

3.6 סטטיסטיקה מספקת

3.6.1 הגדרת סטטיסטי מספיק

ראינו כי קיימות בעיות בהן אומד ML אינו קיים, אינו יחיד או איננו חסר הטיה, וכן קשה להשתמש באומד בייסיאני. נדון כעת במושג של סטטיסטיקה מספקת (שהוצג לראשונה על ידי פישר ב-1922) ונראה כיצד הוא יכול לסייע באמידת פרמטרים. נגדיר סטטיסטי (statistic) כפונקציה של המשתנים המקריים שאינה תלויה באף פרמטר לא ידוע, דהיינו פונקציה של הערכים הנצפים במדגם בלבד (לדוגמא, ממוצע הדגימות הוא סטטיסטי).

דוגמא

נסיין משועמם מטיל מטבע 50 פעם וסופר את מספר התוצאות "עץ", כדי להעריך את ההסתברות ל"עץ" במטבע, אותה נסמן ב- θ . הוא רושם בקפידה את התוצאות המפורשות של כל ההטלות וכן את מספר הפעמים הכולל שקיבל "עץ". נשים לב כי היות וההטלות הן בלתי תלויות ושוות הסתברות, הרי שמרגע שספר את כמות העצים הכוללת במדגם אין אינפורמציה נוספת בתוצאות הספציפיות של ההטלות שביצע. ובמילים אחרות, סך כל מספר הטלות "עץ" מספק לו את כל המידע שקיים במדגם לגבי הפרמטר אותו הוא מעוניין להעריך.

כדי להשתכנע בכך, נשים לב כי אם ידוע לנו שיש k עצים במדגם, הרי שההסתברות לקבל מדגם מסוים המקיים תנאי זה היא זהה לכל המדגמים, והסתברות מותנית זו אינה תלויה ב- θ . במילים אחרות, בתוך קבוצת המדגמים שבהם יצאו k עצים יש לכל מדגם הסתברות שווה שאינה תלויה ב- θ .

באופן פורמלי, ניתן לרשום במפורש

$$(3.33) \quad \Pr \left((X_1, \dots, X_n) = (x_1, \dots, x_n) \mid \sum_i x_i = k \right) = \begin{cases} \frac{1}{\binom{n}{k}} & \text{if } \sum_i x_i = k \\ 0 & \text{otherwise} \end{cases}$$

כלומר הסיכוי לקבל את המדגם המסוים $X^{(n)}$ בהינתן $\sum_i x_i = k$ אינו תלוי ב- θ .

הגדרה: סטטיסטי מספיק (sufficient statistics)

בהינתן מדגם $X^{(n)} = (x_1, \dots, x_n)$ הלקוח מתוך ההתפלגות $P(X^{(n)}|\theta)$ הסטטיסטי $T(X^{(n)})$ נקרא מספיק עבור הפרמטר (הסקלרי) θ אם מתקיים

$$(3.34) \quad P(X^{(n)} | T(X^{(n)}), \theta) = P(X^{(n)} | T(X^{(n)}))$$

מסקנה

על פי נוסחת בייס

$$(3.35) \quad P(\theta | X^{(n)}, T(X^{(n)})) = P(\theta | T(X^{(n)}))$$

כלומר, ההתפלגות של הפרמטר אינה תלויה בכל נקודות המדגם אלא רק בערך של הסטטיסטי. נעיר כי לעתים קרובות הפרמטר בו אנחנו מעוניינים הוא רב ממדי, ואז גם הסטטיסטי המספיק הוא רב ממדי. נדון במקרה זה בסעיף הבא.

אנו רואים אם כן כי כאשר קיים סטטיסטי מספיק עבור פרמטר, הרי שיש לנו מכשיר חזק המאפשר למצות את הידע הרלבנטי מתוך המדגם למספר מצומצם של סטטיסטים. הסטטיסטיקאי המחפש דרכים לאמוד את הפרמטר יכול להתמקד באומדים שהם פונקציה של הסטטיסטי המספיק בלבד.

איך מוצאים סטטיסטים מספיקים? בדוגמא מתחילת הפרק, הראנו במפורש את העובדה כי ההתפלגות המותנית בסטטיסטי אינה תלויה בפרמטר, אך בדרך כלל קשה להראות את אי התלות הזו באופן ישיר. המשפט הבא מספק דרך קונסטרוקטיבית למציאה של סטטיסטים מספיקים.

3.6.2 משפט הפירוק (Factorization criterion)

משפט

בהינתן מדגם $X^{(n)} = (X_1, \dots, X_n)$ הלקוח מתוך ההתפלגות $f(X^{(n)} | \theta)$, סטטיסטי $T(X^{(n)})$ נקרא מספיק עבור הפרמטר θ אם ורק אם לכל ערך של x ולכל ערך של $\theta \in \Omega$ ניתן לפרק ההתפלגות f ולכתוב אותה כ- $f(X^{(n)} | \theta) = h(X^{(n)})g(T, \theta)$, כאשר h היא פונקציה של המדגם בלבד, ו- g תלויה במדגם רק דרך התלות בסטטיסטי T .

הוכחה

נוכיח כאן עבור משתנים מקריים בדידים.

← כיוון ראשון:

נניח כי ניתן לפרק את P כנדרש. לכל ערך t של הסטטיסטי נסמן ב- $A(t)$ את כל המדגמים האפשריים עבורם מתקיים $T(X^{(n)}) = t$. לכל מדגם כזה המקיים $x \in A(t)$ נרשום

$$\begin{aligned}
 \Pr(X = x | T = t, \theta) &= \frac{\Pr(X = x | \theta)}{\Pr(T = t | \theta)} \\
 &= \frac{f(X = x | \theta)}{\sum_{y \in A(t)} f(y | \theta)} \\
 (3.36) \quad &= \frac{h(x)g(t, \theta)}{\sum_{y \in A(t)} h(y)g(t, \theta)} \\
 &= \frac{h(x)}{\sum_{y \in A(t)} h(y)}
 \end{aligned}$$

כאשר השתמשנו בכך שניתן לפרק את ההתפלגות כנדרש, ובכך שאנחנו מתמקדים במדגמים השייכים ל- $A(t)$, דהיינו המדגמים עבורם $T(X^{(n)}) = t$. קיבלנו כי ההתפלגות המותנית אינה תלויה ב- θ .

