CLOUD FOR AI

MODELOS SUPERVIZADOS: REGRESIÓN LOGÍSTICA

VÍCTOR ACEVEDO

Perú

ÍNDICE

NOCIONES INICIALES I

- Dado un conjunto de registros (conjunto de entrenamiento)
  - Cada registro contiene un conjunto de **atributos**, donde uno de ellos es la **clase**.
- Encontrar un modelo para el atributo de clase en función de los valores de los demás atributos.
- Objetivo: Nuevos registros sean asignados a una clase con la mayor precisión posible.
  - Un conjunto de prueba es usada para determinar la precisión del modelo. Usualmente, el conjunto de datos original es dividido en un conjunto de prueba y de entrenamiento, donde el conjunto de entrenamiento es usado para construir el modelo y el de prueba para validarlo.

NOCIONES INICIALES II

- Los modelos de clasificación generan dos tipos de predicciones:
  - Continuas: Usualmente en la forma de una probabilidad (los valores predichos de pertenencia a una clase para un individuo está entre 0 y 1).
  - Categóricas (discretas): Clase predicha.
- Para la mayoría de las aplicaciones prácticas, la predicción de una categoría discreta es necesaria para poder tomar una decisión y es el objetivo de la predicción. Ejemplo: Filtro automático de spam.
- La probabilidad estimada para cada clase puede ser muy útil para medir el ajuste del modelo sobre la clasificación predicha: Un mensaje por e-mail con una probabilidad de ser spam de 0.51 puede ser clasificado de manera similar que otro mensaje con una probabilidad de 0.99
- En algunas aplicaciones el resultado deseado es la probabilidad de pertenecer a una clase, la que será usada como entrada para otros cálculos.

EVALUACIÓN I

Predichos	Observados	
	Eventos	No Eventos
Eventos	TP	FP
No Eventos	FN	TN

Figura: Matriz de confusión para un problema de clasificación con dos clases (eventos y no eventos). Las celdas de la tabla indican el número de verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN)

EVALUACIÓN II

- La métrica más simple es el ratio de la precisión total (o, siendo pesimistas, el ratio de error).
- Este patrón es un indicador de que el modelo tiene una pobre calibración y también desempeño.

FUNDAMENTOS TEÓRICOS I

- En vez de modelar directamente la respuesta Y, los modelos de regresión logística modelan la probabilidad de que Y pertenezca a una categoría en particular.
- Para la data Default, la regresión logística modela la probabilidad de que un cliente incumpla con el pago de la tarjeta de crédito (moroso).
- Por ejemplo, la probabilidad de que sea moroso dado balance puede ser escrita como

$$Pr(default = Yes \mid balance)$$

- Los valores de $Pr(default = Yes|balance)$, que puede abreviarse como π, se encontrarán en el rango entre 0 y 1.
- Por ejemplo, es posible predecir default=yes para aquellos individuos en que $\pi > 0.5$

FUNDAMENTOS TEÓRICOS II

- En la regresión logística, es usada la función logística

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

- La gráfica del lado derecho de la figura anterior muestra el ajuste a un modelo de regresión logística para el conjunto de datos Default.
- Se puede observar que el modelo logístico captura mejor el rango de probabilidades que el modelo de regresión lineal mostrado en el lado izquierdo.
- La probabilidad ajustada promedio en ambos casos es 0.0333, la cual es la misma ue la propoción total de morosos en la data.
- Con alguna manipulaciones básicas se puede obtener

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

FUNDAMENTOS TEÓRICOS III

- El valor $\frac{\pi}{1-\pi}$ es conocido como odds y puede tomar cualquier valor entre 0 y ∞.
- Los valores de odds cercanos a 0 o a ∞ indican probabilidades muy bajas o muy altas de ser morosos, de manera respectiva.
- Por ejemplo, en promedio 1 de 5 personas con un odd de $1/4$ será morosa dado que $\pi = 0.2$ implica que los odds son de $\frac{0.2}{1-0.2} = 1/4$
- Del mismo modo, en promedio nueve de cada diez personas con odds de 9 será morosa, dado que $\pi = 0.9$ implica un odds de $\frac{0.9}{1-0.9} = 9$
- Tomando logaritmos a la anterior ecuación se obtiene

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

La expresión del lado izquierdo es conocido como el logit.

