תרגול 2: הגישה הפרמטרית להערכת פילוגי הסתברות – משערך MLE

1 תקציר התאוריה

<u>סימונים</u>:

X - משתנה מקרי.

$p_X(x)$ - אם X רציף: פונקציית צפיפות ההסתברות של X;
אם X בדיד: פונקציית ההסתברות של X.

$D = \{x_k\}_{k=1}^n$ - דוגמאות (דגימות - *samples*) בלתי תלויות של X.

<u>המטרה</u>:
להעריך את $p_X(x)$ מתוך n הדוגמאות D.

בגישה הפרמטרית באופן כללי, אנו מניחים כי הפילוג הנדרש $p_X(x)$ הינו בעל צורה ידועה, המוגדרת עד כדי וקטור פרמטרים $\theta \in \mathbb{R}^p$, כלומר $p_X(x) = p_X(x \mid \theta)$.

<u>משערך הסבירות המירבית</u> (Maximum Likelihood Estimator) משערך את θ ע״י:

$$\hat{\theta}_{MLE} \triangleq \arg\max_{\theta \in \mathbb{R}^p} p(D \mid \theta) = \arg\max_{\theta \in \mathbb{R}^p} \left\{ \log p(D \mid \theta) \right\},$$

כאשר:

$$p(D \mid \theta) = p(x_1, \ldots, x_n \mid \theta) = \prod_{k=1}^n p_X(x_k \mid \theta).$$

לכן:

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \mathbb{R}^p} \left\{ \sum_{k=1}^n \log p_X(x_k \mid \theta) \right\}.$$

עבור דגימות D נתונות, נהוג לסמן $L(\theta) \triangleq p(D \mid \theta)$. הגודל $L(\theta)$ נקרא <u>פונקציית הסבירות</u>. כמו כן נהוג לסמן את <u>פונקציית הסבירות הלוגריתמית</u> ע״י:

$$l(\theta) \triangleq \log L(\theta) = \sum_{k=1}^n \log p_X(x_k \mid \theta) \triangleq \sum_{k=1}^n l_k(\theta)$$

<u>הערה</u>: משערך MLE הינו משערך לא-בייסיאני, מכיוון שמניח כי θ הינו גודל <u>קבוע</u> ולא משתנה מקרי.

תרגילים

שאלה 1

נתון כי $X \sim N(\mu,\sigma^2)$, כאשר הממוצע μ ושונות σ^2 אינם ידועים. נתונות n דגימות בלתי תלויות של X, $\left\{x_k\right\}_{k=1}^{n}$.

הראו כי: $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu}_{MLE})^2, \quad \hat{\mu}_{MLE} = \frac{1}{n}\sum_{k=1}^{n} x_k$.

פתרון

נסמן $\theta_1 = \mu$, $\theta_2 = \sigma^2$. כלומר $\theta = (\theta_1,\theta_2)^T$.

המטרה: למצוא $\theta = (\theta_1,\theta_2)^T$ הממקסם את $l(\theta)$. כלומר, נדרוש:

$$\frac{\partial l(\theta)}{\theta_1} = 0 \;,\; \frac{\partial l(\theta)}{\theta_2} = 0 .$$

אצלנו:

$$l_k(\theta) \triangleq \ln p(x_k \mid \theta) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

נגזור ונקבל:

$$\frac{d}{d\theta_1} l_k(\theta) = \frac{1}{\theta_2}(x_k - \theta_1)$$
$$\frac{d}{d\theta_2} l_k(\theta) = -\frac{1}{2\theta_2} + \frac{1}{2\theta_2^2}(x_k - \theta_1)^2$$

בסה"כ נקבל את שני התנאים הבאים למקסימום:

$$\begin{cases} \sum_{k=1}^{n}\frac{1}{\theta_2}(x_k - \theta_1) = 0 \\ -\sum_{k=1}^{n}\frac{1}{\theta_2} + \sum_{k=1}^{n}\frac{1}{\theta_2^2}(x_k - \theta_1)^2 = 0 \end{cases}$$

