3. Linear Regression

$\{(y_n, x_n)\}_{n=1}^N \quad x_n \in \mathbb{R}^d, \; y_n \in \mathbb{R}$

What we train?

$$\sum_{n=1}^N (y_n - \langle x_n, \beta \rangle)^2 \cdot w_n \to \min_\beta$$

2. Statistical model? ← with the same optimal

$$r_n | x_n \sim \mathcal{N}(r_n | \mu(x_n), \sigma_n^2)$$

$$\mu(x_n) = \mathbb{E}\, r_n | x_n = \langle x_n, \beta \rangle$$

Maximum Likelihood Estimation (MLE)

$$\sum_{n=1}^N \log \mathcal{N}(y_n | \langle x_n, \beta \rangle, \sigma_n^2) \to \max_\beta$$

$$\sum_{n=1}^N -\frac{1}{2} \frac{1}{\sigma_n^2} (y_n - \langle x_n, \beta \rangle)^2 \to \max_\beta$$

$$\sum_{n=1}^N \frac{1}{\sigma_n^2} (y_n - \langle x_n, \beta \rangle)^2 \to \min_\beta$$

$r_n \in \mathbb{R}$ → $\mathbb{R}$

link function

$\mathcal{N}(y_n | \mu_n, \sigma)$

- $r_n \in \mathbb{N}_+$
- $r_n \in \{0, 1\}$

$$g(\mathbb{E}\, r_n | x_n) = x_n^T \beta \in \mathbb{R}$$

2. Example

y_n is expected number of deaths on the day t_n

- model $\quad y_n = \gamma \exp(\delta t_n)$

$$\underbrace{\log y_n}_{\in \mathbb{R}} = \underbrace{\log \gamma + \delta t_n}_{\text{linear model}}$$

3. Exponential Family of Distributions

$$p_\theta(y) = \exp(\langle \theta, T(y) \rangle - B(\theta))\, h(y)$$

$$B(\theta) = \log \int \exp(\langle \theta, T(y) \rangle)\, h(y)\, dy$$

Because $\int p_\theta(y)\, dy = 1$.

$$\log p_\theta(y) = \langle \theta, T(y) \rangle - B(\theta) + \log h(y)$$

$$\nabla_\theta \log p_\theta(y) = T(y) - \nabla_\theta B(\theta)$$

$$\nabla_\theta B(\theta) = \nabla_\theta \log \int \exp(\langle \theta, T(y) \rangle) h(y)\, dy =$$

$$= \frac{1}{\int \exp(\langle \theta, T(y) \rangle) h(y)\, dy} \int \exp(\langle \theta, T(y) \rangle)\, T(y)\, h(y)\, dy = \mathbb{E}\, T(y).$$

$p_\theta(y)$

$$\nabla_\theta \log p_\theta(y) = T(y) - \mathbb{E}\, T(y)$$

$$\mathbb{E}\, \nabla_\theta \log p_\theta(y) = \mathbb{E}\, T(y) - \mathbb{E}\, T(y) = 0.$$

$$\sum_{n=1}^{N} \log p_\theta(y_n) \to \max_\theta; \quad \sum_{n=1}^{N} \langle \theta, T(y_n) \rangle - N \underbrace{B(\theta)}_{\text{find integral}}$$

Take the gradient

$$\sum_{n=1}^{N} T(y_n) - N \nabla_\theta B(\theta) = 0$$ ← Because problem is convex

$$\frac{1}{N}\sum_{n=1}^{N} T(y_n) = \nabla_\theta B(\theta) = \mathbb{E}\, T(y_n)$$

empirical average — true average

$$g(\mathbb{E}\, T(X_n)) = x_n^T \beta \quad ? \qquad \in \mathbb{R}$$

$$\nabla_\theta B(\theta) = \mathbb{E}\, T(y_n)$$

$$x_n^T \beta = \theta = (\nabla_\theta B)^{-1}(\mathbb{E}\, T(y_n)).$$

- we solve for 1-D,
- we could restrict our desire $T(y) = y$ (mean statistics only)

"mean" — "variance"

$$p_\theta(y) = \exp\left\{ \frac{y \cdot \theta - b(\theta)}{\varphi} + c(y, \varphi) \right\}$$

$$\log p_\theta(y) = \left\{ y \cdot \theta - b(\theta) \right\} \frac{1}{\varphi} + c(y, \varphi)$$

$$\nabla_\theta : \left[y - \nabla_\theta b(\theta) \right] \frac{1}{\varphi}$$

$$\mathbb{E}\, y = \nabla_\theta b(\theta), \quad \theta = (\nabla_\theta b)^{-1}(\mathbb{E}\, y).$$

