פרק 3: יסודות סטטיסטיים – הערכת פילוגי הסתברות

מקור: DHS:2001 פרקים 3.1-3.4 (שיטות פרמטריות) 4.1-4.3 (שיטות א-פרמטריות), HTF:2001 פרק 8.5.1 (עירוב גאוסי).

3.1 הקדמה

ניתן לחלק את מרכיבי הלמידה לשלושה: מידול (modeling), הסקה (inference) ושיערוך מודל (learning, estimation). המרכיב הראשון מקיף את האופן שבו אנו מייצגים בעייה אמיתית בכלים המתמטיים העומדים לרשותינו. לדוגמא, רישום הבעיה ע"י מודל בייסיני הסתברותי, בו יש לנו פילוג על המקורות השונים (קראנו לכך פריור – prior) ופילוג של תצפיות בהנתן כל מקור (הנראות). כמו כן, ניתן לראות את פונקציית השגיאה או המחיר כחלק מן המודל.

המרכיב השני – ההסקה– כולל אוסף של כלים ושיטות להסקת מסקנות מתוך מודל נתון, לדוגמא בהנתן תצפית מהוא מצב העולם המסתבר ביותר? מהו מצב העולם עם תוחלת מחיר נמוכה ככל האפשר? בהנתן מצב מסויים מה התכונות של התצפיות האופייניות לו, וכד'

המרכיב השלישי – שיערוך המודל או הלמידה – בו בעיקר נעסוק בקורס. לרוב התאור המדוייק של המודל אינו אפשרי רק על סמך ידע מוקדם ועלינו להסיקו מתוך נתונים או דגימות. אנו נניח כי נתון מדגם (קבוצת דגימות) של זוגות – זהות המקור ותצפית מתאימה – ונרצה למצוא מודל מיטבי. לעיתים התשובה שונה עבור הגדרות שונות למונח מיטבי, או בפונקציית המחיר בפרט.

הסקה סטטיסטית עוסקת בהערכת ("שיערוך") גודל בעל-עניין מתוך מדידות חלקיות ורועשות (משתנים מקריים) התלויות בגודל זה.

הבעיה היסודית בתחום הינה: <u>הערכת פילוג הסתברות של משתנה מקרי</u> מתוך דגימות של משתנה זה. בבעיה זו ניגע בקצרה בפרק הנוכחי.

בעיות נוספות הקשורות בכך הינן:

- הערכת פרמטרים מסוימים של הפילוג: למשל ממוצע, וריאנס, הסתברות מאורע נתון.
- הערכת מודל: זיהוי מודל הסתברותי הקושר בין פלט x לפלט y, ל סמך דגימות של זוגות $\{x_i, y_i\}$. בבעיה זו נדון בפירוט בהמשך הקורס.

נציין כי תחום השיערוך הסטטיסטי הינה תחום נרחב ובעל תיאוריה מקיפה, ואנו ניגע בו רק "על קצה המזלג". מבוא רחב יותר לנושא ניתן בקורסים "אותות אקראיים", "מבוא לעיבוד אותות אקראיים" (046201), ובקורסי המוסמכים.

ניתן להבחין בין הגישות הבאות להסקה סטטיסטית (ובפרט, להערכת פילוגי הסתברות):

- גישה פרמטרית (שערוך פרמטרים) לעומת גישה א-פרמטרית
- גישה בייסיאנית לעומת הגישה הלא-בייסיאנית ("קלאסית")

בהמשך נתאר בקצרה את גישות אלו וההבדלים ביניהן, ונתמקד מעט יותר בגישה הלא-בייסיאנית.

הבעיה הספציפית שבה נעסוק הינה בעיית שערוך הפילוג הבאה:

- יהי X משתנה מקרי בעל פונקציית פילוג (לא ידועה) $p_X(x)$. מטרתנו להעריך את $p_X(x)$ מתוך n דגימות בלתי תלויות $D=\{x_k\}_{k=1}^n$ של המשתנה המקרי X.

נזכיר כי $p_X(x)$ מסמנת את פונקצית צפיפות ההסתברות כאשר X משתנה רציף, ואת ההסתברות עצמה כאשר X משתנה בדיד.

3.2 שיטות פרמטריות

א. משפחות פרמטריות

בגישה הפרמטרית, אנו מניחים כי הפילוג הנדרש $p_X(x)$ הינו בעל צורה ידועה, המוגדרת עד כדי וקטור פרמטרים θ, כלומר

$$p_X(x)=p_X(x\,|\,\theta)$$

לרוב מדובר בווקטור פרמטרים ממשיים בעל מימד נתון, כלומר $\theta=(\theta_1,\ldots,\theta_P)^T\in\mathbb{R}^P$.

