מערכות לומדות - 046195 הפקולטה להנדסת חשמל
אביב תשע״ט (2019) הטכניון - מכון טכנולוגי לישראל

תרגיל בית 1 – שערוך פרמטרים, הסתברות

תאריך הגשה: 28.04.2019

מוריאל בן משה 304832512, דניאל אנגלסמן 300546173

חלק יבש

שאלה 1

יוסי, בוגר קורס מבוא למערכות לומדות, בחר לפתוח חנות גרעינים ״חכמה״ שם יוכל להביא לידי ביטוי את סט היכולות שרכש במהלך לימודיו. למשל, יוסי שמע ממחקרים כי זמן ההגעה של לקוח לקופה מפולג פואסונית. נסמן ב- $X \sim Poisson(\lambda)$ משתנה אקראי בדיד המציין את זמן ההגעה :

$$p_X(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad x = 0,1,2,....$$

א. המשימה הראשונה שהגדיר יוסי היא שערוך הפרמטר λ. לצורך כך הקליט יוסי את זמני ההגעה (ערכים שלמים) של לקוחות לקופה $\mathrm{D} = \{x_i\}_{i=1}^n$. עזרו ליוסי למצוא את משערך MLE של λ.

(1) רשמו ביטוי עבור פונקציית הסבירות $L(\lambda)$ ופונקציית הסבירות הלוגריתמית $l(\lambda)$.

(2) חשבו במפורש את משערך הסבירות המירבית $\hat{\lambda}_{MLE}$.

1

Let $L(\lambda)$ be likelihood function, and $l(\lambda)$ be log-likelihood function :

$$L(\lambda) = p(D|\lambda) = \prod_{i=1}^{n} p(x_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^x e^{-\lambda}}{x!}$$
$$l(\lambda) = log\, p(D|\lambda) = log\, \prod_{i=1}^{n} \frac{\lambda^x e^{-\lambda}}{x!} = \sum_{i=1}^{n} log\, \frac{\lambda^x e^{-\lambda}}{x!}$$

$\hat{\lambda}_{MLE}$ value is obtained as the maximum / optimum of the likelihood function :

$$\hat{\lambda}_{MLE} = arg\ max\ log\ p(D|\lambda) \quad \rightarrow \quad \frac{d}{d\lambda}\sum_{i=1}^{n} log\ \frac{\lambda^x e^{-\lambda}}{x!} = 0$$

$$\sum_{i=1}^{n}(\frac{f'(\lambda)}{f(\lambda)}) \quad \rightarrow \quad \sum_{i=1}^{n}\frac{\frac{\lambda^x e^{-x}}{x!}(\frac{x}{\lambda}-1)}{\frac{\lambda^x e^{-x}}{x!}} \quad \rightarrow \quad \sum_{i=1}^{n}(\frac{x}{\lambda}-1) = 0 \quad \rightarrow \quad \hat{\lambda}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (1.1)$$

2nd derivation proves local maxima :

$$\frac{d^2}{d^2\lambda}l(\lambda) = -\frac{1}{\lambda^2}\sum_{i=1}^{n} x_i < 0 \quad \forall \quad x_i \in \mathbb{N} \quad \Leftrightarrow \quad arg\ max$$

ב. מעתה, נניח כי λ בעצמו משתנה אקראי שפילוגו :

$$p(\lambda) = \lambda e^{-\lambda}\ , \quad \lambda > 0$$

חשבו את התפלגות בדיעבד (posterior) של λ , כלומר $p(\lambda \mid D)$.

הערה : אין צורך לחשב במפורש אינטגרלים בביטוי הסופי.

$$p(\lambda|D) = \frac{p(D|\lambda)p_0(\lambda)}{p(D)} = \frac{\prod_{i=1}^{n} p(x_i|\lambda)p_0(\lambda)}{\int p(D|\lambda)p_0(\lambda)d\lambda} = \frac{(\prod_{i=1}^{n}\frac{\lambda^x e^{-\lambda}}{x!})\cdot\lambda e^{-\lambda}}{\int p(D|\lambda)p_0(\lambda)d\lambda} \quad (1.2)$$

ג. בחישוב משערך MAP של הסעיף הקודם מתקבל כי :

$$\hat{\lambda}_{MAP} = \frac{1+\sum_{i=1}^{n} x_i}{n+1}$$

(1) חשבו את ההטיה של משערך זה כתלות בפרמטר λ . (רמז : $E[X] = \lambda$).

