con lo anterior y la ecuación 7, se obtiene que

$$\frac{\left\langle w^{(T+1)}, w^* \right\rangle}{||w^*||\ ||w^{(T+1)}||} \geq \frac{T}{B\sqrt{T}R} = \frac{\sqrt{T}}{BR}$$

En otras palabras, se acaba de probar que el coseno del ángulo entre w^* y $w^{(T+1)}$ es al menos $\frac{\sqrt{T}}{RB}$. Por la desigualdad de Cauchy-Schwartz,

$$\left\langle w^{(T+1)}, w^* \right\rangle \leq ||w^*||\ ||w^{(T+1)}|| \quad \Rightarrow \quad \frac{||w^*||\ ||w^{(T+1)}||}{\left\langle w^{(T+1)}, w^* \right\rangle} \geq 1$$

Por lo tanto, $1 \geq \frac{\sqrt{T}}{BR}$. De esta manera, concluimos que

$$T \leq (RB)^2$$

Nota. El algoritmo Perceptrón es seniclo de implementar y garantiza convergencia, sin embargo, depende del valor del parámetro B, el cual puede llegar a tomar exponencialmente grandes en $\mathbb{R}^d$. En tales caso, sería mejor implementar el problema ERM como una solución de programación lineal.

Dimensión VC de los Espacios Binarios

falta ...

1.1.1. Modelos lineales continuos

La regresión lineal es una herramienta estadística común para modelar la relación entre variables explicativas y se respuesta (valores reales). El dominio es un subconjunto χ de $\mathbb{R}^d$ y el conjunto de respuesta o etiquetas es $Y = \mathbb{R}$.

Nos gustaría obtener una función lineal $h : \mathbb{R}^d \to \mathbb{R}$ que mejor aproxime la relación entre nuestras variables. La Figura 2 muestra un ejemplo de un predictor de regresión lineal

para d = 1.

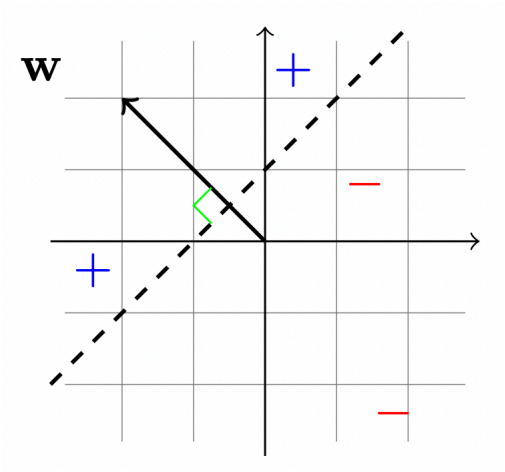

Figura 2: Ejemplo de regresión lineal para $d = 1$

La clase de hipótesis para la predicción por regresión lineal es el conjunto de funciones lineales

$$H_{reg} = L_d = \{x \to \langle w, x \rangle + b : w \in \mathbb{R}^d, b \in \mathbb{R}\}$$

Definimos una función de pérdida para la regresión (*loss function*) denotada por $l(h,(x,y))$.

En clasificación, $l(h,(x,y))$ simplemente indica si $h(x)$ etiqueta correctamente o no, en cambio, en regresión, si deseamos predecir el peso de un bebé de 3kg y obtenemos predicciones de 3.00001 kg y 4 kg, ambas son incorrectas, pero claramente preferiríamos el primero sobre el segundo.

Por ello, necesitamos definir cuánto se penalizará por la discrepancia entre $h(x)$ e y. Una forma común es usar la función de pérdida cuadrática ($squared - lossfunction$), dada por

$$l(h,(x,y)) = (h(x) - y)^2$$

Para esta función de pérdida, la función de riesgo empírica (EMR) se denomina error cuadrático medio ($MeanSquaredError$), dada por

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} (h(x_i) - yi)^2.$$

Hay una variedad de funciones de pérdida que se pueden utilizar, por ejemplo, la función de pérdida de valor absoluto,

$$l(h,(x,y)) = |h(x) - y|.$$

. La regla ERM para la función de pérdida de valor absoluto se puede implementar mediante programación lineal.

Dado que la regresión lineal no es una tarea de predicción binaria, no podemos analizar su complejidad de muestra utilizando la dimensión VC. Un posible análisis de la complejidad es confiar en el "truco de la discretización" Observación 4.1 en el Capítulo 4). PREGUNTAR .

Mínimos cuadrados.

Los mínimos cuadrados (*Least squares*) es el algoritmo que resuelve el problema ERM para la clase H_{reg} con respecto a la función de pérdida cuadrática. Dado un conjunto de entrenamiento S y usando la versión homogénea de L_d, se enfoca en encontrar w que minimice la expresión

$$L_S(h_w) = \frac{1}{m}\sum_{i=1}^{m}(\langle w, x_i\rangle - y_i)^2$$

Para resolverlo, se calcula el gradiente de la función objetivo y se iguala a cero, es decir,

$$\frac{2}{m}\sum_{i=1}^{m}(\langle w, x_i\rangle - y_i)x_i = 0$$

Lo cual es equivalente a

$$\begin{aligned}
&\sum_{i=1}^{m}(\langle w, x_i\rangle - y_i)x_i = 0 \\
\Rightarrow &\sum_{i=1}^{m}(\langle w, x_i\rangle x_i - y_i x_i) = 0 \\
\Rightarrow &\sum_{i=1}^{m}\langle w, x_i\rangle x_i - \sum_{i=1}^{m} y_i x_i = 0 \\
\Rightarrow &\sum_{i=1}^{m}\langle w, x_i\rangle x_i = \sum_{i=1}^{m} y_i x_i
\end{aligned}$$

Notemos que podemos reescribir el problema como

$$Aw = b$$

donde

$$A = \sum_{i=1}^{m} x_i x_i^T \quad \text{y} \quad b = \sum_{i=1}^{m} y_i x_i$$

$$A = \begin{pmatrix} x_1 & x_2 & \dots & x_m \end{pmatrix} \cdot \begin{pmatrix} x_1 & x_2 & \dots & x_m \end{pmatrix}^T$$
$$= \begin{pmatrix} x_1 & x_2 & \dots & x_m \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \text{ (Ver Observación 3.3)}$$

$$\text{y } b = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

Observación 3.3. La matriz A es de $p \times m$, donde cada columna i corresponde al vector de variables x_i con $i \in \{1,2,..,m\}$, además se sabe que x_i tiene p entradas que es el número de variables que determinan la etiqueta de la i-ésima observación. La notación utilizada hace referencia a producto vectorial. Por ejemplo, si $m = 3$ (hubo 3 observaciones) y $p = 2$ (hay dos variables que determinan la etiqueta), entonces

$$A = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \cdot \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix}^T$$
$$= \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$
$$= x_1 \cdot x_1^T + x_2 \cdot x_2^T + x_3 \cdot x_3^T$$
$$= \begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix} \cdot \begin{pmatrix} x_{11} & x_{12} \end{pmatrix} + \begin{pmatrix} x_{21} \\ x_{22} \end{pmatrix} \cdot \begin{pmatrix} x_{21} & x_{22} \end{pmatrix} + \begin{pmatrix} x_{31} \\ x_{32} \end{pmatrix} \cdot \begin{pmatrix} x_{31} & x_{32} \end{pmatrix}$$
$$= x_{11}^2 + x_{12}^2 + x_{21}^2 + x_{22}^2 + x_{31}^2 + x_{32}^2$$

