Modelos de clasificación

La clasificación supervisada es una de las tares que más frecuentemente son llevadas a cabo por los denominados Sistemas Inteligentes. Por lo tanto, un gran número de paradigmas desarrollados bien por la Estadística (Regresión Logística, Análisis Discriminante) o bien por la Inteligencia Artificial (Redes Neuronales, Inducción de Reglas, Árboles de Decisión, Redes Bayesianas) son capaces de realizar las tareas propias de la clasificación.

La clasificación es una subcategoría del aprendizaje supervisado en la que el objetivo es predecir las etiquetas de clase categóricas (discreta, valores no ordenados, pertenencia a grupo) de las nuevas instancias, basándonos en observaciones pasadas.

Hay dos tipos principales de clasificaciones:

Clasificación Binaria: Es un tipo de clasificación en el que tan solo se pueden asignar dos clases diferentes (0 o 1). El ejemplo típico es la detección de email spam, en la que cada email es: spam → en cuyo caso será etiquetado con un 1 ; o no lo es → etiquetado con un 0.

Clasificación Multi-clase: Se pueden asignar múltiples categorías a las observaciones. Como el reconocimiento de caracteres de escritura manual de números (en el que las clases van de 0 a 9).

Codificación de Características

Desde cierto punto de vista, las características predictivas pueden ser clasificadas como **variables cuantitativas** y **variables cualitativas** (frecuentemente denominadas **categóricas**).

Las *variables cuantitativas* son aquellas que se expresan con números y en las que tiene sentido realizar operaciones aritméticas. Básicamente, son estas características en las que se va a basar un algoritmo para entrenar un modelo predictivo.

Por el contrario, las *variables cualitativas o categóricas* son aquellas que expresan cualidades o atributos del elemento al que se refieren y no pueden ser medidas con números o, éstos, si se usan, no justifican la posibilidad de aplicar operaciones aritméticas sobre ellos. Por ejemplo, supongamos que asignamos al estado civil de una persona un número (*1 = soltero*, *2 = casado*, *3 = viudo*, *4 = divorciado*, etc.). No tendría ningún sentido calcular el "valor medio del estado civil" de una población, pues bastaría con asignar otros valores a cada uno de los estados para que el resultado fuese totalmente distinto.

Las variables categóricas pueden, a su vez, ser divididas en dos tipos: **variables categóricas nominales** y **variables categóricas ordinales**.

Las **variables categóricas nominales** son aquellas en las que no hay un orden entre sus valores. En el ejemplo de los estados civiles, no tendría sentido afirmar que los solteros van antes que los

casados y estos antes que los viudos o que los divorciados. Se trata, por lo tanto, de una variable categórica nominal.

Por último, las **variables categóricas ordinales** son aquellas en las que existe un orden implícito en los valores que puede tomar. Por ejemplo: el grado de satisfacción de un consumidor respecto a un cierto producto, variable que podría tomar los valores "*insatisfecho*", "*poco satisfecho*", "*ni satisfecho ni insatisfecho*", "*algo satisfecho*" o "*muy satisfecho*", o por ejemplo la temperatura: "*frío*", "*templado*" o "*caliente*".

En general, las variables categóricas (tanto las nominales como las ordinales) deberán ser *codificadas* o convertidas a valores numéricos para entrenar un algoritmo a partir de ellas.

Codificación mediante enteros

- A cada valor de categoría único se le asigna un valor entero único.
- No es necesario que sean números consecutivos.
- Es fácilmente reversible y podemos obtener la etiqueta a partir del número.
- Existe un orden implícito en los valores de los números que puede no ser representativo de las categorías que quieren representar

Ejemplo: Codificación mediante enteros

Color	**ID**
Rojo	1
Amarillo	2
Azul	3

Para algunas variables o algoritmos esto puede ser suficiente. Pero esta codificación adolece del problema que los números tienen una relación de orden entre sí y los algoritmos de aprendizaje automático pueden ser capaces de comprender y aprovechar esta relación de orden de nuestros valores que representan las categorías.