מהתנאי הראשון מקבלים:

$$\sum_{k=1}^{n} x_k = n\theta_1 \quad \Rightarrow \quad \theta_1 = \hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

מהתנאי השני:

$$\frac{1}{\theta_2} n = \frac{1}{\theta_2^2}\sum_{k=1}^{n}(x_k - \theta_1)^2 \Rightarrow \theta_2 = \sigma^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \theta_1)^2 .$$

<u>שאלה 2</u>

נתונה בעיית סיווג חד-ממדית עם שתי מחלקות ω_1 ו- ω_2 בעלות הסתברויות אפריוריות זהות ($P(\omega_1) = P(\omega_2) = \frac{1}{2}$).

הפילוגים המותנים של התבניות הם $p(x \mid \omega_1) \sim N(0,1)$, $p(x \mid \omega_2) \sim N(1,10^6)$.

מודל: הנח כי הפילוג המותנה של ω_1 **ידוע**, ואילו הפילוג המותנה של ω_2 **לא ידוע**. הוחלט לנחש (**באופן שגוי**) כי $p(x \mid \omega_2) \sim N(\mu,1)$, כאשר התוחלת μ אינה ידועה.

בהנחת המודל הנ"ל, נרצה לשערך את μ מתוך n דוגמאות בלתי תלויות מהמחלקה ω_2, ולהשתמש במסווג הבייסיאני האופטימלי המתאים.

א. נתונות n דגימות (בלתי תלויות) מהפילוג (**האמיתי**) של מחלקה ω_2. מהו ערכו של שערוך $\hat{\mu}_{MLE}$ עבור $p(x \mid \omega_2)$, כאשר נתון כי n גדול מאוד ?

ב. מהם תחומי ההחלטה (הבייסיאניים) המתקבלים כאשר משתמשים בערך שמצאת בסעיף א. (כלומר עבור $p(x \mid \omega_1) \sim N(0,1)$ ו- $p(x \mid \omega_2) \sim N(\hat{\mu}_{MLE},1)$) ?

כעת נשווה את תחומי ההחלטה של המסווג שקיבלנו למסווג אופטימלי הנכון.

ג. מהם תחומי ההחלטה (הבייסיאניים) המתקבלים כאשר משתמשים בפילוגים המותנים הנכונים ($p(x \mid \omega_1) \sim N(0,1)$, $p(x \mid \omega_2) \sim N(1,10^6)$) ?

מטרת הסעיף הבא היא להראות כי אם מניחים מודל שגוי מראש, משערך MLE לא נותן את התוצאות האופטימליות **אפילו בתוך קבוצת המודלים (השגויים) הנתונה.**

ד. נחזור למודל (השגוי): $p(x \mid \omega_2) \sim N(\mu,1)$. על סמך התוצאה של סעיף ג', הציעו ערך חדש ל- μ (השונה מ- $\hat{\mu}_{MLE}$) אשר יתן שגיאה נמוכה יותר מאשר משערך MLE. מהי מסקנתכם לגבי החשיבות של הידע המקדים לגבי המודל ?

<u>פתרון</u>

א. נתון $\{x_k\}_{k=1}^n$, כך ש- $x_k \sim N(1,10^6)$. לכן נקבל כי

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{k=1}^{n} x_k \xrightarrow[n\to\infty]{} 1 .$$

