Regresión lineal multivariada

Datos con n variables m records

dimensiones

Dataframe

	$\vec{y}$	x_1	x_2	...	x_n
1	y_1^1	x_1^1	x_2^1		x_n^1
⋮					
m	y_1^m	x_1^m	x_2^m	...	x_n^m

records

	y	X_1	X_2	...	X_d
1	$y^{(1)}$		$x_2^{(1)}$		$x^{(1)}$
⋮					
n	$y^{(n)}$		$x_2^{(n)}$		

$(x^{(n)})^T$

En notación matricial

target

$$\vec{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}_{n\times 1} \qquad X = \begin{bmatrix} — (x^{(1)})^T — \\ \vdots \\ — (x^{(n)})^T — \end{bmatrix}_{n \times d+1}$$

fila datos entrenamiento

$$x^{(i)T} = (x_0^{(i)}, x_1^{(i)}, \ldots x_d^{(i)})_{1\times d+1}$$

$$\vec{\theta} = \begin{bmatrix} \theta_0 \\ \theta_d \end{bmatrix}_{d+1 \times 1}$$

Modelo $h_\theta(x^{(i)}) = (x^{(i)})^T \theta$

$$= \theta^T x^{(i)}$$

ej $(\underbrace{x_0^{(i)}}_{1}, \underbrace{x_1^{(i)} x_d^{(i)}}) \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_d \end{pmatrix}$

$(\theta_0, \ldots \theta_d) \begin{pmatrix} x_0^{(i)} = 1 \\ \vdots \\ x_d^{(i)} \end{pmatrix}$

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{n} (h_\theta(x^{(i)}) - y^{(i)})^2$$

$$= \frac{1}{2} \sum_{i=1}^{m} (x^{(i)T}\theta - y^{(i)})^2$$

$$= \frac{1}{2} (X\vec{\theta} - \vec{y})^T (X\vec{\theta} - \vec{y})$$

* usa $a^Tb = b^Ta$

$$= \frac{1}{2} \left((X\vec{\theta})^T X\vec{\theta} - (X\vec{\theta})^T\vec{y} - \vec{y}^T X\vec{\theta} + \vec{y}^T y \right)$$

se reescribe $J(\theta) = \frac{1}{2} \left(\vec{\theta}^T X^T X \vec{\theta} - \vec{y}^T (X\vec{\theta}) - \vec{y}^T X\vec{\theta} + \vec{y}^T y \right)$

$$= \frac{1}{2} \left(\vec{\theta}^T (X^T X) \vec{\theta} - 2\vec{y}^T (X\vec{\theta}) + \vec{y}^T y \right)$$

$$= \frac{1}{2} ($$

Minimizamos J con respecto a Θ con

$$\frac{\partial J(\Theta)}{\partial \Theta_j} = 0 \qquad M^T = M$$

termino $y^T y$ no contiene Θ_j

$$\frac{\partial J(\Theta)}{\partial \Theta_j} = \frac{1}{2}\left(\frac{\partial}{\partial \Theta_j}\left(\vec{\Theta}^T \underbrace{(X^T X)}_{M}\vec{\Theta} - 2\underbrace{\vec{y}^T (X}_{N}\vec{\Theta})\right)\right)$$

expresando en coordenadas

$$\frac{\partial}{\partial \Theta_j}\left\{\sum_{ik} \Theta_i M_{ik} \Theta_k - 2\sum_{e} N_e \Theta_e\right\} = \frac{\partial}{\partial \Theta_j}\left\{\sum_{k} \Theta_j M_{jk} \Theta_k - 2 N_j \Theta_j\right\}$$

$$= \frac{\partial}{\partial \Theta_j}\left\{\sum_{k\neq j} \Theta_j M_{jk} \Theta_k + \sum_{i\neq j} \Theta_i M_{ij} \Theta_j + M_{jj}\Theta_j^2 - 2 N_j \Theta_j\right\}$$

$$= \sum_{k\neq j} M_{jk}\Theta_k + 2 M_{jj}\Theta_j + \sum_{i\neq j}\Theta_i \underbrace{M_{ij}}_{M_{ji}} - 2N_j$$

$$= 2\sum_{k\neq j} M_{jk}\Theta_k + 2M_{jj}\Theta_j - 2N_j = 2\left(M\vec{\Theta} - N\right)_j$$

$$= 2\left(X^T X\vec{\Theta} - y^T X\right)_j \iff \frac{\partial}{\partial \Theta_j} J = \left(X^T X\vec{\Theta} - y^T X\right)_j$$