→ כיוון שני:

נניח ש- T הוא סטטיסטי מספיק, אז לכל ערך של t , x ו- θ , ההתפלגות המותנית $P(X | T, \theta)$ אינה תלויה ב- θ ולכן היא מהצורה $P(X | T, \theta) = P(X | T) = h(X)$. כעת, אם נגדיר את g להיות $g(T, \theta) = \Pr(T | \theta)$ אז נקבל $P(X | \theta) = P(X | T, \theta) \cdot P(T | \theta) = h(X)g(T, \theta)$. כנדרש. מכאן שניתן לפרק את P כנדרש.

דוגמא 1: התפלגות נורמלית

יהיו X_1, \dots, X_n דגימות מהתפלגות נורמלית עם שונות σ^2 , ותוחלת μ $X_i \sim N(\mu, \sigma^2)$. אז פונקציית הצפיפות המשותפת היא

$$f(X^{(n)} | \mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

וניתן לכתוב אותה כ-

$$(3.37) \quad f(X^{(n)} | \mu) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^n x_i^2\right) \exp\left(\frac{\mu}{\sigma^2} \sum_{i=1}^n x_i - \frac{n\mu^2}{2\sigma^2}\right)$$

מכאן שפרקנו אותה לגורם שאינו תלוי ב- μ , ולגורם שתלוי ב- μ וב- $\sum_{i=1}^n X_i$ בלבד,

אז אם נגדיר $T(X) = \sum_{i=1}^n X_i$ נקבל ש- T הוא סטטיסטי מספיק עבור התוחלת μ .

דוגמא 2: התפלגות אחידה בקטע סגור

יהי X מ"מ המתפלג אחיד בקטע $[0, \theta]$

$$f(x | \theta) = \begin{cases} \frac{1}{\theta} & \text{for } 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

פונקצית הצפיפות המשותפת של n דגימות היא

$$f(x_n | \theta) = \begin{cases} \frac{1}{\theta^n} & \text{for } 0 \leq x_1, \dots, x_n \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

נשים לב שיש הסתברות אפס לקבל מדגם שבו קיימת דגימה שערכה קטן מאפס $x_i < 0$. לכן מספיק לתאר פירוק עבור המקרה בו כל הדגימות גדולות מאפס.

נגדיר את הסטטיסטי $T(X^{(n)}) = \max(x_1, \dots, x_n)$ וכן נגדיר את הפונקציה הבאה:

$$h(T, \theta) = \begin{cases} 1 & \text{if } T \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

כעת נשים לב כי היות ו- $\max(x_1, \dots, x_n) \leq \theta$ אם ורק אם כל הדגימות מקיימות $x_i \leq \theta$, מתקיים כי

$$f_n(x | \theta) = 1 \cdot \frac{1}{\theta^n} \cdot h(T, \theta)$$

וקיבלנו כי אגף ימין תלוי במדגם רק דרך הסטטיסטי $T = \max(x_1, \dots, x_n)$ ולכן הוא סטטיסטי מספיק.

טענה

אם קיים סטטיסטי מספיק עבור פרמטר θ , אזי אומד ML לפרמטר θ הוא פונקציה של הסטטיסטי המספיק לפרמטר.

הוכחה

θ_{ML} הוא הערך של θ שמביא למקסימום את $f(X^{(n)} | \theta)$, ועל פי משפט הפירוק הוא מביא למקסימום את $g(T, \theta)$, ולכן הוא בהכרח פונקציה של הסטטיסטי ולא של המדגם.

הערה: פעמים רבות ה- MLE יהיה בעצמו סטטיסטי מספיק.

3.6.3 סטטיסטים מספיקים במשותף (Jointly sufficient statistics)

במקרים רבים, ובמיוחד כאשר ההתפלגות מאופיינת על ידי וקטור פרמטרים θ (כמו למשל ההתפלגות הנורמלית המאופיינת על ידי שני פרמטרים), לא נוכל למצוא סטטיסטי מספיק בודד עבור כל הפרמטרים. במקרים אלו דרושים לנו מספר סטטיסטים שביחד מהווים סטטיסטיקה מספקת.

הגדרה

הסטטיסטים T_1, \dots, T_k הם סטטיסטים מספיקים במשותף עבור הפרמטר θ אם ורק אם לכל $x \in \mathbb{R}^n$ ולכל $\theta \in \Omega$ ניתן לפרק את ההתפלגות באופן הבא

$$(3.38) \quad f_n(x | \theta) = g(x)h(T_1, \dots, T_k, \theta)$$

כאשר התלות היחידה של h במדגם היא דרך הסטטיסטים T_1, \dots, T_k .

הגדרה: סטטיסטיים מספיקים מינימליים

הסטטיסטים T_1, \dots, T_k הם סטטיסטים מספיקים מינימליים, אם כל קבוצה אחרת של סטטיסטים מספיקים במשותף היא ממימד גדול או שווה ל- k .

נשים לב כי באופן טריוויאלי, נקודות המדגם עצמו x_1, \dots, x_n מהוות סטטיסטים מספיקים במשותף עבור הפרמטרים, אך ברור כי פתרון כזה אינו מעניין אותנו. בנוסף לכך, כאשר הדגימות הן בלתי תלויות אז סטטיסטי הסדר

$$\min(X), 2^{nd} \min(X), \dots, \max(X)$$

גם הם מהווים סטטיסטים מספיקים במשותף. ניתן להראות כי קיימות התפלגויות (למשל התפלגות קושי) עבורן סטטיסטי הסדר מהווים את הסטטיסטים המספיקים המינימליים ולא ניתן להגיע לקבוצה של סטטיסטים מספיקים שהמימד שלה קטן מממד המדגם.

