Curso

12

MACHINE LEARNING & BIG DATA

Supervised Learning: Logistic Regression

Algoritmos Supervisados

JOSÉ NELSON ZEPEDA DOÑO

Cluster de Estudio: Advanced Analytics

Este material es el resumen de muchos autores que por medio de sus libros y documentos nos ofrecen fuentes riquísimas de conocimiento sobre los temas de Big Data y Machine Learning.
Algunas citas, figuras y tablas pueden ser encontradas de forma textual tal como lo indica el autor en su material original.
Mención especial para la Tesis Doctoral La metodología cuantitativa aplicada al estudio de la reincidencia en menores infractores del Dr. Jacinto Pallarés Mestre, pues su marco teórico es el fundamente de este extracto, muchos textos, conceptos, formulas han sido tomados literalmente de la tesis del Dr. Pallarés.

Nelson Zepeda
MIP • V 1.0
San Salvador El Salvador
Phone 503 79074137 • @nelsonzepeda733

Tabla de Contenido

Capítulo 1

Conceptos Básicos

La regresión es uno de los pilares de la estadística, y data de principios de 1800 gracias a Gauss y Laplace.

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

De acuerdo a Wikipedia, en estadística, 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.

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?
- Etc.

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. Surgió en la década de los 60, su generalización dependía de la solución que se diera al problema de la estimación de los coeficientes. El algoritmo de Walker-Duncan para la obtención de los estimadores de máxima verosimilitud vino a solucionar parte de este problema, pero era de naturaleza tal que el uso de recurso informático era imprescindible.

La identificación del mejor modelo de regresión logística se realiza mediante la comparación de modelos utilizando el coeficiente de verosimilitud, que indica a partir de los datos de la muestra cuanto más probable es un modelo frente a otro.

Si a partir de este modelo no se puede demostrar que un modelo resulta mejor que el otro, se considerará como el más adecuado, el más sencillo.

Pero ante de continuar hablando de regresión logística, partamos por el inicio, recordando los supuestos que debe cumplir una regresión lineal, pues estos nos servirán como fundamente de esta sección en adelante:

- 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.

El supuesto de Homocedasticidad ha sido históricamente uno de los más cuestionados por su incumplimiento en los siguientes casos:

- Variables independientes con frecuencias elevadas con valores extremos
- Utilización de variables no transformadas con valores mayores o muy diferentes a los de las otras variables regresoras.
- Ineficiencia en el proceso de minimos cuadrados en donde la presencia de heterocedasticidad, puede producir un incremento desproporcionado de la varianza estimada.

Estos incumplimientos, han impulsado la evolución hacia los determinados modelos lineales generalizados dentro de los cuales puede incluirse los log-lineales y los logit.

Los Modelos Lineales Generalizados son una generalización del procedimiento Mínimos cuadrados ordinario, que pone en relación la función de distribución de la variable dependiente con las independientes.

Un modelo Log-Lineal está diseñado para analizar dos o más variables categóricas, sin designar a ninguna de ellas como predictoras, el modelo LOGIT es una variante del modelo LOG-LIN en el cual una de las variables se designa como variable dependiente y las demás se utilizan como predictoras o explicativas.

El modelo de Log-Lineales no resulta adecuado para tratar variables predictoras, cualitativas, sin embargo, la regresión logística puede manejar indistintamente variables predictoras tanto categóricas.

Dicho lo anterior, podemos concluir que la regresión logística puede ser considerada un modelo hibrido entre el modelo LOGIT y Mínimos cuadrados pues permite manejar indistintamente variables predictoras, tanto categóricas como continuas.

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.

Esta unificación permitió desarrollar una estrategia general para la estimación de la máxima verosimilitud en todos estos modelos, que consiste en ajustar el modelo propuesto a los datos y estimar sus parámetros.

En 1970 se escribe el manual titulado The Analysis of Binary Data, que contribuye notablemente a la divulgación de la regresión logística. Su uso se extiende notablemente a partir de la década de los 80 y desde entonces hasta la actualidad, se ha convertido en una de las técnicas estadísticas más utilizadas.

