Maestría en Ciencias del Procesamiento de la Información
Universidad Autónoma de Zacatecas

Regresión Logística

Reconocimiento de Patrones

Dr. Gamaliel Moreno Chávez
gamalielmch@uaz.edu.mx

Enero-Julio 2021

Temas

Clasificación

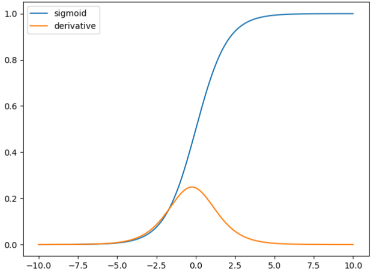
- Regresión: $y = h_{\boldsymbol{\theta}}(\boldsymbol{x})$ con $y \in \mathbb{R}$ continuo
- Clasificación: $y = h_{\boldsymbol{\theta}}(\boldsymbol{x})$ como antes, pero con $y \in \mathbb{C}$ es discreto y posiblemente no ordenado.
- iniciaremos con problema de clasificación binario $y \in \{0, 1\}$
- Ejemplos
  - Clasificación de correo como spam o no spam
  - Paciente con enfermedad o no la presenta
  - Una máquina fallará o no dadas ciertas lecturas de sensores
- Dado $\boldsymbol{x}^{(i)}$, el valor deseado correspondiente y^i se llama etiqueta

Regresión Logística

- Buscar una hipótesis $h(\boldsymbol{x}) \in [0, 1]$
- Elijamos:

$$h_{\boldsymbol{\theta}} = g(\boldsymbol{\theta}^T \boldsymbol{x}) = \sigma(\boldsymbol{\theta}^T \boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \boldsymbol{x}}}$$

con g una función no lineal, es este caso concreto elegido como la función sigmoide, función logística



Función logística $\sigma(z) = \frac{1}{1+e^{-z}}$ y su derivada $\sigma'(z) = \sigma(z)(1 - \sigma(z))$

Regresión Logística. Planteo probabilístico

- Vamos a usar planteo probabilístico para derivar solución
- Supongamos que la hipótesis cumple

$$P(y = 1|\boldsymbol{x}; \boldsymbol{\theta}) = h_{\boldsymbol{\theta}}(\boldsymbol{x})$$

además

$$P(y = 0|\boldsymbol{x}; \boldsymbol{\theta}) = 1 - h_{\boldsymbol{\theta}}(\boldsymbol{x})$$

las cuales se pueden combinar (Bernoulli)

$$P(y|\boldsymbol{x}; \boldsymbol{\theta}) = h_{\boldsymbol{\theta}}(\boldsymbol{x})^{y}(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^{1-y}$$

- La verosimilitud, igual que el caso de regresión con i.i,d es:

$$L(\boldsymbol{\theta}) = P(\boldsymbol{y}|\boldsymbol{X}; \boldsymbol{\theta}) = \prod_{i=1}^{m} P(y^{(i)}|\boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

$$= \prod_{i=1}^{m} h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})^{y^{(i)}}(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}))^{1-y^{(i)}}$$

Regresión Logística. Planteo probabilístico

- Queremos maximizar esta verosimilitud, lo que es más fácil a través de la verosimilitud logarítmica

$$\ell(\boldsymbol{\theta}) = lnL(\boldsymbol{\theta}) = \sum_{i=1}^{m} ln\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})^{y^{(i)}}(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}))^{1-y^{(i)}}\right)$$
$$= \sum_{i=1}^{m} y^{(i)} ln\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})\right) + (1 - y^{(i)}) ln(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}))$$

- Podemos maximizar esto utilizando ascenso de gradiente:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$$

- Si derivamos $\partial \ell(\boldsymbol{\theta})/\partial \theta_j$ llegamos con $\sigma'(z) = \sigma(z)(1 - \sigma(z))$ a

$$\frac{\partial}{\partial \theta_j} \ell(\boldsymbol{\theta}) = \sum_{i=1}^{m} \left(y^{(i)} - h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})\right) \boldsymbol{x}_j^{(i)} \quad \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \sum_{i=1}^{m} \left(y^{(i)} - h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})\right) \boldsymbol{x}^{(i)}$$

- Con lo anterior obtenemos para el ascenso estocástico de gradiente:

$$\theta_j \leftarrow \theta_j + \alpha \left(y^{(i)} - h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) \right) \boldsymbol{x}_j^{(i)} \quad \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \left(y^{(i)} - h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) \right) \boldsymbol{x}^{(i)}$$