3.6.4 משפט PKD

משפט (Pitman, Koopman, Darmois)

קיימים סטטיסטים מספיקים עבור סט הפרמטרים θ אם ורק אם קיימות פונקציות B, A ו- C כך שניתן לרשום את התפלגות המדגם על ידי

$$(3.39) \quad f_n(X^{(n)} | \theta) = \exp\left(A(X^{(n)}) + B(\theta)C(X^{(n)})\right)$$

הוכחה

← כיוון ראשון:

אם קיימת הצגה כנדרש אז מתקיים משפט הפירוק עם הסטטיסטי $C(X^{(n)})$, היות ו-

$$(3.40) \quad f_n(X^{(n)} | \theta) = \exp\left(A(X^{(n)})\right)\exp\left(B(\theta)C(X^{(n)})\right)$$

→ כיוון שני:

כיוון זה קשה יותר להוכחה ונביא כאן את הוכחה של Pitman הנכונה תחת הנחות מסוימות. יהי $T(X_1, \dots, X_n)$ סטטיסטי מספיק עבור הפרמטר θ , אזי לפי משפט הפירוק $f(X^{(n)} | \theta) = h(X^{(n)})g(T, \theta)$ וניתן לרשום את נגזרת הנראות

$$(3.41) \quad \begin{aligned} \frac{\partial}{\partial \theta} \log(f_n(X^{(n)} | \theta)) &= \sum_{i=1}^n \frac{\partial}{\partial \theta} \log(f(x_i | \theta)) \\ &= \sum_{i=1}^n \frac{\partial}{\partial \theta} \log(g(T, \theta)) = k(T, \theta) \end{aligned}$$

כאשר הפונקציה k הפיכה, ניתן (לכל ערך של θ) להציג את T כפונקציה של נגזרת לוג הנראות. דהיינו אם נסמן

$$(3.42) \quad \underline{B}(X_1, \dots, X_n) = \sum_{i=1}^n B(X_i) = \sum_{i=1}^n \frac{\partial \log(f(X_i | \theta))}{\partial \theta}$$

נוכל לרשום

$$(3.43) \quad T(X_1, \dots, X_n) = F(\underline{B}(X_1, \dots, X_n)).$$

מהצבה חזרה בתוך 3.41 נקבל

$$(3.44) \quad \sum_{i=1}^n \frac{\partial \log(f(X_i | \theta))}{\partial \theta} = k(T, \theta) \equiv K(\underline{B}, \theta)$$

כעת לכל X_i בסכום נגזור ונקבל

$$\frac{\partial^2 \log(f(X_i | \theta))}{\partial \theta \cdot \partial X_i} = \frac{\partial K}{\partial \underline{B}}(\underline{B}, \theta) \frac{dB(X_i)}{dX_i}$$

והיות ואגף שמאל תלוי רק ב- X_i , הרי שכך גם אגף ימין. אולם $\frac{\partial K}{\partial \underline{B}}$ סימטרית ב-

X_1, \dots, X_n ולכן בלתי תלויה בהם, ומכאן שניתן לרשום את אגף ימין בצורה

$$\frac{\partial^2 \log(f(X_i | \theta))}{\partial \theta \cdot \partial X_i} = \alpha(\theta) \frac{dB(X_i)}{dX_i}$$

ועל ידי אינטגרציה לפי X_i נקבל

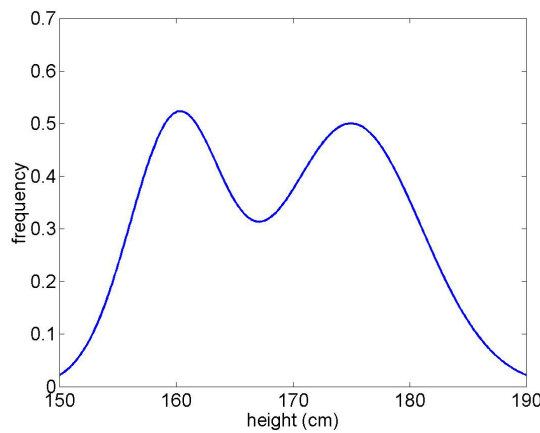
$$(3.45) \quad \frac{\partial \log(f(X_i | \theta))}{\partial \theta} = \alpha(\theta)B(X_i) + \beta(\theta)$$

ואינטגרציה נוספת לפי θ תתן את הצורה האקספוננציאלית כנדרש.

3.7 משתנים חבויים ותערובות של התפלגויות

3.7.1 מבוא

עד כה דנו באמידה של התפלגויות שהן ידועות למעט מספר פרמטרים, כלומר הנחנו כי ידועה לנו ההתפלגות $P(x|\theta)$, והראנו כי אם התפלגות זו היא ממשפחה אקספוננציאלית אזי ניתן למצות את האינפורמציה במדגם באמצעות מספר קטן יחסית של פונקציות של המדגם (סטטיסטיים מספיקים). במציאות, התצפיות עשויות להגיע מתוך התפלגויות שאינן אקספוננציאליות, ומקרה שכיח כזה הוא כאשר הנתונים מגיעים מתוך תערובת של התפלגויות. לדוגמא, ידוע שגובהם של גברים מתפלג נורמלית בקירוב, וכך גם הגובה של נשים, אך ממוצע הגובה של גברים גבוה מזה של נשים. אם נמדוד את התפלגות הגבהים באוכלוסיה, נקבל תערובת של שתי התפלגויות נורמליות שתראה כמו באיור הבא



נשים לב כי אם היתה ידועה לנו האוכלוסיה שאליה שייכת כל דגימה (למשל האם הגובה הוא של גבר או של אישה), הרי שהיינו יכולים לאמוד כל אחת מההתפלגויות בנפרד. אך במקרה בו האינפורמציה הזו חסרה ואין בידינו את התיוג של כל דגימה לא קיים פתרון כללי מלא ויעיל לאמידת הפרמטרים החסרים. עם זאת, בפרק זה נתאר משפחה של אלגוריתמים המבטיחים התכנסות לפתרון אופטימלי מקומי, שבמקרים רבים הוא פתרון טוב בהחלט. נתאר יישום של אלגוריתמים אלו לבעיה נפוצה בה יש למצוא משתנים חבויים: חלוקת דגימות לצברים.