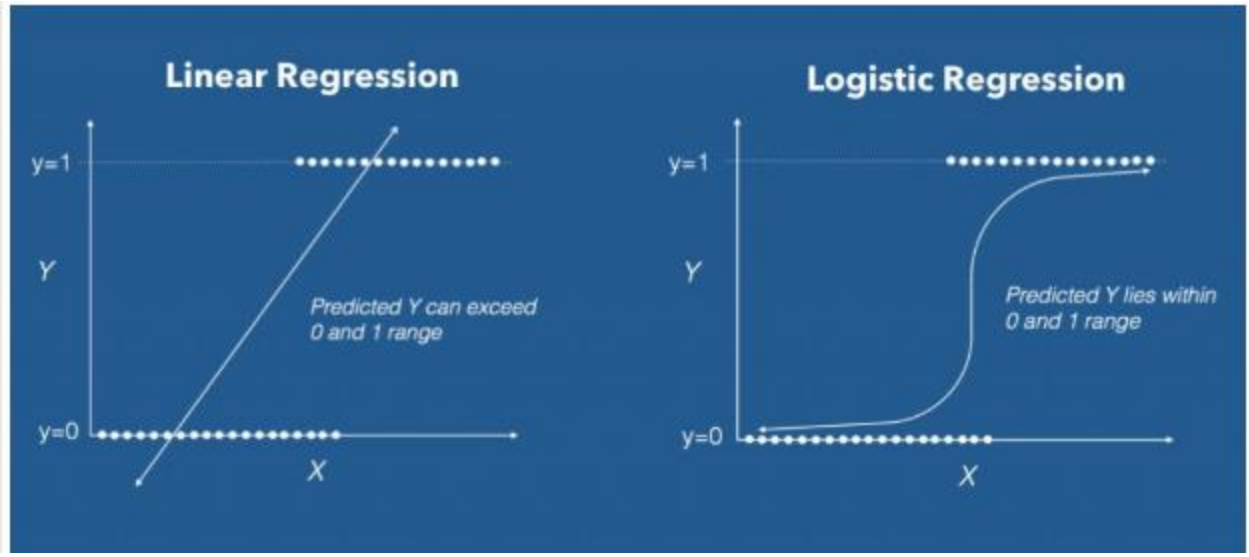

Fig 1-1 Regresión Lineal Vs Regresión Logística

Regresión Logística Binaria Simple y Múltiple

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. Si combinamos la transformación mediante la función logarítmica y la modelamos como una función lineal, se llega a la denominada función logística, expresada en la ecuación:

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

Si evaluamos la función en 3 puntos principales tendremos:

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

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

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

Gráficamente esta función tiene forma de S, y también empezaremos a mencionar que puede adoptar la forma inversa:

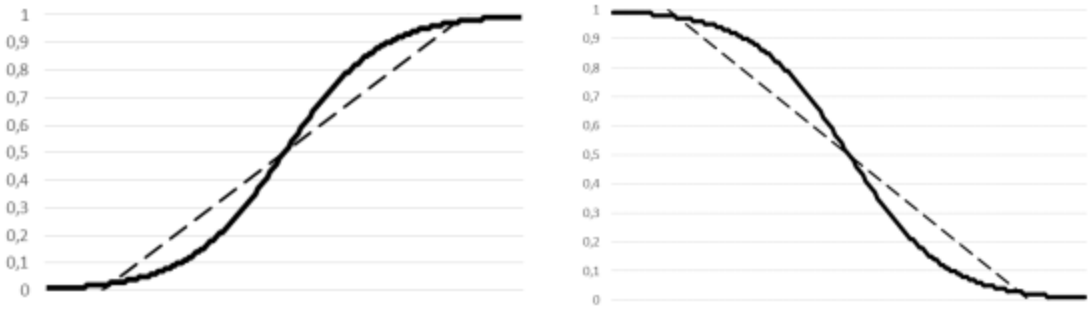

Fig 1-2 Función Logística

Esta función siempre indica una probabilidad de ocurrencia de la variable dependiente (Y) que oscila entre los valores 0 y 1 para cualquier valor de X.

En el caso de una regresión logística binaria simple, es decir con una sola variable independiente, el modelo se define como:

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

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

Indicando que la ocurrencia de un suceso es 3 veces más probable que la no ocurrencia.

Tal como se observa en la fórmula propuesta, se requiere un paso más que corresponde a calcular el logaritmo de los Odds:

$$Odds\ (log\ [P(y=1)/[1-P(y=1)]),$$

Y es aquí en donde estamos aplicando la transformación logística o Logit y llegamos a la ecuación

$$f(Y) = p(Y) = \frac{1}{1+e^{-y}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_i)}}$$

La función F(Y) representa en una escala logarítmica, la relación entre las probabilidades a una de las dos poblaciones de ocurrencia/no ocurrencia del evento.

Ahora bien, hasta el momento el análisis se ha centrado en una sola variable independiente, pero de todos es conocido que los fenómenos son bastante complejos y se suele utilizar más de una sola variable independiente para analizarlos.

El modelo de regresión logística binaria multinomial incorpora más variables independientes y facilita la comprensión de la variación de las respuestas entre los distintos individuos.

Su ecuación es:

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

Para valorar el grado en que un modelo clasifica adecuadamente las observaciones, se puede recurrir a la estimación de la probabilidad conjunta a partir del producto total de todas las probabilidades predichas por el modelo.

A dicha estimación se le conoce como Verosimilitud del Modelo, su ecuación es:

$$V = \hat{P}1 \times \hat{P}2 \ldots \times \hat{P}r \ \times \ (1-\hat{P}r+1) \times (1-\hat{P}r+2) \ \times \ldots (1-\hat{P}s)$$

En la medida que el valor de V se aproxima a 1 mayor eficiencia del modelo.

Otros Aspectos Importantes

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.

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 ya sean ordenadas jerárquicamente o no.

Una regresión logística binaria o binomial se trata de variables de respuesta ordinal y el segundo caso hace alusión a una respuesta categórica.

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