

פרק ד'

שיטות לא פרמטריות

4.1 מבוא

בפרקים הקודמים דנו בהסקה סטטיסטית כאשר המשפחה הפרמטרית ממנה באו התצפיות היתה ידועה. בפועל, במקרים רבים אין לנו ידע כלשהו על צורת ההתפלגות, והנתונים עשויים להגיע מתוך התפלגות שאינה דומה כלל לאחת ההתפלגויות הנפוצות.

ניתן לחלק את מגוון השיטות שאינן מניחות ידע על הצורה הפרמטרית של ההתפלגות לשלושה סוגים. ראשית, שיטות שמבצעות הערכת פונקציית ההתפלגות $P(x|\omega)$ מתוך מדגם, כך שמאוחר יותר ניתן לבצע הכרעות על פי ההתפלגות המשוערת. שנית, קיימות שיטות המדלגות על שלב הערכת הצפיפות, ופונות הישר אל שלב ההכרעה על מצב העולם בנקודה תוך שהן מעריכות את ההתפלגויות האפוסטריוריות $P(\omega|x)$. לבסוף קיימות שיטות לביצוע טרנספורמציות על הדגימות מתוך תקווה שאפשר יהיה להפעיל שיטות פרמטריות על מרחב התכונות לאחר הטרנספורמציה. בפרק הנוכחי נתאר שיטות משני הסוגים הראשונים.

4.2 הערכת פונקציית התפלגויות

4.2.1 הרעיון הבסיסי

הבעיה הבסיסית ביותר הדורשת שימוש בשיטות לא פרמטריות היא הערכת ההתפלגות של דגימות המפולגות על פני המרחב (בלא שידוע לנו דבר על מקורן) עלינו ליחס לכל נקודה במרחב x מספר ממשי $\hat{f}(x)$ המבטא את האומדן להתפלגות $f(x)$ של המקור ממנו נלקחו הדגימות.

הרעיון העומד בבסיס שיטות להערכת התפלגויות הוא פשוט ביותר ומתבסס על העובדה כי ההסתברות P שדגימה x תקבל ערך בקטע S ניתנת על ידי

$$(4.1) \quad P = \int_S p(x) dx$$

מכאן שההסתברות P היא גרסה מוחלקת של פונקציית צפיפות ההתפלגות $p(x)$. אם נתונות לנו n דגימות, אזי ההסתברות ש- k מתוכן יפלו בקבוצה S ניתנת על ידי ההתפלגות הבינומית

$$P_k = \binom{n}{k} P^k (1-P)^{n-k}.$$

התוחלת של k היא כמובן nP וההתפלגות מרוכזת בצמצום סביב התוחלת, כך ש- k/n הוא אומד טוב עבור ההסתברות P , ובעקיפין עבור פונקציית ההתפלגות p . אם ההתפלגות $p(x)$ רציפה והקטע S קצר מספיק כך שההתפלגות $p(x)$ אינה

משתנה כמעט בקטע, אז מתקיים $\int_S p(x') dx' \approx p(x) |S|$ ונוכל לרשום

$$(4.2) \quad p(x) \approx \frac{k/n}{|S|}$$

במקרה של משתנה מקרי רב מימדי, הניתוח נשאר תקף, אך במקום באורך הקטע $|S|$ יש להתחשב בנפח החלון הרב ממדי V .

על מנת להעריך את $p(x)$ ברזולוציה טובה עלינו להקטין את גודל החלון, אבל אז מספר הדגימות בתוכו יקטן והשגיאה בהערכת $p(x)$ בקטע תגדל. נרצה לכן לבנות חלון באופן שיאפשר אופטימיזציה בין רזולוציה לשגיאה. לצורך כך, נגדיר לכל מספר דגימות n חלון S_n שגודלו V_n תלוי במספר הדגימות. לכל מספר סופי של דגימות נקבל שגיאה בהערכת פונקציית הצפיפות, שתלך ותקטן ככל שנגדיל את מספר דוגמאות. הערכה של שגיאה זאת עלולה להיות מסובכת ולא נדון בה כאן. נגדיר את פונקציית הצפיפות המשוערכת על פי n דגימות באופן הבא