3.7.2 אמן פרמטרים במידע חלקי (Expectation Maximization)

אלגוריתמי EM (Dempster et al 1977) הם משפחה של אלגוריתמים למציאת מקסימום נראות תחת אינפורמציה לא מלאה. הניסוח הכללי של הבעיה מאפשר לפתור בעיות מגוונות ביותר על ידי אלגוריתמים דומים, ולהוכיח את התכונותם למקסימום מקומי של הנראות.

הגדרות

יהיו $\mathbf{x} = \{x_1, \dots, x_n\}$ תצפיות נתונות ו- $\mathbf{h} = \{h_1, \dots, h_k\}$ אוסף משתנים חבויים (hidden variables) שערכיהם לא ידועים לנו, ונסמן את המידע המלא ב- $\mathbf{y} = \{\mathbf{x}, \mathbf{h}\}$. אנו מניחים כי ידועה לנו הצורה הכללית של התפלגות התצפיות, עד כדי המידע החבוי והפרמטרים שאותם אנחנו רוצים לאמוד, דהיינו ידועה לנו ההתפלגות $P(x|h, \theta)$. כמו בפרקים קודמים, מטרתנו היא למצוא את ערכי הפרמטרים θ שמביאים למקסימום את הנראות $P(x|\theta)$.

לצורך ההמחשה נחזור לדוגמא של התפלגות הגבהים באוכלוסייה. כאן x יהיו הגבהים שאנחנו מודדים, h יהיו תיוגים המסמנים לכל מדידת גובה האם היא של גבר או אישה, ו- θ יהיה משתנה מקרי בן ארבעה פרמטרים המכיל את התוחלת והשונות של ההתפלגויות הנורמליות של גבהי גברים ושל גבהי הנשים (נניח כי משקלם היחסי באוכלוסייה שווה). הסיכוי $P(x|\theta)$ לצפות בדגימה של גובה x היא ממוצע משוקלל של הסיכויים לצפות ב- x עבור גבר ועבור אישה.

אלגוריתמי EM

המטרה של אלגוריתם EM היא להביא למקסימום את הנראות הנצפית $\log P(\mathbf{x}|\theta)$ אך הוא עושה זאת באופן בלתי ישיר ועובד על הנראות המלאה $\log P(\mathbf{y}|\theta)$. לצורך האלגוריתם מגדירים את פונקצית העזר הבאה המודדת את תוחלת הנראות ביחס להתפלגות המשתנים החבויים, בהנתן התצפיות \mathbf{x} וערכים מסוימים $\hat{\theta}_n$ לפרמטרים

$$Q(\theta, \hat{\theta}_n) = E_{P(\mathbf{h}|\hat{\theta}_n)} [\log P(\mathbf{y}|\theta)] = \int_{\{\mathbf{h}\}} \log P(\mathbf{y}|\theta) P(\mathbf{h}|\hat{\theta}_n) d\mathbf{h}$$

פונקציה זו היא פונקציה של הפרמטרים שאותם אנחנו מחפשים θ , ושל סט ערכים של הפרמטרים $\hat{\theta}_n$ שאותו אנחנו "מנחשים".

האלגוריתם מתבצע באופן איטרטיבי: בתחילה מציעים ניחוש ל- $\hat{\theta}_1$, ובכל שלב אומדים את $\hat{\theta}_{n+1}$ מחדש על בסיס האומד בצעד הקודם $\hat{\theta}_n$, וזאת באופן הבא

E-step: בהנתן הערכים הנוכחיים של הפרמטרים $\hat{\theta}_n$,

חשב את תוחלת הנראות $Q(\theta, \hat{\theta}_n)$.

M-step: חשב את הערך של $\hat{\theta}_{n+1}$ המקיים

$$\hat{\theta}_{n+1} = \arg \max_{\theta} (Q(\hat{\theta}_n, \theta))$$

נשים לב שבשני השלבים, המיצוע והמקסימיזציה נעשים בהנחת אינפורמציה מלאה על $\log P(\mathbf{y}|\theta)$ וחישוב זה פשוט במיוחד כאשר $P(\mathbf{y}|\theta)$ ממשפחה אקספוננציאלית.

התכנסות אלגוריתם EM

נראה כעת כי אלגוריתם EM מתכנס למקסימום לוקלי של הנראות. ראשית נשים לב כי מתקיים

$$P(\mathbf{y}|\theta) = \frac{P(\mathbf{h}, \mathbf{x}, \theta)}{P(\theta)} = \frac{P(\mathbf{h}, \mathbf{x}, \theta)}{P(\mathbf{x}, \theta)} \cdot \frac{P(\mathbf{x}, \theta)}{P(\theta)}$$

$$= P(\mathbf{y}|\mathbf{x}, \theta) \cdot P(\mathbf{x}|\theta)$$

ולכן

$$P(\mathbf{x}|\theta) = \frac{P(\mathbf{y}|\theta)}{P(\mathbf{y}|\mathbf{x}, \theta)}$$

או עבור הלוגריתמים

$$\log P(\mathbf{x}|\theta) = \log P(\mathbf{y}|\theta) - \log P(\mathbf{y}|\mathbf{x}, \theta)$$

כעת נביט על הממוצע של שני האגפים על פני התפלגות המשתנים החבויים $P(\mathbf{y}|\mathbf{x}, \theta) = P(\mathbf{h}|\theta)$, היות ו- $P(\mathbf{x}|\theta)$ לא תלוי ב- \mathbf{h} אז אגף שמאל נשאר ללא שינוי, ונקבל

$$\begin{aligned} \log P(\mathbf{x} | \theta) &= \\ &= \int_{\{\mathbf{h}\}} \log(P(\mathbf{y} | \theta)) P(\mathbf{h} | \hat{\theta}_n) d\mathbf{h} - \int_{\{\mathbf{h}\}} \log P(\mathbf{y} | \mathbf{x}, \theta) P(\mathbf{h} | \hat{\theta}_n) d\mathbf{h} \\ &= Q(\theta, \hat{\theta}_n) - \int_{\{\mathbf{h}\}} \log(P(\mathbf{h} | \theta)) P(\mathbf{h} | \hat{\theta}_n) d\mathbf{h} \end{aligned}$$

