שעור 3 BML - אופטימליות

November 7, 2022

בשעור הקודם הסתכלנו על אוסף הטלות מטבע $D = \{0, 1, ...\}$ וראינו כי $\hat{\theta}^{MLE} = \frac{n_0}{N}$. אם נניח $\theta \sim \mathcal{U}[0,1]$ נקבל $\hat{\theta}^{MMSE} = E[\theta|D] = \frac{n_0+1}{N+2}$. עבור הנתונים $n_0 = 6, N = 10$ נקבל $\hat{\theta}^{MLE} = 0.6, \hat{\theta}^{MMSE} \approx 0.583$. נניח כי יש לנו גם את $\hat{\theta}^{NN}$ שנאמד ע"י רשת נוירונים.
משפט: האומד $\hat{\theta}^{MMSE}$ הוא אופטימלי במובן של BMSE.
הוכחה:

$$BMSE = E_{\theta,D}\left[\|\theta - \hat{\theta}(D)\|^2\right] = \int_{\theta,D} P(D)P(\theta|D)\|\theta - \hat{\theta}(D)\|^2 d\theta dD$$
$$= \int_D P(D)\left[\int_\theta P(\theta|D)\|\theta - \hat{\theta}(D)\|^2 d\theta\right] dD = \int_D P(D) J_D\left(\hat{\theta}\right) dD$$

כאשר $J_D\left(\hat{\theta}\right) = \int_\theta P(\theta|D)\|\theta - \hat{\theta}(D)\|^2 d\theta$. האינטגרל על D הוא בעצם מיצוע על ה-loss.
אם נמצא $\hat{\theta}(D)$ שהוא מינימלי לכל D אז גם השגיאה הממוצעת תהיה מינימלית.

$$\frac{\partial J}{\partial \hat{\theta}} = -\int_\theta P(\theta|D)\cdot 2(\theta - \hat{\theta})d\theta = 0 \to \int_\theta P(\theta|D)\theta d\theta = \int_\theta P(\theta|D)\hat{\theta} d\theta$$

נשים לב כי $\int_\theta P(\theta|D)\theta d\theta$ היא בעצם התוחלת המותנה וכן $\int_\theta P(\theta|D)d\theta = 1$ לכן נקבל:

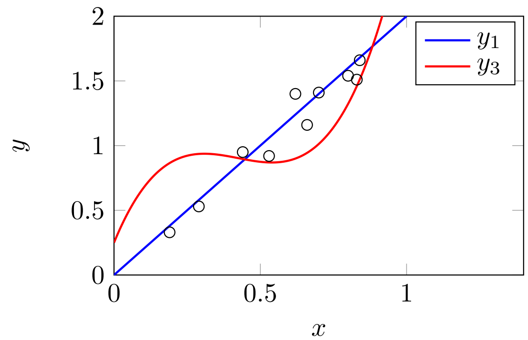
$$E[\theta|D] = \hat{\theta}\int_\theta P(\theta|D)d\theta = \hat{\theta}\cdot 1 \to \hat{\theta} = E[\theta|D]$$

מה קורה אם במקום לחשב את הפרמטר אנחנו רוצים לתת תחזית להטלה הבאה? התחזית הטובה ביותר היא $E\left[d_{N+1}|\left\{d_1, ..., d_N\right\}\right]$, כלומר $\int_\theta P(\theta|D)P(\text{"}0\text{"}|\theta)d\theta$, זו כבר דוגמא לחישוב מסובך.
הערה: לכל פונקציה לינארית של θ (למשל $\alpha = A\theta + b$) התחזית האופטימלית במובן BMSE נתונה לפי $\alpha^{MMSE} = A\theta^{MMSE} + b$.
הוכחה: $E[\alpha|D] = E[A\theta + b|D] = AE[\theta|D] + b = A\theta^{MMSE} + b$

<u>רגרסיה לינארית בייסיאנית</u>

$$y(x) = \sum_n \theta_n h_n(x)$$

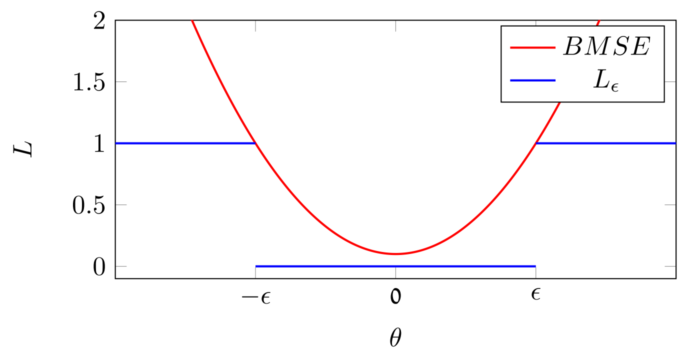
הפונקציות $h_n(x)$ הן קבועות ביחס ל-θ אבל לא בהכרח פונקציות לינאריות של x. למשל ברגרסיה פשוטה $y_1(x) = \theta_0 + \theta_1 x$ הפונקציות הן $h_0 = 1, \quad h_1 = x$. לעומת זאת, ברגרסיה פולינומיאלית ממעלה שלישית הפונקציה היא $y_3(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$ והפונקציות הן $h_0 = 1, \quad h_1 = x, \quad h_2 = x^2, \quad h_3 = x^3$.



