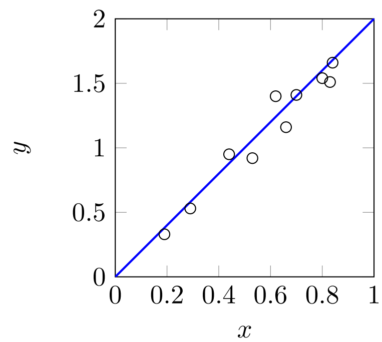
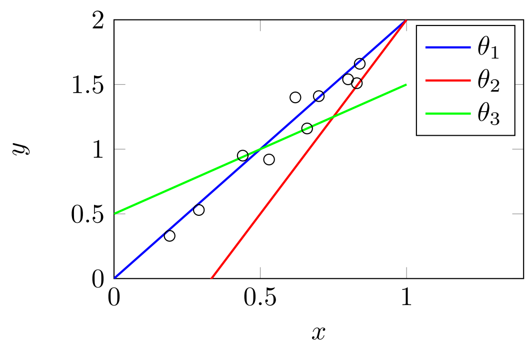
שעור 2 BML - הפילוסופיה הבייסיאנית

November 2, 2022

נניח כי נתון לנו אוסף נקודות $D = \{x_i, y_i\}$. ע"פ הגישה ה-frequentist-ית, הסתברות היא שכיחות יחסית. כלומר, בלמידה קלאסית אנחנו מוצאים את $\theta^* = \arg\min_\theta L(\theta, D)$ ומציירים קו רגרסיה יחיד.



לעומת זאת, לפי הגישה בייסיאנית, הסתברות היא דרגת אמונה (degree of belief). כלומר, בלמידה בייסיאנית יש לנו הרבה קווים ולכל ערך של θ יש לנו את "דרגת האמונה" שהיא $P(\theta|D)$.



דוגמא: יש לנו מטבע והטלנו אותו N פעמים. הפרמטר θ הוא הסיכוי שלנו לקבל 0 והמידע שלנו הוא $D = \{0, 1, 0, ...\}$. נניח כי $n_0 = \#0 = 64, n_1 = \#1 = 36$.

לפי הגישה הקלאסית, נאמוד את θ באמצעות אלגוריתם הנראות המרבית (Maximum Likelihood). פונקצית הנראות היא:

$$P(D;\theta) = \theta \cdot (1-\theta) \cdot \theta \cdot ... = \theta^{n_0} \cdot (1-\theta)^{N-n_0}$$

לפי הגישה הקלאסית, אם נחזור על הניסוי המון פעמים, מספר הפעמים שנקבל את וקטור התוצאות D פרופורציוני ל-$P(D;\theta)$. נראות מרבית - נרצה למצוא את המקסימום של פונקצית הנראות. הפרמטר שממקסם את פונקצית הנראות ממקסם גם את הלוגריתם שלה:

$$\arg\max_{\theta} \theta^{n_0} \cdot (1-\theta)^{N-n_0} = \arg\max_{\theta} n_0 \ln(\theta)(N-n_0)\ln(1-\theta)$$

נגזור ונשווה לאפס:

$$\frac{\partial}{\partial\theta} \ln(P(D;\theta)) = \frac{n_0}{\theta} - \frac{N-n_0}{1-\theta} = 0 \rightarrow \theta = \frac{n_0}{N}$$

את אומד הנראות המרבית (אנ"מ) אנחנו מסמנים $\hat{\theta}(D)$, האומד בהנתן D.
בגישה הבייסיאנית, נחפש את $P(\theta|D)$. לפי חוק בייס,

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

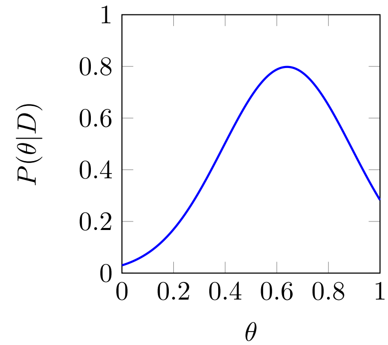
כאשר:

- האיבר $P(\theta)$ הוא הפריור (prior, "קודם" בלטינית) כי זה הסיכוי שמניחים לפני שרואים את הנתונים
- האיבר $P(D|\theta)$ הוא פונקצית הנראות הבייסיאנית
- האיבר $P(\theta|D)$ הוא פונקצית הנראות הפוסטריורית (posterior, "אחרי" בלטינית)

נניח פריור $\theta \sim \mathcal{U}[0,1]$ ואז $P(\theta) = \begin{cases} 1 & 0 \leq \theta \leq 1 \\ 0 & o.w \end{cases}$. הנראות היא $P(D|\theta) = \theta^{n_0} \cdot (1-\theta)^{N-n_0}$ ונקבל:

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

שנראה כמו... (בשעור הבא חישוב מדויק)



מדוע הגישה הבייסיאנית שנויה במחלוקת?