$$(4.3) \quad p_n(x) = \frac{k_n(x)/n}{V_n}$$

כאשר $k_n(x)$ הוא מספר הדגימות שנפלו בחלון בנפח V_n המרוכז סביב הנקודה x . כאשר מספר הדגימות שואף לאינסוף נרצה לקבל הערכה מדויקת של צפיפות ההסתברות בנקודה x (דהיינו נרצה שיתקיים $\lim_{n \rightarrow \infty} p_n(x) = p(x)$) ולצורך כך

נדרוש כי

1. נפח החלון סביב כל נקודה x ישאף לאפס כאשר מספר הדגימות ישאף לאינסוף (על-מנת לקבל הערכה של הצפיפות ברזולוציה אינסופית)

$$\lim_{n \rightarrow \infty} V_n = 0.$$

2. מספר הדגימות בכל חלון הממורכז סביב נקודה x שלה צפיפות חיובית $p(x) \neq 0$ (על-מנת להקטין לאפס את השגיאה בהערכת הצפיפות בתוך החלון)

$$\lim_{n \rightarrow \infty} k_n(x) \rightarrow \infty.$$

3. המספר היחסי של הדגימות בתוך החלון ישאף לאפס (היות והחלון שואף לכסות חלק אפסי מהמרחב, החלק היחסי של הדגימות שיכולות ליפול בתוכו צריך גם-כן לשאוף לאפס)

$$\lim_{n \rightarrow \infty} \frac{k_n}{n} \rightarrow 0.$$

4. היחס בין המספר היחסי של הדגימות בחלון לבין נפח החלון ישאר סופי כלומר V_n ישאף ל-0 כפונקציה לאט יותר מאשר k_n/n (על-מנת לשמור על יציבות בהערכה של $p(x)$ בתהליך של העלאת מספר הדגימות והקטנת החלון)

$$\lim_{n \rightarrow \infty} \frac{k_n/n}{V_n} < \infty.$$

קיימות בחירות שונות של V_n ו k_n המקיימות את הדרישות לעיל. נתאר בפרוט אחת מהן המכונה חלונות פרזן.

4.2.2 חלונות פרזן (Parzen window)

נניח כי הדגימות שלנו הן d ממדיות, ונגדיר חלון ריבועי כקוביה d ממדית בעלת אורך צלע h_n . נגדיר גם פונקציית חלון ריבועית

$$(4.4) \quad \Phi(\mathbf{u}) = \begin{cases} 1 & \text{if } |\mathbf{u}_j| \leq \frac{1}{2} \quad \forall j, j=1,2,\dots,d \\ 0 & \text{otherwise} \end{cases}$$

ובמלים, הפונקציה היא 1 אם המרחק מהראשית בכל אחד מהמימדים קטן מ- $\frac{1}{2}$ והיא 0 אחרת. עבור קוביה בעלת אורך-צלע h_n הממורכזת סביב \mathbf{x} נרשום

$$(4.5) \quad \Phi\left(\frac{\mathbf{x} - \mathbf{x}^{(i)}}{h_n}\right) = \begin{cases} 1 & \text{if } \frac{|\mathbf{x}_j - \mathbf{x}_j^{(i)}|}{h_n} \leq \frac{1}{2} \quad \forall j=1,2,\dots,d \\ 0 & \text{otherwise} \end{cases}$$

מספר הדגימות בתוך הקוביה יהיה נתון לפיכך על ידי

$$(4.6) \quad k_n(\mathbf{x}) = \sum_{i=1}^n \Phi\left(\frac{\mathbf{x} - \mathbf{x}^{(i)}}{h_n}\right)$$