FUNDAMENTOS TEÓRICOS IV

- **Componente aleatorio**: Sean $Y_1, ..., Y_n$ v.a. dicotómicas independientes. Asumiendo que $y_i = 1$ tiene probabilidad π_i y $y_i = 0$ con probabilidad $1 - \pi_i$:

$$y_i \sim Bernoulli\,(\pi_i)$$

- **Componente sistemático**:

$$\eta_i = \beta_1 x_{1i} + ... + \beta_1 x_{pi} = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

donde η_i es denominado como predictor lineal y $x_i = (x_{i1}, \ldots, x_{ip})'$ es un vector de covariables, donde x_{i1} igual a 1 corresponde al intercepto.

- **Función de Enlace:**

$$g(\pi_i) = \eta_i$$

donde $g(.)$ es una función monótona y diferenciable.

FUNDAMENTOS TEÓRICOS V

- Enlace Logit

$$\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

- Enlace Probit

$$\Phi(\pi(x))^{-1} = \beta_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

donde $\Phi(.)$ es la f.d.a. de la normal estándar.

- Enlace log-log complementario (cloglog)

$$\log\{-\log(1-\pi(X))\} = \beta_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

FUNDAMENTOS TEÓRICOS VI

- Enlace Logit

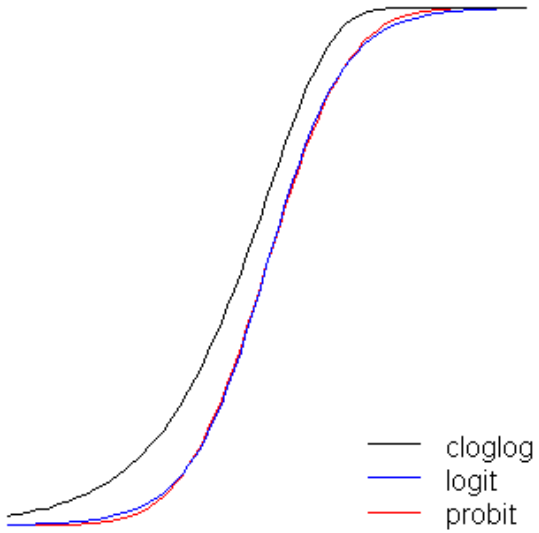
  Este enlace debería ser usado si los errores siguen una distribución logística. Tiene colas ligeramente más largas que probit

- Enlace Probit

  Este enlace debería ser usado si los errores siguen una distribución normal. En estricto debería denominarse modelo normit.

- Enlace log-log complementario (cloglog)

  Es comúnmente usado cuando el evento a predecir es poco frecuente.



