Minería de Datos para el Análisis de Big Data

Por: Carlos Carreño

ccarrenovi@Gmail.com

Abril,2021

Modulo 6 Modelo de Clasificación

- Modelo de regresión logística simple
- Formulación Matemática del Modelo
- Concepto de ODDS
- Transformación Logarítmica de ODDS
- Regla de ODDS y Log
- Ajuste del Modelo
- Evaluación del Modelo
- Interpretación del Modelo
- Predicciones
- Probabilidad y Clasificación
- Modelo de regresión logística multivariable

Modelo de Clasificación y Aplicaciones

a) La probabilidad de estar empleado a los 3 meses (Sí o No) en función del promedio de carrera, el puntaje CENEVAL, el TOEFL, la carrera del egresado, etc.

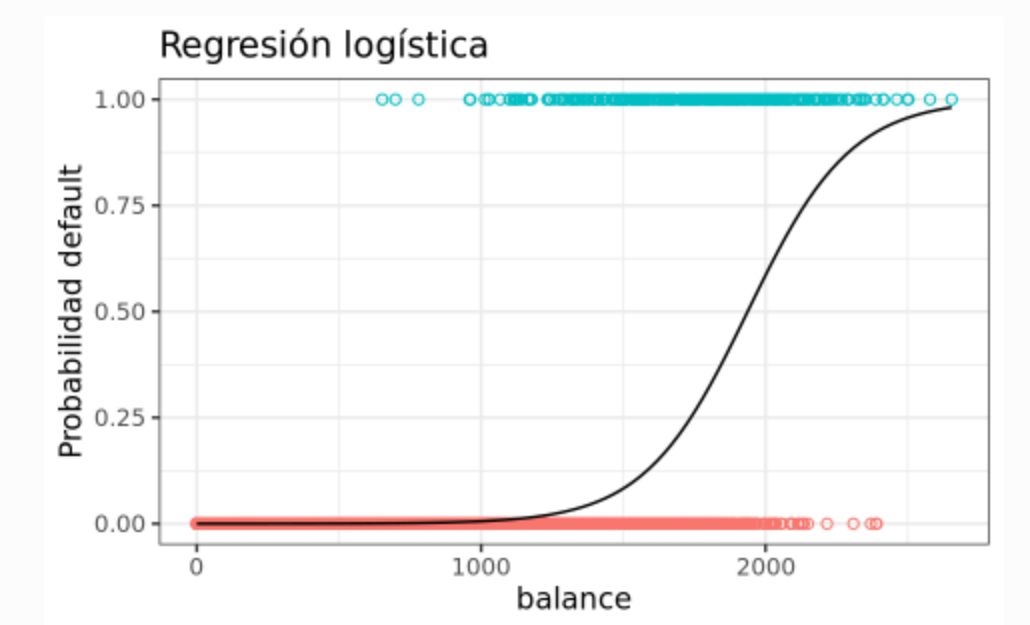
b) La probabilidad de que una persona compre un producto (Sí o No), en función de su edad, ingreso, clase social, género, etc.

c) La probabilidad de que una persona pague un crédito (Sí o No), en función de su ingreso, edad, deudas, nivel socioeconómico, nivel de estudios, estado civil, etc.

d) La probabilidad de que una empresa falle en pagar su deuda (Sí o No) en función del tamaño, antigüedad, giro, etc.

Modelo de Regresión Logística Simple

- La Regresión Logística Simple, desarrollada por David Cox en 1958.
- El método de regresión logística simple, 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.***
- 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 existencia de una relación significativa entre una variable cualitativa con dos niveles y una variable continua se puede estudiar mediante otros test estadísticos tales como t-test o ANOVA.
- la regresión logística permite además calcular la probabilidad de que la variable dependiente pertenezca a cada una de las dos categorías en función del valor que adquiera la variable independiente

Ejemplo: Regresión Logística

- Supóngase que se quiere estudiar la relación entre los niveles de colesterol y los ataques de corazón. Para ello, se mide el colesterol de un grupo de personas y durante los siguientes 20 años se monitoriza que individuos han sufrido un ataque.
- Un *t-test* entre los niveles de colesterol de las personas que han sufrido ataque *vs* las que no lo han sufrido permitiría contrastar la hipótesis de que el colesterol y los ataques al corazón están asociados.
- Si además se desea conocer la probabilidad de que una persona con un determinado nivel de colesterol sufra un infarto en los próximos 20 años, o poder conocer cuánto tiene que reducir el colesterol un paciente para no superar un 50% de probabilidad de padecer un infarto en los próximos 20 años, se tiene que recurrir a la regresión logística.