(משום שעבור כל דגימה שנמצאת מתוך הקוביה הפונקציה תתן "1") והקרוב ה-
 n -י לפונקצית הצפיפות נתון על ידי

$$(4.7) \quad P_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{V_n} \Phi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right)$$

כאשר נפח הקוביה הוא $V_n = h_n^d$. קל לראות כי $P_n(x) \geq 0$ (משום ש $\Phi(\mathbf{u}) \geq 0$)
 וכי

$$\int P_n(\mathbf{x}) d\mathbf{x} = 1$$

כלומר $P_n(\mathbf{x})$ עונה על הדרישות הבסיסיות מפונקצית צפיפות-התפלגות.
 נגדיר כעת

$$\delta_n(\mathbf{x}) = \frac{1}{V_n} \Phi\left(\frac{\mathbf{x}}{h_n}\right)$$

ואז ניתן לכתוב

$$P_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \delta_n(\mathbf{x} - \mathbf{x}_i)$$

הערך של $P_n(\mathbf{x})$ יהיה תלוי בדגימות הספציפיות שלקחנו. כתוצאה מכך $P_n(\mathbf{x})$
 הוא בעצם משתנה מיקרי. נבדוק כעת מהי התוחלת של $P_n(\mathbf{x})$ בהנחה שהדגימות

מפולגות i.i.d. על-פי צפיפות (לא-ידועה) $P(\mathbf{x})$

$$(4.8) \quad \begin{aligned} \langle P_n(\mathbf{x}) \rangle &= E[P_n(\mathbf{x})] = \frac{1}{n} \sum_{i=1}^n E\left[\frac{1}{V_n} \Phi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right)\right] \\ &= \int \frac{1}{V_n} \Phi\left(\frac{\mathbf{x} - \mathbf{u}}{h_n}\right) P(\mathbf{u}) d\mathbf{u} = \int \delta(\mathbf{x} - \mathbf{u}) P(\mathbf{u}) d\mathbf{u} \end{aligned}$$

כלומר קונבולוציה של פונקצית הצפיפות עם פונקצית החלון. (פונקצית החלון היא
 ה"גרעין" ("kernel") של הקונבולוציה) בגבול $n \rightarrow \infty$ הפונקציה $\delta_n(\mathbf{x} - \mathbf{u})$
 שואפת לפונקצית-דלתה הממוקדת סביב \mathbf{x} . אם רציפה ב- \mathbf{x} ומספר הדגימות
 שואף לאינסוף אנו מקבלים, לפיכך

$$(4.9) \quad P_n(\mathbf{x}) \rightarrow \int \delta(\mathbf{x} - \mathbf{u}) P(\mathbf{u}) d\mathbf{u} = P(\mathbf{x}).$$

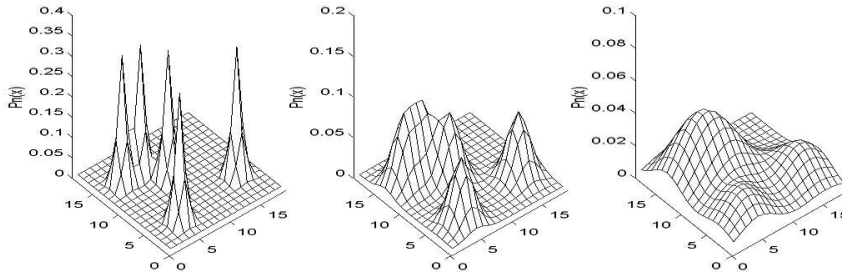
השימוש בפונקצית-חלון ריבועית אינו בהכרח אופטימלי, שכן המשקל שמקבלת
 נקודה בהערכת הצפיפות משתנה באופן לא-רציף מ- 1 ל- 0 כאשר אנו "יוצאים
 מהקוביה". ניתן להגדיר פונקצית חלון אחרת, למשל גאוסיאן

$$\Phi_s(\mathbf{u}) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}|\mathbf{u}|^2\right]$$