לשם פשטות נתייחס מעתה ל"פרמטר θ". ההנחה לגבי צורת הפילוג מתקבלת כמובן מתוך ידע מוקדם ו/או ניתוח ראשוני של המידע.

לדוגמא:

- $X\sim N(\mu,\Sigma)$ כאשר Σ מטריצת קווריאנס ידועה. במקרה זה $\theta=\mu\in\mathbb{R}^d$ ו- d הוא מימד הווקטור X.

- $X \sim N(\mu,\Sigma)$, כאשר μ ו- Σ אינן ידועות. המקרה זה θ כולל את האיברים המתאימים מתוך מטריצות אלו. (עקב סימטריה מספיקים $d(d+1)/2$ פרמטרים לתאר את Σ, וניתן להגבילם כך שהמטריצה תהיה חיובית מוגדרת).
- $X \sim \exp(\theta)$: משתנה אקראי (סקלרי) בעל פילוג אקספוננציאלי, דהיינו $p_X(x)=\theta^{-1}e^{-x/\theta}$ (עבור $x \geq 0$), כאשר θ הינו פרמטר חיובי.
- $X \sim Bern(\theta)$: X הינו משתנה בינארי כך ש: $P\{X=0\}=1-\theta$, $P\{X=1\}=\theta$. הפרמטר θ מוגבל כמובן לתחום $[0,1]$.

הערכת פילוג ההסתברות שקולה עתה להערכת הפרמטר θ, מתוך המדידות D. הבעיה האחרונה מכונה <u>שערוך פרמטרים</u> (parameter estimation) בספרות ההנדסית, או <u>אמידת פרמטרים</u> בספרות הסטטיסטית. ה<u>משערך</u> $\hat{\theta}=\hat{\theta}(D)$ הינו למעשה פונקציה (הנתונה לבחירתנו) של המידע הנמדד D.

<u>ב. הגישה הבייסיאנית</u>

נבדיל עתה בין הגישה הבייסאנית ללא-בייסיאנית:

- בגישה הבייסיאנית הסתברויות מייצגות דרגות של בטחון בערך המשתנה, ולאו דווקא גבול של ממוצעים בניסוי. אנו מניחים כי הפרמטר (הווקטורי) θ (כמו כל גורם אחר בבעיה) הינו משתנה מקרי. בפרט, נניח כי הפרמטר מפולג לפי איזה פילוג א-פריורי $p(\theta)$. נחשוב על פילוג זה כעל דרגת בטחון בזהות הערך של הפרמטר לפני שצפינו בדגימות כלשהו, כל דגימה חדשה משנה את דרגות הבטחון בערכים השונים. בפרט, לאחר שדגמנו תצפיות נקבל פילוג בדיעבד (א-פוסטריורי) $p(\theta\,|\,D)$.
- בגישה הלא-בייסיאנית, אנו רואים את הפרמטר θ כגודל קבוע ודטרמיניסטי, ובפרט נמנעים מהנחות סטטיסטיות כלשהן לגביו. אנו מחפשים את המודל הכי טוב שמסביר את התצפיות.

הגישה הבייסיאנית מגדירה את הפילוג-בדיעבד (posterior distribution) $p(\theta\,|\,D)$, כאשר D הינו המידע הנצפה.

הגישה המלאה קובעת כי אין לבחור מודל אחד מתוך הפילוג-בדיעבד, ויש לעשות בכל שימוש האפשרויות. בפרט, אם לדוגמא עלינו לחשב את ההסתברות לתצפית חדשה מסויימת x יש לקחת בחשבון את כל הפרמטרטים השונים θ ולמצע עליהם, בפרט:

$$p(x|D)=\int p(x,\theta|D)d\theta$$

נפרק את האינטגרנד ע"י נוסחת ההסתברות השלמה $p(x,\theta|D)=p(x|\theta,D)p(\theta|D)$ ומההנחה שהמודל הוא פרמטרי יחד עם הנחת אי-התלות נקבל כי $p(x|\theta,D)=p(x|\theta)$. ולכן, סה"כ

$$p(x|D)=\int p(x|\theta)p(\theta|D)d\theta$$

ראינו דוגמא אחת לשימוש שעושים בפילוג-בדיעבד: הערכת נראות של דוגמא חדשה.