En el ejemplo de los colores, es bastante claro que en los colores no hay un orden y por lo tanto, nuestro modelo podría ser sensible al orden de los valores, lo cual no es válido. Este tipo de consecuencias puede dar como resultado un rendimiento inadecuado o resultados inesperados.

Codificación One-Hot

Para variables categóricas donde no existe una relación de orden, la codificación mediante enteros no suele ser adecuada. En estos casos, se puede aplicar una codificación especial donde se agrega una nueva variable binaria (con valores verdadero o falso) para cada valor de categoría posible.

La codificación One-Hot es un método para etiquetar a qué clase pertenecen los datos y la idea es asignar 0 a toda la dimensión, excepto 1 para la clase a la que pertenecen los datos.

En el ejemplo de los colores, si hubiera solo 3 categorías, con 3 variables binarias, podemos representar la codificación asignando un "1" en la variable binaria asociada con el color correspondiente y valores "0" para el resto de los colores.

Ejemplo: Codificación One-Hot

Id	Color
1	Rojo
2	Azul
3	Rojo
4	Amarillo
5	NULL

Id	Rojo	Azul	Amarillo
1	1	0	0
2	0	1	0
3	1	0	0
4	0	0	1
5	0	0	0

1. Regresión logística

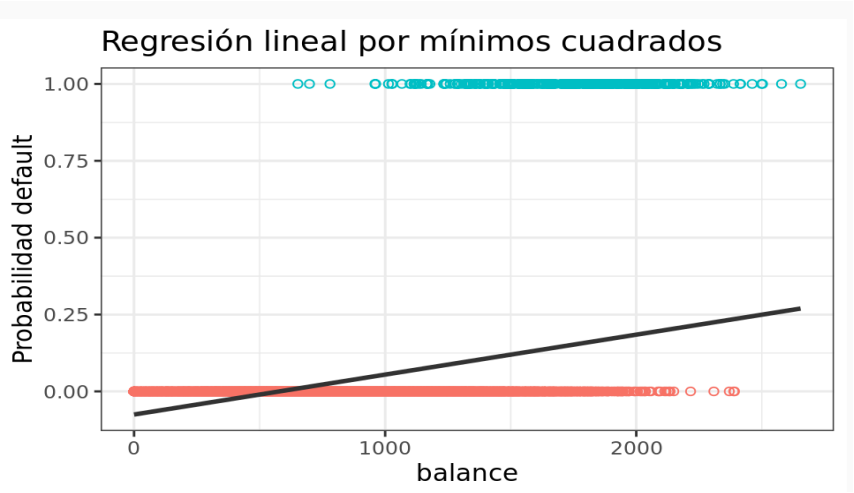
La Regresión Logística Simple, desarrollada por David Cox en 1958, es un método de regresión que permite estimar la probabilidad de una variable cualitativa binaria en función de una variable cuantitativa. Una de las principales aplicaciones de la regresión logística es la de clasificación binaria, en el que las observaciones se clasifican en un grupo u otro dependiendo del valor que tome la variable empleada como predictor. Por ejemplo, clasificar a un individuo desconocido como hombre o mujer en función del tamaño de la mandíbula. Es importante tener en cuenta que, aunque la regresión logística permite clasificar, se trata de un modelo de regresión que modela el logaritmo de la probabilidad de pertenecer a cada grupo. La asignación final se hace en función de las probabilidades predichas.

La regresión logística es un algoritmo de clasificación (a pesar de su nombre) simple, pero potente. Funciona muy bien en clases linealmente separables y se puede extender a clasificación multiclase.

¿Por qué regresión logística y no regresión lineal?

Si una variable cualitativa con dos niveles se codifica como 1 y 0, matemáticamente es posible ajustar un modelo de regresión lineal por mínimos cuadrados ($h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x$). El problema de esta aproximación es que, al tratarse de una recta, para valores extremos del predictor, se obtienen valores de Y menores que 0 o mayores que 1, lo que entra en contradicción con el hecho de que las probabilidades siempre están dentro del rango [0,1].