Si A es invertible, la solución está dada por

$$w = A^{-1} b$$

En caso de que A no es invertible, se puede demostrar fácilmente que si las instancias de entrenamiento no generan el espacio $\mathbb{R}^d$, entonces A es no invertible. Sin embargo, podemos encontrar una solución al sistema $Aw = b$ porque b está en el rango de A. Como A es simétrtica, se puede utilizar sus descomposición en eigenvalores dada por

$$A = VDV^T$$

donde D es una matriz diagonal y V es una matriz ortonormal, es decir $V^T \times V = I$.

Se define D^+ como la matriz diagonal tal que

$$D^+_{i,i} = \begin{cases} 0 & \text{si } D_{i,i} = 0 \\ \frac{1}{D_{i,i}} & \text{si } D_{i,i} \neq 0 \end{cases}$$

De esta forma, se definen

$$A+ = VD^+V^T \quad \text{y} \quad \hat{w} = A^+b$$

Sea v_i la i-ésima columna de la matriz V. Entonces

$$\begin{aligned} A\hat{w} &= AA^+b \\ &= (VDV^T)(VD^+V^T)b \\ &= (VD)(V^TV)(D^+V^T)b \\ &= (VD)I(D^+V^T)b \\ &= (VD)(D^+V^T)b \\ &= \sum_{i\in 1,..,m:D_{i,i}\neq 0} v_i v_i^T b \end{aligned}$$

Esto imlpica que $A\hat{w}$ es la poryección de b sobre el conjunto generado por los vectores v_i donde $D_{i,i} \neq 0$. Como el generado por $x_1, ..., x_m$ es el mismo que el de los v_i, y b es generado por los x_i, entonces obtenemos que $A\hat{w} = b$.

Regresión lineal para tareas de regresión polinomial .

Algunas tareas de aprendizaje requieren predictores no lineales, como los predictores polinómicos, como en la Figura 3 , en la cual el conjunto de entrenamiento se ajusta mejor usado un predictor polinomial de tercer grado que usando una función lineal.

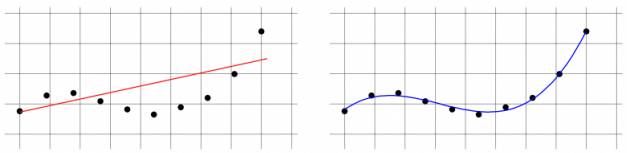

Figura 3: Regresión lineal vs Regresión polinomial

En esta sección nos enfocaremos en la clase de polinomios de una dimensión de grado n, esto es

$$H^n_{poly} := \{x \to p(x) = a_0 + a_1x + a_2x_2 + + a_nx_n\}$$

donde $(a_0, ..., a_n)$ es un vector de coeficientes de tamaño $n+1$. Observemos que el conjunto dominio es $\chi = \mathbb{R}$ porque es un polinomio unidimensional y $Y = \mathbb{R}$ conjunto de respuesta.

Una forma de resolverlo es por medio de reducción del problema que se ha visto en precios capítulos Para mover el problema de regresión polinomial a un problema de regresión lineal, definimos la función $\phi : \mathbb{R} \to \mathbb{R}^{n+1}$ tal que para $x \in \mathbb{R}$,

$$\phi(x) = (1, x, x^2, ..., x^n)$$

. De esta manera, $p(\phi(x) = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n = \langle a, \phi(x) \rangle$ y se puede encontrar el vector óptimo de coeficientes a utilizando el algoritmo de Mínimos cuadrados.

1.1.2. Modelos lineales discretos

Introducción.

El problema de clasificación en dos grupos puede abordarse introduciendo una variable ficticia binaria para representar la pertenencia de una observación a uno de los dos grupos. Por ejemplo, si se desea discriminar entre créditos que se devuelven o que presentan problemas para su cobro, puede añadirse a la base de datos una nueva variable y que tome el valor 0, cuando el crédito se devuelve sin problemas Y valor 1 en otro caso. El problema de discriminación es equivalente a la previsión del valor de la variable ficticia Y. Si el valor previsto está más próximo a 0 que a 1, clasificaremos al elemento en la primera población. En otro caso, lo haremos en la segunda. Supongamos que Y viene explicada por un conjunto de variables m variables $X_1, X_2, \ldots, X_m$. Por otra parte, podemos pensar en utilizar un modelo de regresión lineal múltiple para explicar el comportamiento de la variable Y, es decir:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m + e,$$

Observación. La variable del error e viene siendo el bias y $\beta_0, ..., \beta_m$ la $w \in \mathbb{R}^m$ que deseamos estimar utilizando la notación del espacio L_d.

Bajo el supuesto habitual de que $E(e) = 0$, y suponiendo conocidos los valores que toman las variables explicativas (observaciones), tendremos que:

$$E[Y|X_1, \ldots, X_m] = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m$$

Observemos que por ser Y una variable binaria (i.e.: sólo podrá tomar los valores 0 y 1), siempre se cumplirá que:

$$E(Y) = 0 \cdot P(Y = 0) + 1 \cdot P(Y = 1) = P(Y = 1), \text{ entonces}$$

$$E(Y|X_1, \ldots, X_m) = P(Y = 1|X_1, \ldots, X_m) = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m = Y - e,$$

para $X_1, \ldots, X_m$ conocidos. Observar que esta expresión nos viene a decir que podemos expresar la variable dependiente binaria Y como la probabilidad de "éxito"más un término de perturbación, es decir:

$$Y = P(Y = 1|X_1, \ldots, X_m) + e = E[Y|X_1, \ldots, X_m] + e.$$

Sin embargo, este modelo inicial no será válido para explicar el comportamiento de variables dependientes binarias, pues presenta varios problemas:

1. El error e ya no será una variable aleatoria continua (como ocurría en el los modelos lineales de regresión múltiple (MLRM)), sino que será una variable aleatoria discreta, puesto que, conocidos los valores de las variables explicativas, e sólo puede tomar dos valores determinados. Por tanto, e ya no se distribuirá de forma normal (uno

de los supuestos básicos del MRLM). Si bien este supuesto no resulta estrictamente necesario para aplicar mínimos cuadrados ordinarios (MCO), sí es fundamental a la hora de realizar cualquier tipo de inferencia posterior sobre el modelo (intervalos de confianza para los parámetros estimados, contrastes de hipótesis, etc.).