ואז השערוך יהיה קונבולוציה של פונקצית הצפיפות האמיתית עם גאוסיאן. פונקצית
 חלון כזאת מאפשרת שערוך חלק יותר ורועש פחות, אולם העלות החישובית של

השימוש בה היא גבוהה (חישוב של גאוסיאן מול בדיקת תנאי). ההחלטה על שימוש בפונקציית חלון כזאת תהיה "trade-off" בין העלות של הגדלת מספר הדגימות ושימוש בפונקציית חלון ריבועית לבין העלות החישובית של השימוש בגאוסיאן ("רוחב החלון" במקרה זה יהיה סטית-התקן של הגאוסיאן)

את נפח החלון נגדיר בדרך כלל על ידי $V_n = \frac{V_1}{\sqrt{n}}$ (V_1 שרירותי) בחירה זו תבטיח כי הדרישות שדרשנו לעיל מתקיימות.



איור 4.1: הערכת פונקציית התפלגות דו ממדית על ידי חלונות גאוסיאנים בשלושה רחבים שונים

4.2.3 קביעה דינמית של גודל החלון

אחת הבעיות בחלונות Parzen היא שגודלו של החלון (עבור n נתון) קבוע בכל המרחב. כתוצאה מכך אנו יכולים למצוא את עצמנו עם חלון גדול מדי באזורים בעלי צפיפות גבוהה (מפסידים רזולוציה) וחלון קטן מידי באזורים בעלי צפיפות נמוכה (מאבדים דיוק). דרך טובה להתגבר על כך היא להגדיר את מספר הדגימות בחלון, k_n , כפונקציה של n (למשל $k_n = \sqrt{n}$), ולהגדיל את החלון סביב הנקודה x עד שמקבלים בדיוק k_n דוגמאות בחלון

$$(4.10) \quad P_n(x) = \frac{k_n/n}{V_n(x)} \Rightarrow V_n(x) \approx \frac{1}{\sqrt{n}} \cdot \frac{1}{P(x)}$$

גם כאן V_n הוא מהצורה $\frac{V_1}{\sqrt{n}}$ אלא שהפעם V_1 נקבע עבור כל x על פי הדגימות האמפיריות ולא באופן שרירותי.

4.3 הכרעה בשיטות לא פרמטריות

הסעיף הקודם דן בשיטות לא פרמטריות לאמידת פונקציית התפלגות, כך שלאחר אמידת פונקציית ההתפלגות ניתן להשתמש בה בבעיות הכרעה. לדוגמא, נניח כי נתונות לנו תצפיות $X^{(n)} = (x_1, \dots, x_n)$ וכל תצפית מתויגת ומשויכת לאחת מבין C קבוצות (מצבי עולם) $\omega_1, \dots, \omega_C$, דהיינו נתונים לנו התייגים $Y^{(n)} = (y_1, \dots, y_n)$. כעת מוצגת לנו תצפית חדשה x ומבקשים אותנו לשייך אותה לאחת מהקבוצות, דהיינו למצוא את התייג הנכון y . שיטה אחת לפתרון הבעיה היא לאמוד את התפלגות התצפיות בכל אחת מהקבוצות, ואז לבצע הכרעה למשל בשיטת מקסימום אפוסטריורי. שיטה אחרת שאותה נתאר כאן, היא להעריך את ההסתברות לקבל את התצפית x בכל אחת מהקבוצות ישירות מתוך המדגם.

4.3.1 הערכת הסתברויות אפוסטריוריות

נניח כאמור שיש לנו סט של n דוגמאות מתויגות במרחב באחת מבין C אפשרויות לתייג כל דגימה (C מצבי עולם $\{\omega_1, \dots, \omega_C\}$). על מנת להעריך את הצפיפות של כל אחת מהקבוצות בנקודה x במרחב, נוכל לבנות חלון V סביב הנקודה x , ונניח כי תפסנו k דוגמאות בתוך החלון, מתוכן k_i היו מסוג ω_i . הערכה של ההתפלגות המשותפת תהיה

$$P_n(x, \omega_i) = \frac{k_i/n}{V}$$

ומהגדרת ההסתברות המותנית

$$(4.11) \quad P_n(\omega_i | x) \equiv \frac{P_n(\omega_i | x)}{\sum_{i=1}^C P_n(\omega_i | x)} = \frac{k_i}{k}$$

כלומר החלק היחסי של הדוגמאות מסוג i בחלון מתאים להסתברות של ω_i בחלון.