En el siguiente ejemplo se modela la probabilidad de fraude por impago (default) en función del balance de la cuenta bancaria (balance).



Al tratarse de una recta, si por ejemplo, se predice la probabilidad de default para alguien que tiene un balance de 10000, el valor obtenido es mayor que 1.

Para evitar estos problemas, la regresión logística transforma el valor devuelto por la regresión lineal ($h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x$) empleando una función cuyo resultado está siempre comprendido entre 0 y 1. Existen varias funciones que cumplen esta descripción, una de las más utilizadas es la función logística (también conocida como función sigmoide), $\sigma(x)$:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

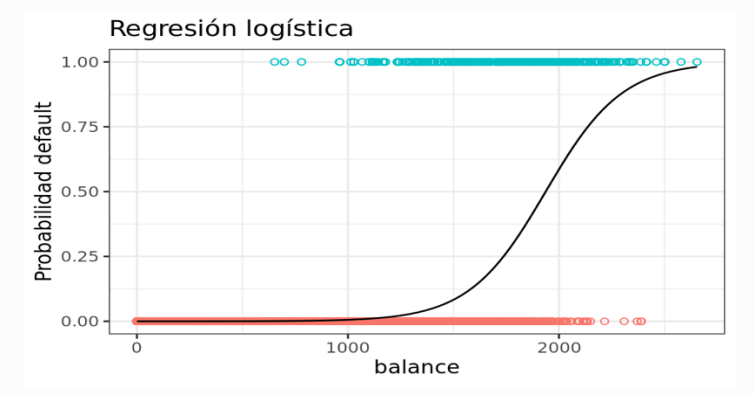
Para valores de x muy grandes positivos, el valor de e^{-x} es aproximadamente 0 por lo que el valor de la función sigmoide es 1. Para valores de x muy grandes negativos, el valor e^{-x} tiende a infinito por lo que el valor de la función sigmoide es 0. Si $\sigma\big(h_{\boldsymbol{\theta}}(x)\big)$ se obtiene que:

$$S \equiv P(Y = k|X = x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x)}}$$

donde $P(Y = k|X = x)$ puede interpretarse como: la probabilidad de que la variable cualitativa Y adquiera el valor k (el nivel de referencia, codificado como 1), dado que el predictor X tiene el valor x.

Esta función, puede ajustarse de forma sencilla con métodos de regresión lineal si se emplea su versión logarítmica, obteniendo lo que se conoce como log of odds

$$ln\left(\frac{P(Y=k|X=x)}{1-P(Y=k|X=x)}\right) = h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x$$



La regresión logística es un método lineal clásico para la clasificación binaria. Los modelos predictivos de clasificación de problemas son aquellos que requieren la predicción de una etiqueta de clase (por ejemplo, 'rojo', 'verde', 'azul') para un conjunto dado de variables de entrada. La clasificación binaria se refiere a aquellos problemas de clasificación que tienen dos etiquetas de clase, por ejemplo verdadero / falso o 0/1.

La regresión logística tiene mucho en común con la regresión lineal, aunque la regresión lineal es una técnica para predecir un valor numérico, no para problemas de clasificación. Ambas técnicas modelar la variable de destino con una línea (o hiperplano, dependiendo del número de dimensiones de entrada. En la regresión lineal los datos se ajustan a la línea, que pueden ser usados para predecir una nueva cantidad, mientras que la regresión logística se ajusta a una línea que mejor separada las dos clases.

Los datos de entrada se denota como ***x*** con n ejemplos y la salida se denota ***y*** con una salida para cada entrada. La predicción del modelo para una entrada dada se denota como $\hat{\boldsymbol{y}}$.

$$\hat{\boldsymbol{y}} = modelo(\boldsymbol{x}; \boldsymbol{\theta})$$

El modelo se define en términos de parámetros denominados coeficientes ($\boldsymbol{\theta}$), donde hay un coeficiente por entrada y un coeficiente adicional que proporciona la intercepción o sesgo.