2. El término de perturbación no cumple la hipótesis de homoscedasticidad (la varianza de dicho término no es constante). Esto se da por que $Var(Y|X_1,\ldots,X_m) = P(Y = 1|X_1,\ldots,X_m) * [1 - P(Y = 1|X_1,\ldots,X_m)]$ varía para cada observación i de la muestra. Debido a este problema, MCO no serán eficientes, por lo que resultará necesario recurrir a la estimación por mínimos cuadrados generalizados (MCG).

3. Se corre el riesgo de predecir valores de Y menores que 0 y mayores que 1.

4. Finalmente, la expresión $P(Y = 1|X_1,\ldots,X_m) = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m$ nos dice que la probabilidad del suceso de "éxito"viene determinada por una combinación lineal de variables. De ello se deduce que la variación en $P(Y = 1|X_1,\ldots,X_m)$ causada por cambios en alguna de las variables explicativas es constante (y, por tanto, independiente del valor actual de dicha variable explicativa), lo cual es una hipótesis muy poco realista, pues si se perturba la variable X_i, digamos X_i', entonces la variación viene dada por :

$$\begin{aligned}
\frac{\delta P(Y = 1|X_1,\ldots,X_m)}{\delta X_i} &= \frac{P(Y = 1|X_1,\ldots X_i,\ldots,X_m) - P(Y = 1|X_1,\ldots X_i',\ldots,X_m)}{X_i - X_i'} \\
&= \frac{[\beta_0 + \ldots + \beta_i X_i + \ldots + \beta_m X_m] - [\beta_0 + \ldots + \beta_i X_i' + \ldots + \beta_m X_m]}{X_i - X_i'} \\
&= \frac{\beta_i [X_i - X_i']}{X_i - X_i'} \\
&= \beta_i
\end{aligned}$$

Para evitar las inconsistencias anteriores se han desarrollado modelos no lineales, los cuales tratan de resolver los problemas anteriores. La idea consiste utilizar un modelo de la forma

$$Y = f(\beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m) + e,$$

donde f es la función real que depende de la expresión lineal $\beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m$. Con el nuevo modelo, y razonando de forma similar al caso del modelo lineal, se cumplirá:

$$E[Y|X_1,\ldots,X_m] = P(Y = 1|X_1,\ldots,X_m) = f(\beta_1 + \beta_1 X_1 + \ldots + \beta_m X_m).$$

Ahora bien, ¿qué tipo de función f estamos buscando?: obviamente, f deberá ser distinta de la función identidad (para evitar los problemas 3 y 4). La clase de funciones no decrecientes, acotadas entre cero y uno, es la clase de las funciones de distribución, por lo que el problema se resuelve tomando como f cualquier.

Regresión Logística.

La regresión logística se utiliza para tareas de clasificación. Trabajaremos en la familia de funciones h que mapean $\mathbb{R}^d$ en el intervalo $[0,1]$.

Se puede interpretar entonces $h(x)$ como la probabilidad de que la etiqueta de x sea 1.Así, la hipótesis asociada con la regresión logística H_{sig} es la composición de la función sigmoid $\phi_{sig} : \mathbb{R} \to [0,1]$ sobre la clase de funciones L_d, es decir,

$$H_{sig} = \phi_{sig} \circ L_d = \{x \to \phi_{sig}(\langle w,x \rangle) : w \in \mathbb{R}^d\}.$$

La función sigmoid usada en regresión logística es la función logic o logística, definida como

$$\phi_{sig}(z) = \frac{1}{1+exp(-z)}.$$

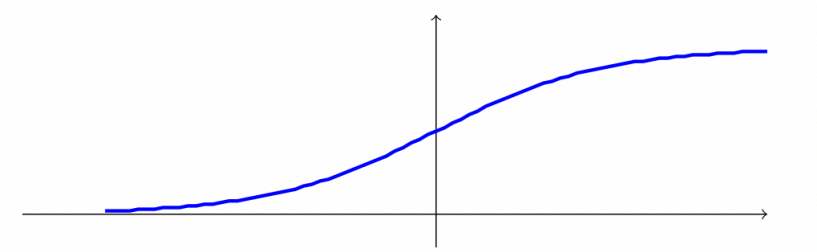

Figura 4: Función logic

Notemos que cuando $\langle w,x \rangle$ es muy grande, entonces $\phi_{sig}(\langle w,x \rangle)$ es cercano a 1, mientras que para valores muy pequeños se aproxima a 0 (Figura 4).

Observación. Recordemos que en el Espacio Binario (Halfspace) (Sección 1) la predicción correspondiente a w es $\phi_{sign}(\langle w,x \rangle)$, por lo tanto las predicciones en regresión logística respecto al espacio binario son muy similares cuando $|\langle w,x \rangle|$.

Cuanto $|\langle w,x \rangle|$ es cercano a 0, $\phi_{sig}(\langle w,x \rangle) \approx \frac{1}{2}$.
En este caso, la regresión logística es incierta respecto a la etiqueta así que determina que es $\phi_{sign}(\langle w,x \rangle)$ con una probabilidad un poco mayor al 50%. En cambio, el espacio binario da una predicción determinista, 1 o -1 incluso si $|\langle w,x \rangle|$ toma un valor cercano a cero.

Definiremos una función de perdida, es decir, que tan mal predice $h_w(x) \in [0,1]$ con $h_w \in H_{sig}$dada su verdadera etiqueta $y \in \{-1,+1\}$. De esta manera, deseamos que $h_w(x)$ (probabilidad de que la etiqueta sea 1) tome valores grandes cuando $y = 1$ y que $1 - h_w(x)$ (probabilidad de que la etiqueta sea -1) sea grande si $y = -1$. Notemos que

$$
\begin{aligned}
1 - h_w(x) &= 1 - \phi_{sig}(\langle w,x\rangle) \\
&= 1 - \frac{1}{1+exp(-\langle w,x\rangle)} \\
&= \frac{exp(-\langle w,x\rangle)}{1+exp(-\langle w,x\rangle)} \\
&= \frac{1}{exp(\langle w,x\rangle)} \cdot \frac{1}{1+exp(-\langle w,x\rangle)} \\
&= \frac{1}{1+exp(\langle w,x\rangle)}
\end{aligned}
$$

