تخمین پارامترهای مدل

$$P(D|\theta) = P(\{A_i\}_{i=1}^{N}|\theta) = \prod_{i=1}^{N} P(A_i|\theta)$$

$$= \prod_{i=1}^{N}\prod_{c=1}^{4} \theta_i^{a_{ic}} \quad , \quad A_i \rightarrow \text{one-hot}$$ بردار کد شده

برای به دست آوردن $\hat{\theta}_{mle}$ ، $NLL(\theta)$ را تشکیل می‌دهیم:

$$NLL(\theta) = -\log \prod_{i=1}^{N}\prod_{c=1}^{4} \theta_c^{a_{ic}} = -\sum_{i=1}^{N}\sum_{c=1}^{4} a_{ic}\log\theta_c$$

حال مسئله بهینه‌سازی زیر را داریم:

$$\min_{\theta} \quad NLL(\theta)$$
$$\text{s.t.} \quad \sum_{c=1}^{4}\theta_c = 1$$

$$L = -\sum_{i=1}^{N}\sum_{c=1}^{4} a_{ic}\log\theta_c + \lambda\left(\sum_{c=1}^{4}\theta_c - 1\right)$$

$$\frac{\partial L}{\partial \theta_k} = -\sum_{i=1}^{N}\sum_{c=1}^{4} a_{ic}\overbrace{\frac{\partial \log\theta_c}{\partial\theta_k}}^{\frac{1}{\theta_k}\delta_{kc}} + \lambda\left(\sum_{c=1}^{4}\overbrace{\frac{\partial}{\partial\theta_k}\theta_c}^{\delta_{kc}} - 1\right)$$

$$= -\sum_{i=1}^{N}\frac{a_{ik}}{\theta_k} + \lambda = 0 \Rightarrow \theta_k = \frac{\sum_{i=1}^{N} a_{ik}}{\lambda}$$

$$\frac{\partial L}{\partial \lambda} = 0 \Rightarrow \sum_{c=1}^{4}\theta_c = 1 \Rightarrow \sum_{c=1}^{4}\frac{\sum_{i=1}^{N} a_{ic}}{\lambda} = 1 \Rightarrow \frac{N}{\lambda} = 1 \Rightarrow \lambda = N$$

$$\Rightarrow \theta_k = \frac{\sum_{i=1}^{N} a_{ik}}{N} \quad , \quad k = 1, 2, \ldots, 4$$

ب:

$$P(\theta|D) = \frac{P(D|\theta)\,P(\theta)}{P(D)}$$

$$= \frac{\prod_{i=1}^{N}\prod_{c=1}^{4}\theta_i^{a_{ic}}\;\frac{1}{B(\alpha)}\prod_{j=1}^{4}\theta_j^{\alpha_j-1}}{P(D)}$$

$$= \frac{\frac{1}{B(\alpha)}\prod_{c=1}^{4}\theta_i^{N_i}\;\prod_{j=1}^{4}\theta_j^{\alpha_j-1}}{P(D)}$$

$$= \frac{1}{B(\alpha)P(D)}\prod_{c=1}^{4}\theta_c^{(N_c+\alpha_c-1)} \sim Dir(\alpha+m)$$

$$, \quad m = \begin{bmatrix} N_1 \\ N_2 \\ \vdots \\ N_4 \end{bmatrix}$$

براساس مشخصات توزیع دیریچله می‌دانیم:

$$Mode[X_i] = \frac{\alpha_i - 1}{\alpha_0 - k}$$

در توزیع پسین داریم:

$$\hat{\theta}_{map} = \begin{bmatrix} \frac{\alpha_1+N_1-1}{\alpha_0+N-4} \\ \vdots \\ \frac{\alpha_4-N_4-1}{\alpha_0+N-4} \end{bmatrix} \quad , \quad \alpha_0 = \sum_{i=1}^{4}\alpha_i$$

ج: یک توزیع پیشین که $\hat{\theta}_{map}$ خواسته شده را ایجاد می‌کند برابر است با:

$$\alpha = \begin{bmatrix} 1 \\ \checkmark \\ \checkmark \\ \checkmark \\ \checkmark \\ \checkmark \end{bmatrix} \Rightarrow \hat{\theta}_{map} = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$

تئوری تصمیم‌گیری

الف)

$$E_{p(y|x)}\left[\ell(a=-1,y)\right] = \sum_{y_i \in \{\pm 1\}} p(y_i|x)\,\ell(a=-1,y_i)$$

$$= P_0 \lambda_e + (1-P_0)\,0 = P_0 \lambda_e$$

$$E_{p(y|x)}\left[\ell(a=+1,y)\right] = \sum_{y_i \in \{\pm 1\}} p(y_i|x)\,\ell(a=+1,y_i)$$

$$= P_0\, 0 + (1-P_0)\lambda_e = (1-P_0)\lambda_e$$

$$E_{p(y|x)}\left[\ell(a=0,y)\right] = \sum_{y_i \in \{\pm 1\}} p(y_i|x)\,\ell(a=0,y_i)$$

$$= P_0 \lambda_r + (1-P_0)\lambda_r = \lambda_r$$

ب)

$$a^* = \begin{cases} +1 & \begin{cases} (1-P_0)\lambda_e \leqslant P_0\lambda_e \Rightarrow P_0 \geqslant 1/2 \\ (1-P_0)\lambda_e \leqslant \lambda_r \Rightarrow P_0 \geqslant 1 - \lambda_r/\lambda_e \end{cases} \\ -1 & \begin{cases} P_0\lambda_e \leqslant (1-P_0)\lambda_e \Rightarrow P_0 \leqslant 1/2 \\ P_0\lambda_e \leqslant \lambda_r \Rightarrow P_0 \leqslant \frac{\lambda_r}{\lambda_e} \end{cases} \\ 0 & \text{otherwise} \end{cases}$$

$$\Rightarrow a^* = \begin{cases} +1 & P_0 \geqslant \max\left\{1/2,\ 1-\frac{\lambda_r}{\lambda_e}\right\} \\ -1 & P_0 \leqslant \min\left\{1/2,\ \frac{\lambda_r}{\lambda_e}\right\} \\ 0 & \text{otherwise} \end{cases}$$

