CARRERA DE INGENIERÍA DE SISTEMAS

Regresión Logística

ESTADÍSTICA APLICADA
PROF. MSC. VILMA S. ROMERO ROMERO
5TO CICLO | 30-06-2022

UNIVERSIDAD DE LIMA

Introducción

Muchas veces cuando realizamos un análisis de datos nos enfrentamos con el hecho de que la variable que deseamos predecir es cualitativa o categórica.

- Predecir si el cliente adquirirá una tarjeta de crédito dadas sus características financieras y demográficas.
- Predecir si un email es spam en base a su origen, caracteres, imágenes o información del encabezado.
- Predecir el nivel académico de una persona (Secundaria, Universitaria, Postgrado) según una foto proporcionada.
- Predecir si un paciente presenta cierta enfermedad en base a los síntomas y signos vitales examinados.

¿Qué es Regresión Logística?

- Herramienta para construir modelos cuando se tiene una variable de respuesta categórica con dos niveles.
- Es un tipo de Modelo Lineal Generalizado (GLM), que permite en la modelación variables respuesta que tienen una distribución diferente a la distribución normal.
  - GLM puede ser visto como un enfoque de modelamiento en dos etapas.
  - 1. Modelar la variable respuesta usando una distribución de probabilidad: Binomial, Poisson.
  - 2. Modelar el parámetro de la distribución usando un conjunto de predictores.
- La RL modela la probabilidad de que la variable respuesta Y pertenezca a una categoría particular en función de las variables predictoras.
  - La presencia de un predictor aumenta (o disminuye) la probabilidad de un resultado determinado en un porcentaje específico.

UNIVERSIDAD DE LIMA

Modelo de Regresión Logística

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \qquad (1)$$

- $X = (X_1, \ldots, X_k)$ son los predictores.
- k es la cantidad de predictores
- X_j representa el `j-ésimo` predictor.
- β_j son los parámetros desconocidos a estimar.

$$P(Y = 1 \mid X) = p(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k}} \qquad (2)$$

UNIVERSIDAD DE LIMA

Modelo de Regresión Logística

- La proporción $\frac{p(X)}{1-p(X)}$ es llamada *odds*.
  - El *odds a favor de que ocurra un evento* se define como la probabilidad de que ocurra el evento dividida por la probabilidad de que no ocurra.
  - Puede tomar cualquier valor entre 0 e ∞.
  - Valores cercanos a 0 indican muy bajas probabilidades de éxito.
  - Valores cercanos a ∞ indican muy altas probabilidades de éxito.
- $log(\frac{p(X)}{1-p(X)})$ recibe el nombre de *log-odds* o *logit*.

Estimación

- Los coeficientes de regresión $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ son estimados mediante Máxima Verosimilitud, considerando la distribución Binomial de las respuestas Y_i para generar estimadores óptimos.
  - MLE (Maximum Likelihood Estimation) encuentra las estimaciones de parámetros que maximizan la función log-likelihood (logaritmo de la función de verosimilitud).
  - El proceso de maximización se realiza empleando métodos numéricos iterativos.

Interpretación de los coeficientes

- El incremento en una unidad del valor de X_j, manteniendo las demás variables constantes, genera un cambio en el *log-odds* en β_j.
- El **odds-ratio** de una variable independiente representa el cambio en los odds debido al cambio en una unidad de la variable independiente manteniendo constantes todas las demás variables independientes.

$$\boxed{\text{Odds-Ratio}(X_j) = \frac{\text{Odds tras cambio en una unidad de } X_j}{\text{Odds originales}} = e^{\beta_j}}$$

  - $OR = 1 \Rightarrow$ Indica un efecto cero. Los odds para ambos eventos son los mismos.
  - $OR > 1 \Rightarrow$ Indica un incremento en odds. La variable independiente tiene un impacto positivo en la probabilidad de que ocurra el evento.
  - $OR < 1 \Rightarrow$ Indica una disminución en odds.

Interpretación de los coeficientes

En términos generales,

- Si β_j es positivo, un aumento de X_j estará asociado con un incremento de $p(X)$.
- Si β_j es negativo, un aumento de X_j estará asociado con una disminución de $p(X)$.

Significancia del modelo global

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1 : \text{Al menos un } \beta_j \text{ no es igual a } 0$$

- Una comparación de la devianza nula (*Null deviance*) y la devianza residual (*Residual deviance*) se utiliza para probar la hipótesis nula global.
  - *Devianza*
    - Utilizada para evaluar el ajuste del modelo.
    - Similar a la SCResidual en una regresión lineal.
    - Cuanto menor es el valor de la devianza, el ajuste es mejor.
  - *Devianza Nula*, similar a la SCResidual en una regresión lineal cuando solo se ajusta una media general ($\sim$ considerando solo el intercepto).
  - *Devianza Residual*, similar a la SCResidual en una regresión lineal cuando se ajusta el modelo completo.

UNIVERSIDAD DE LIMA

Significancia del modelo global

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1 : \text{Al menos un } \beta_j \text{ no es igual a } 0$$

- Se utiliza una **prueba de razón de verosimilitud** (*Likelihood Ratio Test*) para esta prueba anidada que sigue una distribución χ^2_q bajo la H_0 verdadera.
  - χ^2_q es una distribución chi-cuadrado con q grados de libertad, donde q será el número de covariables en el modelo completo.

$$\chi^2_0 = \text{Null deviance} - \text{Residual deviance}$$

  - **Decisión**:
$$\text{p-value} = P(\chi^2_q > \chi^2_0) \Rightarrow \textbf{Rechazar } H_0 \text{ si p-value} \leq \alpha$$

Significancia individual de los coeficientes

Después de rechazar la hipótesis nula global, se puede considerar pruebas Z individuales para los predictores.

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

- Evaluar la contribución individual de cada una de las variables predictoras.
  - Evalúa si los coeficientes $\hat{\beta}_j$ para cada variable predictora son significativamente distintos de cero.
  - Si es distinto de cero $\Rightarrow$ variable predictora realiza una contribución significativa al modelo para predecir la respuesta.
- Se utiliza el estadístico de Wald (Prueba Z) basado en la normalidad asintótica de los $\hat{\beta}_j$.

$$Z_0 = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

- **Decisión**: **Rechazar** H_0 si p-value $\leq \alpha$.

UNIVERSIDAD DE LIMA

Recursos Adicionales I

Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D., and Cochran, J. J. (2017). *Statistics for Business and Economics*. Cengage Learning, 13 edition.

Diez, D., Barr, C., and Çetinkaya-Rundel, M. (2015). *OpenIntro Statistics*. 3 edition. `openintro.org`.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning: with Applications in R*. Springer.