גישה שניה עושה שימוש בפילוג-בדיעבד לצורך קביעת מודל. אחד היתרונות הוא האפשרות "לבחון" את המודל ולהשתמש בו להסביר תופעות בעולם האמיתי וללמוד ממנו על העולם האמיתי. לדוגמא, להסביר מנגנוניים פיסיולוגיים-ביולוגיים מתוך מודל לחיזוי קיום מחלה בתנאים שונים.

מתוך פילוג זה ניתן לגזור משערכים שונים, שניים מקובלים הינם:

- משערך התוחלת המותנית: $\hat{\theta}_{MMSE}=E(\theta|D)$

  משערך זה מביא למינימום את השגיאה הריבועית הממוצעת (ראו תרגיל):

  $$E((\hat{\theta}-\theta)^2|D)\rightarrow \min$$

  ולפיכך הוא נקרא גם Minimum Mean Square Error (MMSE) Estimator.

- משערך Maximum a-Posteriori (MAP):

  $$\hat{\theta}_{MAP}=\arg\max_{\theta\in\Theta} p(\theta|D)$$

  במשערך זה כבר נתקלנו בהקשר של סיווג בייסאני אופטימאלי.

  ((גרף))

הגישה הבייסיאנית מאפשרת להגדיר משערכים אופטימליים ביחס לקריטריוני ביצועים בעלי משמעות ברורה. חסרונותיה:

- קושי חישובי: חישוב $E(\theta|D)$ ואף $p(\theta|D)$ הינו קשה, פרט למקרים מיוחדים.
- קושי עקרוני: בחירת הפילוג האפריורי $p(\theta)$ אינה בהכרח ברורה מאליה, ולעתים אף הגדרת θ כמשתנה מקרי הינה חסר משמעות. יתר על כן, לעיתים גם לפילוג האפריורי יש פרמטרים, ובאופן עקרוני עלינו לקבוע פילוג עליהם וחוזר חלילה.

ג. הגישה הלא-בייסיאנית – משערך הסבירות המירבית

הגישה האמפירית רואה במודל כלי להסברת תופעות הסתברותיות בעולם (לדוגמא רולטה או קוביה). בגישה זו, הנקראת לרוב גישה לא-בייסאנית, המשערך $\hat{\theta}$ נבחר כפתרון לבעיית

אופטימיזציה או על-ידי כללים היוריסטיים. המשערך הנפוץ ביותר ממשפחה זו הינו משערך הסבירות המירבית (Maximum Likelihood Estimator, MLE) המוגדר באופן הבא :

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} p(D\,|\,\theta)$$

נשים לב כי משערך זה אינו מסתמך כלל על הפילוג של הפרמטר θ, ולפיכך אינו דורש התייחסות לפרמטר כגודל אקראי. למשערך הסבירות המירבית מספר תכונות חיוביות כגון :

- התכנסות (במובן מתאים) לפרמטר הנכון כאשר מספר המדידות גדל.
- חישוב פשוט יותר ממשערכים אחרים
- תוצאות המתיישבות עם האינטואיציה (במקרים בהם קיימת נוסחה סגורה)

<u>הגדרות נוספות</u> : עבור מידע D נתון, הגודל $L(\theta)=p(D\,|\,\theta)$ הינו פונקציה של הפרמטר θ ונקרא פונקציית הסבירות. הגודל $\log L(\theta)=\log p(D\,|\,\theta)$ הינו פונקצית הסבירות הלוגריתמית. משערך הסבירות המירבית נתון לפיכך על ידי :

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} L(\theta) \equiv \arg\max_{\theta\in\Theta}\{\log L(\theta)\}$$

<u>מדידות בלתי תלויות</u> : נרשום את משערך הסבירות המירבית באופן מפורש יותר למקרה שבו $D=\{x_k\}_{k=1}^n$ הינה סדרת דגימות (או מדידות) בלתי תלויות של המשתנה המקרי X. במקרה זה :

$$L(\theta)=p(D\,|\,\theta)=p(x_1,\ldots,x_n\,|\,\theta)=\Pi_{k=1}^n p_X(x_k\,|\,\theta)$$

ולכן

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta}\Pi_{k=1}^n p_X(x_k\,|\,\theta) \equiv \arg\max_{\theta\in\Theta}\sum\nolimits_{k=1}^n \log p_X(x_k\,|\,\theta)$$

<u>דוגמא</u> :

נניח כי פילוג תצפית הינו נורמלי עם שונות קבועה דהיינו $p\left(x|\mu\right)\sim N\left(\mu,\sigma^2\right)$ כאשר σ^2 ידוע.