Por lo tanto, deseamos que cualquier función de perdida crezca monótonamente con

$$\frac{1}{1+exp(y\langle w,x\rangle)}$$

, lo cuál es equivalente a que crezca monótonamente con $1+exp(-y\langle w,x\rangle$.
Si lo anterior se cumple, entonces :

i) La función de pérdida aumentaría cuando $1-h_w(x)$ (proba de que $y=-1$) es grande y se le asignara la etiqueta $y=1$, pues $1+exp(-y\langle w,x\rangle)$ sería tal que

$$
\begin{aligned}
1+exp(-y\langle w,x\rangle) &= \frac{1}{1+exp(y\langle w,x\rangle)} \\
&= \frac{1}{1+exp(1\cdot\langle w,x\rangle)} \\
&= \frac{1}{1+exp(\langle w,x\rangle)} \\
&= 1-h_w(x) \text{ que toma un valor grande}
\end{aligned}
$$

de esta manera, como $1-h_w(x)$ es grande, entonces es más probable que la etiqueta real fuera $y=-1$, si se le hubiese asignado $y=+11$ (como se desarrolló en la ecuación), al ser la función de pérdida monótonamente creciente con $1+exp(-y\langle w,x\rangle = 1-h_w(x)$, entonces la función de pérdida tomaría un valor grande (que es lo que deseamos, pues se le estaría asignando una etiqueta incorrecta en este caso).

ii) Del mismo modo, la función de pérdida aumentaría cuando $h_w(x)$ (proba de que $y=+1$) es grande y se le asignara la etiqueta $y=-1$, pues $1+exp(-y\langle w,x\rangle)$ sería tal que

$$
\begin{aligned}
1 + exp(-y\langle w,x\rangle) &= \frac{1}{1+exp(y\langle w,x\rangle)} \\
&= \frac{1}{1+exp(-1\cdot\langle w,x\rangle)} \\
&= \frac{1}{1+exp(-\langle w,x\rangle)} \\
&= h_w(x) \text{ que toma un valor grande}
\end{aligned}
$$

Como $h_w(x)$ es grande, entonces es más probable que la etiqueta real fuera $y = +1$, si se le hubiese asignado $y = -1$, al ser la función de pérdida monótonamente creciente con $1 + exp(-y\langle w,x\rangle = h_w(x)$, entonces la función de pérdida tomaría un valor grande (que es lo que deseamos, pues se le estaría asignando una etiqueta incorrecta).

Así, en la regresión logística, la función de perdida penaliza h_w basado en el logaritmo natural de $1 + exp(-y\langle w,x\rangle)$, esto es

$$l(h_w,(x,y)) = log(1 + exp(-y\langle w,x\rangle)$$

Finalmente, dado un conjunto de entrenamiento $S = \{(x_1,y_1),\ldots,(x_m,y_m)\}$, el problema ERM asociado con la regresión logística es w_{sig} donde w_{sig} es tal que para toda $w \in \mathbb{R}^d$,

$$\frac{1}{m}\sum_{i=1}^{m} log(1 + exp(-y_i\langle w_{sig},x_i\rangle) \leq \frac{1}{m}\sum_{i=1}^{m} log(1 + exp(-y_i\langle w,x_i\rangle)$$

Otro enfoque, es que como acabamos de ver, una posible solución a las inconsistencias que presentaba el modelo de probabilidad lineal para explicar el comportamiento de una variable dependiente binaria es usar un modelo Logit de la forma:

$$Y = f(\beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m) + e,$$

donde f es la función logística, es decir:

$$f(z) = \frac{\exp(z)}{1+\exp(z)} = \frac{1}{1+\exp(-z)}.$$

Por lo tanto tenemos que

$$E[Y|X_1,\ldots,X_m] = P(Y = 1|X_1,\ldots,X_m) = \frac{1}{1+\exp-(\beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m)}.$$

El modelo puede ser linealizado utilizando la simple transformación

$$\ln\Big(\frac{P(Y=1|X_1,\ldots,X_m)}{1-P(Y=1|X_1,\ldots,X_m)}\Big)=\beta_0+\beta_1X_1+\ldots+\beta_mX_m,$$

comúnmente llamada *transformación logit*. De esta ecuación notamos que

$$O=\frac{P(Y=1|X_1,\ldots,X_m)}{1-P(Y=1|X_1,\ldots,X_m)}=\exp(\beta_0+\beta_1X_1+\ldots+\beta_mX_m)=\exp(\beta_0)\cdot\prod_{i=1}^{m}\exp(\beta_j)^{X_j},$$

donde $\exp(\beta_0)$ y $\exp(\beta_j)$ se les llama *odds (0)* o *ratios de probabilidades* , estos valores indican cuánto se modifican las probabilidades por unidad de cambio en las variables X.

Supongamos que consideramos dos elementos que tienen valores iguales en todas las variables menos en una. Sean $(x_{i1},\ldots,x_{ih},\ldots,x_{im})$ el vector de valores para el primer elemento y $(x_{j1},\ldots,x_{jh},\ldots,x_{jm})$ para el segundo, y todas las variables son las mismas en ambos elementos menos en la variable h donde $x_{ih}=x_{jh}+1$. Entonces, el odds ratio para estas dos observaciones es: $\frac{O_i}{O_j}=e^{\beta_h}$ e indica cuánto se modifica el ratio de probabilidades cuando la variable x_j aumenta en una unidad. Se deduce también que un coeficiente β_i cercano a cero, equivalentemente, un odds-ratio cercano a uno significará que cambios en la variable explicativa X_i asociada no tendrán efecto alguno sobre la variable dependiente Y.

Los parámetros $\beta_0,\beta_1,\ldots,\beta_m$ son estimados por máxima verosimilitud. Para una muestra aleatoria $y_1,\ldots,y_n$ de una distribución Bernoulli con distribución $P(y_i=0|X_1,\ldots,X_m)=1-p_i$ y $P(y_i=1|X_1,\ldots,X_m)=p_i$, la función de verosimilitud es

$$L(\beta_0,\beta_1,\ldots,\beta_m)=f(y_1,\ldots,y_n;\beta_0,\beta_1,\ldots,\beta_m)=\prod_{i=1}^{n}f_i(y_i;\beta_0,\beta_1,\ldots,\beta_m)=\prod_{i=1}^{n}p_i^{y_i}(1-p_i)^{1-y_i}.$$

De lo cual se obtiene que

$$\ln L(\beta_0,\beta_1,\ldots,\beta_m)=\sum_{i=1}^{n}y_i(\beta_0+\sum_{j=1}^{m}\beta_jx_{ij})-\sum_{i=1}^{n}\ln(1+e^{\beta_0+\sum_{j=1}^{m}\beta_jx_{ij}}). \tag{9}$$

Diferenciando con respecto a $\beta_0,\ldots,\beta_m$ e igualando a cero tenemos que

$$\text{Para } \beta_0 \quad \sum_{i=1}^{n}y_i=\sum_{i=1}^{n}\frac{1}{1+e^{-\hat{\beta}_0-\sum_{j=1}^{m}\hat{\beta}_jx_{ij}}},$$

$$\text{y para toda } \beta_j \text{ con } j=1,\ldots,m \quad \sum_{i=1}^{n}y_ix_{ij}=\sum_{i=1}^{n}\frac{1}{1+e^{-\hat{\beta}_0-\sum_{j=1}^{m}\hat{\beta}_jx_{ij}}}.$$

Estas ecuaciones se resuelven iterativamente para $\hat{\beta}=(\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_m)$, de la misma manera denotemos a $\hat{Y}$ como el vector estimado bajo el modelo logístico usando los estimadores $\hat{\beta}$.