Formulación Matemática del Modelo de Regresión Logística

- la regresión logística transforma el valor devuelto por la regresión lineal (β0+β1X) 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 o logit):

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

- 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.

...

- Sustituyendo la x de la ecuación sigmoideo , por la función lineal (β0+β1X) se obtiene que:

$$P(Y = k|X = x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X)}} =$$

$$\frac{1}{\frac{e^{\beta_0+\beta_1 X}}{e^{\beta_0+\beta_1 X}} + \frac{1}{e^{\beta_0+\beta_1 X}}} =$$

$$\frac{1}{\frac{1+e^{\beta_0+\beta_1 X}}{e^{\beta_0+\beta_1 X}}} =$$

$$\frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

- Donde ***Pr(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

...

- ***Pr(Y=k|X=x)***, 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

$$ln(\frac{p(Y=k|X=x)}{1-p(Y=k|X=x)}) = \beta_0 + \beta_1 X$$

Concepto de ODDS o razón de probabilidad, ratio de ODDS y logaritmo de ODDS

- En regresión logística se modela la probabilidad de que la variable respuesta *Y* pertenezca al nivel de referencia 1 en función del valor que adquieran los predictores, mediante el uso de *LOG of ODDs*.
- Los ODDs o razón de probabilidad de verdadero se definen como el ratio entre la probabilidad de evento verdadero y la probabilidad de evento falso p/q.

Ejemplo: ODDS

- Supóngase que la probabilidad de que un evento sea verdadero es de *0.8*, por lo que la probabilidad de evento falso es de *1 - 0.8 = 0.2*.
- En este caso los *ODDs* de verdadero son *0.8 / 0.2 = 4*, lo que equivale a decir que se esperan 4 eventos verdaderos por cada evento falso.

Transformación Logarítmica de los ODDS

- La trasformación de probabilidades a ODDs es monotónica, si la probabilidad aumenta también lo hacen los ODDs, y viceversa.
- El rango de valores que pueden tomar los ODDs es de $[0,\infty]$.
- Dado que el valor de una probabilidad está acotado entre [0,1] se recurre a una trasformación **logit** (existen otras) que consiste en el logaritmo natural de los ODDs.
- Esto permite convertir el rango de probabilidad previamente limitado a [0,1] a $[-\infty,+\infty]$.

p	odds	Log(odds)
0.001	0.001001	-6.906755
0.01	0.010101	-4.59512
0.2	0.25	-1.386294
0.3	0.4285714	-0.8472978
0.4	0.6666667	-0.4054651
0.5	1	0
0.6	1.5	0.4054651
0.7	2.333333	0.8472978
0.8	4	1.386294
0.9	9	2.197225
0.999	999	6.906755

Regla de ODDS y Log

- Los *ODDs* y el *logaritmo de ODDs* cumplen que:
  - Si p(verdadero) = p(falso), entonces odds(verdadero) = 1
  - Si p(verdadero) < p(falso), entonces odds(verdadero) < 1
  - Si p(verdadero) > p(falso), entonces odds(verdadero) > 1
  - A diferencia de la probabilidad que no puede exceder el 1, los *ODDs* no tienen límite superior.
  - Si odds(verdadero) = 1, entonces logit(p) = 0
  - Si odds(verdadero) < 1, entonces logit(p) < 0
  - Si odds(verdadero) > 1, entonces logit(p) > 0
  - La transformación *logit* no existe para p = 0

Ajuste del Modelo

- Una vez obtenida la relación lineal entre el logaritmo de los ODDs y la ***variable predictora X***, se tienen que estimar los parámetros ***β0*** *y* ***β1***. La combinación óptima de valores será aquella que tenga la máxima verosimilitud (***maximum likelihood ML***), es decir ***el valor de los parámetros β0 y β1*** con los que se ***maximiza la probabilidad*** de obtener los datos observados.
- Otra forma para ajustar un modelo de regresión logística es empleando descenso de gradiente.

