Medidas de divergencia en modelos de predicción binaria

Alex Pérez
alex.perezt@epn.edu.ec
Escuela Politécnica Nacional
Quito - Ecuador

Diego Huaraca
diego.huaraca@epn.edu.ec
Escuela Politécnica Nacional
Quito - Ecuador

Resumen

El presente artículo describe a detalle la aplicación que presentan ciertas medidas de divergencia en la construcción de modelos de predicción binaria. La importancia del siguiente estudio radica en el desarrollo de un criterio de selección de predictores mediante el análisis de la divergencia generada por cada uno de ellos entre los grupos dados por la variable dependiente binaria.

1. Introducción

En las últimas décadas el análisis estadístico de datos ha despertado un gran interés en varios campos. La resolución de problemas de clasificación en entidades financieras mediante la utilización de modelos de predicción binaria es cada vez más frecuente debido a que constituyen una herramienta técnica de gran utilidad en los procesos de toma de decisiones. En concreto, tales modelos han sido aplicados exitosamente en problemas de clasificación de clientes, otorgamientos de productos financieros, etc. considerando su historial crediticio y su hábito de pago, de ahí la importancia de construir un criterio fundamentado estadísticamente que permita seleccionar los predictores con el mayor poder de predicción, puesto que en la actualidad las variables son incluidas intuitivamente dependiendo de la experiencia del modelador, en ocasiones omitiendo predictores importantes y en otras generando modelos con un bajo rendimiento de clasificación.

2. Medidas de divergencia

Las medidas de divergencia se consideran como medidas cuantitativas de discriminación entre dos poblaciones, caracterizadas por sus respectivas distribuciones de probabilidad, éstas medidas proveen valiosa información sobre la calidad de los datos a emplearse en el desarrollo de modelos de predicción binaria.

En nuestro trabajo mostramos dos de las medidas de divergencia más utilizadas: Kolmogorov Smirnov & Anderson Darling.

2.1. Kolmogorov-Smirnov (KS)

El test KS para dos muestras es una prueba de bondad de ajuste mediante la cual se contrasta la hipótesis de si dos muestras aleatorias independientes provienen de distribuciones continuas idénticas; es una prueba del tipo no paramétrico debido que no es necesario realizar suposiciones a priori sobre la distribución de los datos.

El test KS para dos muestras independientes constrasta las hipótesis:

$$\begin{cases} H_0 : F_1(x) = F_2(x), & \forall x \\ H_1 : F_1(x) \neq F_2(x) \end{cases} \tag{1}$$

donde: F_1 y F_2 son las funciones de distribución de las muestras aleatorias (continuas) $X \sim \{x_1, x_2, \ldots, x_{n_1}\}$ y $Y \sim \{y_1, y_2, \ldots, y_{n_2}\}$ de tamaño n_1 y n_2 respectivamente.

El estadístico empleado para constrastar la hipótesis nula H_0 está basado en la utilización de las funciones de distribución empíricas $\hat{F}_1(x)$ y $\hat{F}_2(x)$ de X y Y, y su valor es obtenido mediante la expresión:

$$\text{KS} = \max_x |\hat{F}_1(x) - \hat{F}_2(x)| \tag{2}$$

La hipótesis nula H_0 se rechaza si el estadístico KS

es mayor que el valor crítico KS_α, para un nivel de significancia α dado. [Massey, 1951] presenta una tabla de valores críticos para diferentes tamaños de muestra.

A partir de (2), podemos decir que el estadístico KS es la distancia máxima entre $\hat{F}_1$ y $\hat{F}_2$, y que su valor oscila entre 0 y 1, donde valores cercanos a 0 indican que la distribuciones de X y Y son idénticas, y valores cercanos a 1 indican que las mismas difieren, es por esta razón, que el estadístico KS es utilizado en la práctica como una medida de divergencia entre las distribuciones de dos variables aleatorias continuas.

2.2. Anderson Darling (AD)

El test AD para dos muestras independientes es en general una prueba más potente que el test KS, debido que el estadístico KS se construye como la diferencia máxima entre las funciones de distribución empíricas sin considerar el comportamiento en las colas de las distribuciones.