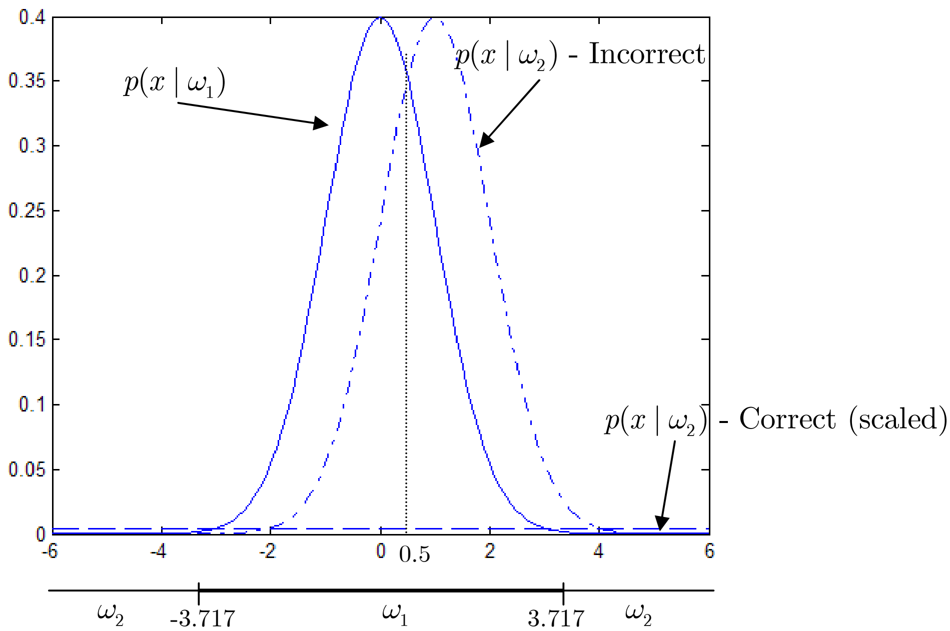
ע"פ חוק המספרים הגדולים.

ב. במקרה זה יש לנו שני פילוגים גאוסיאנים עם שונות זהה: $p(x \mid \omega_1) \sim N(0,1)$ ו- $p(x \mid \omega_2) \sim N(1,1)$ (ושתי קטגוריות שוות הסתברות) ולכן נקודת ההחלטה תהיה באמצע בין שתי התוחלות: $x^* = 0.5$.

ג. כאן יש לפתור את המשוואה הבאה:

$$\frac{1}{\sqrt{2\pi\cdot 1}}\exp[-\frac{1}{2}x^2]=\frac{1}{\sqrt{2\pi\cdot 10^6}}\exp[-(x-1)^2\,/\,(2\cdot 10^6)]$$

נומרית מקבלים: $x^* = \pm 3.717$.



ד. במקרה של המודל שגוי, יש לנו שני פילוגים נורמליים בעלי שונות זהה. לכן גבול ההחלטה היחיד האפשרי הוא נקודה בודדת, כפי שראינו בסעיף א. על סמך הסעיף הקודם, נכוונן את μ כך שנקבל נקודת החלטה ב- $x = 3.717$ (או $x = -3.717$). במקרה כזה, האי-הסכמה בין המסווג המתקבל לבין המסווג האופטימלי היא הקטנה ביותר.

מכיוון שנקודת ההחלטה היא באמצע בין התוחלות, עלינו לפתור: $(0+\mu)/2 = 3.717$. נקבל אם כן: $\mu = 7.43$.

נבדוק כי בחירה שלנו אכן מקטינה את הסתברות השגיאה הממוצעת:

הסתברות השגיאה של המסווג המחליט על ω_2 כאשר $x \geq x^*$ נתונה ע"י

$$\begin{aligned} P(error) &= P(\omega_1)P(error \mid \omega_1) + P(\omega_2)P(error \mid \omega_2) \\ &= \frac{1}{2}\int_{x^*}^{\infty} p(x \mid \omega_1)dx + \frac{1}{2}\int_{-\infty}^{x^*} p(x \mid \omega_2)dx \\ &= \frac{1}{2}(1-\Phi(x^*)) + \frac{1}{2}\Phi\left(\frac{x^*}{10^3}\right) \end{aligned}$$

כאשר Φ היא פונקציית ההתפלגות של משתנה גאוסי סטנדרטי ($N(0,1)$).

שימו לב כי משתמשים בפילוג האמיתי של ω_2 לחישוב הסתברות השגיאה.

