Modelos lineales generalizados

Reconocimiento de Patrones (2024-2)

Julio Waissman

¿Que vamos a ver?

- Regresión lineal, el caso más simple
- Regresión logística, otro caso
- Vamos generalizando la idea
- La familia exponencial de distribuciones
- Modelos lineales generalizados

Regresión lineal

Asumimos que:

$$
\begin{aligned}
\mathcal{D} &= \{(x^{(1)}, y^{(1)}), \ldots, (x^{(M)}, y^{(M)})\} \\
y &= h(x; w, b) + e, \quad x, w \in \mathbb{R}^n, b \in \mathbb{R} \\
y &= x^T w + b + e, \quad e \sim \mathcal{N}(0, \sigma) \text{ i.i.e.}
\end{aligned}
$$

por lo que:

$$y \sim \mathcal{N}(x^T w + b, \sigma) \text{ i.i.d.}$$

y

$$\hat{y} = E[y|x; w, b] = x^T w + b$$

La receta secreta está en i.i.d.

Tenemos $\mathcal{D}$, un conjunto de datos i.i.d. de la variable aleatoria y.

$$\Pr(y^{(1)}, \ldots, y^{(M)} | x^{(1)}, \ldots, x^{(M)}; w, b) = \prod_{i=1}^{M} \Pr(y^{(i)} | x^{(i)}; w, b)$$

Los mejores valores de w y b son tales que:

$$w^*, b^* = \arg\max_{w,b} \Pr(y^{(1)}, \ldots, y^{(M)} | x^{(1)}, \ldots, x^{(M)}; w, b)$$

Máximo de verosimilitud

$$w^*, b^* = \arg\max_{w,b} \prod_{i=1}^{M} \Pr(y^{(i)}|x^{(i)}; w, b)$$

lo que nos lleva a maximizar el logaritmo de la verosimilitud:

$$\begin{aligned} w^*, b^* &= \arg\max_{w,b} \sum_{i=1}^{M} \log \Pr(y^{(i)}|x^{(i)}; w, b) \\ &= \arg\max_{w,b} \sum_{i=1}^{M} \log \left| \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^{(i)} - \hat{y}^{(i)})^2}{2\sigma^2} \right) \right| \end{aligned}$$

Máximo de verosimilitud

$$
\begin{aligned}
w^*, b^* &= \arg\max_{w,b} \sum_{i=1}^{M} \log \left| \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^{(i)} - \hat{y}^{(i)})^2}{2\sigma^2} \right) \right| \\
&= \arg\max_{w,b} K \sum_{i=1}^{M} -\frac{(y^{(i)} - \hat{y}^{(i)})^2}{2\sigma^2} \\
&= \arg\min_{w,b} \frac{K}{2\sigma^2} \sum_{i=1}^{M} (y^{(i)} - \hat{y}^{(i)})^2 \\
&= \arg\min_{w,b} \frac{1}{2M} \sum_{i=1}^{M} (y^{(i)} - \hat{y}^{(i)})^2
\end{aligned}
$$

que es la función de costo de la regresión lineal bajo el criterio de mínimos cuadrados (MSE).

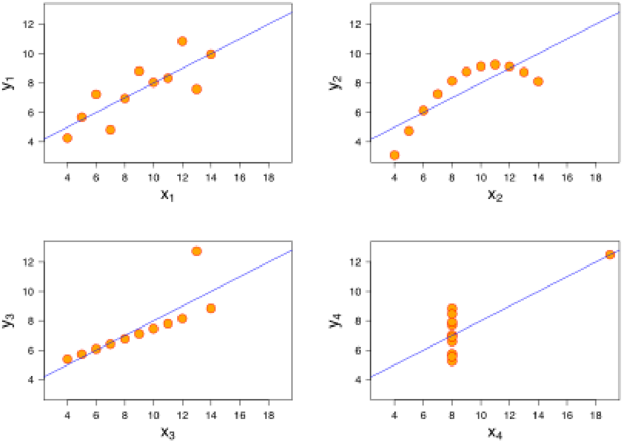
Regresión lineal

- Se asume que $y = x^T w + b + e$, con $e \sim \mathcal{N}(0, \sigma)$
- σ constante en todo el dominio Creíble?
- Se puede verificar si los residuales son normales
- ¿Y si no se cumplen las hipótesis?

