אופטימליות הגישה הבייזיאנית

<u>תזכורת:</u> עבור דאטה-סט (נתונים) D

בגישה הקלאסית מנסים למצוא את הפרמטר θ

שממזער פונקציית הפסד נתונה $\hat{\theta} = \arg\min_{\theta} L(\theta, D)$

בגישה הבייזיאנית מנסים למצוא התפלגות על פני

θ בהינתן D ← $P(\theta | D)$

<u>דוגמא-</u> עבור 1000 הטלות מטבע $D = \{X_i\} = (0, 1, 1, 0, \dots)$

נתבונן ב- 3 אלג' אפשריים

1) $\hat{\theta}_{MLE}(x) = \frac{h_0}{1000}$

2) $\hat{\theta}_2(x) = \frac{\sum_{i=1}^{10} X_i}{1000}$

3) רשת נוירונים עם 20 שכבות קונבולוציה ושכבה אחת fully-conected

כדי למדוד את הצלחת האלגוריתמים, אפשר להשתמש

ב-$BMSE(\hat{\theta})$ (Bayesian mean squared error)

$$BMSE(\hat{\theta}) = E_{\theta^{GT},D} \|\hat{\theta}(D) - \theta^{GT}\|^2$$

$$= \int_{\theta^{GT},D} P(\theta,D)\cdot\|\hat{\theta}(D) - \theta^{GT}\|^2 \, d\theta \, dD$$

<u>במילים:</u>

נחזור על זה המון פעמים ונחשב ממוצע:

1) נגריל θ^{GT} לפי $P(\theta)$

2) נגריל דוגמאות D לפי $P(D|\theta^{GT})$

3) נחשב $\|\hat{\theta}(D) - \theta^{GT}\|^2$

<u>משפט:</u> האלגוריתם $\hat{\theta}(X) = \mathbb{E}(\theta|D)$ הוא האלגוריתם האופטימלי מבחינת BMSE

נזכיר - $E(\theta|D) = \int_{\theta} \theta \cdot P(\theta|D) \, d\theta$

הוכחה-

$$BMSE = E_{\theta, D} \|\hat{\theta}(D) - \theta\|^2$$

$$= \int_{\theta, D} P(\theta, D) \cdot \|\hat{\theta}(D) - \theta\|^2 \, d\theta \, dD$$

$$= \int_{\theta, D} P(D) \cdot P(\theta | D) \cdot \|\hat{\theta}(D) - \theta\|^2 \, d\theta \, dD$$

נסמן- $C(\hat{\theta}, D)$

$$= \int_{D} P(D) \cdot \overbrace{\int_{\theta} P(\theta | D) \cdot \|\hat{\theta}(D) - \theta\|^2 \, d\theta}^{C(\hat{\theta}, D)} \, dD$$

נראה שלכל D אנחנו בוחרים באלגוריתם שממזער את ה-BMSE.
לכן זה יהיה נכון רק עבור הממוצע של ה-D

נגזור את $C(\hat{\theta}, D)$ לפי $\hat{\theta}$ כדי למצוא מינימום

$$\frac{\partial C}{\partial \hat{\theta}} = \int_{\theta} P(\theta | D) \cdot 2\left(\hat{\theta}(D) - \theta\right) d\theta$$

נשווה לאפס-

$$2 \cdot \int_{\theta} P(\theta | D) \cdot \hat{\theta}(D) \, d\theta - 2 \cdot \int_{\theta} P(\theta | D) \cdot \theta \, d\theta = 0$$

$$\underbrace{\int_{\theta} P(\theta | D) \, \hat{\theta}(D) \, d\theta}_{\hat{\theta}(D)} = \underbrace{\int_{\theta} P(\theta | D) \, \theta \, d\theta}_{E(\theta | D)}$$