ולשם הקיצור נסמן

$$L(x | \theta) = Q(\theta, \hat{\theta}_n) - H(\theta, \hat{\theta}_n)$$

$$L(x | \theta_{n+1}) \geq L(x | \theta_n) \quad \text{טענה:}$$

הוכחה

ההפרש בין הנראויות במצבים עוקבים יהיה

$$\begin{aligned} L(x | \theta_{n+1}) - L(x | \theta_n) &= \\ &= [Q(\theta_n, \theta_{n+1}) - Q(\theta_n, \theta_n)] - [H(\theta_n, \theta_{n+1}) - H(\theta_n, \theta_n)] \end{aligned}$$

נוכיח תחילה כי

$$Q(\theta_n, \theta_{n+1}) \geq Q(\theta_n, \theta_n)$$

ולאחר מכן כי

$$H(\theta_n, \theta_{n+1}) \leq H(\theta_n, \theta_n)$$

אם נוכיח כל אחד מן החלקים הללו ביחד יוכח אי השוויון, ובכך הוכחה הטענה המהותית של אלגוריתם ה-EM. ההוכחה של החלק הראשון (העלייה של הפונקציה Q ככל שמתקדמות האיטרציות של האלגוריתם) נובעת ישירות מתוך ההגדרה של שלב המקסימיזציה (M-step).
על מנת להוכיח את החלק השני נרשום

$$\begin{aligned} H(\theta_n, \theta_{n+1}) - H(\theta_n, \theta_n) &= \\ &= E_{k(y|\theta_n)} \log \left[\frac{P(y | x, \theta_{n+1})}{P(y | x, \theta_n)} \right] = \\ &= -E_{k(y|\theta_n)} \log \left[\frac{P(y | x, \theta_n)}{P(y | x, \theta_{n+1})} \right] = \\ &= -D[P(y | x, \theta_n) \| P(y | x, \theta_{n+1})] \leq 0 \end{aligned}$$

$$L(x|\theta_{n+1}) - L(x|\theta_n) = \underbrace{(Q(\theta_n|\theta_{n+1}) - Q(\theta_n|\theta_n))}_{\geq 0} - \underbrace{(H(\theta_n|\theta_{n+1}) - H(\theta_n|\theta_n))}_{\leq 0} \geq 0$$

הוכחנו אם כן כי הנראות $P(x|\hat{\theta}_n)$ גדלה בכל צעד של אלגוריתם EM ולכן בהכרח מתכנסת למקסימום מקומי כנדרש.

כפי שניתן לראות, הוכחת ההתכנסות לעיל מתקיימת בתנאים כלליים ביותר. על מנת להדגים שימוש מעשי באלגוריתם ממשפחת EM, נתאר כעת את אחת הבעיות החשובות ביותר בלמידה לא מונחית – בעיית החלוקה לצברים. נפתח את הפתרון לבעיית החלוקה לצברים ונראה כי הוא שקול לאלגוריתם EM.

3.7.3 חלוקה לצברים של תערובת התפלגויות: K-means

בעיית החלוקה לצברים היא בעיה קלאסית של למידה לא מונחית: נתונות לנו דגימות המגיעות מתוך תערובת של התפלגויות, ואנחנו צריכים להכריע לגבי כל דגימה לאיזו התפלגות היא שייכת. המשתנים החבויים במקרה זה הם התיגו של כל אחת מהדגימות לצבר המתאים לה. בסעיף הנוכחי נתאר אלגוריתם נפוץ לפתרון הבעיה, המשיג ביצועים טובים במקרה שההתפלגויות הן נורמליות.

תיאור האלגוריתם

נניח כי נתונות לנו דגימות המגיעות מתערובת של c התפלגויות נורמליות בעלות שוניות זהות, אך התוחלות של ההתפלגויות אינן ידועות. אלגוריתם איטרטיבי למציאת התוחלות יהיה

1. בחר ערכים התחלתיים לתוחלות המבוקשות $\hat{\mu}_1, \dots, \hat{\mu}_c$.
2. סווג את n הדגימות למחלקות הקרובות אליהן ביותר.
3. חשב מחדש את הממוצעים של כל אחת מן המחלקות על פי הדגימות המרכיבות אותן
4. בדוק אם הערך הממוצע של המרחק השתנה (ירד) - אם כן חזור לשלב מספר 2 אחרת סיים.

לאלגוריתם המתואר קיימות שתי גרסאות עיקריות. על פי הגרסה הרכה (**soft clustering**) מסווגים את הדגימות למחלקות על פי ההסתברות שלהן ושלב השערוך (שלב 3 באלגוריתם) מבוצע על ידי ממוצע משוקלל של הדגימות על פי הסתברויות אלו. על פי הגרסה השניה, ה"קשיחה" (**hard clustering**), משייכים באופן דטרמיניסטי את הדגימות למחלקות שלהן (דהיינו בוחרים את המחלקה הכי סבירה θ^* , ואז קובעים $P(x|\theta^*)=1$).

נעבור כעת לתיאור אלגוריתם מסוג EM לבעיית החלוקה לצברים, ונראה כיצד מתקבל פתרון השקול לגרסה הרכה של אלגוריתם K-means.