נתחיל דווקא במשערך הנראות המירבית, האחרון שראינו. קל לראות ע"י חישוב ישיר כי,

$$L(\theta)=\prod_{k=1}^n p_X(x_k\,|\,\theta)=\prod_{k=1}^n \tfrac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x_k-\mu)^2}{2\sigma^2}\right)$$

$$\log L(\theta) = C - \frac{1}{\sigma^2}\sum_{k=1}^{n}\left(x_k - \mu\right)^2$$

על ידי גזירה לפי μ נקבל:

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

קיבלנו, כי המשערך המיטבי, המביא למקסימום את הנראות, הינו ממוצע הדגימות או התצפיות.

נעבור כעת לשערוך בייסיאני.

ראשית, עלינו להגדיר פילוג אפריורי, אנו נניח כי גם פילוג זה הוא נורמלי,

$$p(\mu) \sim N\left(\mu_0, \sigma_0^2\right)$$

הפילוג הפריורי אומר לנו כי ההערכה הטובה היותר לתוחלת של פונקציית הנראות הוא $\mu = \mu_0$ ואי-הודאות בהערכה זו היא בסדר-גודל של σ_0. (הבחירה דווקא בפילוג נורמלי מקלה על החישובים שלהלן, אותם נערוך באופן אנליטי מלא.) נניח לכן מדגם $D = \{x_1 \ldots x_n\}$ על-לפי $p\left(x|\mu\right) \sim N\left(\mu, \sigma^2\right)$. הפילוג בדיעבד נתון ע"י

$$p\left(\mu|D\right) \propto p\left(D \mid \mu\right) p\left(\mu|\mu_0\right) \propto \prod_{i=1}^{n} p\left(x_i \mid \mu\right) p\left(\mu|\mu_0\right)$$

נציב את ההגדרות ונקבל

$$p(\mu \mid D) \propto \prod_{i=1}^{n} e^{-\left(\frac{x_i - \mu}{2\sigma}\right)^2} e^{-\left(\frac{\mu - \mu_0}{2\sigma_0}\right)^2}$$

$$\propto e^{-\left(\frac{n}{\sigma^2} + \frac{2}{\sigma_0^2}\right)\frac{\mu^2}{2} + \left(\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i + \frac{\mu_2}{\sigma_0}\right)\mu}$$

כאשר סימן ה"פרופרציוני ל" מסתיר קבועים כפליים שאינם תלויים ב- μ. קל לראות כי זהו פילוג נורמלי, נסמן את התוחלת שלו ב- μ_n ואת סטיית התקן שלו ב- σ_n. ע"י השוואת מקדמים נקבל כי

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \qquad \mu_n = \frac{\left(n\sigma_0^2\right)\hat{\mu}_{MLE} + \left(\sigma^2\right)\mu_0}{n\sigma_0^2 + \sigma^2}$$

כלומר התוחלת היא ממוצע משוקלל של משערך הנראות המירבית $\hat{\mu}_{MLE}$ והתוחלת ע"פ הפילוג האפריורי μ_0. המשקל של משערך הנראות המירבית גבוה כאשר יש מספר גבוה של תצפיות (n גדול) (ואז אין טעם או צורך להתחשב בפילוג האפריורי) או כאשר אי-הודאות של הפילוג הא-פריורי גבוהה (σ_0 גדול). המשקל של המשערך הא-פריורי גדול כאשר אי-הודאות המובנת במודל גבוהה (σ גדול).

מידת אי-הודאות, המכומתת ע"י סטיית-התקן σ_n, קטנה כאשר יש מספר גדול n של תצפיות, או לפחות אחד מתוך מידות-אי הודאות (σ_0 או σ) קטן.

מתוך פילוג זה נוכל להסיק, לדוגמא, את משערך ה- Maximum a-Posteriori (MAP). הוא שווה בדיוק ל- μ_n.

לבסוף, נוכל לחשב את הפילוג על פני התצפיות בהנתן המדגם, הוא נתון ע"י

$$p(x|D) = \int p(x|\mu)p(\mu|D)d\mu$$

נציב את הגדלים המתאימים, נחשב את האינטגרל ונקבל

$$p(x|D) \sim N\left(\mu_n, \sigma^2 + \sigma_n^2\right)$$

הפילוג מזכיר את הפילוג-בדיעבד, לשניהם תוחלת זהה. אולם לפילוג שקיבלנו על פני התצפיות שונות גבוהה יותר של $\sigma^2 + \sigma_n^2$ מול σ_n^2 עבור הפילוג בדיעבד. ניתן לאמר, כי הודאות בפילוג התצפיות בהנתן המדגם נובע משילוב של אי-הודאות σ_n^2 עבור פרמטר התוחלת μ בפילוג בדיעבד, ושל אי-הודאות σ^2 של תצפית בהנתן תוחלת זו.