(2) עבור איזה ערך של λ המשערך יהיה בלתי מוטה?

$$bias(\lambda) = E[\hat{\lambda}] - \lambda = E[\frac{1+\sum_{i=1}^{n} x_i}{n+1}] - \lambda = \frac{1}{n+1} + \frac{E[X]}{n+1} - \lambda$$

$$\Rightarrow \quad = \frac{1+n\lambda}{n+1} - \lambda \quad \Rightarrow \quad bias(\lambda) = \frac{1-\lambda}{1+n} \quad (1.3)$$

$$Unbiased: \quad \lambda = 1 \quad (1.4)$$

שאלה 2

נתונות סדרה של n דגימות של משתנה מקרי דו ממדי המפולג גאוסית עם פרמטרים

$$\boldsymbol{\mu} = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}, \boldsymbol{\Sigma}_1 = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} \sigma^2 & \frac{1}{2}\sigma^2 \\ \frac{1}{2}\sigma^2 & \sigma^2 \end{pmatrix}$$

שני סטטיסטיקאים בוחרים לשערך את $\boldsymbol{\mu}$ באמצעות MLE. סטטיסטיקאי א' מניח הנחת

אי-תלות על המשתנים, ולכן משתמש במטריצת ה- covariance Σ_1.

לעומתו, סטטיסטיקאי ב' יודע כי ישנו קשר מוקדם בין שני הפרמטרים הנמדדים, ועל כן

מניח פילוג בעל covariance Σ_2 .

1. חשבו את המשערכים של סטטיסטיקאי א' ו- ב'.

2

$$\hat{\mu}_{MLE} = arg\ max\ log\ p(D|\mu^{(1)}, \mu^{(2)}) = arg\ max\ log\ p(D|\mu^{(1)}) \cdot p(D|\mu^{(2)})$$

$$\Rightarrow \quad log \prod_{i=1}^{n} (\frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}}) \big(e^{-\frac{1}{2}(x-\mu^1)^T \Sigma_1^{-1}(x-\mu^1)} e^{-\frac{1}{2}(x-\mu^2)^T \Sigma_1^{-1}(x-\mu^2)} \big)$$

$$\Rightarrow \quad -log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{2} (x_i^{(j)} - \mu^{(j)})^T (x_i^{(j)} - \mu^{(j)}) \quad \backslash \ \frac{d}{d\mu^{(j)}} \qquad (2.1)$$

$$\underline{\underline{Case \quad \#1}} \quad \Rightarrow \quad \hat{\mu}_{MLE}^{(j)} = \frac{1}{n} \sum_{i=1}^{n} x_i^{(j)}$$

The second case will be slightly different since the covariance matrix' non-diagonals entries are non-zeros, and thus cannot be treated directly as scalars :

$$|\Sigma_2| = \frac{3}{4}\sigma^4, \qquad \Sigma_2^{-1} = \frac{2}{3\sigma^2} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix};$$

Let us develop the inside term (solely) :

$$\sum_{j=1}^{2}(x_i^{(j)}-\mu^{(j)})^T\Sigma_1^{-1}(x_i^{(j)}-\mu^{(j)})=\frac{2}{3\sigma^2}\cdot\begin{bmatrix}x^{(j)}-\mu^{(j)} & x^{(j)}-\mu^{(j)}\end{bmatrix}\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}\begin{bmatrix}x^{(j)}-\mu^{(j)}\\ x^{(j)}-\mu^{(j)}\end{bmatrix}$$

For simplicity we can work with scalars inside terms, using the shorthand notation :

$$a=x^{(1)}-\mu^{(1)},\quad b=x^{(2)}-\mu^{(1)}$$

$$\Rightarrow\frac{2}{3\sigma^2}\cdot\begin{bmatrix}a & b\end{bmatrix}\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}\begin{bmatrix}a\\ b\end{bmatrix}=\frac{2}{3\sigma^2}\cdot\begin{bmatrix}2a-b & -a+2b\end{bmatrix}\begin{bmatrix}a\\ b\end{bmatrix}\Rightarrow\frac{2\cdot 2}{3\sigma^2}\cdot\left(a^2-ab+b^2\right)$$