3.7.4 חלוקה לצברים של תערובת התפלגויות - גישת ML

כפי שתיארנו לעיל, בבעיית החלוקה לצברים נתונות לנו דגימות המגיעות מתוך תערובת של התפלגויות, ואנחנו צריכים להכריע לגבי כל דגימה לאיזו התפלגות היא שייכת. כצפוי, איכות הפתרון והסיבוכיות החישובית הנדרשת להשגת תלויות בידע המוקדם ובאילוצים הקיימים על הבעיה. נטפל בתחילה במקרה בו קיימות המגבלות הבאות

1. הדגימות מגיעות מהתפלגות בעלת מספר ידוע של רכיבים c , שאותם נסמן w_1, \dots, w_c .
2. הצורה הפרמטרית של ההסתברות בכל רכיב $P(x|\omega_j, \theta_j)$ ידועה ואקספוננציאלית (עבור $j=1, \dots, c$).
3. הגדלים היחידים שאינם ידועים הוא הערכים של וקטורי הפרמטרים $\theta_1, \dots, \theta_c$.

הדגימות נוצרות באופן הבא: ראשית נבחר רכיב התערובת ממנו נלקחות הדגימות ω_j , בהסתברות $P(\omega_j)$. לאחר מכן נבחרות הדגימות על פי ההתפלגות $P(x|\omega_j, \theta_j)$. בסיכומו של דבר ההסתברות לקבל אוסף דגימות x בהנתן הפרמטרים θ היא

$$p(x|\theta) = \sum_{j=1}^c p(x|\omega_j, \theta_j) P(\omega_j)$$

פונקצית ההתפלגות זאת נקראת צפיפות התערובת mixture density.

מטרת תהליך החלוקה לצברים הוא להשתמש בדגימות אשר נלקחו מפונקצית פילוג המתוארת לעיל ולשערך מתוכן את ווקטור הפרמטרים הבלתי ידוע θ . על מנת שניתן יהיה לבצע פעולה זאת על פונקציות ההתפלגות להיות "ברת-זיהוי" (identifiable), כאשר פונקציה היא ברת זיהוי אם $\theta \neq \theta'$ אז קיים x כך ש- $p(x|\theta) \neq p(x|\theta')$. דוגמא לפונקציה אשר אינה ברת זיהוי היא התערובת הבינומיאלית הבאה

$$P(x|\theta) = \frac{1}{2} \theta_1^x (1-\theta_1)^{1-x} + \frac{1}{2} \theta_2^x (1-\theta_2)^{1-x}$$

$$= \begin{cases} \frac{1}{2}(\theta_1 + \theta_2) & \text{if } x=1 \\ 1 - \frac{1}{2}(\theta_1 + \theta_2) & \text{if } x=0 \end{cases}$$

ניתן לראות שפונקציה זו תלויה למעשה רק בסכום של שני הפרמטרים ולכן לא ניתן להפריד ביניהם אלא להסיק רק על סכומם. למשל אם אנחנו יודעים ש- $P(x=1|\theta) = 0.6$ ולכן נדע כי $P(x=0|\theta) = 0.4$. אנחנו אמנם יכולים לקבוע את הפונקציה $P(x|\theta)$ אולם לא נוכל לקבוע את θ . המירב שנוכל לדעת הוא שהסכום של $\theta_1 + \theta_2$ הוא 1.2. אנו נדון מכאן ואילך רק בפונקציות אשר הן ברות זיהוי.

שערוך ניראות מירבית

נניח כי נתון לנו אוסף $X^{(n)} = \{x_1, \dots, x_n\}$ של דגימות ללא סימון אשר נלקחו בצורה בלתי תלויה מתוך צפיפות התערובת הבאה

$$P(x|\theta) = \sum_{j=1}^c P(x|\omega_j, \theta_j) P(\omega_j)$$

הנראות של הדגימות היא על פי הגדרה

$$P(X^{(n)}|\theta) = \prod_{k=1}^n P(x_k|\theta)$$

והנראות המרבית תתקבל עבור הערך של הפרמטרים $\theta = (\theta_1, \dots, \theta_c)$ אשר יביא למקסימום את הפונקציה $P(X^{(n)}|\theta)$. נסמן ב- l את הלוגריתם של פונקציית הנראות

$$l = \sum_{k=1}^n \log p(x_k|\theta)$$

וב- $\nabla_{\theta_i} l$ את וקטור הנגזרת הכיוונית (הגרדיאנט) של l ביחס ל θ_i

$$\nabla_{\theta_i} = \frac{\partial}{\partial \theta_i}$$

אז

$$\nabla_{\theta_i} l = \sum_{k=1}^n \frac{1}{p(x_k|\theta)} \nabla_{\theta_i} \left[\sum_{j=1}^c p(x_k|\omega_j, \theta_j) P(\omega_j) \right]$$

נניח כי האלמנטים של θ_i ו- θ_j הם בלתי תלויים אם $i \neq j$ ונשתמש בנוסחת בייס

$$P(\omega_i|x_k, \theta) = \frac{p(x_k|\omega_i, \theta_i) P(\omega_i)}{p(x_k|\theta)} \quad \forall i$$

$$\frac{1}{p(x_k|\theta)} = \frac{P(\omega_i|x_k, \theta)}{P(\omega_i) p(x_k|\omega_i, \theta)}$$

אז נוכל לכתוב את הגרדיאנט של לוגריתם הנראות בצורה

$$\nabla_{\theta_i} l = \sum_{k=1}^n P(\omega_i|x_k, \theta) \nabla_{\theta_i} \log p(x_k|\omega_i, \theta)$$

וחישוב הנראות המרבית יוביל אותנו להשוות את הגרדיאנט ל-0 דהיינו לקבל את האומד $\hat{\theta}$ עבור θ

$$\sum_{k=1}^n P(\omega_i|x_k, \hat{\theta}) \nabla_{\theta_i} \log p(x_k|\omega_i, \hat{\theta}) = 0 .$$

נדגים כעת את הפתרון הכללי שמצאנו למקרה של תערובת של שתי התפלגויות נורמליות.

