MACHINE LEARNING: ALGORITMOS SUPERVISADOS

Regresión Logística

Instructor: José Nelson Zepeda

San Salvador, Mayo 2019

Regresión Logística

Concepto

Objetivo

Metodología

Definiciones

Conceptos Básicos

Historia

El termino regresión fue introducido en 1889 por Francis Galton en su escrito Natural Inheritance, en donde aparece el término regresión hacia la media.

Fue en la década de los 70s que se dio toda la unificación de toda la teoría existente sobre regresión lineal, regresión logística y regresión de Poisson, así como el análisis de varianza.

la regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras

Historia

La regresión logística, nos permitirá contestar preguntas como:

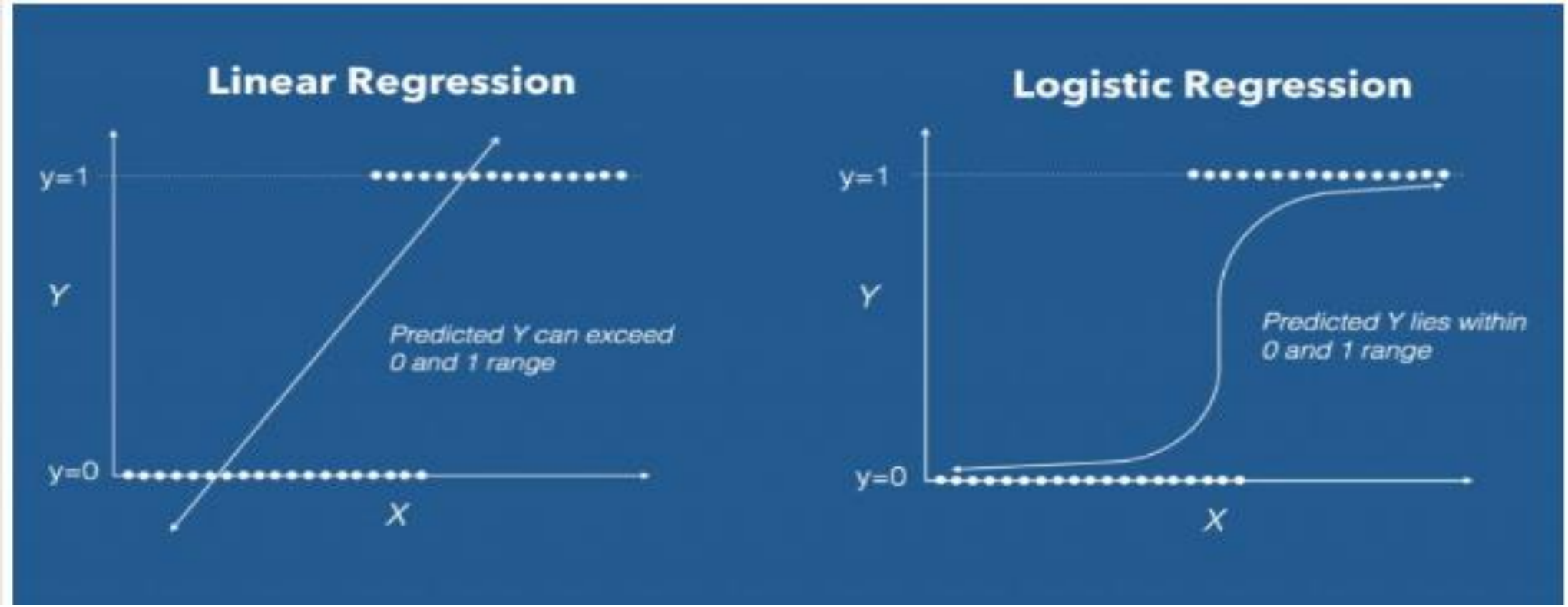
- ¿Se puede saber si un cliente se ira dado que está llamando con cierta frecuencia a atención al cliente?
- ¿Es posible saber anticipadamente si un paciente de un hospital padecerá alguna enfermedad tropical?

La regresión logística es especialmente útil en particular cuando solo hay 2 posibles repuestas (la variable de salida es dicotómica), que suele ser el caso más común.