Por ejemplo, un problema con las entradas **x** con *n* variables $x_1, \ldots, x_i, \ldots, x_n$ tendrá coeficientes $\theta_1, \ldots, \theta_i, \ldots, \theta_n, \theta_0$. Una entrada dada se predice como la suma ponderada de las entradas para el ejemplo y los coeficientes.

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta_1 x_1 + \cdots + \theta_i x_i + \cdots + \theta_n x_n + \theta_0$$

El modelo también se puede describir usando álgebra lineal, con un vector para los coeficientes ($\boldsymbol{\theta}$) y una matriz para los datos de entrada (**x**).

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = [x_1 \quad \cdots \quad x_i \quad \cdots \quad x_n \quad | \quad 1] \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_i \\ \vdots \\ \theta_n \\ - \\ \theta_0 \end{bmatrix}$$

$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{A\theta}$, donde $\boldsymbol{A}$ es la matriz de características de tamaño $q \times (n+1)$ y $\boldsymbol{\theta}$ es un vector de coeficientes de tamaño $(n+1) \times 1$

Hasta ahora, esto es idéntico a la regresión lineal y es insuficiente como la salida será un valor real en lugar de una etiqueta de clase. En cambio, el modelo aplasta la salida de esta suma ponderada usando una función no lineal para asegurar que las salidas son un valor entre 0 y 1. La función logística, $\sigma(x)$, (también llamado el sigmoide) se utiliza para este escenario. En el caso de regresión logística, x se reemplaza con la suma ponderada, $\sigma(h_{\boldsymbol{\theta}}(\boldsymbol{x}))$:

$$\boldsymbol{S} = \frac{1}{1 + e^{-h_{\boldsymbol{\theta}}(\boldsymbol{x})}}$$

La salida ($\boldsymbol{S}$) se interpreta como una probabilidad de una función de distribución de probabilidad Binomial para la clase marcada como 1, si las dos clases en el problema están etiquetados 0 y 1. Tenga en cuenta que la salida, al ser un número entre 0 y 1, se puede interpretar como una probabilidad de pertenencia a la clase de etiquetado 1.

Regresión logística y estimación de máxima verosimilitud

Podemos enmarcar el problema de ajustar un modelo de aprendizaje automático como el problema de la estimación de la densidad de probabilidad.

Específicamente, la elección del modelo y los parámetros del modelo se conoce como una hipótesis de modelado $h_{\boldsymbol{\theta}}(\boldsymbol{x})$, y el problema implica encontrar $h_{\boldsymbol{\theta}}(\boldsymbol{x})$ que mejor explica los datos **x**. Podemos, por lo tanto, encontrar la hipótesis de modelado que maximiza la función de probabilidad.

Para utilizar la máxima probabilidad, debemos asumir una distribución de probabilidad. En el caso de la regresión logística, se asume una distribución de probabilidad binomial para la muestra de datos, donde cada ejemplo es un resultado de un ensayo Bernoulli. La distribución de Bernoulli tiene un solo parámetro: la probabilidad de un resultado exitoso (**S**).

$P(y = 1 \mid \boldsymbol{x};\ \boldsymbol{\theta}) = \boldsymbol{S}^y$
$P(y = 0 \mid \boldsymbol{x};\ \boldsymbol{\theta}) = [\boldsymbol{1} - \boldsymbol{S}]^{1-y}$

La distribución de probabilidad que se usa con mayor frecuencia cuando hay dos clases es la distribución binomial. Esta distribución tiene un solo parámetro, **S**, que es la probabilidad de un evento o una clase específica. La Función de probabilidad para la distribución de Bernoulli

$$P(y \mid \boldsymbol{x};\ \boldsymbol{\theta}) = \boldsymbol{S}^{y}\ [\mathbf{1} - \boldsymbol{S}]^{1-y}$$

$$L(\boldsymbol{\theta}) = P(\boldsymbol{y} \mid \boldsymbol{x};\ \boldsymbol{\theta}) = \prod_{p=1}^{q} P\left(y_p \mid x_{p,i};\ \boldsymbol{\theta}\right) = \prod_{p=1}^{q} S_p^{y_p}\left[\mathbf{1} - S_p\right]^{1-y_p}$$