Para medir la potencia de ajuste del vector estimado $\hat{\beta}$ para el modelo logístico se utiliza la *devianza* o también llamada como la *devianza residual o pseudoresiduos*, esta se puede interpretar como la suma de los errores cuadrados en regresión múltiple lineal. La devianza se define como menos dos veces el logaritmo natural de la log-verosimilitud de los valores ajustados.

$$D = -2\ln L(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m) = -2\sum_{i=1}^{n} y_i(\hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_{ij}) + 2\sum_{i=1}^{n} \ln(1 + e^{\hat{\beta}_0 + \Sigma_{j=1}^{m} \hat{\beta}_j x_{ij}}).$$

Entre más pequeña sea ésta mejor es el ajuste. Para medir la significancia (bajo la hipótesis de que $\beta_j = 0$ $(j = 0, 1, \ldots, m)$) de las variables del modelo se utiliza la prueba univariada de Wald, la cual se obtiene de la siguiente manera

$$W_j = \frac{\hat{\beta}_j}{\hat{\mathrm{SE}}(\hat{\beta}_j)} \qquad j = 0, 1, \ldots, m$$

donde W_j se distribuye como una normal estándar, entonces si $\alpha = P(|z| > W_j)$ es menor a un nivel de significancia α' se rechaza la hipótesis nula de que $\beta_j = 0$ $(j = 0, 1, \ldots, m)$, donde α es el p-value.

Para calcular los intervalos de confianza de los estimadores se usa la siguiente expresión

$$\hat{\beta}_j \pm z_{1-\alpha/2} \hat{\mathrm{SE}}(\hat{\beta}_j) \qquad j = 0, 1, \ldots, m.$$

La matriz de var-cov del estimador $\hat{\beta}$ es $\hat{\mathrm{Var}}(\hat{\beta}) = (X^\top V X)^{-1}$, donde

$$X = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{nm} \end{bmatrix}, \quad \text{y} \quad V = \begin{bmatrix} \hat{y}_1(1-\hat{y}_1) & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \hat{y}_n(1-\hat{y}_n) \end{bmatrix}.$$

Y podemos identificar a $\hat{SE}(\hat{\beta}_j)$ como la raíz cuadrada del j-ésimo elemento de la diagonal de la matriz $\hat{\mathrm{Var}}(\hat{\beta})$, es decir, $\hat{\mathrm{SE}}(\hat{\beta}_j) = \sqrt{\hat{\mathrm{Var}}(\hat{\beta})_{jj}}$.

La prueba multivariada de Test para la hipótesis nula de que cada uno de los $m+1$ coeficientes β sea igual a cero es

$$W = \hat{\beta}^\top \left[\hat{\mathrm{Var}}(\hat{\beta})\right]^{-1} \hat{\beta} = \hat{\beta}^\top \left(X^\top V X\right) \hat{\beta},$$

el cual tiene una distribución chi-cuadrada con $m+1$ grados de libertad, donde el $p-value$ se calcula como $p-value = P[\chi_{p+1} \geq W]$, y si este es menor a un nivel de significancia α' entonces se rechaza la hipótesis nula. Si se requiere hacer la prueba para sólo $h < m+1$ coeficientes únicamente se tienen que eliminar las β que no se requiera probar y la fila y columna respectiva de la matriz $(X^\top V X)$. Para medir la potencia del ajuste del modelo se utilizan comúnmente los estadísticosχ^2 (Estadístico de Pearson) y D (Devianza)

$$\chi^2 = \sum_{j=1}^{n} \frac{\left[y_i - \left(1 + \exp(-[\hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_{ij}]) \right)^{-1} \right]^2}{\left(1 + \exp(-[\hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_{ij}]) \right)^{-1} \left[1 - \left(1 + \exp(-[\hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_{ij}]) \right)^{-1} \right]}$$

$$D = -2 \sum_{i=1}^{n} y_i (\hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_{ij}) + 2 \sum_{i=1}^{n} \ln(1 + e^{\hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_{ij}}).$$

Si n es suficientemente largo, en ambos grupos, ambos estadísticos se distribuyen aproximadamente como una $\chi^2_{n-(m+1)}$. Por otra parte, si n es grande y n_1 o n_2 $(n_1 + n_2 = n)$ es pequeño, entonces el uso de estos estadísticos para medir la potencia del ajuste puede ser peligroso, por lo que para valores grandes de χ^2 o D no es evidencia suficiente de falta de ajuste.

Entonces para un nivel de tolerancia específico α_T rechazamos la hipótesis de que el modelo no se distribuye como una $\chi^2_{n-(m+1)}$, si $p - value_D = P[\chi_{n-(m+1)} \leq D]$ o $p - value_{\chi^2} = P[\chi_{p+1} \leq \chi^2]$, es menor que α_T, es decir rechazamos si $\{p - value_D, p - value_{\chi^2}\} \leq \alpha_T$.

Bondad de ajute para la Regresión Logística.

Existen distintos métodos y estad´ısticas para probar la bondad de ajuste de los modelos de regresión logística. Entre ellos están las **Tablas de clasificación.** En ellas e calculan las probabilidades ajustadas $\hat{\pi}_i$ y para cada caso i se obtienen las predicciones (o clasificaciones), "´exito" o "fallo" ("positivo" o "negativo"), dependiendo de si $\hat{\pi}_i$ es mayor o menor que cierto umbral.

Con ello se obtiene la tabla de clasificación 1

	Positivos	**Negativos**
Positivos	VP	FP
Negativos	FN	VN

Cuadro 1: Tabla de clasificación. *Predicciones* (en itálicas) y **Observaciones** (en negritas)

donde

- VP : = Verdaderos positivos
- FN = falsos negativos
- VN:= verdaderos negativos
- FP:= falsos positivos

La utilidad del modelo se resumen con la medida de la *Sensibilidad* y la *Especificidad*.