<u>נסכם עד כה</u> : חישבנו שלושה משערכים בהנתן בעייה שכל מרכיביה בעלי פילוג נורמלי. פילוג הנראות המירבית קובע ערך אחד ויחיד לפרמטרים המתאימים (דהיינו תוחלת). הגישה הבייסיאנית מחשבת פילוג אל פני הפרמטרים הלא ידועים (דהיינו הפילוג בדיעבד). ממנו ניתן לגזור משערכים אחרים (לדוגמא MAP) או לחשב ישירות את פונקציית הנראות בהנתן המדגם, בה ניתן לעשות שימוש ישיר.

<u>דוגמאות נוספות לחישוב משערך נראות מירבית ומשערך מוטה :</u>

1. הערכת הממוצע של פילוג נורמלי : $X \sim N(\mu,\Sigma)$ הוא וקטור גאוסי d-מימדי, כאשר Σ מטריצת קווריאנס ידועה, אולם הממוצע μ אינו ידוע. לפיכך $\theta = \mu \in \mathbb{R}^d$. ע"י חישוב ישיר נקבל :

$$L(\theta) = \prod_{k=1}^{n} p_X(x_k \mid \theta) = \prod_{k=1}^{n} \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp(-(x_k - \mu)^T \Sigma^{-1}(x_k - \mu))$$

$$\log L(\theta) = C - \sum_{k=1}^{n} (x_k - \mu)^T \Sigma^{-1} (x_k - \mu)$$

על ידי גזירה לפי μ נקבל:

$$\hat{\mu}_{MLE} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

2. הערכת ממוצע ווריאנס של פילוג נורמלי: נניח עתה כי גם מטריצת הקווריאנס אינה ידועה. החישוב במקרה זה שהינו מעט מסובך יותר נותן (תרגיל):

$$\hat{\mu}_{MLE} = \frac{1}{n} \sum_{k=1}^{n} x_k, \quad \hat{\Sigma}_{MLE} = \frac{1}{n} \sum_{k=1}^{n} (x_k - \hat{\mu})(x_k - \hat{\mu})^T$$

3. פילוג ברנולי: $X \sim Bern(\theta)$, כאשר $0 \le \theta \le 1$. גם במקרה זה נקבל (תרגיל):

$$\hat{\theta}_{MLE} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

<u>הטיה (bias)</u>: תכונה רצויה של משערך $\hat{\theta}$ עבור פרמטר θ הינה כי $E(\hat{\theta} \mid \theta) = \theta$. נשים לב כי התוחלת בביטוי זה הינה ביחס למדידה D, דהיינו

$$E(\hat{\theta} \mid \theta) = E(\hat{\theta}(D) \mid \theta) = \int \hat{\theta}(D) p(D|\theta) dD$$

ההפרש $b(\theta) = E(\hat{\theta} \mid \theta) - \theta$ נקרא ההטיה של המשערך. משערך שעבורו $b(\theta) = 0$ ייקרא משערך בלתי-מוטה.

<u>דוגמאות</u>:

1. (המשך):
$$E(\hat{\mu}_{MLE} \mid \mu) = E(\frac{1}{n}\sum_{k=1}^{n} x_k \mid \mu) = \mu$$

משערך זה הינו בלתי-מוטה.

2. (המשך):
$$E(\hat{\Sigma}_{MLE} \mid \Sigma, \mu) = \ldots = \frac{n-1}{n}\Sigma$$

משערך זה הינו מוטה. על מנת לקבל משערך בלתי מוטה "מתקנים" לעיתים את $\hat{\Sigma}_{MLE}$ על ידי חלוקה ב- $(n-1)$ במקום n, דהיינו: $\hat{\Sigma} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \hat{\mu})(x_k - \hat{\mu})^T$. גודל זה נקרא קווריאנס המדגם.

3.3 שיטות לא-פרמטריות

בגישה הלא-פרמטרית לא מניחים צורה מסוימת עבור פילוג ההסתברות המבוקש. ערך פונקצית הצפיפות $p_X(x)$ בנקודה רצויה x יתקבל כעיקרון על ידי אינטרפולציה מנקודות מדגם סמוכות. כלומר, אנו נחשב את הגודל $p\left(x\middle|D\right)$ ישירות, ללא פירוק פרמטרי (ראינו לעיל $p(x|D) = \int p(x|\theta)p(\theta|D)d\theta$).