$\theta = x^T \beta$

Poisson Regression

$y_n \in \mathbb{N}_+$

$$p(y_n) = \frac{\mu^{y_n}}{y_n!} \exp(-\mu)$$

$$\log p(y_n) = y_n \cdot \log \mu - \mu - \log(y_n!)$$

$\theta = \log \mu, \quad \mu = \exp(\theta)$

$(x_n, y_n) \quad y_n \in \mathbb{N}_+ \quad$ MLE

$$\sum_n \log p(y_n) \to \max$$

$$\sum_n y_n \cdot \log \mu - \mu \to \max$$

$$\sum_{n=1}^{N} y_n \cdot \theta - \exp(\theta) \to \max_\theta$$

$\theta = x_n^T \beta \quad$ $$\sum_{n=1}^{N} y_n \cdot (x_n^T \beta) - \exp(x_n^T \beta) \to \max_\beta$$

∇_β: $$\sum_{n=1}^{N} y_n \cdot x_n - \exp(x_n^T \beta) x_n$$

∇_β: $$\sum_{n=1}^{N} \underbrace{\{y_n - \exp(x_n^T \beta)\}}_{= \theta} x_n$$

$y_n - \mathbb{E}\, y_n | x_n$

↓ features

$$\beta^{k+1} = \beta^k + \alpha \nabla_\beta = \beta^k + \alpha \sum_{n=1}^{N} (y_n - \mathbb{E}\, y_n(x_n)) x_n$$

Classification Problem

Bernoulli Distribution

$y_n \in \{0, 1\}$

$$p(y_n) = p^{y}(1-p)^{(1-y)}, \quad p \in (0,1)$$

p: probability of success $P(y=1)$

$$\log p(y_n) = y_n \log p + (1-y_n)\log(1-p) =$$

$$= y_n \cdot \log \frac{p}{1-p} + \log(1-p)$$

$\log \frac{p}{1-p}$: θ ; $\log(1-p)$: ?

$$\theta = \log \frac{p}{1-p}, \quad \exp(\theta) = \frac{p}{1-p} = \frac{1}{\frac{1}{p} - 1}$$

$$\exp(-\theta) + 1 = \frac{1}{p}$$

$$p = \frac{1}{\exp(-\theta)+1}$$

Sigmoid function

$x = 1 \times 5 \quad x = 2$

$0 \times 5, \ 1 \times 10$

$P(y=1) = \frac{10}{15} = \frac{2}{3}.$

$\left(\frac{2}{3}\right) = \frac{1}{\exp(-\theta)+1}$

$$y \cdot \theta - \log(1 + \exp(\theta)) \quad \partial\theta$$

$$\mathbb{E}\, y = \frac{\exp(\theta)}{1+\exp(\theta)} = \frac{1}{\exp(-\theta)+1}$$

$$\log \frac{\mathbb{E}y}{1-\mathbb{E}y} = \theta$$

logit function

inverse of sigmoid

$\sigma^{-1}(\frac{2}{3}) = \theta$.

$\mathbb{E}y|x=1 = 0$

$\mathbb{E}y|x=2 = 1$

$$\log \frac{\mathbb{E}y}{1-\mathbb{E}y} = x_n^T \beta$$

likelihood of Bernulli distribution.

MLE

$$\sum_{n=1}^{N} y_n \theta_n - \log(1+\exp(\theta_n))$$

$\theta_n = x_n^T \beta$

$$\sum_{n=1}^{N} y_n x_n^T \beta - \log(1+\exp(x_n^T \beta))$$

∂_β:

$$\sum_{n=1}^{N} y_n x_n - \frac{\exp(x_n^T \beta)}{1+\exp(x_n^T \beta)} x_n = 0$$

$$\sum_{n=1}^{N} \left(y_n - \frac{\exp(x_n^T \beta)}{1+\exp(x_n^T \beta)}\right) x_n$$

$$\frac{\exp(x_n^T \beta)}{1+\exp(x_n^T \beta)} = \frac{1}{1+\exp(-x_n^T \beta)}$$

$(y_n - \mathbb{E}y_n(x_n))\, x_n$

y_n: observ.

$\mathbb{E}y_n(x_n)$: conditional expe.

$\mathbb{E}\, y_n | x_n$

$p(y) = \mathcal{N}(y | \mu, \sigma^2)$ $\quad \mathbb{E}\, y = \mu$

$p(y|x) = \mathcal{N}(\tilde{y} | x^T\beta, \sigma^2)$ $\quad \mathbb{E}\, \tilde{y} = x^T\beta$

$\mathbb{E}\, y|x = x^T\beta$

$\mathbb{E}\, y = \mathbb{E}_x \mathbb{E}\, y|x = \mathbb{E}_x x^T\beta = (\mathbb{E}_x x)^T\beta.$