Evaluación del Modelo

se usa típicamente dos métodos: (Se calcula el ***p-value*** la significancia del modelo en conjunto)

- ***Wald chi-square***: está muy expandido pero pierde precisión con tamaños muéstrales pequeños.
- ***Likelihood ratio***: usa la ***diferencia*** entre la ***probabilidad*** de obtener los valores observados con el ***modelo logístico*** creado y las probabilidades de hacerlo con un modelo ***sin relación*** entre las variables.

Nota:
Para determinar la significancia individual de cada uno de los predictores introducidos en un modelo de regresión logística se emplea el estadístico Z y el test ***Wald chi-test***. En R, este es el método utilizado para calcular los ***p-values*** que se muestran al hacer ***summary()*** del modelo.

Interpretación del modelo

- A diferencia de la regresión lineal, en la que β1 se corresponde con el cambio promedio en la variable dependiente Y debido al incremento en una unidad del predictor X, en regresión logística, ***β1 indica el cambio en el logaritmo de ODDs*** debido al incremento de una unidad de X, o lo que es lo mismo, multiplica los ODDs por e^β1.
- Dado que ***la relación entre p(Y) y X no es lineal***, β1 no se corresponde con el cambio en la probabilidad de Y asociada con el incremento de una unidad de X. Cuánto se incremente ***la probabilidad de Y*** por unidad de X ***depende*** del valor de X, es decir, ***de la posición en la curva logística*** en la que se encuentre.

Condiciones del Modelo

- **Independencia**: las observaciones tienen que ser independientes unas de otras.
- **Relación lineal entre el logaritmo natural de *ODDs* y la variable continua**: patrones en forma de U son una clara violación de esta condición.
- **La regresión logística no precisa de una distribución normal** de la variable continua independiente.
- **Número de observaciones**: no existe una norma establecida al respecto, pero se recomienda entre 50 a 100 observaciones.

Interpretación de Coeficientes

En las estimaciones de la regresión logística, los coeficientes miden el cambio en el logaritmo de la razón de probabilidad de éxito vs fracaso (conocida como razón par, odd) cuando X se incrementa en una unidad.

Los coeficientes no se interpretan, porque están relacionados con el logaritmo de la razón Odd, se interpreta su exponencial, esto es la razón Odd:

$$\ln\left(\frac{\pi i}{1-\pi i}\right) = \beta_0 + \beta_1 X_i$$

$$\frac{\pi i}{1-\pi i} = e^{\beta_0+\beta_1 X_i} = = e^{\beta_0} e^{\beta_1 X_i}$$

e^{β_j} Es la razón par de la variable Xj

Ejemplo: Interpretación de Coeficientes

Ejemplos:
Y = Paga un crédito una persona (Sí, No)
X1 = Ingresos (miles de pesos)
X2 = Estado civil (casado o soltero)
X3 = Género (Hombre, Mujer)

Variable	Coeficiente	Razón odd	Interpretación
Ingreso	0.03634	exp(0.03634) = 1.037	Por cada mil adicional de ingresos, la probabilidad de que pague se incrementa en 3.7%
Estado Civil - Casado	0.71320	exp(0.71320) = 2.04	El casado tiene el doble de probabilidad de pagar que el soltero.
Género - Hombre	-0.04320	exp(-0.04320)= 0.96	El hombre tiene un 4% menos de probabilidad de pagar el crédito que las mujeres.

Predicciones

- Una vez estimados los coeficientes del modelo logístico, es posible conocer la probabilidad de que la variable dependiente pertenezca al nivel de referencia, dado un determinado valor del predictor. Para ello se emplea la ecuación del modelo:

$$\hat{p}(Y = 1|X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}}$$

Predicciones

- En R se emplea la función **glm()** con **family="binomial"** para ajustar modelos de regresión logística. Esta función predice por defecto el log(ODDs) de la variable respuesta. Para obtener las probabilidades P(y=1) hay que aplicar la ecuación previa, donde el valor $e^{\hat{\beta}_0+\hat{\beta}_1 X}$ es el log(ODDs) devuelto por el modelo. Otra opción es indicar el argumento **type="response"** en la función **predict().**

Matriz de Confusión

- Es una matriz que permite calcular la probabilidad de aciertos del modelo

Podemos usar las observaciones con las que se estimó el modelo para hacer validar el porcentaje de observaciones clasificadas correctamente, sin embargo es mejor hacerlo con una muestra de datos adicional, no usada en el modelo.