Podemos actualizar la función de probabilidad transformándola en una función de probabilidad logarítmica y finalmente, podemos sumar la función de probabilidad en todos los ejemplos en el conjunto de datos para maximizar la probabilidad, $ln\left(L(\boldsymbol{\theta})\right)$:

$$ln\left(L(\boldsymbol{\theta})\right) = \sum_{p=1}^{q} y_p\ log\left(S_p\right) + \left(1 - y_p\right) log\left(1 - S_p\right)$$

Es una práctica común minimizar una función de costo para problemas de optimización; por lo tanto, podemos invertir la función para minimizar la probabilidad negativa, $-ln\left(L(\boldsymbol{\theta})\right)$. El cálculo negativo de la función de probabilidad logarítmica para la distribución de Bernoulli es equivalente a calcular la *función de entropía cruzada binaria* para la distribución de Bernoulli.

funcion de cosot binaria

$$\boldsymbol{\mathcal{L}}(\boldsymbol{\theta}) = -ln\left(L(\boldsymbol{\theta})\right) = -\sum_{p=1}^{q} y_p\ ln\left(S_p\right) + \left(1 - y_p\right) ln\left(1 - S_p\right)$$

A diferencia de la regresión lineal, no hay una solución analítica para resolver este problema de optimización. Como tal, se debe utilizar un algoritmo de optimización iterativo, esto es, ya no podemos escribir la MLE en forma cerrada. En su lugar, necesitamos usar un algoritmo de optimización para calcularlo.

La función $\boldsymbol{\mathcal{L}}(\boldsymbol{\theta})$ proporciona información para ayudar en la optimización (específicamente, se puede calcular una matriz hessiana), lo que significa que se pueden usar procedimientos de búsqueda eficientes que explotan esta información, como el algoritmo BFGS (y sus variantes). Para esto, necesitamos derivar el gradiente y la matriz hessiana.

$$\frac{\partial \boldsymbol{\mathcal{L}}(\boldsymbol{\theta})}{\partial \theta_i} = -\sum_{p=1}^{q}\left(\frac{y_p}{S_p} - \frac{1 - y_p}{1 - S_p}\right)\frac{\partial S}{\partial h_{\boldsymbol{\theta}}(\boldsymbol{x})}\frac{\partial h_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_i}$$

$$\frac{\partial \boldsymbol{\mathcal{L}}(\boldsymbol{\theta})}{\partial \theta_i} = -\sum_{p=1}^{q}\left(y_p - S_p\right)\frac{\partial h_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_i}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_i} = -\sum_{p=1}^{q} (y_p - S_p)\, x_{p,i}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_0} = -\sum_{p=1}^{q} (y_p - S_p)$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\boldsymbol{A}'\boldsymbol{e}$$ gradiemnte

donde $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{S}$

2. Regresión logística multinomial (regresión softmax)

La regresión logística múltiple es una extensión de la regresión logística simple. Se basa en los mismos principios que la regresión logística simple (explicados anteriormente) pero ampliando el número de predictores. Los predictores pueden ser tanto binarios como categóricos.

$$h_{p,\mathrm{j}}(\boldsymbol{x}) = \theta_{1,j}x_{p,1} + \cdots + \theta_{i,j}x_{p,i} + \cdots + \theta_{n,j}x_{p,n} + \theta_{0,j}$$

$$h_{p,\mathrm{j}}(\boldsymbol{x}) = [x_{p,1} \quad \cdots \quad x_{p,i} \quad \cdots \quad x_{p,n} \quad | \quad 1] \begin{bmatrix} \theta_{1,j} \\ \vdots \\ \theta_{i,j} \\ \vdots \\ \theta_{n,j} \\ - \\ \theta_{0,j} \end{bmatrix} = A\Theta$$