3.7.5 פתרון נראות מרבית עבור תערובת שתי התפלגויות נורמליות

נתונות לנו תצפיות מתוך תערובת של התפלגויות נורמליות רב מימדיות שעבורן ידוע לנו הכל פרט לווקטור התוחלות ברכיבי התערובות השונים μ_i (כלומר, מטריצת השונות המשותפת Σ_i ידועה והפרמטרים שיש לאמוד θ_i הם התוחלות). ניתן לכתוב את לוגריתם הנראות לכל רכיב בתערובת כ-

$$\begin{aligned} \log P(x | \omega_i, \mu_i) &= \\ &= -\log \left[(2\pi)^{d/2} |\Sigma_i|^{1/2} \right] - \frac{1}{2} (x - \mu_i)^t \Sigma_i^{-1} (x - \mu_i) \end{aligned}$$

ולכן

$$\nabla_{\mu_i} \log p(x | \omega_i, \mu_i) = \Sigma_i^{-1} (x - \mu_i)$$

לפי הנוסחה שפיתחנו קודם לכן עבור נראות מרבית ניתן לכתוב כי

$$\sum_{k=1}^n P(\omega_i | x_k, \hat{\mu}) \Sigma_i^{-1} (x_k - \hat{\mu}_i) = 0$$

אחרי מכפלה במטריצה Σ_i וארגון גורמים מחדש נקבל

$$\hat{\mu}_i = \frac{\sum_{k=1}^n P(\omega_i | x_k, \hat{\mu}) x_k}{\sum_{k=1}^n P(\omega_i | x_k, \hat{\mu})}$$

הנוסחה המתקבלת מראה כי השערוך עבור μ_i הוא שקלול של ממוצע הדגימות על פי ההסתברות שדגימה מסוימת שייכת לתערובת המשוערכת. אולם למרבה הצער לא ניתן להשתמש בצורה אנליטית ישירה בנוסחה המתקבלת על מנת לקבל ביטוי סגור לפרמטרים המבוקשים, וזאת משום שאם נציב

$$P(\omega_i | x_k, \hat{\mu}) = \frac{p(x_k | \omega_i, \hat{\mu}) P(\omega_i)}{\sum_{j=1}^c p(x_k | \omega_j, \hat{\mu}) P(\omega_j)}$$

ו- $p(x | \omega_i, \mu_i) \propto N(\hat{\mu}_i, \Sigma_i)$ נקבל ביטוי סבוך של משוואות לא-ליניאריות, אשר בדרך כלל אין להן פתרון יחיד.

פתרון אפשרי לבעיה יהיה להתחיל עם ניחוש התחלתי של הפרמטרים μ_i^0 ולהמשיך באלגוריתם איטרטיבי בו

$$\mu_i^{t+1} = \frac{\sum_{k=1}^n P(\omega_i | x_k, \mu_i^t) x_k}{\sum_{k=1}^n P(\omega_i | x_k, \mu_i^t)}$$

האלגוריתם האמור משפר בכל צעד את לוגריתם הנראות (מתוך מה שהראינו לעיל) ובסופו של דבר חייב להתכנס. בפועל נכריז על התכנסות אם בשתי איטרציות עוקבות לא חל שיפור משמעותי בלוגריתם הנראות.

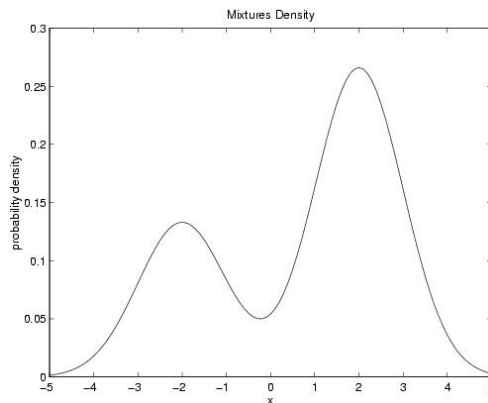
נשים לב כי קיבלנו את גרסה הרכה של אלגוריתם K-means שתואר בתחילת הפרק. הניתוח לעיל מראה כי ניתן לבצע תהליך דומה עבור הערכת השוניות, ולהשתמש באלגוריתם איטרטיבי המעריך בכל צעד את התוחלות ואת השוניות עד להתכנסות. אלגוריתמים אלו ידועים במספר וריאציות ושמות: VQ, Iso-Data, GMM (Gaussian Mixture Models), (Vector Quantization).

דוגמא

נסתכל על התערובת הבאה של שתי התפלגויות נורמליות

$$p(x | \mu_1, \mu_2) = \frac{1}{3\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x - \mu_1)^2\right] + \frac{1}{3\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x - \mu_2)^2\right]$$

צפיפות ההתפלגות של הפונקציה מופיעה באיור הבא

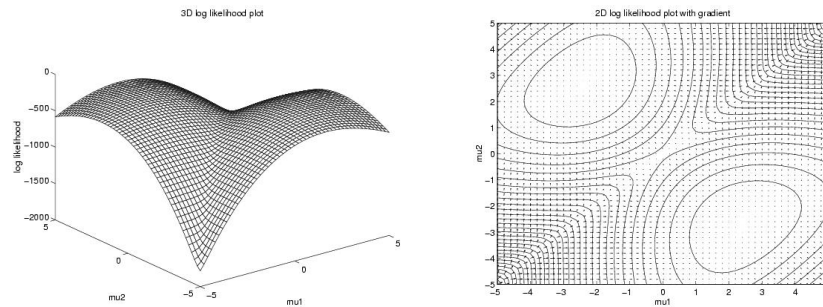


איור 2: התפלגות תערובת של גאוסיאנים

אם נסתכל על לוגריתם הנראות של הפונקציה

$$l(\mu_1, \mu_2) = \sum_{k=1}^n \log p(x_k | \mu_1, \mu_2)$$

ונצייר אותו עבור אוסף של דגימות (איור 3)



איור 3: לוג הנראות כפונקציה של פרמטר התוחלות

מתוך הציורים לעיל ניתן להסיק את המסקנות הבאות:
 קיימת נקודת התכנסות עבור המרחב הנתון ברביע השמאלי העליון אשר מייצגת בצורה מקורבת את הפתרון הנכון של הבעיה $\mu_1 = 2, \mu_2 = -2$.
 קיימת נקודת התכנסות חזקה נוספת והיא זאת הנמצאת ברביע הימני התחתון אשר מייצגת בצורה מקורבת את הפתרון הבא של הבעיה $\mu_1 = 2, \mu_2 = -2$. כלומר היפוך הפרמטרים (מכיוון שהפרמטרים מוגדרים היטב ואין להחליף ביניהם זהו פתרון שגוי). ניתן לראות שכל נקודות ההתחלה מתחת למשולש הראשי יתכנסו לפתרון זה.