4.3.2 טוב שכן קרוב (K-Nearest Neighbors)

חסרון אפשרי של השיטה לעיל הוא (שוב) השימוש בגודל חלון קבוע. כדי לקבוע גודל חלון המשתנה דינמית ניתן לבחור חלון בעל התפלגות אחידה המכסה את k התצפיות הקרובות ביותר לתצפית החדשה x שעבורה עלינו לבצע הכרעה. באופן טבעי (ומתוך מחווה למכרם חורי) שיטה זו נקראת k-Nearest Neighbors. הגרסה הפשוטה ביותר היא כמובן עבור $k=1$, דהיינו כלל השכן הקרוב ביותר, שאז מתייגים נקודת מדגם באותו תיוג של הנקודה הקרובה לה ביותר. מסתבר כי כלל פשוט זה הוא מוצלח למדי.

טענה

אם נסמן ב- E^* את השגיאה של כלל הכרעה בייסיאני (שהוא כלל הכרעה אופטימלי), אז השגיאה של כלל השכן הקרוב ביותר מתכנסת בגבול של מדגם גדול לערך המקיים E_1

$$(4.12) \quad E_1 \leq E^* \left(2 - \frac{C}{C-1} E^* \right)$$

כאשר C הוא מספר הקבוצות (מצבי העולם).

הוכחה

יהי X_1 השכן הקרוב ביותר לדגימה X_0 . ונניח כי X היא דגימה מקבוצה C_0 בעוד ש- X_1 הוא מקבוצה C_1 . נטען ראשית ללא הוכחה כי בגבול של מדגם גדול, ההסתברות $P(\omega_i | X_1)$ קרובה ל- $P(\omega_i | X_0)$. ובאופן יותר מדויק

$$(4.13) \quad E[P(\omega | X_0) - P(\omega | X_1)] \rightarrow 0$$

כעת ההסתברות לשגיאה, היא הסיכוי שהדגימות באות מקבוצות שונות

$$(4.14) \quad \begin{aligned} \Pr(C_0 \neq C_1 | X_0 = x) &= \\ &= \sum_{i \neq j} P(\omega_i | x) E(P(\omega_j | X_1) | X_0 = x) \end{aligned}$$

או על פי הקירוב 4.13,

$$E \left[\Pr(C_0 \neq C_1 | X_0 = x) - \sum_{i \neq j} P(\omega_i | X_0) P(\omega_j | X_0) \right] \rightarrow 0$$

כלומר בגבול של מדגם גדול מתקיים

$$(4.15) \quad E_1 = \sum_{i \neq j} P(\omega_i | X_0) P(\omega_j | X_0) = 1 - \sum_i P(\omega_i | X_0)^2$$

כעת הסיכון המותנה הבייסיאני R^* הוא $R^* = 1 - \max_i (P(\omega_i | X_0))$ ונניח לצורך

הסימון כי מצב העולם המביא אותו למינימום הוא k , דהיינו $R^* = 1 - P(\omega_k | X_0)$.

