Algoritmos de regresión

Curso de aprendizaje automático para el INE

Alberto Torres Barrán

2019-04-01

Regresión lineal

La salida y es continua, $y \in \mathbb{R}$

- Regresión lineal (*MSE* o *RSS*)

$$\min_{w} \; ||y - \mathrm{X}w||_2^2$$

- Regresión *ridge* (*MSE* + regularización l_2):

$$\min_{w} \; ||y - \mathrm{X}w||_2^2 + ||w||_2^2$$

Regresión logística

La salida y es discreta, $y \in \{0, 1\}$

- Regresión logística (*log-loss*)

$$\min_w \; -(y^T \log[\sigma(\mathrm{X}w)] + (1-y)^T \log[1-\sigma(\mathrm{X}w)])$$

- ¿Regresión logística ridge?

GLMs

- Generalización de la regresión lineal que permite distribuciones de errores distintas de la distribución normal.
- Componentes:
  - Distribución de Y_i con media μ_i
  - Predictor lineal,

    $$g(\mu_i) = w^T x_i$$

    donde $g(\cdot)$ es la función de media
- La función de media proporciona la relación entre la media de la distribución y el predictor lineal
- El inverso de la función de media, $g^{-1}(\cdot)$ se conoce con el nombre de **función de enlace**

Ejemplo: distribución binomial

- La regresión logística es un caso particular de GLM donde la distribución de Y es la binomial
- La función de media es la logística,

$$\mu = g^{-1}(w^T x_i) = \frac{1}{1 + \exp(-w^T x_i)}$$

- La función de enlace es la inversa de la anterior,

$$w^T x_i = g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$$

- Para cada distribución, hay una función de enlace "canónica" que es la que se usa habitualmente

Ejemplo: distribución de Poisson

- Esta distribución está indicada cuando queremos modelizar una variable de salida entera y no real (por ej. conteos)
- Función de media

$$\mu = \exp(w^T x_i)$$

- Función de enlace

$$w^T x_i = \ln(\mu)$$

- Otras distribuciones posibles son la Gamma, Exponencial, Multinomial, etc.

GLMs en R

- La función para ajustar modelos lineales generalizados es `glm()`
- Tiene los mismos argumentos principales que `lm()`, pero además tenemos que especificar la distribución de la variables dependiente con el parámetro `family`
- Por defecto se usa la función de enlace ""canónica", pero esto se puede modificar (ver ayuda)
- Implementa el algoritmo IRLS (Newton-Raphson), que se puede generalizar para cualquier GLM donde la distribución pertenece a la familia exponencial

Ejemplo: regresión logística

```
fit <- glm(Species ~ Petal.Length, data=iris, family=binomial)
```