במקרה של מסווג המבוסס על $\hat{\mu}_{MLE}$ מקבלים כי $x^* = 0.5$, וע"י הצבה לנוסחה הנ"ל $P(error) \cong 0.4$.

לעומת זאת, ע"י שימוש ב-$\mu = 7.43$ מקבלים כי $x^* = 3.717$, ו-$P(error) \cong 0.25$!

כפי שניתן לראות, משערך MLE במקרה של מודל שגוי לא נותן שגיאה מינימלית אפילו בקבוצת המודלים השגויים. הנחת מודל שגוי גורמת לבעיה אשר לא ניתן להתגבר עליה, אפילו אם יהיה לנו מספר אינסופי של דוגמאות.

<u>שאלה 3</u>

נתונה בעיית סיווג עם שתי מחלקות ω_1 ו- ω_2 בעלות הסתברויות אפריוריות זהות ($P(\omega_1) = P(\omega_2) = 1/2$).

נתון כי הקלט $X = (X_1, \ldots, X_d)^T$ הינו וקטור בינארי ממימד d, כאשר כל רכיב בווקטור זה מוגרל באופן בת״ס.

נסמן את ההסתברויות לקבלת 1 באחד הרכיבים ע״י

$$q_i \triangleq \mathbb{P}\{X_i = 1 \mid \omega_2\} = 1 - \mathbb{P}\{X_i = 0 \mid \omega_2\}$$
$$p_i \triangleq \mathbb{P}\{X_i = 1 \mid \omega_1\} = 1 - \mathbb{P}\{X_i = 0 \mid \omega_1\}$$

נתון כי:

$$p_i = p \quad , q_i = 1 - p \,.$$

א. נניח כי הווקטור $x = (x_1, \ldots, x_d)^T$ הוגרל מקטגוריה ω_1. הראו כי משערך הסבירות המירבית עבור הפרמטר p (על סמך הדוגמא הבודדת הנ״ל) נתון ע״י

$$\hat{p}_{MLE} = \frac{1}{d} \sum_{i=1}^{d} x_i \,.$$

נגדיר באמצעות $T \triangleq 1/d \sum_{i=1}^{d} X_i$ את החלק היחסי של אחדים במשתנה מקרי X.

ב. הראו כי אם $p > 0.5$ המסווג הבייסיאני מסווג ל-ω_1 כאשר $T > 0.5$, ול-ω_2 אחרת. **תזכורת:** בשאלה 2 בתרגול 2 הראנו כי המסווג הבייסיאני מסווג ל-ω_1 אם

$$\sum_{i=1}^{d} \left\{ X_i \ln\left(\frac{p_i}{q_i}\right) + (1 - X_i) \ln\left(\frac{1 - p_i}{1 - q_i}\right) \right\} > \ln\left(\frac{p(\omega_2)}{p(\omega_1)}\right)$$

ג. תארו את ההתנהגות של T עבור $d \to \infty$ ואת הקשר ל-$\hat{p}_{MLE}$. הסבירו מדוע, ע״י הגדלת מספר המאפיינים (=הרכיבים של X) לאינסוף, ניתן להגיע לסיווג ללא שגיאה על סמך סט לימוד שמכיל דוגמא בודדת.

<u>פתרון</u>

א. נקבל בדומה לשאלה 2 מתרגול 2:

$$p(x \mid \omega_1) = \prod_{i=1}^{d} p(x_i \mid \omega_1) = \prod_{i=1}^{d} p^{x_i}(1-p)^{1-x_i}$$

לכן פונקציית הסבירות הלוגריתמית:

$$l(p) = \ln p(x \mid \omega_1) = \sum_{i=1}^{d}\left[x_i \ln p + (1-x_i)\ln(1-p)\right].$$