> *Todos los modelos son incorrectos, pero algunos son útiles*
>
> George Box

Hay que tener cuidado

Anscombe's quartet



Ahora la regresión logística

Asumimos que:

$$
\begin{aligned}
\mathcal{D} &= \{(x^{(1)}, y^{(1)}), \ldots, (x^{(M)}, y^{(M)})\} \\
y &= h(x; w, b) + e, \quad x, w \in \mathbb{R}^n, b \in \mathbb{R} \\
y &= \sigma(x^T w + b) + e
\end{aligned}
$$

y asumimos que:

$$
y \sim \text{Bernoulli}(\rho) \text{ i.i.d., donde } \rho = \sigma(x^T w + b)
$$

y

$$
\hat{y} = E[y|x; w, b] = \rho = \sigma(x^T w + b)
$$

La receta secreta está en i.i.d.

Tenemos $\mathcal{D}$, un conjunto de datos i.i.d. de la variable aleatoria y.

$$\Pr(y^{(1)}, \ldots, y^{(M)} | x^{(1)}, \ldots, x^{(M)}; w, b) = \prod_{i=1}^{M} \Pr(y^{(i)} | x^{(i)}; w, b)$$

Los mejores valores de w y b son tales que:

$$w^*, b^* = \arg\max_{w,b} \Pr(y^{(1)}, \ldots, y^{(M)} | x^{(1)}, \ldots, x^{(M)}; w, b)$$

Máximo de verosimilitud

$$w^*, b^* = \arg\max_{w,b} \prod_{i=1}^{M} \Pr(y^{(i)}|x^{(i)}; w, b)$$

lo que nos lleva a maximizar el logaritmo de la verosimilitud:

$$\begin{aligned} w^*, b^* &= \arg\max_{w,b} \sum_{i=1}^{M} \log \Pr(y^{(i)}|x^{(i)}; w, b) \\ &= \arg\max_{w,b} \sum_{i=1}^{M} \log \left| \rho^{y^{(i)}} (1-\rho)^{1-y^{(i)}} \right| \end{aligned}$$

Máximo de verosimilitud

$$
\begin{aligned}
w^*, b^* &= \arg\max_{w,b} \sum_{i=1}^{M} \log \left| \rho^{y^{(i)}} (1-\rho)^{1-y^{(i)}} \right| \\
&= \arg\max_{w,b} \sum_{i=1}^{M} y^{(i)} \log |\rho| + (1-y^{(i)}) \log |1-\rho| \\
&= \arg\min_{w,b} \sum_{i=1}^{M} -y^{(i)} \log |\hat{y}^{(i)}| - (1-y^{(i)}) \log |1-\hat{y}^{(i)}|
\end{aligned}
$$

que es la función de costo de la regresión logística de mínimo de entropía.

Aquí hay un patrón...

$$\mathcal{D} = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(M)}, y^{(M)})\}$$

y asumimos que:

$$\Pr[y|x; w, b] = f(x^T w + b)$$

y la estimación la hacemos encontrando:

$$\hat{y} = E[y|x; w, b]$$

Familia exponencial de distribuciones

Una familia de distribuciones es exponencial si la función de densidad de probabilidad (o masa) se puede escribir como:

$$\Pr[y;\eta] = f(y;\eta) = h(y)\exp(-A(\eta))\exp(\eta^T T(y))$$

donde:

- $h(y)$ es una función no negativa (medida de base)
- $A(\eta)$ es una función (*función de partición*)
- $\eta = (\eta_1, \ldots, \eta_n)$ son los parámetros naturales
- $T(x) = (T_1(x), \ldots, T_n(x))$ son las estadísticas suficientes

La familia exponencial es la familia de distribuciones más importante en estadística

La función de enlace (*link function*)

- $\eta = (\eta_1, \ldots, \eta_n)$ son los parámetros naturales
- $\theta = (\theta_1, \ldots, \theta_n)$ son los parámetros de la distribución
- Existe una función ϕ tal que $\eta = \phi(\theta)$

$$\eta = (\phi_1(\theta), \ldots, \phi_n(\theta))$$

- ϕ es la función de enlace
- ϕ^{-1} es la función de enlace inversa

Ejemplo simple de la familia exponencial

Vamos a ver que pasa con $\mathcal{N}(\mu, \sigma^2)$, σ fija

$$
\begin{aligned}
\Pr[y; \mu] &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2 - 2y\mu + \mu^2}{2\sigma^2}\right) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2}{2\sigma^2}\right) \exp\left(\frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) \\
&= h(y) \exp(-A(\eta)) \exp(\eta^T T(y))
\end{aligned}
$$

continua ejemplo

Para este modelo, tenemos que:

- $h(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2}{2\sigma^2}\right)$
- $\theta = \mu = \sigma^2\eta$
- $\eta = \frac{\mu}{\sigma^2}$
- $A(\mu) = \frac{\mu^2}{2\sigma^2} = A(\eta) = \frac{\sigma^2\eta^2}{2}$
- $T(y) = y$

Modelo lineal generalizado

Tenemos de evidencia:

$$\mathcal{D} = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(M)}, y^{(M)})\}$$

y asumimos una distribución de la familia exponencial:

$$\Pr[y|x; w, b] = f(y; \eta) = h(y) \exp(-A(\eta)) \exp(\eta^T T(y))$$

donde:

$$\eta = x^T w + b$$

y la estimación la hacemos encontrando:

$$\hat{y} = E[y|x; w, b]$$

Retomando la regresión logística

$$
\begin{aligned}
\Pr[y|x; w, b] &= \mathrm{Bernoulli}(\rho) \\
&= \rho^y (1-\rho)^{1-y} \\
&= \exp(y \log \rho + (1-y) \log(1-\rho)) \\
&= \exp(y(\log \rho - \log(1-\rho)) + \log(1-\rho)) \\
&= \exp(\log(1-\rho)) \exp(y \log \frac{\rho}{1-\rho}) \\
&= h(y) \exp(-A(\eta)) \exp(\eta^T T(y))
\end{aligned}
$$

Seguimos con la regresión logística

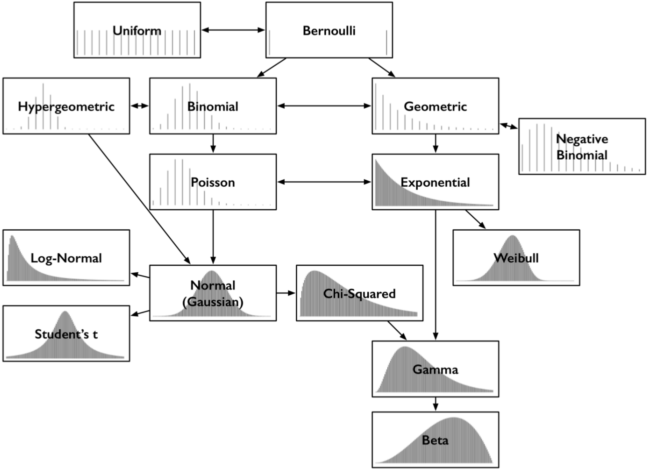
- $h(y) = 1$
- $A(\eta) = -\log(1-\rho)$
- $\eta = \log \frac{\rho}{1-\rho}$
- $\theta = \rho = \frac{1}{1+e^{-\eta}}$
- $T(y) = y$

Y sabemos que:

- $E[y|x; w, b] = \rho$
- $\eta = x^T w + b$

Por lo que:

$$\hat{y} = E[y|x; w, b] = \rho = \frac{1}{1 + \exp(-(x^T w + b))}$$



Algunas de ellas se pueden consultar en Wikipedia.

Ejercicio

Si asumimos que tenemos $y \in \{p_1, \ldots, pk\}$ y consideramos un modelo lineal generalizado basado en una distribución categórica:

- ¿Cuales son los parámetros θ?
- ¿Cuales son los parámetros naturales η?
- ¿Cual es la función de enlace y de enlace inversa?
- ¿Cuál es el vector de estadísticas suficientes $T(y)$?
- ¿Cual es el valor de $\hat{y}$?
- Deriva la función de costo usando máxima verosimilitud

¿Y como los usamos en python?

- Para regresión lineal, logística y categórica (softmax) es mejor usar los modelos de Scikit-Learn específicos.
- Scikit-Learn tiene una implementación de de modelos lineales generalizados, aunque con pocas distribuciones (solo las distribuciones de la familia de Tweedie).
- Statsmodels tiene una implementación más completa de modelos lineales generalizados. Aunque es más compleja de usar, trata de parecerse a R.
- R no tiene competencia en este rubro. Si lo tuyo son los modelos estadísticos como los GLM con distribuciones extrañas, R es la solución.