Para constrastar la hipótesis (1) por medio del estadístico (AD) asumiremos nuevamente dos funciones de distribución F_1 y F_2 correspondientes a dos muestras aleatorias (continuas) $X \sim \{x_1, x_2, \ldots, x_{n_1}\}$ y $Y \sim \{y_1, y_2, \ldots, y_{n_2}\}$ de tamaño n_1 y n_2 respectivamente. El estadístico de contraste se obtiene mediante la expresión:

$$AD = \frac{n1 * n2}{N} \int_{-\infty}^{\infty} \frac{[F_1(x) - F_2(x)]^2}{H_N(x)(1 - H_N(x))} dH_N(x) \quad (3)$$

donde:

$$N = n_1 + n_2$$

$$H_N(x) = \frac{n_1 F_1(x) - n_2 F_2(x)}{N}$$

[Scholz and Stephens, 1987] proponen utilizar la siguiente fórmula computacional como una aproximación de (3):

$$AD = \frac{1}{N} \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{N-1} \frac{[NM_{ij} - jn_i]^2}{j(N-j)}, \quad \text{con } k = 2 \quad (4)$$

para lo cual ordenamos los valores de las variables X y Y conjuntamente con la finalidad de obtener una muestra ordenada $Z_1 < Z_2 < \ldots < Z_N$, y finalmente definir a M_{ij} como el número de observaciones de la muestra i que son menores o iguales que Z_j.

3. Implementación

Una vez que hemos descrito las medidas de divergencia, el siguiente paso consiste en implementar un algoritmo que permita calcularlas automáticamente, usando R[1] y *R Analytic Flow*[2]

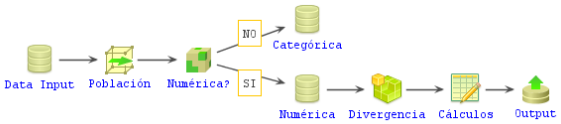

Figura 1: Flujo implementado

El flujograma de la figura anterior ejecuta todas las tareas necesarias para el cálculo de las medidas KS y AD, la implementación se la realizó de tal manera que facilite al usuario trabajar con volumenes de datos a gran escala sin la necesidad de tener conocimientos previos de R, el flujo generado puede ser obtenido directamente del repositorio GitHub[3].

Una versión detallada de la implementación, así como un instructivo sobre la utilización del flujo se pueden encontrar en RPubs[4].

4. Resultados

Con la finalidad de evaluar los resultados de la implementación se trabajó con una base de 15105 observaciones y 15 variables. La variable a predecir *var.dep* es una variable binaria que analiza la activación (realizar al menos un consumo) de una nueva tarjeta de crédito TC, por un individuo, en una ventana de tiempo determinada, definida como sigue:

$$var.dep = \begin{cases} 1, & \text{si se activa la TC,} \\ 0, & \text{caso contrario.} \end{cases}$$

[1] http://www.r-project.org
[2] http://www.ef-prime.com/index_en.html
[3] https://github.com/DimatEpn/Flujograma-Medidas-Divergencia
[4] https://rpubs.com/DimatEpn/90586

Al ejecutar el algoritmo (Figura 1) se obtuvieron los siguientes resultados para los 7 primeros predictores:

Variable	KS	AD
mean.tic	0.9561	0.5497
consumos	0.9175	1.0000
sd.tic	0.9175	0.5285
tc.apert3m	0.3374	0.1521
score	0.2931	0.0945
cupo.tot	0.2525	0.0639
porc.cupo	0.2447	0.0772

Figura 2: Divergencia

Si analizamos la variable tiempo inter-consumo promedio (*mean.tic*), observamos un KS de $0,9561$ (cercano a 1) y AD de $0,5497$, lo cual indica que este predictor genera una divergencia considerable entre los grupos *activa* y *no activa*, esto resulta lógico pues se esperaría que mientras el tiempo inter-consumo promedio sea menor, el individuo sea más propenso a activar la nueva tarjeta de crédito. Un análisis similar se podría plantear para las variables restantes, por ejemplo, número de consumos realizados (*consumos*), desviación estándar del tiempro inter-consumo (*sd.tic*), número de tarjetas apertudas los últimos 3 meses (*tc.apert3m*), etc. se esperaría que éstas variables generen una divergencia considerable entre los grupos puesto que analizan el patrón de uso y consumo de TC que registra el individuo.

5. Conclusiones

1. Dentro de la construcción de modelos de predicción binaria una de las problemáticas, a las cuales el especialista enfrenta consiste en establecer a priori una cota mínima de divergencia que generará el modelo final, mediante la metodología planteada se establece que dicha cota está dada por la máxima divergencia que presentan las variables predictoras del modelo.

2. La metodología descrita genera un criterio de selección de predictores, permitiendo reducir tiempo y recursos en la construcción de modelos de este tipo. Esto a su vez nos da una idea de la calidad y tipo de información con la cual se trabaja.

6. Bibliografía

- Arnold, T. and Emerson, J. (2011). Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions. *The R Journal*, 3:34-35.

- Engmann, S. and Cousineau, D. (2011). Comparing Distributions: The Two-Sample Anderson-Darling Test as an Alternative to the Kolmogorov-Smirnof Test. *Journal of Applied Quantitative Methods*, 6.

- Massey, F. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 68-78.

- R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

- Scholz, F. W. and Stephens, M. A. (1987). K-Sample Anderson-Darling Tests. *Journal of the American Statistical Association*, 82:918Ű919.