הסבר - Max Likelihood

תמנע קליינמן

כתבתי את הסיכום הזה לעצמי, ייתכן וכנראה שיש בו טעויות. מוזמנים לכתוב לי - timna.kleinman@mail.huji.ac.il

חלקים בכתום מסמנים קטע שהוא לא קריטי בשביל להבין, ולא קשור ישירות ל-Max Likelihood אבל לדעתי יעזרו לגבש תמונה כללית

1 טרמינולוגיה

נסמן -

- $x_1, ..., x_n$ הן הדגימות.
- θ היא המשתנה אותו אנו מנסים לאמוד.

נזכיר את חוק בייס -

$$\overbrace{P(\theta|x_1, ..., x_n)}^{\text{posterior}} = \frac{\overbrace{P(x_1, ..., x_n|\theta)}^{\text{likelihood}} \cdot \overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(x_1, ..., x_n)}_{(\text{constant normalization})}}$$

נסביר את ההיגיון בכל אחד מהשמות -

Prior - זה מידע מקדים שיש לנו על ההסתברות שהפרמטר הזה הוא האמיתי. לדוג' - יש לנו מטבע שקיבלנו מהסופר, ואנחנו מנסים לאמוד את ההסתברות שלו ליפול על עץ. כלומר $\hat{\theta}$ יהיה אומד להסתברות לקבל עץ. אזי, מאחר וזה סתם מטבע שקיבלנו מהסופר, ובלי קשר לדגימות שלנו (זריקות מטבע) בהסתברות גבוהה $\theta \simeq 0.5$, ובהסתברות ממש ממש נמוכה $\theta = 0$.[1] במילים אחרות $P(\theta = 0.5)$ הוא גדול, ו-$P(\theta = 0)$ הוא קטן מאוד. זהו ה-prior.

Likelihood - עבור θ מסויימת מה הסיכוי (What is the likelihood) שהיינו מקבלים את הדגימות $x_1, ..., x_n$?

Postertior - אחרי (post) שראינו את הדגימות, מה הסיכוי שזו אכן ה-θ?

[1] כלומר לא נקבל עץ אף פעם

הערה 1.1. שימו לב להבדל בין השאלה של likelihood לבין posterior: likelihood מסתכל על θ ספציפית שכבר נבחרה, ואז שואל - האם בכלל סביר שהיינו מקבלים את הדגימות האלו? posterior אומר שנתונות לי כבר הדגימות, מה הסיכוי ש-θ היא הנכונה? מאחר ו-θ היא הפרמטר שמעניין אותנו, נוכל לכתוב

$$\overbrace{P\left(\theta|x_1,...,x_n\right)}^{\text{posterior}} \propto \overbrace{P\left(x_1,...,x_n|\theta\right)}^{\text{likelihood}} \cdot \overbrace{P\left(\theta\right)}^{\text{prior}}$$

כלומר, ה-posterior פרופורציונלי ל-prior, רק שמצטרפת אליו גם הסבירות מלכתחילה ש-θ היא הנכונה. נחזור לדוגמה שלנו עם זריקות המטבע - נניח וזרקנו מטבע 3 פעמים, ויצא לנו 3 פעמים פלי. אז, ה-likelihood יגיד לנו שהכי סביר שייצא לנו 3 פעמים פלי כאשר $\theta = 0$, כלומר הסיכוי לקבל עץ הוא 0. לכן, אם היינו בוחרים לפי ה-likelihood היינו בוחרים $\theta = 0$. ה-posterior לעומת זאת, אומנם כן "יבין" ש-$\theta = 0$ נשמע הכי סביר בהינתן הדגימות, אבל גם ישקלל לתוך זה את העובדה שממש לא סביר ש$\theta = 0$ (ה-prior). לכן, יכול להיות שה-posterior "ייבחר" ב-θ קטנה אומנם, אבל לא 0 כי זה ממש ממש לא סביר.

2 אמידה

מטרה: לאמוד את θ בהינתן $x_1,...,x_n$ דגימות.

דרך א' (MAP): אינטואיטיבית, בהמשך לחלק של הטרמינולוגיה, מרגיש הכי הגיוני לבחור את $\hat{\theta}$ ע"י בחירת θ שהיא בעלת הסתברות posterior הכי גבוהה. כלומר

$$\begin{aligned}\hat{\theta}_{\text{MAP}} &= \arg\max_{\theta}\left(P\left(\theta|x_1,...,x_n\right)\right) \overset{\text{Bayes}}{=} \arg\max_{\theta}\left(\frac{P\left(x_1,...,x_n|\theta\right)\cdot P\left(\theta\right)}{P\left(x_1,...,x_n\right)}\right) \\ &\overset{\text{Doesn't depend on }\theta}{\overset{P(x_1,...,x_n)}{=}} \arg\max_{\theta}\left(P\left(x_1,...,x_n|\theta\right)\cdot P\left(\theta\right)\right) \overset{\text{בת"ל}}{=} \arg\max_{\theta}\left(\prod_{i=1}^{n} P\left(x_i|\theta\right)\cdot P\left(\theta\right)\right) \\ &\overset{\text{מונוטוניות }log}{=} \arg\max_{\theta}\left(\log\left(\prod_{i=1}^{n} P\left(x_i|\theta\right)\cdot P\left(\theta\right)\right)\right) \overset{\text{תכונות log}}{=} \arg\max_{\theta}\left(\sum_{i=1}^{n}\log\left(P\left(x_i|\theta\right)\right)+\log\left(P\left(\theta\right)\right)\right)\end{aligned}$$