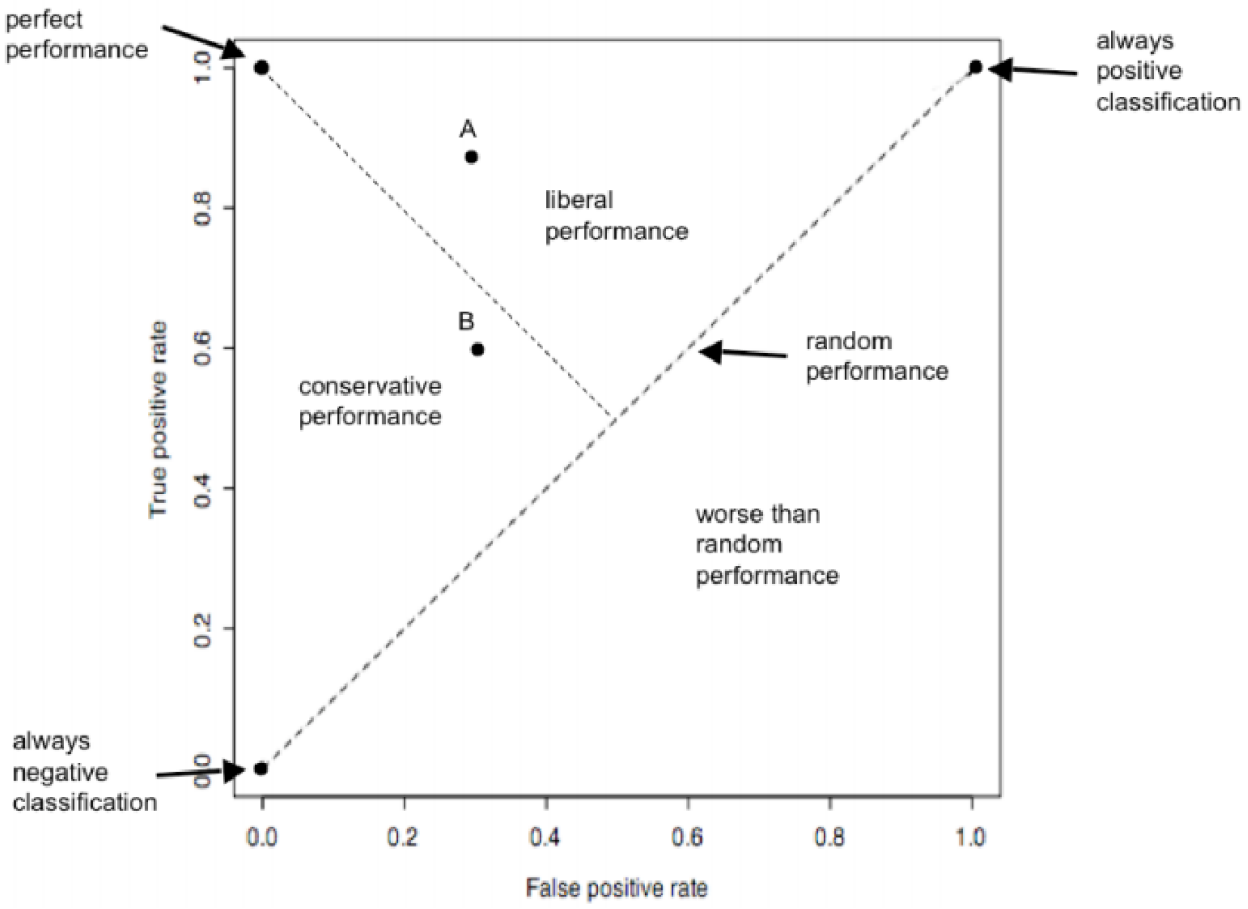
CURVA ROC I

Una curva ROC se construyen en base a:

- La sensibilidad (S), definida como $S = \frac{TP}{TP+FN}$; es decir, la proporción de objetos correctamente clasificados como éxitos e, informalmente, conocidos como la proporción de verdaderos positivos.
- La especificidad (E), definido como $S = \frac{TN}{FP+TN}$; es decir, la proporción de objetos correctamente clasificados como fracasos.

La curva ROC no es sino la gráfica de $1 - E = \frac{n_{12}}{n_{.2}}$ o proporción de falsos positivos en el eje de las abscisas frente a la sensibilidad S o proporción de verdaderos positivos en el eje de las ordenadas, para diferentes valores del punto de corte $c \in [0, 1]$.

CURVA ROC II


perfect performance
always positive classification
A
liberal performance
B
random performance
conservative performance
worse than random performance
always negative classification
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

BIBLIOGRAFÍA

- Green, William H. (2003). *Econometric Analysis, fifth edition*. Prentice Hall.
- Hosmer, David W.; Stanley Lemeshow (2000). *Applied Logistic Regression, 2nd ed*. New York; Chichester, Wiley.

- Micromaster Data Science, EDX plataforma virtual
- Apuntes Clases de Maestría Universidad Nacional Agraria La Molina

GRACIAS
POR SU ATENCIÓN