$$\Rightarrow\frac{d}{d\mu^{(j)}}\Big(\sum_{i=1}^{n}\frac{4}{3\sigma^2}\cdot\big((x^{(1)}-\mu^{(1)})^2-(x^{(1)}-\mu^{(1)})(x^{(2)}-\mu^{(2)})+(x^{(2)}-\mu^{(2)})^2\big)\Big)=0$$

$$\Rightarrow\frac{4}{3\sigma^2}\cdot\sum_{i=1}^{n}\big(\{-2x^{(1)}+2\mu^{(1)}+x^{(2)}-\mu^{(2)}\}+\{x^{(1)}-\mu^{(1)}-2x^{(2)}+2\mu^{(2)}\}\big)$$

$$\Rightarrow\frac{4}{3\sigma^2}\cdot\sum_{i=1}^{n}\sum_{j=1}^{2}(-x^{(j)}+\mu^{(j)})=0\quad\underline{\underline{Case\quad\#2}}\quad\Rightarrow\quad\hat{\mu}_{MLE}=\frac{1}{n}\sum_{i=1}^{n}x_i^{(j)}$$

2. כעת, בשל רמת בטחון גבוהה, הסטטיסטיקאים קובעים את התוחלת של המשתנה הראשון כ- μ_1, כלומר $\boldsymbol{\mu}=\begin{pmatrix}\mu_1\\ \mu^{(2)}\end{pmatrix}$ והמשתנה שברצוננו לשערך הינו $\mu^{(2)}$. חשבו את המשערכים של שני הסטטיסטיקאים.

Let us check whether there's a difference between both models :

$$\mu=\begin{bmatrix}\mu_1\\ \mu^{(2)}\end{bmatrix}\Rightarrow\quad -log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{2}(x_i^{(j)}-\mu^{(j)})^T(x_i^{(j)}-\mu^{(j)})\quad\Big\backslash\frac{d}{d\mu^{(2)}}$$

$$\Rightarrow\quad -log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}((x_i^{(1)}-\mu_1)^2+(x_i^{(2)}-\mu^{(2)}))\quad\Big\backslash\frac{d}{d\mu^{(2)}}$$

$$\underline{\underline{Case\quad\#1}}\quad\Rightarrow\quad\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i^{(2)}-\mu^{(2)})=0\quad\Rightarrow\quad\hat{\mu}_{MLE}^{(2)}=\frac{1}{n}\sum_{i=1}^{n}x_i^{(2)}$$

And now the second statistical estimator :

$$\Big(\sum_{i=1}^{n}\frac{4}{3\sigma^2}\cdot\big((x^{(1)}-\mu_1)^2-(x^{(1)}-\mu_1)(x^{(2)}-\mu^{(2)})+(x^{(2)}-\mu^{(2)})^2\big)\Big)\setminus\frac{d}{d\mu^{(2)}}=0$$

$$\Rightarrow\quad\frac{4}{3\sigma^2}\sum_{i=1}^{n}\big((x^{(1)}-\mu_1)-2(x^{(2)}-\mu^{(2)})\big)\quad\Rightarrow\quad\sum_{i=1}^{n}(x_i^{(1)}-2x_i^{(2)})+n(2\mu^{(2)}-\mu_1)=0$$

$$\underline{\underline{Case\quad\#2}}\quad\Rightarrow\quad\hat{\mu}^{(2)}_{MLE}=\frac{1}{2}\mu_1+\frac{1}{2n}\sum_{i=1}^{n}(x_i^{(1)}-2x_i^{(2)})$$

3. הסבירו את ההבדל בין התוצאה בסעיף א' והתוצאה בסעיף ב'.

At first, we had no idea regarding the nature of expected values. Although $Cov_{\#2}(x)$ assume prior correlation, the results kept independent and subtracted each other. Afterwards, when the expected value was assumed be known a-priori, case #1 kept as is. But ase #2 showed different results where the prior relation affected.

4. בדקו האם קיימת הטייה (bias) למשערכים שמצאתם.