Supuestos

- Normalidad: Los valores de la variable se ajustan a la distribución normal
- Linealidad: La relación entre la variable dependiente e independiente es lineal
- No colinealidad: Ninguna de las variables independientes puede ser función exacta de otras variables indpendientes, en caso contrario se produce colinealidad o multicolinealidad.
- Independencia: Los errores de medición de las variables regresoras son independientes entre sí.
- Homocedasticidad: La varianza de error es constante en todo el rango de medición
- Distribución Normal: Los errores del modelo siguen una distribución normal.

Comparación



Regresión Logística Binaria

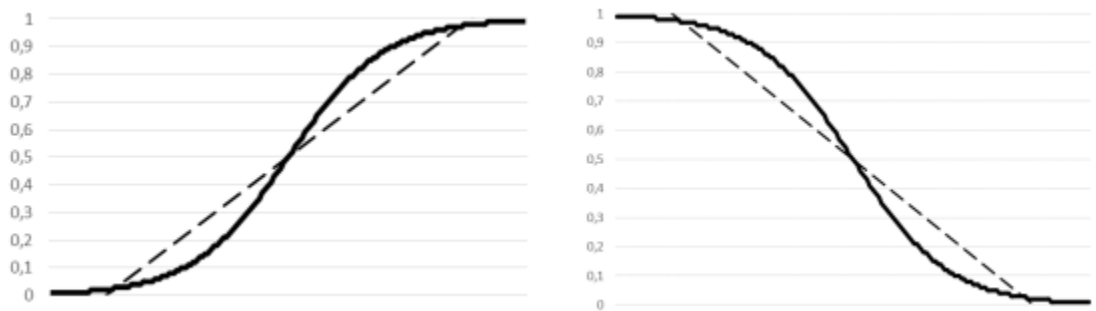
La popularidad de la regresión logística se debe, en parte, a que se basa en la función logaritmo natural que puede aplicarse únicamente a valores en el intervalo (0, infinito), pero de ella se obtiene cualquier número real (igual que una recta). Además tiene la propiedad de ser una función monótona creciente.

$$f(Y) = \frac{1}{1+e^{-y}}$$

$Y = -\infty;$ $\quad f(-\infty) = \frac{1}{1+e^{-(-\infty)}} = \frac{1}{1+e^{\infty}} = \frac{1}{1+\infty} = 0$

$Y = 0;$ $\quad f(0) = \frac{1}{1+e^{0}} = \frac{1}{1+1} = 0{,}5$

$Y = \infty;$ $\quad f(\infty) = \frac{1}{1+e^{-\infty}} = \frac{1}{1+0} = 1$



Odds

$$f(Y) = log\left[\frac{P(y=1)}{1-P(y=1)}\right] = \beta_0 + \beta_x$$

Si vemos la ecuación el argumento que recibe la función log equivale a una razón entre probabilidades, este término es popularmente conocido como "Odds".

Por ejemplo si P(Y=1)= 0.75, los Odds serán 0.75/0.25=3

$$f(Y) = p(Y) = \frac{1}{1+e^{-y}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}\ )}$$

Otros Aspectos

Colinealidad

La colinealidad es un problema potencial, común tanto a la regresión lineal múltiple como a la regresión logística multinomial. Ocurre cuando dos variables independientes están altamente correlacionadas. Una situación más extrema se da cuando una variable independiente puede explicarse como combinación de las otras dando lugar a la llamada multicolinealidad.

La forma más sencilla de detectarla es por medio de la matriz de correlaciones de las variables independientes. Cuando alguno de esos valores es superiores al 80%, esas variables pueden provocar problemas de multicolinealidad.

Otros Aspectos

Extensiones de la Regresión Logística

La regresión logística multinomial hace referencia a la aplicación de la regresión logística cuando la variable dependiente tiene más de dos categorías

Si la variable independiente incluye más de dos categorías, dichas variables deben incluirse en el modelo como variables dummy, de modo que si consta de K categorías, se crean K-1 variables dummy asociadas a dicha variable nominal.

$$(Y_1, Y_2) - \begin{cases} (1,0) \text{ si } Y = 1 \\ (0,1) \text{ si } Y = 2 \\ (0,0) \text{ si } Y = 3 \end{cases}$$

Bibliografía

R Data Analysis Cookbook

by Kuntal Ganguly,2017

Tesis Doctoral La metodología cuantitativa aplicada al estudio de la reincidencia en menores infractores

by Jacinto Pallarés,2017