$$h_{\Theta}(\boldsymbol{x}) = A\Theta$$

donde $A = [x_{p,1} \quad \cdots \quad x_{p,i} \quad \cdots \quad x_{p,n} \quad | \quad 1]$ de tamaño $q \times (n+1)$ y Θ de tamaño (n+1) x m, definida como:

$$\Theta = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,j} & \cdots & \theta_{1,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \theta_{i,1} & \cdots & \theta_{i,j} & \cdots & \theta_{i,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \theta_{n,1} & \cdots & \theta_{n,j} & \cdots & \theta_{n,m} \\ \theta_{0,1} & \cdots & \theta_{0,j} & \cdots & \theta_{0,m} \end{bmatrix}$$

$$S_{p,j} = softmax\left(h_{\Theta}(\boldsymbol{x})\right) = \frac{exp\left(h_{\theta_{:,j}}(x_{p,:})\right)}{\sum_{k=1}^{m} exp\left(h_{\theta_{:,k}}(x_{p,:})\right)}$$

$$L(\Theta) = \prod_{p=1}^{q}\prod_{j=1}^{m} S(j = c|\, x_{p,:};\, \theta_{:,j})^{y_p} = \prod_{p=1}^{q}\prod_{j=1}^{m}\left[\frac{exp\left(h_{\theta_{:,j}}(x_{p,:})\right)}{\sum_{k=1}^{m} exp\left(h_{\theta_{:,k}}(x_{p,:})\right)}\right]^{y_p}$$

$$\ln L(\Theta) = \sum_{p=1}^{q}\sum_{j=1}^{m}\{y_p = j\} \ln\left[\frac{exp\left(h_{\theta_{:,j}}(x_{p,:})\right)}{\sum_{k=1}^{m} exp\left(h_{\theta_{:,k}}(x_{p,:})\right)}\right]$$

$$\ln L(\boldsymbol{\Theta}) = \sum_{p=1}^{q}\sum_{j=1}^{m}\{y_p = j\}\left[h_{\theta_{:,j}}(x_{p,:}) - \ln\sum_{k=1}^{m} exp\left(h_{\theta_{:,k}}(x_{p,:})\right)\right]$$

$$\mathcal{L}(\boldsymbol{\Theta}) = -ln\left(L(\boldsymbol{\Theta})\right) = -\sum_{p=1}^{q}\sum_{j=1}^{m}\{y_p = j\}\left[h_{\theta_{:,j}}(x_{p,:}) - \ln\sum_{k=1}^{m} exp\left(h_{\theta_{:,k}}(x_{p,:})\right)\right]$$

$$\mathcal{L}(\boldsymbol{\Theta}) = -ln\left(L(\boldsymbol{\Theta})\right) = -\sum_{p=1}^{q}\sum_{j=1}^{m}\{y_p = j\}\, log\left(S_{p,j}\right)$$ programar

$$\frac{\partial\mathcal{L}(\boldsymbol{\Theta})}{\partial\theta_{i,j}} = -\sum_{p=1}^{q}\sum_{j=1}^{m}(y_p - S_{p,j})x_{p,i}$$

$$\frac{\partial\mathcal{L}(\boldsymbol{\Theta})}{\partial\theta_{0,j}} = -\sum_{p=1}^{q}\sum_{j=1}^{m}(y_p - S_{p,j})$$

$$\boldsymbol{\frac{\partial\mathcal{L}(\Theta)}{\partial\Theta} = -A'e}$$

donde

$$\boldsymbol{e = Y - S}$$

$$\boldsymbol{\frac{\partial\mathcal{L}(\Theta)}{\partial\Theta} = A'(S - Y)}$$

Resumen: Regresión Logística

Predictor polinomial multivariable de orden superior

$\boldsymbol{h}_{\Theta}(\boldsymbol{x}) = \boldsymbol{A\Theta}$, donde $\boldsymbol{A}$ es la matriz de diseño de tamaño $q \times \varrho$ y $\boldsymbol{\Theta}$ de tamaño $\varrho \times m$.