The first 3 terms :

$$bias(\hat{\mu}^{(j)}_{MLE})=\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}x_i^{(j)}]-\mu^j=\mu^j-\mu^j=0$$

But the last one :

$$bias(\hat{\mu}^{(2)}_{MLE})=\mathbb{E}\big[\frac{1}{2}\mu_1+\frac{1}{2n}\sum_{i=1}^{n}(x_i^{(1)}-2x_i^{(2)})\big]=\frac{1}{2}\mu_1+\frac{1}{2}\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}x_i^{(1)}-\frac{2}{n}\sum_{i=1}^{n}x^{(2)}]$$

$$bias(\hat{\mu}^{(2)}_{MLE})=\frac{1}{2}\mu_1+\frac{1}{2}\mu_1-\mu_2=\mu_1-\mu_2\qquad(2.2)$$

3

שאלה 3

נתונות n דגימות $x_1,..,x_n$ IID מתוך פילוג המוגדר ע"י הצפיפות

$$f(x;\theta)=\binom{N}{x}\theta^x\cdot(1-\theta)^{N-x}$$

כאשר $x\in\{0,1,...,N\}$ ו- $0<\theta<1$

1. רשמו את פונקציית ה-log-likelihood .

$$l(\theta)=log\, p(D|\theta)=log\,\prod_{i=1}^{n}\binom{N}{x_i}\theta^{x_i}\cdot(1-\theta)^{N-x_i}=\sum_{i=1}^{n}log\,\binom{N}{x_i}\theta^{x_i}\cdot(1-\theta)^{N-x_i}$$

$$\Rightarrow\quad\sum_{i=1}^{n}\left[log\,\binom{N}{x_i}\right]+log(\theta^{\sum_{i=1}^{n}x_i})+log(1-\theta)^{\sum_{i=1}^{n}N-x_i}$$

$$\Rightarrow\quad l(\theta)=\sum_{i=1}^{n}\left[log\,\binom{N}{x}\right]+\sum_{i=1}^{n}x_i\cdot log(\theta)+(nN-\sum_{i=1}^{N}x_i)\cdot log(1-\theta)\qquad(3.1)$$

3.2 Calculate its ML estimator ($\hat{\theta}_{MLE}=\frac{d}{d\theta}l(\theta)=0$):

$$\Rightarrow\quad\sum_{i=1}^{N}x_i(\frac{1}{\theta})-(\frac{nN-\sum_{i=1}^{N}x_i}{1-\theta})\to\sum_{i=1}^{N}x_i(\frac{1}{\theta(1-\theta)})-\frac{nN}{1-\theta}=0$$

$$\Rightarrow\quad\hat{\theta}_{MLE}=\frac{1}{n\cdot N}\sum_{i=1}^{N}x_i\qquad(3.2)$$

3.3 Describe its connection to Bernoulli ML estimator : The probability mass function of Bernoulli distribution is given by -

$$f(k;p)=\begin{cases}p & \text{if } k=1,\\ q=1-p & \text{if } k=0.\end{cases}\quad\longrightarrow\quad p^k(1-p)^{1-k}\quad\text{for } k\in\{0,1\}$$

For $n\cdot N$ consecutive experiments we would get the same term.

3.4 Given $\theta \in U[0, \frac{1}{4}]$ calculate the corresponding MAP estimator :

$$\hat{\theta}_{MAP} = arg\ max\ p(D|\theta) \cdot p_0(\theta)$$

$$p(D|\theta) \cdot p_0(\theta) = \prod_{i=1}^{n} [\binom{N}{x_i} \theta^{x_i} \cdot (1-\theta)^{N-x_i}] \cdot \frac{1}{4} I_{\theta\in[0,0.25]} \quad \backslash\, log(\circ)$$

$$\Rightarrow [\sum_{i=1}^{n}(log \binom{N}{x}) + \sum_{i=1}^{n} x_i \cdot log(\theta) + (nN - \sum_{i=1}^{N} x_i) \cdot log(1-\theta) + log(\frac{1}{4})] \cdot I_{\theta\in[0,0.25]} \quad (3.3)$$

Indicator I indicates that results exist only in $\theta \in [0, \frac{1}{4}]$. The rest leads similarly to **Eq. 3.2** (MLE), but having maximal value of must be at one of the ends, due to its monotony.

3.5 Compare $\hat{\theta}_{MLE}$ with $\hat{\theta}_{MAP}$:

As seen, both contain the likelihood function $p(D|\theta)$ with slight changes, since the prior exists between $\theta \in [0, \frac{1}{4}]$ and thus $\hat{\theta}_{MAP}$ will provide more accurate results in that range.

$-fin-$