נזכיר כי מטרתנו להעריך את פונקציית צפיפות ההסתברות $p_X(x)$ של משתנה מקרי X מתוך סדרת מדידות בלתי תלויות של X: $D = \{x_k\}_{k=1}^{n}$. הדיון פה הוא במשתנה רציף, אולם ניתן להכלילו בקלות למשתנים בדידים.

<u>א. הצפיפות האמפירית והחלקתה</u>

נגדיר את פונקצית הצפיפות האמפירית באופן הבא:

$$\hat{p}_{\delta}(x) = \frac{1}{n}\sum_{k=1}^{n}\delta(x - x_k)$$

כאשר $\delta(x)$ פונקצית ההלם של דירק. זו נותנת משקל שווה לכל מדידה, המרוכז כולו בנקודה בה התקבלה המדידה.

מתוך פונקצית הצפיפות האמפירית ניתן להעריך גם את ההסתברות של מאורע כלשהו:

$$\hat{P}(A) = \frac{1}{n}\sum_{k=1}^{n} I\{x_k \in A\}$$

פונקצית הצפיפות האמפירית אינה מהווה הערכה טובה לפונקצית הצפיפות, עקב החוסר בהכללה (או החלקה): בכל נקודה (או סט) שבה לא התקבלה מדידה, הערכת הצפיפות תהיה אפסית. לפיכך נדרשת החלקה כלשהי של המדידות שהתקבלו. נתאר מספר שיטות פשוטות לכך.

ב. היסטוגרמה

נחלק את מרחב המצב של X, שיסומן S_X, למספר אזורים (או תאים) זרים: $S_X = \bigcup_{j=1}^{J} R_j$. בכל תא R_j נעריך את פונקצית הצפיפות על ידי מספר הדגימות היחסי באותו תא:

$$\hat{p}_X(x) = \frac{\hat{P}(R_j)}{v(R_j)}, \quad x \in R_j$$

כאשר

$$\hat{P}(R_j) = \frac{N_j}{n} = \frac{1}{n}\sum_{k=1}^{n} I(x_k \in R_j)$$

הינו מספר הדגימות היחסי בתא, ואילו $v(R_j)$ הינו נפח התא: $v(R_j) = \int_{R_j} dx^{(1)} \cdots dx^{(d)}$.

בדקו כי אכן זוהי פונקציית צפיפות (דהיינו אי-שלילית ומנורמלת ליחידה, 1) (תרגיל)

גישת ההיסטוגרמה מאפשרת הערכה קלה של פונקצית הצפיפות כולה (ולא רק בנקודה x מסוימת), ומקובלת מאוד להצגה גרפית של פונקצית הצפיפות במקרה החד מימדי.

הערות:

- לפי הגדרתה, פונקצית הצפיפות המתקבלת בגישת ההיסטוגרמה תהיה בלתי-רציפה בגבולות התאים R_j. בהמשך נדון בשיטות אינטרפולציה חלקות יותר.
- בחירת גודל התא היא משתנה קריטי הקובע את דיוק השיטה. תאים גדולים מדי מבצעים מיצוע של פונקצית הצפיפות בתחום רחב, בעוד תאים קטנים מדי עשויים לכלול מספר קטן של דגימות, כך שההערכה של $\hat{p}_X(x)$ תהיה רועשת (וריאנס גבוה). ניגוד זה מהווה ביטוי נוסף של בעיית ההטיה-לעומת-השונות (bias-variance tradeoff) שהזכרנו בהקשר לבחירת סדר המודל. עבור n גבוה, בחירה סבירה של מספר התאים היא בסדר גודל של $\sqrt{n}$. מעבר לכך לא ניכנס פה לדיון מפורט בשאלה זו.

ג. איטרפולציה עם פונקצית חלון

תהי $\phi(z)$ פונקציה כלשהי על מרחב X. נגדיר

$$\hat{p}_\phi(x) = \frac{1}{n}\sum_{k=1}^{n}\phi(x_k - x)$$

ניתן לראות ביטוי זה כהחלקה (אינטרפולציה) של פונקציות ההלם בנקודות x_i. ניתן גם לוודא כי $\hat{p}_\phi = \hat{p}_\delta * \phi$ (קונוולוציה של הפילוג האמפירי עם פונקצית החלון). תכונות סבירות הנדרשות מ- $\phi(z)$ הינן כלהלן:

1. $\phi(z) \geq 0$
2. $\phi(z)$ חיובי בסביבת הראשית, ו- $\phi(z) \to 0$ כאשר $\| z \| \to \infty$.
3. $\int \phi(z)dz = 1$ (מדוע?)

פונקציה כזו מכונה פונקצית חלון, או חלון פרזן (Parzen Window).