קיים פתרון שלישי לבעיה - אם כי פתרון בלתי יציב - הפתרון נמצא בראשית הציורים ואפשר לראות שם נקודת אוקף. אם מתחילים את החיפוש בדיוק בנקודה הנמצאת על האלכסון הראשי תהיה התכנסות לנקודה זו. פתרון זה נובע מכך שאם הערכים ההתחלתיים של הפרמטרים $\hat{\mu}_1(0) = \hat{\mu}_2(0)$ אז $P(\omega_i | x_k, \hat{\mu}_1(0), \hat{\mu}_2(0)) = P(\omega_i)$ והשימוש בנוסחת השערוך תיתן בכל אחד מן השלבים את התוחלת של כל הדגימות עבור $\hat{\mu}_1$ ו- $\hat{\mu}_2$.

תרגילים

1. יהי $X^{(n)} = (x_1, \dots, x_n)$ מדגם אקראי מתוך הפילוגים הבאים:

א. פואסוני $P(x|\theta) = \frac{\theta^x}{x!} e^{-\theta}$

ב. אקספוננציאלי $P(x|\theta) = \frac{1}{\theta} e^{-x/\theta}$

ג. $P(x|\theta) = \frac{1}{2} \exp[-|x - \theta|]$

ד. $P(x|\theta) = \theta \cdot x^{\theta-1} \quad 0 \leq x \leq 1$

מצאו את אומד ה-ML בכל אחד מהמקרים.

2. מצאו אומד ML עבור הפרמטר הדו ממדי (μ, σ) של משתנה מקרי נורמלי.

3. מחיר האומדן.

א. הראו כי עבור $\lambda(\hat{\theta}|\theta) = (\hat{\theta} - \theta)^2$ האומד הבייסיאני האופטימלי בהנתן

מדגם נתון $X^{(n)}$ הוא התוחלת המותנית $\hat{\theta} = E(\theta | X^{(n)}) = \int_{-\infty}^{\infty} \theta P(\theta | X^{(n)}) d\theta$.

ב. הראו כי עבור פונקצית המחיר $\lambda(\hat{\theta}|\theta) = |\hat{\theta} - \theta|$ האומד האופטימלי עבור

מדגם נתון $X^{(n)}$ יהיה החציון של ההתפלגות $P(\theta | x^{(n)})$, כלומר אותו

ערך של θ שעבורו $\int_{-\infty}^{\hat{\theta}} P(\theta | x^{(n)}) d\theta = \int_{\hat{\theta}}^{\infty} P(\theta | x^{(n)}) d\theta$.

4. מצאו את התוחלת המותנית בהנתן המדגם $X^{(n)} = \{2, 2, 4, 3, 5\}$ עבור התפלגות פואסונית.

א. הניחו התפלגות אפרירית אחידה.

ב. הניחו התפלגות אפרירית אקספוננציאלית, והשוו לתוצאות סעיף א.

5. הראו כי עבור מדגם i.i.d x_1, x_2, \dots, x_n המפולג

אקספוננציאלית $\forall x \geq 0 \quad f(x|\theta) = \frac{1}{\theta} \exp(-x/\theta)$ מתקיים כי התפלגות

Inverse Gamma היא ה-conjugate prior. מצאו מהו $\pi(\theta | x_1, x_2, \dots, x_n)$ ומהי ההתפלגות השולית $f(x)$. נזכיר כי התפלגות Inverse Gamma עבור

$$f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha x^{(\alpha+1)}} \exp^{-\frac{1}{x\beta}} \quad \forall x \geq 0 \quad \beta \geq 0 \quad \alpha \geq 0$$

6. חשבו את אינפורמציות-פישר עבור התפלגות ברנולי עם סיכוי הצלחה θ .

$$7. \text{ הראו כי אם } P \equiv P(x|\theta) \text{ אז } E\left[\left(\frac{\partial \log P}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \log P}{\partial \theta^2}\right]$$

8. הראו כי עבור משתנה מקרי רב מימדי, מטריצת אינפורמציות פישר מקיימת

$$J_n(\theta_i, \theta_j) = -E\left[\frac{\partial^2 \log P}{\partial \theta_i \partial \theta_j}\right]$$

9. חשבו את מטריצת אינפורמציות פישר עבור התפלגות נורמלית עם תוחלת μ וסטיית תקן σ .

10. הראו כי ההתפלגות הנורמלית שייכת למשפחה האקספוננציאלית.

11. מצאו סטטיסטים מספיקים עבור ההתפלגויות מתרגיל 1 (אם הם אכן קיימים) והראו כי אומדי ה-ML מתרגיל 1 הם פונקציות של הסטטיסטים המספיקים.

12. הוכיחו כי עבור סטטיסטי מספיק T כל פונקציה הפיכה $Z=u(T)$ שלא תלויה ב- θ גם היא סטטיסטי מספיק.

13. בהנתן ס"מ $T_1 = u_1(x_1, \dots, x_n)$ ופרמטר θ . הראו כי הפילוג של סטטיסטי אחר, $T_2 = u_2(x_1, \dots, x_n)$ בהינתן T_1 אינו תלוי ב- θ .

תרגיל מחשב

כתבו תכנית לצבירה (Clustering) סטטיסטית של נקודות ב R^d ל c צברים גאוסיאנים (צפיפות התפלגות בעלת c תערובות) בעלי ממוצעים $\hat{\mu}_1, \dots, \hat{\mu}_c$, מטריצות קווריאנס אלכסוניות $\sum_k = \sigma_k \cdot I$, תוך שימוש באלגוריתם Isodata. הפעילו את התכנית עבור $c=1,2,\dots,5$ גאוסיאנים. הניחו כי ההסתברות לכל תערובת היא זהה.