פרק ב

הסקה סטטיסטית

2.1 על בעיית ההסקה הסטטיסטית

הסקה סטטיסטית (statistical inference) מטפלת במצב בו יש לנו נתונים שנוצרו מתוך התפלגות שאינה ידועה לנו, ועלינו לנתח אותם ולהסיק מסקנות לגביהם ולגבי ההתפלגות שיצרה אותם. במילים אחרות, ברבות מהבעיות הסטטיסטיות בהן נדון, קיימות כמה התפלגויות אפשריות שיצרו נתונים מסוימים (ובבעיות אמיתיות, מספר אינסופי של התפלגויות אפשריות כאלו), ואנחנו מנסים ללמוד על ההתפלגויות האלו, להסיק על תכונות מסוימות שלהן, ולקבוע את הסבירות שכל אחת מההתפלגויות האלו היא זו שיצרה את הנתונים בפועל.

בפרק הנוכחי נתאר את הגישה הבייסיאנית להסקה סטטיסטית. כדי להבהיר את המושגים הבסיסיים, נתמקד במקרה הפשוט בו יש מספר קטן של התפלגויות אפשריות שיצרו את הנתונים.

2.2 הסקה והכרעה בייסיאנית

תורת ההסקה הבייסיאנית היא גישה סטטיסטית לקבלת החלטות בתנאי אי ודאות. גישה זו מבוססת על ההנחה כי הידע הרלבנטי להחלטה מבוטא בצורה הסתברותית וכי כל ההסתברויות הרלוונטיות ידועות. המודל הפורמלי להכרעה בייסיאנית מבוסס על חמישה מרכיבים שיתוארו להלן.

2.2.1 תמונת העולם בגישה הבייסיאנית

הדוגמא הקאנונית המשמשת לתיאור תמונת העולם הבייסיאנית, היא אדם היוצא מהבית ביום חורפי ומתלבט האם לקחת עמו מטריה. נניח לשם הפשטות כי קיימות מבחינתו שתי אפשרויות בלבד: יהיה יום גשום או לא. מצד אחד הוא חושש להרטב אם לא ייקח מטריה ויהיה גשום, ומצד שני אם ייקח מטריה ביום ללא גשם, ייסחב אתה שלא לצורך. האדם מציץ מהחלון ורואה עננים שחורים וכבדים, ולכן מחליט שהסיכון לגשם גובר, ומחליט לקחת מטריה. תיאור פורמלי של הבעיה במונחים בייסיאנים מתבסס על המרכיבים הבאים:

קבוצת מצבי העולם האפשריים $\Omega=\{\omega_i\}$

"מצבי העולם" מוגדרים כך שידיעת מצב העולם מספקת לנו מידע הסתברותי מקסימלי: ידועות לנו ההתפלגויות שיצרו את התצפיות. מצבי העולם השונים הם זרים $\omega_i \cap \omega_j = \phi$ וממצים $\bigcup \omega_j = \Omega$.

תצפיות $X=\{x_1,\ldots,x_n\}$

אלו הם הנתונים שיש בידינו ומהם אנחנו מנסים להסיק מהו מצב העולם. בדרך כלל לא נוכל להסיק בוודאות מתוך התבוננות בתצפיות מהו מצב העולם.

מודל הסתברותי של העולם $P=\{P_0(\omega_i),P(X|\omega_i)\}$

על פי הגישה הבייסיאנית אנו מניחים כי יש לנו ידע הסתברותי מפורש על העולם. ידע זה כולל הסתברויות א-פריוריות $P_0(\omega_i)$ על הסיכוי להמצא במצב עולם ω_i, והסתברויות מותנות לערכי התצפיות X בהינתן מצב עולם נתון $P(x_j|\omega_i)$.

פעולות אפשריות $A=\{\alpha_1,\ldots,\alpha_k\}$

קבוצת הפעולות מביניהן עלינו לבחור. לכל פעולה נקבע מחיר (ראה הפריט הבא) התלוי במצב העולם, ונשאף כמובן לבחור בפעולה המתאימה ביותר למצב העולם.

מחיר לכל פעולה $\Lambda=\{\lambda(\alpha_k,\omega_i)\}$

לתוצאות של הפעולות שלנו יש, כידוע, מחיר, וזה נקבע על פי מצב העולם. פעולה שאינה מתאימה למצב העולם בו אנחנו נמצאים תלווה בדרך כלל בקנס (מחיר בעל ערך חיובי), ופעולה מתאימה תלווה ברווח עבורנו (מחיר אי שלילי). המחיר של פעולה α_k במצב עולם ω_i יסומן ב- $\lambda(\alpha_k|\omega_i)$, ואת מטריצת המחירים נסמן ב- $\Lambda=\{\lambda(\alpha_k|\omega_i)\}$.

בדוגמת המטריה שתיארנו קודם, הרי שישנם שני מצבי עולם אפשריים (יש או אין גשם), ונניח כי שכיחותם של ימי הגשם בחורף ידועה. ישנן גם שתי פעולות אפשריות (לקחת מטריה או לא), ולשתיהן מחירים שונים כתלות בשאלה האם ירד גשם או לא. התצפית (עננים גשם) משנה את ההערכה על ההסתברות שגשם אכן ירד, ומשפיעה על ההחלטה לקחת מטריה.