על מנת להדגיש את חשיבות רוחב החלון, נהוג לרשום אותו כפרמטר. במקרה כזה:

$$\phi_h(z) = \frac{1}{h^d}\phi\left(\frac{z}{h}\right)$$

כאשר $\phi(z)$ פונקצית החלון הבסיסית. עתה נקבל

$$\hat{p}_h(x) = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{h^d}\phi\left(\frac{x_k - x}{h}\right).$$

פונקציות חלון מקובלות כוללות:

1. חלון ריבועי –

$$\phi(z) = I\{| z^{(j)} | \leq \tfrac{1}{2},\ \ j = 1,\ldots,d\} = \begin{cases} 1: & | z^{(j)} | \leq \tfrac{1}{2},\ \ j = 1,\ldots,d \\ 0: & \text{otherwise} \end{cases}$$

2. חלון גאוסי –

$$\phi(z) = \frac{1}{(2\pi)^{d/2}}\exp(-\tfrac{1}{2}\| z \|^2)$$

הערות:

1. איטרפולציית חלון נותנת הערכה חלקה יותר מהיסטוגרמה. בפרט, כאשר נבחר חלון ϕ רציף, פונקציית הצפיפות המשוערכת $\hat{p}_\phi(x)$ תהיה אף היא פונקציה רציפה של x.
2. עבור חלון ריבועי, ניתן לראות כי $\hat{p}_\phi(x)$ הינה פשוט המספר המנורמל של הדגימות שהתקבלו באיזור ריבועי מסביב לנקודה x.
3. אינטרפולציית חלון מקובלת במיוחד כאשר נדרשת הערכת הצפיפות בנקודות מסוימות (לעומת הערכת הפונקציה כולה).
4. בדומה לקביעת גודל האזורים בשיטת ההיסטוגרמה, בחירת רוחב החלון מהווה גורם מרכזי הקובע את איכות התוצאה. אפשרות מעניינת הינה בחירה אדפטיבית של גודל החלון: למשל, עבור חלון ריבועי, ניתן לבחור את h כך שיכלול מספר נתון (k) של דגימות מסביב לנקודה הנחקרת x. זוהי גישת k-NN, שכבר נתקלנו בה בבעיית הסיווג. בחירה סבירה הינה $k = O(\sqrt{n})$.

3.4 עירוב גאוסי

במקרים רבים הפילוג המשוערך $p_X(x)$ הינו מולטי-מודלי, כלומר מרוכז במספר אזורים במרחב.

מודל מקובל עבור פונקציית הצפיפות במקרה זה הינו מודל עירוב (Mixture), ובמיוחד העירוב הגאוסי (Gaussian Mixture):

$$p_X(x \mid \theta) = \sum_{j=1}^{J} w_j N(x; \mu_j, \Sigma_j)$$

כאשר

$$N(x;\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp(-\tfrac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))$$

על מנת ש- p_X אכן תהיה צפיפות הסתברות נדרוש: $\sum_j w_j = 1$, $w_j \geq 0$.

<u>אינטרפרטציה</u>: ניתן לפרש צורה זו של פונקצית הפילוג כבחירה דו-שלבית של הדגימה x:

(1) בשלב הראשון נבחר האינדקס j, בהסתברות w_j.

(2) בשלב שני נבחר x, לפי הפילוג הגאוסי $N(\mu_j, \Sigma_j)$.

<u>בעיית השיערוך</u>: הפרמטרים אותם עלינו לשערך הינם: $\theta=\{w_j,\mu_j,\Sigma_j\}_{j=1}^J$. משערך הסבירות המירבית (MLE) מוגדר באופן הרגיל:

$$\hat{\theta}_{MLE}=\arg\max_{\theta\in\Theta} p(D\,|\,\theta)=\arg\max_{\theta\in\Theta}\sum_{k=1}^{n}\log p_X(x_k\,|\,\theta)$$

לרוע המזל, בעיית האופטימיזציה היא בעיה קשה. ראשית, לא ניתן לקבל ביטוי סגור עבור הפרמטרים. יתר על כן, פונקציית הסבירות $L(\theta)=p(D\,|\,\theta)$ היא פונקציה לא קעורה, עם נקודות מקסימום מקומיות מרובות, דבר המקשה על הפעלת האלגוריתמים הסטנדרטיים לאופטימיזציה נומרית.

הגישה המקובלת לפתרון בעיית אופטימיזציה זו היא אלגוריתם דו-שלבי, כלהלן.