En forma de vector $\quad \nabla_\Theta J = X^T X\vec{\Theta} - y^T X$

Igualando a cero $\quad X^T X\vec{\Theta} = y^T X = X^T y$

Despejando $\quad \vec{\Theta} = \underbrace{(X^T X)^{-1} X^T}_{X^+} y$

X^+ Moore-Penrose inverse

Dados datos X y $\vec{y}$ por multiplicación matricial obtenemos Θ parametros que minimizan coste de la regresión lineal ... sin iteraciones

Punto de vista probabilístico sobre ML

Datos tienen incertidumbre inherente x ejemplo
1) Datos aleatorios (quantum) o errores aleatorios (clásico)
2) Variables no observadas que influencian el mecanismo
3) Discretización en la toma de datos

Utilizamos conceptos de proba y estadística v.a

- Independencia $P(X=x, Y=y) = P(X=x)\,P(Y=y)$
- Def proba condicional $P(Y=y \mid X=x) = \dfrac{P(Y=y, X=x)}{P(X=x)}$
- Regla de producto
$$P(x^{(1)}, \dots, x^{(n)}) = p(x^{(1)}) \prod_{i=2}^{n} p(x^{(i)} \mid x^{(1)}, \dots, x^{(i-1)})$$
- Regla de Bayes
$$p(x \mid y) = \frac{p(y \mid x)}{p(y)}\, p(x)$$
$p(x)$ ← prior knowledge w/o evidence

$p(x \mid y)$ ← updated knowledge

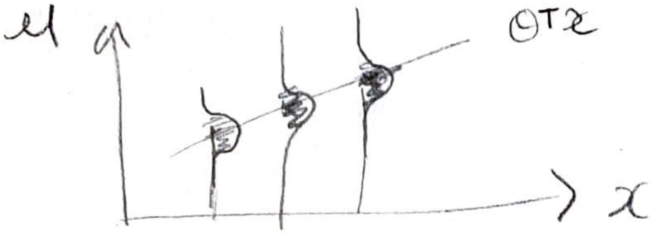
- Regla de suma $p(y) = \sum_x p(y \mid x)\, p(x)$

Problema ML regresión

$$\begin{cases} y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)} \\ \epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2) \end{cases} \circledast$$

Hay una relación lineal y la incertidumbre contenida en $\epsilon^{(i)}$

$\circledast$ quiere decir que $p(\epsilon^{(i)}) = \dfrac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(\epsilon^{(i)})^2}{2\sigma^2}}$



en cada medida todos los valores de y son posibles siguiendo distribución de proba dado θ como parámetro

$$p(y^{(i)} \mid x^{(i)}; \theta) \qquad y^{(i)} \mid x^{(i)}; \theta \sim \mathcal{N}(\theta^T x^{(i)}, \sigma^2)$$

verosimilitud

Función de θ := Likelihood function

$\mathcal{L}(\theta) = p(\vec{y} \mid X; \theta)$ para escoger los mejores parámetros (dan máximo de $\mathcal{L}$) por ind de x, y, ε

$$\mathcal{L}(\theta) = p(\vec{y} \mid X; \theta) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta)$$

$$= \prod_{i=1}^{n} \mathcal{N}(y_i \mid x_i^T \theta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}}$$

Buscamos $\theta_{MLE} = \arg\max_\theta p(\vec{y} \mid X; \theta)$

$$= \arg\max_\theta \log p(\vec{y} \mid X; \theta)$$

función creciente

log prod → sum logs

$$= \arg\max_\theta \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}}$$

$$= \arg\max_\theta \left\{ \underbrace{n \log \frac{1}{\sqrt{2\pi}\sigma}}_{\text{cte}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y^{(i)} - \theta^T x^{(i)})^2 \right\}$$

$$\theta_{MLE} = \arg\max_\theta \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y^{(i)} - \theta^T x^{(i)})^2 \right\}$$

$$= \arg\min_\theta J(\theta)$$

← distancia cuadrática

Regresión lineal = MLE con ruido gaussiano