La *Sensibilidad* es la frecuencia relativa de predecir un evento como positivo cuando el evento observado es positivo, es decir,

$$Sensibilidad = \frac{VP}{VP+FN}$$

Por otro lado, la *Especificidad* es la frecuencia relativa de predecir un evento como negativo cuando el evento observado es negativo, de esta forma,

$$Especificidad = \frac{VN}{VN+FP}$$

Así, lo ideal sería que ambas medidas fueran cercanas a 1.

También, existen las **Curvas ROC** (Receiver Operating Characteristic) las cuales grafican la sensibilidad y especificidad para cada umbral. Tradicionalmente, en el eje horizontal se grafica $1 especificidad$, y en el eje vertical se grafica la $sensibilidad$.

Con esta orientación de los ejes, un valor del eje x cercano a cero (alta especificidad) generalmente implica un valor del eje y bajo (baja sensibilidad), y viceversa. Todas las curvas ROC comienzan en el punto (0, 0), terminan en el punto (1, 1) y son monótonas crecientes. Un modelo que predice adecuadamente resulta en una curva ROC que crece rápidamente a 1: cuanto más cercana esté a la curva a la parte superior izquierda, mejor serán sus predicciones. Generalmente se calcula el área debajo de la curva ROC, y ésta es una medida de la capacidad predictiva del modelo.

Regresión Probit.

Otra posible solución a las inconsistencias que presentaba el modelo de probabilidad lineal - para explicar el comportamiento de una variable dependiente binaria- es usar un modelo Probit (también llamado modelo Normit) de la forma:

$$Y = f(\beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m) + e,$$

f es la función de distribución de una normal estándar, es decir

$$f(z) = \int_{-\infty}^{z} \frac{\exp(-t^2/2)}{\sqrt{2\pi}} dt.$$

Por lo tanto tenemos que

$$E[Y|X_1,\ldots,X_m] = P(Y=1|X_1,\ldots,X_m) = \int_{-\infty}^{\beta_0+\beta_1 X_1+\ldots+\beta_m X_m} \frac{\exp(-t^2/2)}{\sqrt{2\pi}} dt.$$

En este tipo de modelos no resulta posible interpretar directamente las estimaciones de los parámetros β, ya que son modelos no lineales. Lo que haremos en la practica es tomar en

cuenta el signo de los estimadores. Si el estimador es positivo, significará que incrementos en la variable asociada causan incrementos en $P(Y=1|X_1,\ldots,X_m)$ (aunque desconocemos la magnitud de los mismos). Por el contrario, si el estimador muestra un signo negativo, ello supondrá que incrementos en la variable asociada causaran disminuciones en $P(Y=1|X_1,\ldots,X_m)$.

Podemos ver que $P(Y=1|X_1,\ldots,X_m)=\Phi(X\beta^\top)$ con $X=(1,X_1,\ldots,X_m)_{n\times(m+1)}$, $\beta=(\beta_0,\ldots,\beta_m)$, y $\Phi(\cdot)$ la función de distribución acumulativa de una normal estándar. Los parámetros β son estimados por máxima verosimilitud. Supongamos que tenemos una muestra de n observaciones independientes y_i y x_{ij} $(j=1,\ldots,m+1)$, con $x_{i1}=1$ para todo $i=1,\ldots,n$, entonces la log-verosimilitud conjunta es

$$\ln L(\beta)=\ln\Big\{\prod_{i=1}^{n}\Phi(x_i\beta^\top)^{y_i}(1-\Phi(x_i\beta^\top))^{1-y_i}\Big\}=\sum_{i=1}^{n}\Big(y_i\ln\Phi(x_i\beta^\top)+(1-y_i)\ln(1-\Phi(x_i\beta^\top))\Big).$$

De igual forma que en el modelo logit, para maximizar esta log-verosimilitud se tiene que derivar con respecto β igualarse a cero y resolver vía optimización, para β_j $(j=0,\ldots,m)$ tenemos

$$\sum_{i=1}^{n}\frac{y_i\varphi(x_i\beta^\top)}{\Phi(x_i\beta^\top)}=\sum_{i=1}^{n}\frac{(1-y_i)\varphi(x_i\beta^\top)}{1-\Phi(x_i\beta^\top)},$$

donde $\varphi(\cdot)$ es la función de densidad de una normal estándar.

Una vez obtenido los estimadores $\hat{\beta}$, vía optimización, para medir la potencia del ajuste de esta estimación, se utiliza la devianza $-2\ln L(\hat{\beta})$, entre más pequeña sea ésta mejor es el ajuste. Para medir la significancia (bajo la hipótesis de que $\beta_j=0$ $(j=0,1,\ldots,m)$) de las variables del modelo se utiliza la prueba univariada de Wald, la cual se obtiene de la siguiente manera

$$W_j=\frac{\hat{\beta}_j}{\hat{\mathrm{SE}}(\hat{\beta}_j)}\qquad j=0,1,\ldots,m$$

donde W_j se distribuye como una normal estándar, entonces si $\alpha=P(|z|>W_j)$ es menor a un nivel de significancia α' se rechaza la hipótesis nula de que $\beta_j=0$ $(j=0,1,\ldots,m)$, donde α es el p-value.
Para calcular los intervalos de confianza de los estimadores se usa la siguiente expresión

$$\hat{\beta}_j\pm z_{1-\alpha/2}\hat{\mathrm{SE}}(\hat{\beta}_j)\qquad j=0,1,\ldots,m.$$

La matriz de var-cov del estimador $\hat{\beta}$ es $\hat{\mathrm{Var}}(\hat{\beta})=(X^\top VX)^{-1}$, donde

$$X=\begin{bmatrix}1 & x_{11} & \ldots & x_{1m}\\ \vdots & \vdots & \vdots & \vdots\\ 1 & x_{n1} & \ldots & x_{nm}\end{bmatrix},\quad \text{y}\quad V=\begin{bmatrix}\frac{\varphi^2(x_1\beta^\top)}{\hat{y}_1(1-\hat{y}_1)} & 0 & \ldots & 0\\ \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \ldots & \frac{\varphi^2(x_n\beta^\top)}{\hat{y}_n(1-\hat{y}_n)}\end{bmatrix},$$

donde $\varphi(z)=\exp(-\frac{z^2}{2})/\sqrt{2\pi}$.Podemos identificar a $\hat{SE}(\hat{\beta}_j)$ como la raíz cuadrada del

j-ésimo elemento de la diagonal de la matriz $\hat{\text{Var}}(\hat{\beta})$, es decir, $\hat{\text{SE}}(\hat{\beta}_j) = \sqrt{\hat{\text{Var}}(\hat{\beta})_{jj}}$.La prueba multivariada de Test para la hipótesis nula de que cada uno de los $m+1$ coeficientes β sea igual a cero es

$$W = \hat{\beta}^\top \left[\hat{\text{Var}}(\hat{\beta})\right]^{-1} \hat{\beta} = \hat{\beta}^\top \left(X^\top V X\right) \hat{\beta},$$

el cual tiene una distribución chi-cuadrada con $m+1$ grados de libertad, donde el $p-value$ se calcula como $p-value = P[\chi_{m+1} \geq W]$, y si este es menor a un nivel de significancia α' entonces se rechaza la hipótesis nula. Si se requiere hacer la prueba para sólo $h < m+1$ coeficientes únicamente se tienen que eliminar las β que no se requiera probar y la fila y columna respectiva de la matriz $(X^\top V X)$.