נחזור לדוגמא- $\theta \sim U[0,1]$ ולכן $P(\theta)=\begin{cases}1 & \theta\in[0,1]\\ 0 & else\end{cases}$

$$P(D|\theta)=\prod_{i=1}^{1000}\begin{cases}\theta & x_i=0\\ 1-\theta & x_i=1\end{cases}=\theta^{n_0}\cdot(1-\theta)^{n_1}$$

$$P(\theta|D)=\frac{P(D|\theta)}{P(D)}\cdot P(\theta)=\frac{\theta^{n_0}(1-\theta)^{n_1}}{\int_{\theta'}\theta'^{n_0}(1-\theta')^{n_1}d\theta'}$$

ולחישוב התוחלת

$$\mathbb{E}(\theta|D)=\int_{\theta}\theta\cdot P(\theta|D)\,d\theta$$

$$=\int_{\theta}\theta\cdot\frac{\theta^{n_0}(1-\theta)^{n_1}}{\int_{\theta'}\theta'^{n_0}(1-\theta')^{n_1}d\theta'}\,d\theta$$

בקטע לא מקרי, האינטגרל הנ"ל ניתן לחישוב בעזרת התפלגות דיריכלה.

התפלגות דיריכלה- על וקטור בסימפלקס (וקטור θ שכל איבריו חיוביים וסכום האיברים 1).

ההתפלגות $\theta\sim Dirichlet(\alpha)$ מוגדרת לפי

$$P(\theta)=\frac{1}{B(\alpha)}\cdot\prod_{k}\theta_k^{\alpha_k-1}$$

כאשר $\alpha_0 := \sum_k \alpha_k$ $\quad B(\alpha) := \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\alpha_0)}$

התוחלת של θ_k היא $E(\theta_k) = \frac{\alpha_k}{\alpha_0}$

דוגמה:

1) $\alpha = (1,1,1,1)$ (התפלגות יוניפורמית מעל הסימפלקס)

$$P(\theta) = \frac{1}{B(\alpha)} \cdot \prod_k \theta_k^0 = \frac{1}{B(\alpha)} \cdot 1$$

2) $\alpha = (10^6, 10^6, \dots, 10^6)$

$$P(\theta) = \frac{1}{B(\alpha)} \cdot \prod_k \theta_k^{10^6}$$

בהתפלגות זו, אם אחת הקואורדינטות של θ שונה מ- 1 ההסתברות היא פרקטית 0.

נחזור לחישוב של $P(\theta|D)$. נסמן $\theta = \theta_0$, $(1-\theta) = \theta_1$

$$P(\theta|D) = \frac{\theta^{h_0}(1-\theta)^{h_1}}{\int_{\theta'} \theta'^{h_0}(1-\theta')^{h_1} d\theta'} = \frac{1}{B} \cdot \theta_0^{h_0} \cdot \theta_1^{h_1}$$

כלומר מצאנו ש- $\hat{\theta}|D \sim \text{Dirichlet}(h_0+1, h_1+1)$

ולפי ההתפלגות הנ"ל מצאנו $\mathbb{E}(\theta|D) = \left(\frac{h_0+1}{2}, \frac{h_1+1}{2}\right)$

הגדרות-

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

$P(\theta)$ - "Prior Probability"

$P(D|\theta)$ - "Likelihood"

$P(D)$ - "Evidence"

$P(\theta|D)$ - "Posterior Probability"

$\hat{\theta}^{MMSE}(D)=E(\theta|D)$ - אלגוריתם המחשב תוחלת מותנה

$\hat{\theta}^{MAP}(D)=\arg\max_{\theta} P(\theta|D)$ - "Maximum a posteriore"

$\hat{\theta}^{MLE}(D)=\arg\max_{\theta} P(D|\theta)$ - "Maximum Likelihood Estimator"

הגדרה- Conjugate Prior

אם $P(\theta|D)$ שייך לאותה משפחה פרמטרית כמו $P(\theta)$, אז $P(\theta)$ נקרא Conjugate Prior (פריור צמוד) של $P(D|\theta)$.

דוגמא- אם $P(\theta|D)$ התפלגות מולטינומיאלית, אז $P(\theta)$ דיריכלה היא CP של $P(D|\theta)$.