אלגוריתם 1:

1. בחר ניחוש התחלתי לפרמטרים $\theta=\{\hat{w}_j,\hat{\mu}_j,\hat{\Sigma}_j\}$.
2. שיוך המדידות: שייך כל דגימה x_k לגאוסיאן j הסביר ביותר, לפי:

$$j_k=\arg\max_{1\le j\le J}\left\{\hat{w}_j N(x_k;\hat{\mu}_j,\hat{\Sigma}_j)\right\}$$

3. עדכון פרמטרי הגאוסיאנים: עדכן את הפרמטרים $\hat{w}_j,\hat{\mu}_j,\hat{\Sigma}_j$ עבור כל גאוסיאן j לפי המדידות ששויכו לגאוסיאן זה:

$$\hat{w}_j=\frac{n_j}{\sum_i n_i},\qquad \text{where}\quad n_j=\sum_{k=1}^{n} I\{j_k=j\}$$

$$\hat{\mu}_j=\frac{1}{n_j}\sum_{k=1}^{n} I\{j_k=j\}x_k$$

$$\hat{\Sigma}_j=\frac{1}{n_j}\sum_{k=1}^{n} I\{j_k=j\}(x_k-\hat{\mu}_j)(x_k-\hat{\mu}_j)^T$$

4. חזור על שלבים 2,3 עד להתכנסות (כלומר: עד שלא מתקבל שינוי בשלב 2).

הערות:

1. ניתן לראות את שלב 2 כסיווג בייסיאני (MAP) של כל דגימה x_k לאחת מהמחלקות $\{\omega_j\}_{j=1}^J$, כאשר $P(\omega_j)=\hat{w}_j$, $p(x\,|\,\omega_j)=N(x;\hat{\mu}_j,\hat{\Sigma}_j)$.
2. ניתן להראות כי בכל שלב של אלגוריתם זה מגדיל (או לפחות לא מקטין) את פונקצית הסבירות. עם זאת התכנסות אינה מובטחת, והיא תהיה בד"כ למקסימום מקומי.

3. אלגוריתם האישכול K-means (כאשר K=J) הינו למעשה מקרה פרטי של אלגוריתם זה, עם $\Sigma_k \equiv I$.

אלגוריתם 2 (EM) : גירסה משופרת של אלגוריתם 1 נמנעת משיוך מוחלט של x_k למחלקה אחת j_k, ומחליפה אותו בשיוך הסתברותי. כלומר, שלב 2 מוחלף על ידי :

$$q_{kj} = \frac{\hat{w}_j N(x_k;\hat{\mu}_j,\hat{\Sigma}_j)}{\sum_j \hat{w}_j N(x_k;\hat{\mu}_j,\hat{\Sigma}_j)} \quad (= p(\omega_j \mid x_k))$$

ושלב 3 מוחלף בהתאם על ידי

$$\hat{w}_j = \frac{n_j}{\sum_i n_i}, \qquad \text{where} \quad n_j = \sum_{k=1}^{n} q_{kj}$$

$$\hat{\mu}_j = \frac{1}{n_j}\sum_{k=1}^{n} q_{kj} x_k$$

$$\hat{\Sigma}_j = \frac{1}{n_j}\sum_{k=1}^{n} q_{kj}(x_k-\hat{\mu}_j)(x_k-\hat{\mu}_j)^T$$

הערות :

1. האלגוריתם המשופר הינו מקרה פרטי של אלגוריתם EM (Expectation-Maximization) שהוא אלגוריתם סטנדרטי לחישוב משערך MLE בבעיות מורכבות.
2. התכנסות אלגוריתם זה לנקודת מכסימום מקומית הינה מובטחת.
3. למודל העירוב הגאוסי קיים גם יישום ישיר לבעיית הסיווג, ראה DHS פרק 10.4.

3.5 אנטרופיה

אחד המדדים למידת האקראיות של מידת הסתברות היא האנטרופיה. עבור מידת הסתברות נתונה מעל מרחב סופי Ω כאשר p_ω היא ההסתברות עבור כל ω ב- Ω אנו מגדירים את האנטרופיה כ :

$$H(p) = -\sum_{\omega\in\Omega} p_\omega \log_2(p_\omega)$$

כאשר $0\log 0$ מוגדר כ- 0.

כמה עובדות על אנטרופיה :

1. $0 \le H(p) \le \log_2(|\Omega|)$

2. $H(p)=0$ אם ורק אם p מנוונת.
3. $H\left(p\right)=\log_2\left(\left|\Omega\right|\right)$ אם ורק אם p אחידה.

בהמשך הקורס נשתמש באנטרופיה כדי להעריך את מידת האקראיות של מידת הסתברות.