1. פילוסופית - ההסתכלות על θ כעל משתנה מקרי במקום פרמטר קבוע, מה שמאפשר לנו לדבר על $P(\theta)$ ועל $P(D|\theta)$.
2. הנחות חזקות -למה θ מתפלג אחיד ולא למשל בטא? אנחנו נדרשים לספק את $P(\theta)$ והתוצאות שנקבל תלויות בו בצורה חזקה.
3. בעית החישוב - צריך לחשב אינטגרלים (כמו $\int_\theta P(\theta)P(D|\theta)d\theta$) וחלקם בלתי נעימים בעליל.

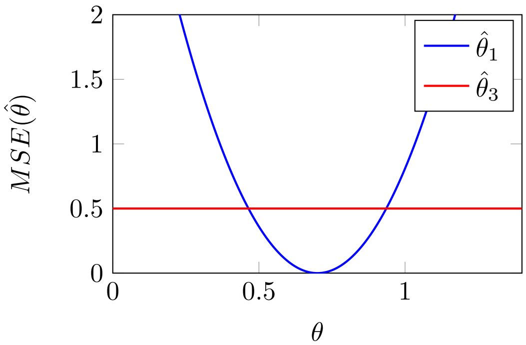
מדוע בכל זאת נרצה להשתמש בשיטה הבייסיאנית?

1. מאפשרת אי-ודאות
2. אופטימליות
3. חוסר פרמטריות

אז איך בעצם מעריכים אופטימליות של שיטה? חיפשנו לאמוד את θ, יש לנו מספר אפשרויות:

1. אנ"מ שראינו $\hat{\theta}_1(D)=\frac{n_0}{N}$
2. אנ"מ נוסף $\hat{\theta}_2(D)=\frac{1}{2}\cdot\frac{n_0}{N}$
3. אומד קבוע $\hat{\theta}_3(D)=\frac{1}{2}$

ביחס לאיזשהי θ שהיא groud truth, אפשר להגדיר פונקצית הפסד $L(\hat{\theta},\theta)=(\hat{\theta}-\theta)^2$. נניח $\theta=0.7$ ו-$D=\{0,0,0,0,0,0,0,1,1,1\}$ ונוכל לחשב את L עבור כל אחד מהאומדים הנ"ל. בבעיית שערוך קלאסית נסתכל על $MSE(\hat{\theta}(D))=E_D\left[\|\theta-\hat{\theta}(D)\|^2\right]$, אבל זו פונקציה של θ - כלומר לכל $\theta,\hat{\theta}$ נקבל ערך שונה של MSE ולכל נקודה יש משערך אופטימלי אחר:



הגדרה : עבור משערך בייסיאני,

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

משפט (הוכחה בשבוע הבא) : האלגוריתם האופטימלי מבחינת BMSE הוא $\hat{\theta}^{MMSE} = E[\theta|D]$.
האלגוריתם מחשב את $P(\theta|D)$ ואז מחזיר את התוחלת ("מרכז הכובד") שלה, שהיא $E[\theta|D]$.
האופטימליות כאן היא מבחינת שגיאה ריבועית אבל גם תלויה בבחירת הפריור.
בחזרה לדוגמא :

$$\theta \sim \mathcal{U}[0,1], \qquad P(D|\theta) = \theta^{n_0} \cdot (1-\theta)^{N-n_0}, \qquad P(\theta|D) = \frac{1}{P(D)} \theta^{n_0} \cdot (1-\theta)^{N-n_0}$$

הגדרה : משתנה מקרי מתפלג דיריכלה (Dirichlet) ומסומן $\theta \sim D(\alpha)$ כאשר α הוא וקטור שכל אבריו חיוביים ומתקיים

$$P(\theta) = \frac{1}{B(\alpha)} \prod_k \theta_k^{\alpha_k - 1}, \qquad B(\alpha) = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\alpha_0)}, \qquad \alpha_0 = \sum_k \alpha_k$$

התוחלת של ההתפלגות הזו היא $E[\theta_k] = \frac{\alpha_k}{\alpha_0}$. בדוגמא יש לנו את $P(\theta|D) = \frac{1}{P(D)}\theta^{n_0} \cdot (1-\theta)^{N-n_0}$ אז נוכל לסמן $\alpha = (n_0+1, N-n_0+1)$, כלומר $\alpha_0 = N+2$ ובסך הכל נקבל $E[\theta|D] = \frac{n_0+1}{N+2}$.