		Valores observados	
		Sí Paga	No Paga
Valores Predichos	Sí Paga	320	23
	No Paga	53	250

Porcentaje de clasificación correcta = (320+250)/(320+23+53+250) = 88.2%

Se busca que el punto de corte máximize el porcentaje de observaciones clasificadas correctamente de éxito (Si Paga) versus fracaso (No paga).

Convertir probabilidad en clasificación

- Una de las principales aplicaciones de un modelo de regresión logística es clasificar la variable cualitativa en función de valor que tome el predictor. Para conseguir esta clasificación, es necesario establecer un ***threshold*** (umbral o punto de corte) de probabilidad a partir de la cual se considera que la variable pertenece a uno de los niveles.
- Por ejemplo:
  - Se puede asignar una observación al grupo 1 si $\hat{p}(Y = 1|X) > 0.5$ y al grupo 0 si de lo contrario.

Regresión Logística Múltiple

- 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 continuos como categóricos.

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

$$logit(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

- El valor de la probabilidad de *Y* se puede obtener con la inversa del logaritmo natural:

$$p(Y) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_i x_i}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_i x_i}}$$

- A la hora de evaluar la validez y calidad de un modelo de regresión logística múltiple, se analiza tanto el modelo en su conjunto como los predictores que lo forman. Se considera que el modelo es útil si es capaz de mostrar una mejora respecto al modelo nulo, el modelo sin predictores. Existen 3 test estadísticos que cuantifican esta mejora mediante la comparación de los residuos: *likelihood ratio*, *score* y *Wald test*

Prueba de Significancia

- Ho: El modelo no es explicativo (B1=B2=B3=…=Bi=0)
- H1: El modelo es significativo (Al menos un Bi es distinto de 0)

Estadístico = (Error del modelo con las variables explicativas) – (Error del modelo sin las variables explicativas)

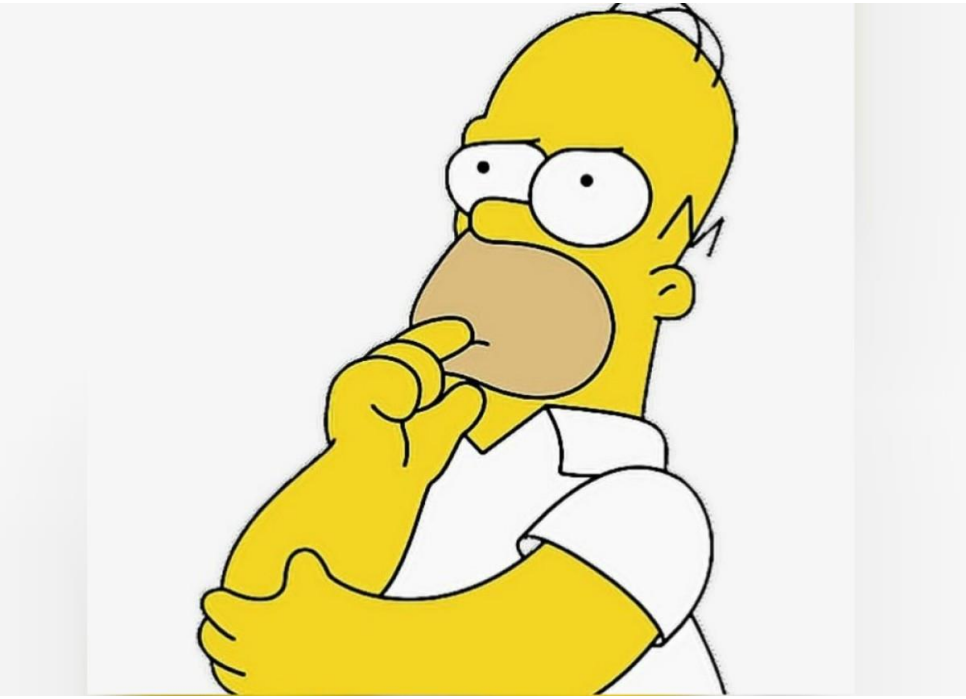
El estadístico tiene una distribución Ji-Cuadrada con k grados de libertad (k = al total de variables explicativas)

Rechazar Ho a favor de Ha sí

Estadístico >= Ji-Cuadrada,α,k o bien si α >= Valor P

Preguntas

- Alguna pregunta?



Demo

- Estudiar el caso de clasificación para determinar si un cliente caerá en ***default*** (atraso en sus compromisos de pago), teniendo en cuenta el ***balance*** (promedio mensual del saldo en su tarjeta de crédito.)