Función de Probabilidad Logística

$$\boldsymbol{S} = logistic\big(\boldsymbol{h}_{\Theta}(\boldsymbol{x})\big) = \frac{1}{1 + e^{-\boldsymbol{h}_{\Theta}(\boldsymbol{x})}}$$

Término de error

$$\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{S}$$

Función de entropía cruzada binaria

$$\mathcal{L}(\boldsymbol{\Theta}) = -\sum_{p=1}^{q}\sum_{j=1}^{m}\big[y_{p,j}\ ln\big(S_{p,j}\big) + \big(1 - y_{p,j}\big)ln\big(1 - S_{p,j}\big)\big]$$

Función gradiente de entropía cruzada binaria

$$\nabla_{\boldsymbol{\Theta}}\mathcal{L} = -\boldsymbol{A}'\boldsymbol{e}$$

Función Hessiana de entropía cruzada binaria

$$\nabla^2_{\boldsymbol{\Theta}}\mathcal{L} = -\boldsymbol{A}'\boldsymbol{W}\boldsymbol{A}$$

donde $\boldsymbol{W} = diag[\boldsymbol{S}\odot(\boldsymbol{1} - \boldsymbol{S})]$ es la derivada de la función de probabilidad Logística

Método de Newton

$$\boldsymbol{v} = \boldsymbol{A\Theta}^{(t)} + \boldsymbol{W}^{\dagger}(\boldsymbol{Y} - \boldsymbol{S})$$

$$\boldsymbol{\Theta}^{(\boldsymbol{t+1})} = (\boldsymbol{A}'\boldsymbol{W}\boldsymbol{A})^{\dagger}(\boldsymbol{A}'\boldsymbol{W})\boldsymbol{v}$$

donde † es la pseudoinversa de Moore-Roose

Resumen: Regresión Softmax

Predictor polinomial multivariable de orden superior

$h_\Theta(\boldsymbol{x}) = A\Theta$, donde A es la matriz de diseño de tamaño $q \times \varrho$ y Θ de tamaño $\varrho \times m$.

Función de Probabilidad Softmax

$$S_{p,j}\left(y_p = j \mid \boldsymbol{x}, \boldsymbol{\theta}\right) = softmax\left(h_\Theta(\boldsymbol{x})\right) = \frac{exp\left(h_{\theta_{:,j}}\left(x_{p,:}\right)\right)}{\sum_{k=1}^{m} exp\left(h_{\theta_{:,k}}\left(x_{p,:}\right)\right)}$$

Término de error

$$\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{S}$$

Función de entropía cruzada categórica

$$H\left(S_p, Y_p\right) = -\sum_{j=1}^{m} \mathbb{I}\{Y_p = j\}\, log\left(S_{p,j}\right)$$

$$\mathcal{L}(\boldsymbol{\Theta}) = \frac{1}{q}\sum_{p=1}^{q} H\left(S_p, Y_p\right)$$

Función gradiente de entropía cruzada categórica

$$\nabla_{\boldsymbol{\Theta}}\mathcal{L} = -\boldsymbol{A}'\boldsymbol{e}$$

Función Hessiana de entropía cruzada categórica

$$\nabla^2_{\boldsymbol{\Theta}}\mathcal{L} = -\boldsymbol{A}'\boldsymbol{W}\boldsymbol{A}$$

donde $\boldsymbol{W} = diag[\boldsymbol{S} \odot (\boldsymbol{1} - \boldsymbol{S})]$ es la derivada de la función de probabilidad softmax

Método de Newton

$$\boldsymbol{v} = \boldsymbol{A}\boldsymbol{\Theta}^{(t)} + \boldsymbol{W}^{\dagger}(\boldsymbol{Y} - \boldsymbol{S})$$

$$\boldsymbol{\Theta}^{(t+1)} = (\boldsymbol{A}'\boldsymbol{W}\boldsymbol{A})^{\dagger}(\boldsymbol{A}'\boldsymbol{W})\boldsymbol{v}$$