مشتق‌گیری نسبت به پارامترها:

$$NLL(W) = -\sum_{c=1}^{C} y_c \log \mu_c$$

فرض کنید $W = \begin{bmatrix} \hat{w}_1^T \\ \vdots \\ \hat{w}_C^T \end{bmatrix}$. در اینصورت

$$\frac{\partial NLL(W)}{\partial \hat{w}_j} = \frac{\partial}{\partial \hat{w}_j}\left(-\sum_c y_c \log \mu_c\right) = -\sum_c y_c \underbrace{\frac{\partial}{\partial \hat{w}_j} \log \mu_c}_{= \frac{\partial}{\partial \mu_c} \log \mu_c \times \frac{\partial \mu_c}{\partial \hat{w}_j}}$$

$$= -\sum_c \frac{y_c}{\mu_c} \frac{\partial \mu_c}{\partial \hat{w}_j} \quad (1)$$

حال $\frac{\partial \mu_c}{\partial \hat{w}_j}$ را به دست می‌آوریم:

$$\frac{\partial \mu_c}{\partial \hat{w}_j} = \sum_k \frac{\partial \mu_c}{\partial a_k} \overbrace{\frac{\partial a_k}{\partial \hat{w}_j}}^{\delta_{kj} x} = \frac{\partial \mu_c}{\partial a_j} x \quad (2)$$

حال $\frac{\partial \mu_c}{\partial a_j}$ را به دست می‌آوریم:

$$\frac{\partial \mu_c}{\partial a_j} = \begin{cases} \frac{\partial}{\partial a_j} \frac{\exp(a_j)}{\sum_k \exp(a_k)} = \mu_j(1-\mu_j) = \mu_c(1-\mu_j) & j = c \\ \frac{\partial}{\partial a_j} \frac{\exp(a_c)}{\sum_k \exp(a_k)} = -\mu_c \mu_j & j \neq c \end{cases}$$

$$= \mu_c(\delta_{jc} - \mu_j) \quad (3)$$

با استفاده از ۱، ۲ و ۳:

$$\frac{\partial NLL(W)}{\partial \hat{w}_j} = -\sum_c \frac{y_c}{\mu_c} x \mu_c (\delta_{jc} - \mu_j)$$

$$= -x\left(\sum_c y_c \delta_{jc} - \sum_c y_c \mu_j\right) = x(\mu_j - y_j)$$

حال اگر رابطه را برای تمام سطرها تکرار کنیم، داریم:

$$\frac{\partial NLL(W)}{\partial W} = \begin{bmatrix} (\mu_1 - y_1)\underline{x} \\ \vdots \\ (\mu_C - y_C)x \end{bmatrix} = (\mu - y)^T x \in \mathbb{R}^{C \times D}$$

رگرسیون لاجستیک باینری :

الف)

$$P(D|w) = \prod_{n=1}^{N} Ber(y_n|\mu_n) = \prod_{n=1}^{N} \mu_n^{y_n}(1-\mu_n)^{1-y_n}$$

$$\Rightarrow -\log P(D|w) = -\sum_{n=1}^{N}\left[y_n \log\mu_n + (1-y_n)\log(1-\mu_n)\right]$$

ب) برای جداپذیر خطی بودن لازم است :

$$\begin{cases} a_n > 0 & \text{if } y_n = 1 \\ a_n < 0 & \text{if } y_n = 0 \end{cases} \equiv \begin{cases} \mu_n > 1/2 & y_n = 1 \\ \mu_n < 1/2 & y_n = 0 \end{cases}$$

در شرایط فوق تمام داده‌های پایگاه داده درست طبقه‌بندی شده، چون مرز تصمیم‌گیری در رگرسیون لاجستیک باینری، خطی است در این شرایط مسأله جداپذیر خطی خواهد بود.

ج) فرض کنید $\tilde{w}$ شرط جداپذیری خطی را دارد. حال بردار $\hat{w} = g\tilde{w}$ را در نظر بگیرید.

بردار $y_k = 1$ را در نظر بگیرید. چون $\tilde{w}$ شرط جداپذیری را دارد بنابراین $\tilde{a}_k > 0$.

در نتیجه:

$$\hat{a}_k = \hat{w}^T x = g\tilde{w}^T x = g\tilde{a}_k > \tilde{a}_k \overset{\sigma(\cdot)}{\Rightarrow} \sigma(g\tilde{a}_k) > \sigma(\tilde{a}_k)$$

$$\Rightarrow \hat{\mu}_k > \tilde{\mu}_k \Rightarrow -\left[y_n \log\tilde{\mu}_k + (1-y_n)\log(1-\tilde{\mu}_k)\right] \geq -\left[y_n \log\hat{\mu}_k + (1-y_n)\log(1-\hat{\mu}_k)\right]$$

به طور مشابه برای نمونه‌های $y_k = 0$ نیز می‌توان نشان داد رابطه فوق برقرار است پس :

$$-\sum_{n=1}^{N}\left[y_n \log\hat{\mu}_n + (1-y_n)\log(1-\hat{\mu}_n)\right] \leq -\sum_{n=1}^{N}\left[y_n \log\tilde{\mu}_n + (1-y_n)\log(1-\tilde{\mu}_n)\right]$$

$$\Rightarrow NLL(\hat{w}) \leq NLL(\tilde{w})$$