

שעור 1 BML - מבוא

October 31, 2022

נושאי הקורס:

1. הפילוסופיה הבייסיאנית
 2. האופטימליות של גישה בייסיאנית
 3. בחירת מודלים בייסיאנית
 4. רגרסיה באינסוף פרמטרים
 5. שימוש ב-Random Fourier Features
 6. תהליכים גאוסיים
 7. עירוב גאوسی (Gaussian Mixture)
 8. מונטה קרלו ודגימת גיבס (Gibbs Sampler)
 9. ניתוח צבירים גאوسی (Gaussian Clustering)
 10. למידה עמוקה בייסיאנית (Gaussian Deep Learning)
- בנושאים 1-7 נניח פילוגים גאוסים, אחרי כן נעבוד עם פילוגים נוספים.

מהי למידה חישובית? "מאפשרת קבלת החלטות על בסיס נתונים".
הגדרה אפשרית - יש לנו מחלקת השערות H , פונקצית הפסד L ונתונים מתויגים $D = \{x_i, y_i\}$.
המטרה שלנו היא למצוא פונקציה מיפוי f_θ שממזערת את $\arg \min_{f_\theta \in H} L(f_\theta, D)$.
חסרונות של ההגדרה הזו:

- מנין מגיעה המחלקה H ?
- מנין מגיעה הפונקציה L ?
- למה פונקציה יחידה? למה צריך לתת רק פרדיקציה אחת?
- חסר מדד של אי ודאות
- מאד מוטל לכיוון של למידה מפקחת (supervised).

מה המשמעות של "בייסיאני"? הכוונה היא לדברים שמשמשים בחוק בייס:

$$P(A|B) = \frac{P(A, B)}{P(B)} = \frac{P(A)P(B|A)}{P(B)}$$

נקרא גם "חוק ההסתברות ההפוכה" (Law of Inverse Probability), כי קל ואינטואיטיבי לחשב את $P(A|B)$ אבל לפעמים קשה ולא אינטואיטיבי לחשב את הכיוון השני (כלומר $P(B|A)$). **דוגמא:** מה הסיכוי להדבק בקורונה בהנתן בדיקות מהירות? ידוע כי הבדיקות אמינות מאד, נניח 95% דיוק סימטרי.

נגדיר מאורעות: "pos", "neg" הפלט של הבדיקה; "+", "-" המצב האמיתי של הנבדק. הנחנו סימטריה ולכן נקבל:

$$P(pos|+) = P(neg|-) = 0.95, \quad P(pos|-) = P(neg|+) = 0.05$$

נרצה לדעת מהו הסיכוי של נבדק אקראי להיות חולה למרות שהתוצאה שלו שלילית:

$$P(+|neg) = \frac{P(+)P(neg|+)}{P(neg)}$$

את $P(neg)$ נחשב לפי נוסחת ההסתברות השלמה:

$$P(B) = \sum_A P(A)P(B|A)$$

כלומר $P(neg) = P(+)P(neg|+) + P(-)P(neg|-)$. נניח כי $P(+) = 0.4$ ונקבל:

$$P(+|neg) = \frac{P(+)P(neg|+)}{P(neg)} = \frac{0.4 \cdot 0.05}{0.4 \cdot 0.05 + 0.6 \cdot 0.95} = \frac{2}{59} \approx 3.4\%$$

לעומת זאת לו היינו לוקחים $P(+) = 0.99$ אז היינו מקבלים $P(+|neg) \approx 1$.

אז מהי בעצם BML?

בהנתן נתונים D או פונקציה שנרצה למצוא f , נחשב את $P(f|D)$ - כלומר נרצה למצוא את הפונקציה f הנכונה.

דוגמא 1 - רגרסיה לינארית:

יש לנו נתונים של ציוני שנה א' מול ציוני הפסיכומטרי של אותם תלמידים. בגישה הקלאסית, H יהיה אוסף הפונקציות הלינאריות ואז נחפש פונקציה יחידה, $f = \arg \min_{f \in H} \sum_i (b_i - f(x_i))^2$. לעומת זאת, בגישה בייסיאנית נחשב את הסיכוי לקבל את כל אחת מהפונקציות האפשריות.

דוגמא 2 - clustering:

במקום הגישה הקלאסית שמחפשת חלוקה אופטימלית ל- k קבוצות, ניתן דרגת ודאות לכל מיני חלוקות על ערכים שונים של k . יתרונות השיטה הבייסיאנית:

1. מאפשרת אי-ודאות
 2. אופטימליות (תחת הנחות)
 3. לא פרמטרית
- חסרונות השיטה הבייסיאנית :
1. דורשת הנחות מוקדמות (פריור)
 2. לעתים החישוב אינו אפשרי