Para medir la potencia del ajuste del modelo se utilizan comúnmente estos estadísticos

$$\text{Estadístico de Pearson} \quad \chi^2 = \sum_{j=1}^{n} \frac{(y_i - \Phi(x_i\hat{\beta}^\top)^2}{\Phi(x_i\hat{\beta}^\top)(1-\Phi(x_i\hat{\beta}^\top))} \qquad \text{y}$$

$$\text{Devianza} \quad D = -2\sum_{i=1}^{n} \left(y_i \ln \Phi(x_i\hat{\beta}^\top) + (1-y_i)\ln(1-\Phi(x_i\hat{\beta}^\top))\right).$$

Si n es suficientemente largo, en ambos grupos, ambos estadísticos se distribuyen aproximadamente como una $\chi^2_{n-(m+1)}$. Por otra parte, si n es grande y n_1 o n_2 $(n_1+n_2=n)$ es pequeño, entonces el uso de estos estadísticos para medir la potencia del ajuste puede ser peligroso, por lo que para valores grandes de χ^2 o D no es evidencia suficiente de falta de ajuste.

Entonces para un nivel de tolerancia específico α_T rechazamos la hipótesis de que el modelo no se distribuye como una $\chi^2_{n-(m+1)}$, si $p-value_D = P[\chi_{n-(m+1)} \leq D]$ o $p-value_{\chi^2} = P[\chi_{p+1} \leq \chi^2]$, es menor que α_T, es decir rechazamos si $\{p-value_D, p-value_{\chi^2}\} \leq \alpha_T$.

1.1.3. Modelos lineales Generalizados

Los modelos lineales generalizados (MLG) son aquellos en los que una variable, la cual se conoce como *dependiente* o *respuesta*, se explica por medio de la combinación lineal de otras variables, que son llamadas *independientes*, *explicativas* o como *covariables*, cuando se trata de categorías se le conoce como *factores*.

Los MLG son una extensión de los modelos lineales y se caracterizan por

- *Componente aleatorio*. Un vector de n observaciones, digamos

$$y = (y_1, ..., y_n)$$

, donde y_i es la realización de la variable aleatoria Y_i. El vector de v.a independientes $Y = (Y_1, ..., Y_n)$ es tal que toda Y_i pertenece a la familia exponencial y

$$E[Y_i] = \mu$$

- *Componente sistemática*. El conjunto de covariables $\{x_1, ..., x_p\}$ forma un predictor lineal dado por

$$\eta = \sum_{j=1}^{p} x_i B_j$$

o bien, en notación matricial, sea $X \in M_{n\times p}(\mathbb{R})$, es decir, una matriz de $n \times p$, cuya $j-sima$ columnas vienen dadas por la covariable $x_j = (x_{1j}, ..., x_{nj})$ y $B \in_{p\times 1} M(\mathbb{R})$ el vector de parámetros desconocidos cuyo j-ésimo renglón está dado por β_i, entonces

$$\begin{aligned}\eta &= \begin{pmatrix} x_1 & x_2 & ... & x_p \end{pmatrix} \times \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \\ &= X\beta\end{aligned}$$

- *Liga o función de enlace*: La función de enlace relaciona los componentes aleatorio y sistemático:

$$\eta_i = g(\mu_i)$$

La familia exponencial y los MLG.

La familia exponencial está estrechamente relacionada con los MLG. Sea Y una va.a con distribución perteneciente a la familia exponencial, entonces su función de densidad está dada por

$$f_Y(y) = c(y,\psi)exp\left(\frac{y\theta - a(\theta)}{\psi}\right)$$

Donde

- θ es el parámetro canónico
- ψ es el parámtero de dispersión
- $a:\mathbb{R}\to\mathbb{R}$ es una función de θ
- $c:\mathbb{R}^2\to\mathbb{R}$ es una función de y y ψ

Algunas de las distribuciones más conocidas como la Binomial, la Geométrica, la Binomial negativa, la Poisson, la Gama, la Normal y la Beta, pertenecen a la familia exponencial.

Ejemplo. La distribución Binomial pertenece a la familia exponencial. Sea $Y \sim Bin(n,\pi)$, con $n \in \mathbb{N}$ y $\pi \in (0,1)$. Entonces la función de masa de probabilidad está dada por

$$\begin{aligned}
P(Y=y) &= \binom{n}{y}\pi^y(1-\pi)^{n-y} \\
&= \binom{n}{y}\pi^y(1-\pi)^{-y}\cdot\frac{1}{(1-\pi)^{-n}} \\
&= \binom{n}{y}\left(\frac{\pi}{1-\pi}\right)^y\cdot\frac{1}{\frac{1}{(1-\pi)^n}} \\
&= \binom{n}{y}\left(\frac{\pi}{1-\pi}\right)^y\cdot\frac{1}{\left(\frac{1-\pi+\pi}{1-\pi}\right)^n} \\
&= \binom{n}{y}\left(\frac{\pi}{1-\pi}\right)^y\cdot\frac{1}{(1+\frac{\pi}{1-\pi})^n} \\
&= \binom{n}{y}exp\left\{ln\left(\frac{\left(\frac{\pi}{1-\pi}\right)^y}{(1+\frac{\pi}{1-\pi})^n}\right)\right\} \\
&= \binom{n}{y}exp\left\{ln\left(\frac{\pi}{1-\pi}\right)^y - ln\left(1+\frac{\pi}{1-\pi}\right)^n\right\} \\
&= \binom{n}{y}exp\left\{\frac{y\cdot ln\left(\frac{\pi}{1-\pi}\right) - n\cdot ln\left(1+\frac{\pi}{1-\pi}\right)}{1}\right\}
\end{aligned}$$

Se proponen :

$$\theta = ln\left(\frac{\pi}{1-\pi}\right) \quad ,\psi = 1, \quad c(y,\psi) = \binom{n}{y}, \quad \text{, y } a(\theta) = n\cdot ln\left(1+\frac{\pi}{1-\pi}\right) = n\cdot ln\left(1+e^{\theta}\right)$$

Por lo tanto, la distribución Binomial pertenece a la familia exponencial. Algunas distribuciones de la familia exponencial son la Binomial, Poisson, Normal, Gamma, Gaussiana inversa y Binomial negativa.