כדי לקבל אומדים בייסיאנים נחשב את $\theta^{MMSE} = E[\theta|D]$ (ההנחה היא שהפונקציות $h_0, ..., h_n$ ידועות מראש). איך נחשב? נראה בשבוע הבא.
איך ניתן תחזית יחידה עבור דגימה עם x נתון? נשים לב כי לכל x, $y(x) = h^T(x)\theta$ ולכן $y^{MMSE}(x) = h^T(x)\theta^{MMSE}$.
שתי הנחות שלקחנו ביחס לאופטימליות של θ^{MMSE}:

1. הפריור $P(\theta)$ ידוע ונכון
2. האופטימליות היא ביחס ל-BMSE. למה דווקא שגיאה ריבועית? מסיבות היסטוריות, כנראה כי היא קמורה ונוחה לגזירה

נגדיר פונקצית הפסד חדשה בשם Hit or Miss מהצורה $L_\epsilon = \begin{cases} 0 & \|\theta - \hat{\theta}\| < \epsilon \\ 1 & o.w \end{cases}$. בחד מימד ועבור $\epsilon = 0.5$ הפונקציה נראית כמו

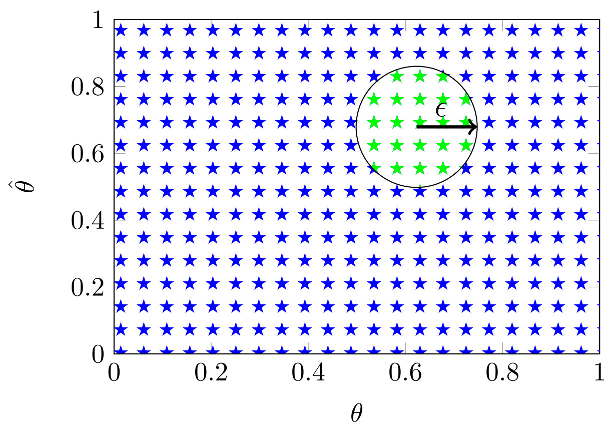


נגדיר אומד חדש - Maximum A Posteriori: $\theta^{MAP} = \arg\max_\theta P(\theta|D)$.
משפט: המשערך האופטימלי במובן Hit or Miss (כלומר ממזער את $E_{\theta,D}\left[L_\epsilon\left(\theta,\hat{\theta}\right)\right]$) הוא θ^{MAP}.

הוכחה:

$$E_{\theta,D}\left[L_\epsilon\left(\theta,\hat{\theta}\right)\right]=\int_{\theta,D}P(D)P(\theta|D)L_\epsilon\left(\theta,\hat{\theta}\right)d\theta dD$$
$$=\int_D P(D)\left[\int_\theta P(\theta|D)L_\epsilon\left(\theta,\hat{\theta}\right)d\theta\right]dD=\int_D P(D)J_D\left(\hat{\theta}\right)dD$$

כאשר $J_D\left(\hat{\theta}\right)=\int_\theta P(\theta|D)L_\epsilon\left(\theta,\hat{\theta}\right)d\theta$. הפונקציה $L_\epsilon\left(\theta,\hat{\theta}\right)$ נראית כמו:



בעצם יש לנו כאן ספרה ברדיוס ϵ מסביב לנקודה $P(\theta|D)$: בתוך הספרה ערך הפונקציה הוא 0 (ירוק) ומחוץ לה הערך הוא 1 (כחול), לכן נקבל:

$$\int_D P(D)J_D\left(\hat{\theta}\right)dD=1-\int_{|\theta-\hat{\theta}|<\epsilon}P(\theta|D)d\theta\approx 1-B_\epsilon P(\theta|D)$$

כאשר B_ϵ הוא נפח הספרה. הערך שיביא את $J_D\left(\hat{\theta}\right)$ למינימום הוא $\theta^{MAP}=\arg\max_\theta P(\theta|D)$ ולכן θ^{MAP} אופטימלי במובן L_ϵ.

אם נחזור לדוגמת המטבעות, צריך לחשב את $P(d_{11}=0|D)$ ואת $P(d_{11}=1|D)$ במובן Hit or Miss ואז מחזירים את הערך שממקסם את $P(\theta|D)$, כלומר 0 או 1 - מי שההסתברות שלו גבוהה יותר.

הערה: אם הפריור על θ הוא קבוע $P(\theta)=c$ (למשל $\theta\sim\mathcal{U}[0,1]$) אז

$$P(\theta|D)=\frac{P(\theta)P(D|\theta)}{P(D)}=\frac{c\cdot P(D|\theta)}{P(D)}\rightarrow\arg\max_\theta P(\theta|D)=\arg\max_\theta P(D|\theta)\rightarrow\theta^{MAP}=\theta^{MLE}$$

אולם ייתכן כי θ^{MMSE} יהיה שונה.

הערה: אם $P(\theta|D)$ גאוסי אז $\theta^{MAP}=\theta^{MMSE}$ (נכון עבור כל התפלגות יונימודאלית וסימטרית).