- Es exactamente la misma expresión que para los mínimos cuadrados ordinarios (OLS), a pesar de que $h_{\boldsymbol{\theta}}(\cdot)$ es distinta

Modelos lineales generalizados

- Hemos visto dos tipos de problemas
  - Regresión lineal con mínimos cuadrados: $y|\boldsymbol{x};\boldsymbol{\theta} \backsim \mathcal{N}(\boldsymbol{\mu}\sigma^2)$
  - Regresión logística: $y|\boldsymbol{x};\boldsymbol{\theta} \backsim Ber(\phi)$
- ¿Por qué en ambos problemas llegamos a la misma regla de actualización de $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \left(y^{(i)} - h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) \right) \boldsymbol{x}^{(i)} \text{ ?}$$

- Ambos métodos pertenecen a los Modelos Lineales Generalizados (GLM)

Familias de distribuciones

- Podemos ver las distribuciones con sus parámetros como familias:
  - $Ber(\phi) : P(y = 1; \theta) = \theta$
  - $\mathcal{N}(\boldsymbol{\mu}, \sigma^2) : p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
- Esto es: cada instancia de parámetros produce una distribución particular
- Mostraremos que estos son casos especiales de la familia exponencial

La familia exponencial

- La familia exponencial incluye las distribuciones que se pueden expresar:
$$p(y; \boldsymbol{\eta}) = b(y) exp(\boldsymbol{\eta}^T \boldsymbol{T}(y) - a(\boldsymbol{\eta}))$$
- η es el parámetro natural (o canónico) de la distribución
- $\boldsymbol{T}(y)$ es el estadístico suficiente. En casos aquí, se cumple $\boldsymbol{T}(y) = y$
- $a(\boldsymbol{\eta})$ es la función de partición logarítmica
- Usualmente $e^{-a(\boldsymbol{\eta})}$ es constante de normalización
- Elección fija de $\boldsymbol{T}, a$ y b define una familia (o conjunto) de distribuciones parametrizadas con $\boldsymbol{\eta}$

Caso de distribución de Bernoulli

$$p(y;\boldsymbol{\eta}) = b(y)exp(\boldsymbol{\eta}^T \boldsymbol{T}(y) - a(\boldsymbol{\eta}))$$

Distribución de Bernoulli

$$\begin{aligned} p(y;\phi) &= \phi^y(1-\phi)^{1-y} = exp(ln(\phi^y(1-\phi)^{1-y})) \\ &= exp(yln\phi + (1-y)ln(1-\phi)) \\ &= exp\left(\underbrace{\left(ln\left(\frac{\phi}{1-\phi}\right)\right)}_{\eta}\underbrace{y}_{\boldsymbol{T}(y)} + \underbrace{ln(1-\phi)}_{-a(\boldsymbol{\eta})}\right) \end{aligned}$$

con lo que se deriva el parámetro natural

$$\boldsymbol{\eta} = \eta = ln\left(\frac{\phi}{1-\phi}\right)$$

Caso de distribución de Bernoulli

- Despejando ϕ en términos de η resulta en:

$$\phi = \frac{e^{\eta}}{1 + e^{\eta}} = \frac{1}{1 + e^{-\eta}}$$

  es la función logística que usamos anteriormente

- Además

$$\begin{aligned} T(y) &= y \\ a(\eta) &= -ln(1 - \phi) = ln(1 + e^{\eta}) \\ b(y) &= 1 \end{aligned}$$

Caso de distribución normal

- Para la interpretación probabilística de regresión lineal, la variaza σ^2 no tuvo efecto
- Vamos a simplificar caso asumiendo varianza $\sigma^2 = 1$
- Para la distribución gaussiana tenemos:

$$\begin{aligned} p(y;\mu) &= \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}(y-\mu)^2\right) \\ &= \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}y^2\right) exp\left(\mu y - \frac{1}{2}\mu^2\right) \end{aligned}$$

familia gaussiana $p(y;\boldsymbol{\eta}) = b(y) exp(\boldsymbol{\eta}^T \boldsymbol{T}(y) - a(\boldsymbol{\eta}))$

Caso de distribución normal

$$
\begin{aligned}
\eta &= \mu \\
\boldsymbol{T}(y) &= y \\
a(\boldsymbol{\eta}) &= \frac{\mu^2}{2} = \frac{\eta^2}{2} \\
b(y) &= \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}y^2\right)
\end{aligned}
$$

Test

Test

Test

Test

Test

Test

Test

Test

Test