Observación. El modelo lineal clásico, o lineal normal es uno de los más importantes y utilizados de los MLG. Es crucial para para el estudio de los demás modelos pertenecientes a la clase de MLG.

Veremos 3 formas de los MLG: para respuestas categóricas, para datos de conteo y para respuestas continuas.

MLG para respuestas categóricas

Recordemos que las variables categóricas toman los valores de un número posible de categorías. Existen dos tipos de variables categóricas: las variables cuyas categorías tienen un orden natural (ordinal) y las que no lo tienen (nominal).

Para variables binarias, digamos $y = 0$ o $y = 1$. Si π es la probabilidad de que $y = 1$, entonces $y \sim B(1, \pi)$. El MLG Bernoulli (Binomial con n = 1) es

$$Y \sim B(1, \pi), g(\pi) = X\beta$$

Recordemos que la proporción $\pi/(1\pi)$ se llama odds o momios, e indica proporcionalmente cuanto más probable es la ocurrencia del evento comparada con la no-ocurrencia. Para las respuestas binarias tenemos la regresión logística o los modelos Probit que se vieron anteriormente.

Cuando todas las variables explicativas son categóricas es posible expresar un conjunto de datos en forma agrupada. Un grupo consiste de todos los casos con los mismos valores de las variables explicativas y puede corresponder a un conjunto de riesgos homogéneo. En el caso de una respuesta binaria, una vez que los datos están agrupados, la respuesta observada es el número de eventos que ocurren en cada grupo. La notación es la siguiente:

- $m :=$ número de grupos.
- $n_i :=$ número de casos en el grupo i
- $y_i :=$ número de eventos ocurridos en el grupo i.
- $\pi_i :=$ probabilidad de que el evento ocurra para un caso en el grupo i
- $n :=$ tamaño de la muestra, $n = \sum_{i=1}^{m} n_i$

Por lo tanto, y_i es el número de ocurrencias del evento, de un máximo de n_i, donde la probabilidad de ocurrencia del evento es π_i.

Por lo tanto, la respuesta observada tiene una distribución, $y_i B(n_i, \pi_i)$, donde la probabilidad π_i se modela como una función de las variables explicativas.

Ahora bien, si tuviéramos una respuesta categórica con r categorías, para la respuesta se define Para la respuesta se definen $r1$ variables respuesta indicadoras y_j, con $j \in \{1, ..., r1\}$ y definimos

$$y_j = \begin{cases} 1 & \text{si la respuesta está en el nivel } j \\ 0 & \text{en otro caso} \end{cases}$$

. De esta manera, la respuesta

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_{r-1} \end{pmatrix}$$

multivariada. Los modelos con respuesta nominal y ordinal entran en la clase de los MLG multivariados, usando la familia exponencial multivariada. Dadas n observaciones independientes de la respuesta y, es de interés obtener el número de veces que la categoría j ocurre.

falta , falta : ...

- Respeus ordinal
- Modelo logístico o de momios porporcionales modelo log-log complementario acumulado modelo porbit acumulado
- respuesta nominal

para datos de conteo : regresión Poison regresión binomial negativa para repsuestas continuas : regresión gamma, regresión gaussiana inversa

1.2. Algoritmos de regularización y estabilidad

El nuevo paradigma de aprendizaje que presentamos en este capítulo se denomina Minimización de pérdida regularizada (Regularized Loss Minimization) o RLM para abreviar. En RLM minimizamos la suma del riesgo empírico L_s con una función de regularización R.

Intuitivamente, la función de regularización mide la complejidad de las hipótesis, es como un estabilizador del algoritmo de aprendizaje. Intuitivamente, un algoritmo se considera estable si un ligero cambio en su entrada no cambia mucho su salida. Definiremos formalmente la noción de estabilidad (lo que entendemos por “cambio leve de entrada” y por “no cambia mucho la salida”) y demostraremos su estrecha relación con la capacidad de aprendizaje.

Finalmente, mostraremos que el uso de la norma l_2 al cuadrado como función de regularización estabiliza todos los problemas de aprendizaje.

Minimización de pérdidas regularizadas

La Minimización de Pérdida Regularizada (RLM) es una regla de aprendizaje en la que minimizamos conjuntamente el riesgo empírico y una función de regularización. Formalmente, una función de regularización es un mapeo $R : \mathbb{R}^d \to \mathbb{R}$ y ka función que miniza la perdida regularizada nos da w que minimice la expresión

$$L_s(W) + R(w)$$

La minimización de pérdidas regularizada comparte similitudes con los algoritmos de longitud mínima de descripción y la minimización de riesgos estructurales. Intuitivamente, la "complejidad" de las hipótesis se mide por el valor de la función de regularización y el algoritmo equilibra el riesgo empírico bajo con hipótesis "más simples" o "menos complejas".

Una de las función de regularización sencilla está dada por

$$R(w) = \lambda ||w||^2$$

donde $\lambda > 0$ es un escalar y la norma $||\cdot|| : \mathbb{R}^d \to \mathbb{R}$ es la de l_2 dada por

$$||w|| = \sqrt{\sum_{i=1}^{2} w_i^2}$$

Por lo tanto, al regla de aprendizaje está dada por

$$A(S) = w \text{ que minimice } (L_s(w) + \lambda ||w||^2)$$

Este tipo de función de regularización se conoce como egularización de Tikhonov.

Regresión Ridge

Aplicando la regla RLM con regularización de Tikhonov a la regresión lineal con la pérdida al cuadrado, obtenemos la siguiente regla de aprendizaje:

$$arming_{w \in \mathbb{R}^d} \left(\lambda ||w||_2^2 + \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} (\langle w, x_i \rangle - y_i)^2 \right)$$

Realizando una regreisón linea con la ecuación anteriore se le denomina regressión ridge.

Para resolver la ecuación, comparamos el gradiente de la ecuación a cero y obtenemos un

conjunto de ecuaciones lineales

$$(2\lambda mI + A)w = b$$

donde I es la matriz identidad y A, b son definidos como se definieron con anterioridad en la página 11 y 12.

Como Aes definida como una matriz semipositiva y la atriz $2\lambda mI + A$ tiene sus eigenvalores acotados por debajo por $2\lambda m$, entonces la matriz es invertible y la solución está dada por :

$$w = (2\lambda I + A)^{-1} b$$

1.3. Árboles de decisión