על פי אי שוויון קושי שוורץ נוכל לרשום

$$(C-1) \sum_{i \neq k} P(\omega_i | x)^2 \geq \left[\sum_{i \neq k} P(\omega_i | x) \right]^2 = R^*(X)^2$$

$$(4.16) \quad (C-1) \sum_i P(\omega_i | x)^2 \geq R^*(X)^2 + (C-1)(1 - R^*(X)^2)$$

$$1 - \sum_i P(\omega_i | x)^2 \leq 2R^*(X)^2 - \frac{C}{(C-1)} R^*(X)^2$$

ואם נמצע את שני האגפים נקבל תוך שימוש ב-
 $E(Y^2) = \text{Var}(Y) + E^2(Y) \geq E^2(Y)$, את אי השוויון הנדרש

$$(4.17) \quad E_1 \leq 2E^* - \frac{C}{(C-1)} E[R^*(X)^2] \leq 2E^* - \frac{C}{(C-1)} (E^*)^2.$$

עבור כלל ההכרעה ל- k שכנים קרובים יותר קיימות תוצאות אנליטיות מפורטות בעיקר למקרה של שני מצבי עולם (שני תיוגים אפשריים). באופן אינטואיטיבי, כלל הכרעה המסתמך על שני שכנים אינו טוב יותר מכלל המסתמך על שכן אחד, היות ולצורך הכרעת רוב שני השכנים צריכים להיות זהים. מצד שני, במקרה של אי הסכמה בין שני השכנים, אנחנו יכולים להחליט שלא להחליט ולטעון שאיננו יודעים את התיוג של הדגימה. על מנת לסכם את החסמים על השגיאות נביא את הטענה הבאה ללא הוכחה.

טענה: השגיאה של כלל k שכנים קרובים

במקרה של שני מצבי עולם, נסמן ב- E_k את השגיאה בשימוש בכלל ההכרעה של k שכנים קרובים ביותר בגבול של מדגם גדול, כאשר שיוון בהצבעות נפתר באופן מקרי, ונסמן ב- E_k' את השגיאה כאשר שיוון בהצבעה נפתר באי החלטה. אזי מתקיים

$$(4.18) \quad \begin{aligned} E'_2 &\leq E'_4 \leq \dots \leq E'_{2k} \leq E^* \leq E_{2k} = \\ &= E_{2k-1} \leq \dots \leq E_2 = \\ &= E_1 = 2E'_2 \end{aligned}$$

תרגילים

יהיו התפלגויות אחידות בתוך שני כדורי יחידה המרוחקים זה מזה 10 יחידות (למשל) כאשר $P_0(w_1) = P_0(w_2) = \frac{1}{2}$. יהיה $X^{(n)} = (x_1, \dots, x_n)$ סט של n דוגמאות מתוגות (ומר אנו יודעים עבור כל אחת מהן אם היא באה מתוך w_1 או w_2) ו- $X^{(k)} = (x_1', \dots, x_k')$ הם k השכנים הקרובים ביותר של x . על פי כלל k השכנים הקרובים ביותר אנו מחליטים אם דוגמא לא מתוגת, x , שייכת למצב עולם w_1 או w_2 על פי הרוב מתוך k השכנים הקרובים ביותר.

א. הראו (על סמך שיקולים קומבינטורים פשוטים למדי) כי עבור k אי זוגי סיכוי השגיאה הממוצע הוא

$$P_n(\text{error}) = \frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \binom{n}{j}$$

ב. הראו כי במקרה זה כלל השכן-הקרוב-ביותר (קובעים את השיוך של נקודה לפי השיוך של השכן הקרוב ביותר) הוא בעל סיכוי שגיאה קטן יותר מאשר כלל "k השכנים הקרובים ביותר" עבור $k > 1$.

תרגיל מחשב

כתבו תכנית מחשב המשתמשת בחלון Parzen על מנת לשערך את הצפיפות $P_n(x)$ של נקודות במישור הנוצרות על ידי שני גאוסיאנים. השתמשו בפונקצית חלון ריבועית ובפונקצית חלון גאוסיאנית על מנת לשערך את

$$P_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{V_n} \Phi_s \left(\frac{x - x^{(i)}}{h_n} \right)$$

עבור n שונים, למשל $n=1, 16, 256, 1024$ כאשר $h_n = h_1 / \sqrt{n}$ ו- $h_1 = 0.25, 1, 4$.