הנגזרת והשוואתה לאפס נותנת:

$$\frac{d}{dp}l(p) = \frac{1}{p}\sum_{i=1}^{d}x_i - \frac{1}{1-p}\sum_{i=1}^{d}(1-x_i) = 0$$

$$\Rightarrow \quad (1-p)\sum_{i=1}^{d}x_i = p\left(d - \sum_{i=1}^{d}x_i\right)$$

$$\Rightarrow \quad \boxed{\hat{p}_{MLE} = \frac{1}{d}\sum_{i=1}^{d}x_i}$$

ב. נשתמש בתוצאה מתרגול 2 – נחליט על ω_1 אם:

$$\sum_{i=1}^{d}\left\{X_i \ln\left(\frac{p}{1-p}\right) + (1-X_i)\ln\left(\frac{1-p}{p}\right)\right\} > 0$$

$$2\ln\left(\frac{p}{1-p}\right)\sum_{i=1}^{d}X_i > d\ln\left(\frac{p}{1-p}\right)$$

$$p > 0.5 \quad \Rightarrow \quad \frac{1}{d}\sum_{i=1}^{d}X_i > \frac{1}{2}$$

ג. מחוק המספרים הגדולים נקבל כי כאשר מספר המאפיינים (מימד) שואף לאינסוף, יהיה:

$$(*)\, T \to \begin{cases} p, & \text{if } X \text{ is drawn from } \omega_1 \\ 1-p, & \text{if } X \text{ is drawn from } \omega_2 \end{cases}.$$

הסבר:

$$\mathbb{E}\{T \mid \omega_1\} = \frac{1}{d}\sum_{i=1}^{d}\mathbb{E}\{X_i \mid \omega_1\} = \frac{1}{d}dp = p,$$

ובאופן דומה לגבי ω_2. מכאן עבור $d \to \infty$ ניתן לראות כאן שני דברים:

(1) $\hat{p}_{MLE} \to p$.

(2) ע״י שימוש בנוסחה (*) מסווג בייסיאני על סמך דוגמא בודדת מהווה מסווג חסר שגיאה. בעצם ניתן לסכם ולהגיד כי עבור $d \to \infty$ המסווג:

$T > 0.5 \Leftarrow$ בחר ב- ω_1
$T \le 0.5 \Leftarrow$ בחר ב- ω_2

הינו מסווג עם הסתברות שגיאה 0.

שאלה 4

יהי X ווקטור אקראי בינארי ממימד $d=2$, כלומר: $X=(X_1,X_2)^T \in \{0,1\}^2$.

בשאלה זו נדגים את מדד **האנטרופיה** על סמך שני פילוגים שונים של X.

א. יהי $p(x_1,x_2)$ הפילוג האחיד על $\{0,1\}^2$, כלומר:

$x_1 \backslash x_2$	0	1
0	$\frac{1}{4}$	$\frac{1}{4}$
1	$\frac{1}{4}$	$\frac{1}{4}$

מהי האנטרופיה של p ?

ב. יהי $q(x_1,x_2)$ הפילוג הבא:

$x_1 \backslash x_2$	0	1
0	$\frac{1}{2}$	0
1	0	$\frac{1}{2}$

מהי האנטרופיה של q ? מהי מסקנתכם לגבי השפעת הקורלציה בין רכיבי הווקטור על האנטרופיה ?

פתרון 4

א. האנטרופיה של p נתונה ע"י:

$$H(p) = -\sum_{x\in\{0,1\}^2} p(x)\log p(x) = 4\cdot\left(-\frac{1}{4}\log\frac{1}{4}\right) = \log 4 = 2$$

ב. האנטרופיה של q נתונה ע"י:

$$H(q) = -\sum_{x\in\{0,1\}^2} q(x)\log q(x) = 2\cdot\left(-\frac{1}{2}\log\frac{1}{2}\right) = \log 2 = 1$$

בפרט רואים כי למרות של-p ול-q אותם פילוגים שוליים, האנטרופיה של q קטנה יותר מהאנטרופיה של p. המסקנה היא כי הקורלציה בין רכיבי הווקטור מקטינה אנטרופיה.