עם זאת, ישנן שתי סיבות עיקריות (שאני מכירה) לכך שנעדיף להשתמש ב-likelihood -

- ה-prior הרבה פעמים לא ידוע לנו (וכך גם ה-posterior), ולכן לא נוכל כלל לחשב את הביטוי הרצוי.
- אם נסתכל על החלק האחרון בפיתוח, נשים לב כי אנו סוכמים על הדגימות. לכן, אם ניקח "ממש הרבה" דגימות, ה-prior יהפוך לזניח ונוכל להשתמש ב-likelihood בלבד.

הערה 2.1. כל הנושא הזה של כן או לא prior מתקשר לגישות שונות - frequentist vs. bayesian שאליהן לא ניכנס כאן.

דרך ב' (Max Likelihood): כאמור למעלה, ה-likelihood הוא $P\left(x_1,...,x_n|\theta\right)$. **חשוב לשים לב!** כי ה-likelihood הוא פונקציה של θ - הדגימות נתונות לנו, ואנחנו שואלים "עבור θ ספציפית, כמה סביר שהיינו מקבלים את הדגימות הנתונות?". לכן נסמן -

$$\mathcal{L}\left(\theta|x_1,...,x_n\right) = P\left(x_1,...,x_n|\theta\right)\left(\underbrace{= f_{\theta}\left(x_1,...,x_n\right) = f\left(x_1,...,x_n|\theta\right)}_{\text{במקרה הרציף}}\right)$$

כעת, נרצה לבחור את θ בעלת ה-likelihood הגבוה ביותר -

$$\hat{\theta}_{ML}=\arg\max_{\theta}\left(P\left(x_1,...,x_n|\theta\right)\right)\overset{\text{i.i.d}}{=}\arg\max_{\theta}\left(\prod_{i=1}^{n}P\left(x_i|\theta\right)\right)\overset{log\text{ מונוטונית}}{=}\arg\max_{\theta}\left(\log\left(\prod_{i=1}^{n}P\left(x_i|\theta\right)\right)\right)$$
$$\overset{log\text{ מונוטונית}}{=}\arg\max_{\theta}\left(\sum_{i=1}^{n}\log P\left(x_i|\theta\right)\right)$$

הערה 2.2. לא היינו חייבים להוסיף את ה-log, אך פעמים רבות הוא עוזר ומקל מאוד על החישובים.

לבסוף, נזכיר כי על מנת למצוא את המקסימום של הביטוי נרצה לגזור אותו ולהשוות ל-0 (ואם אנחנו פדנטיים אז גם לוודא שהוא מקסימום).

דוגמה 2.3. (מהשיעור) נניח שמשפחת ההתפלגויות ידועה לנו והיא ההתפלגות הנורמלית. אנחנו מחפשים את μ. כלומר $x_1,...,x_n\overset{\text{i.i.d}}{\sim}\mathcal{N}\left(\mu,\sigma^2\right)$ ו-$\theta:=\mu$ עבור איזושהי σ^2. נזכיר כי ההתפלגות הנורמלית היא מהצורה - $\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$.

עתה, נשתמש ב-$\hat{\theta}_{ML}$ שחישבנו קודם (עבור $\theta:=\mu$) -

$$\hat{\mu}_{ML}=\arg\max_{\theta}\left(\sum_{i=1}^{n}\log f_{\theta}\left(x_i\right)\right)=\arg\max_{\theta}\left(\sum_{i=1}^{n}\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right)\right)\right)$$
$$\overset{\frac{1}{\sqrt{2\pi\sigma^2}}\text{ constant}}{=}\arg\max_{\theta}\left(\sum_{i=1}^{n}\log\exp\left(-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right)\right)=\arg\max_{\theta}\left(-\sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right)$$
$$\overset{\frac{1}{2\sigma^2}\text{ constant}}{=}\arg\max_{\theta}\left(-\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right)=\arg\min_{\theta}\left(\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right)$$

עכשיו, על מנת למצוא את המינימום של הפונקציה ביחס ל-μ נוכל לגזור אותה (ביחס ל-μ) ונקבל

$$\frac{\partial\left(\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right)}{\partial\mu}=\sum_{i=1}^{n}\frac{\partial\left(x_i-\mu\right)^2}{\partial\mu}=\sum_{i=1}^{n}2\left(x_i-\mu\right)\overset{!}{=}0$$
$$\sum_{i=1}^{n}\left(x_i-\mu\right)=\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\mu=\sum_{i=1}^{n}x_i-n\mu\overset{!}{=}0$$
$$\hat{\mu}_{ML}=\frac{1}{n}\sum_{i=1}^{n}x_i$$