Resumen Académico: *Logistic Regression in Data Analysis: An Overview* (Maalouf, 2011)

Resumen generado por Carlos con ayuda de ChatGPT

Resumen general

El artículo de Maher Maalouf (2011) es una revisión detallada de la regresión logística (RL) como técnica central para problemas de clasificación binaria. Se abordan tanto los fundamentos teóricos del modelo como estrategias computacionales y estadísticas para mejorar su rendimiento, en especial en contextos con desbalance de clases o datos de alta dimensión.

Modelo base de regresión logística

La RL modela la probabilidad de un evento binario $y_i \in \{0, 1\}$ en función de un vector de predictores x_i mediante:

$$p_i = \frac{1}{1 + e^{-x_i \beta}} \tag{1}$$

La verosimilitud del modelo es:

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i} \tag{2}$$

Y su log-verosimilitud:

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)] \tag{3}$$

Derivadas: Gradiente y Hessiano

- Gradiente:

$$\nabla_\beta \ell(\beta) = X^T (\mathbf{y} - \mathbf{p}) \tag{4}$$

- Hessiano:

$$\nabla^2_\beta \ell(\beta) = -X^T V X, \quad \text{donde } V = \text{diag}(p_i(1 - p_i)) \tag{5}$$

Regularización

Para evitar el sobreajuste, se añade un término de penalización L2 (ridge):

$$\ell_\lambda(\beta) = \ell(\beta) - \frac{\lambda}{2}\|\beta\|^2 \tag{6}$$

- Gradiente regularizado:

$$\nabla_\beta \ell_\lambda(\beta) = X^T(\mathbf{y} - \mathbf{p}) - \lambda\beta \tag{7}$$

- Hessiano regularizado:

$$\nabla^2_\beta \ell_\lambda(\beta) = -X^T V X - \lambda I \tag{8}$$

Algoritmo IRLS (Iteratively Reweighted Least Squares)

Una técnica común para estimar los parámetros del modelo es IRLS, que utiliza pesos $v_i = p_i(1 - p_i)$ y variables ajustadas z_i:

$$z_i = x_i\hat{\beta} + \frac{y_i - p_i}{v_i} \tag{9}$$

En cada iteración, se resuelve:

$$(X^T V X + \lambda I)\hat{\beta}^{(c+1)} = X^T V z^{(c)} \tag{10}$$

Este método es eficiente para bases de datos de tamaño moderado.

Algoritmo CG (Conjugate Gradient)

En problemas a gran escala, se recomienda el método del gradiente conjugado:

- Se inicializa el residuo $r^{(0)} = b - A\beta^{(0)}$.
- Se actualizan las direcciones de búsqueda y pasos óptimos iterativamente.
- Permite resolver sistemas lineales sin invertir matrices.

Es especialmente útil cuando $X^T V X$ es grande o disperso.

Correcciones para eventos raros

- **Ajuste del intercepto:** basado en la tasa real de eventos:

$$\tilde{\beta}_0 = \hat{\beta}_0 - \ln\left(\frac{1-\tau}{\tau} \cdot \frac{y}{1-y}\right) \tag{11}$$

- **Ponderación:** modifica la verosimilitud con pesos:

$$\ell(\beta|y, X) = \sum_{i=1}^{n} w_i \ln\left(\frac{e^{x_i\beta}}{1 + e^{x_i\beta}}\right) \tag{12}$$

Conclusiones clave

- La regresión logística es robusta y se adapta bien a diferentes contextos de datos.
- Las técnicas de regularización y los métodos numéricos como IRLS y CG la hacen escalable.
- Las correcciones para eventos raros mejoran la inferencia en muestras sesgadas.
- Es una herramienta base para modelos más complejos como regresión multinomial o clasificación ordinal.