הגישה הבייסיאנית לקבלת החלטות דורשת שיהיו בידיכם הן ההסתברות האפריוריות (שכיחות ימי הגשם), והן ההסתברויות המותנות (מה ההסתברות לעננות כבדה ביום גשום). למרות שמידע כזה אינו ידוע בדרך כלל במפורש לכל אדם, הרי שאין מניעה עקרונית לאסוף אותו, כך שהאזרח התמים יוכל לשמור על בגדיו יבשים במינימום מאמץ.

נפנה כעת לדון באסטרטגיה הנכונה לקבלת החלטות בגישה הבייסיאנית.

2.2.2 הכרעה בייסיאנית

בהינתן בעיית ההכרעה הבייסיאנית $\{\Omega,X,P,A,\Lambda\}$, נרצה לבחור את הפעולה האופטימלית שכדאי לנקוט אם אנו רואים תצפית x_j. לצורך כך, ננסה כעת להגדיר פונקצית החלטה דטרמיניסטית $\alpha : X \to A$ המתאימה לכל תצפית x_j פעולה אופטימלית α_k. עד כה הגדרנו מחיר לפעולות בהינתן מצב העולם, אך מה שנתון לנו בפועל הן התצפיות ולכן עלינו לשקלל את מחירי הפעולות בהתאם להסתברויות של מצבי העולם השונים, כפי שהן מושפעות מהתצפיות שברשותנו.

הסיכון המותנה

כדי למצוא פונקצית החלטה דטרמיניסטית אופטימלית נגדיר את הסיכון המותנה (Conditional Risk) לביצוע פעולה α_k בהינתן שראינו תצפית x_j

$$R(\alpha_k \mid x_j) \equiv \sum_{\omega_i \in \Omega} \lambda(\alpha_k \mid \omega_i) P(\omega_i \mid x_j) \ , \qquad (2.1)$$

ואת ההסתברות האפוסטריורית להימצא במצב עולם ω_i נחשב תוך שימוש בנוסחת בייס (סעיף 1.2.2)

$$P(\omega_i \mid x_j) = \frac{P(x_j \mid \omega_i)}{P(x_j)} P_0(\omega_i) = \frac{P(x_j \mid \omega_i)}{\sum_{\omega_t} P(x_j \mid \omega_t) P_0(\omega_t)} P_0(\omega_i)$$

הסיכון הכולל

בהינתן אסטרטגיית הכרעה הקובעת באיזו פעולה ננקוט עבור כל תצפית, ניתן לחשב את הסיכון הכולל של שימוש בפונקציה כזו. הסיכון הכולל הוא ממוצע הסיכונים על פני התצפיות האפשריות:

$$R[\alpha(x)] \equiv \sum_j R(\alpha(x_j) \mid x_j) P(x_j) \qquad (2.2)$$

ובמקרה הרציף

$$R[\alpha(x)] \equiv \int_X R(\alpha(x_j) \mid x_j) P(x_j) dx$$

משפט: פונקצית ההכרעה האופטימלית

פונקצית ההכרעה $\alpha^*(x)$ המביאה למינימום את הסיכון הכולל תהיה הפונקציה המביאה למינימום את הסיכון המותנה לכל תצפית אפשרית. במלים אחרות, פונקצית ההכרעה האופטימלית קובעת לכל תצפית x את הפעולה בעלת הסיכון המותנה הקטן ביותר. ובאופן פורמלי: בהינתן x הכרע α^* אם לכל $\alpha'\neq\alpha^*$ מתקיים $R(\alpha^*|x)\leq R(\alpha'|x)$.

הוכחה

לכל $\alpha'\neq\alpha^*$ ולכל x מתקיים $R(\alpha^*|x)\leq R(\alpha'|x)$, ולכן מתקיים $\sum_i R(\alpha^*(x_i)|x_i)P(x_i)\leq\sum_i R(\alpha'(x_i)|x_i)P(x_i)$ ולכן $R[\alpha^*(x)]\leq R[\alpha'(x)]$ כנדרש.

פונקצית δ כפונקצית מחיר

הסיכון מקבל משמעות פשוטה כאשר פונקצית המחיר מקבלת ערך אפס אם בחרנו נכונה וערך 1 אם שגינו $\lambda(\alpha_k|\omega_i)=1-\delta_{ki}$. עם פונקצית המחיר הזו, אנחנו משלמים מחיר רק אם טעינו, ולכן הסיכון המותנה הוא פשוט הסיכוי לטעות

$$R(\alpha_k|x_j)=\sum_i(1-\delta_{ki})P(\omega_i|x_j)=\sum_{i\neq k}P(\omega_i|x_j)$$

והסיכון הכולל יהיה הסיכוי הכולל לטעות (עבור כל התצפיות האפשריות). כלל ההכרעה האופטימלי במקרה זה גם הוא פשוט -

"בחר את מצב העולם הסביר ביותר בהנתן x"

ובאופן פורמלי בחר $\alpha(x)=\alpha_k$ כך ש- $P(w_k|x)$ יהיה מקסימלי.

2.2.3 שני מצבי עולם

הכרעה בייסיאנית אופטימלית

ראינו כי ההכרעה הבייסיאנית האופטימלית מתבצעת על ידי בחירת אסטרטגיית-פעולה שהיא בעלת הסיכון המותנה הנמוך ביותר. במקרה שקיימים רק שני מצבי עולם, ושתי פעוות אפשריות, אסטרטגיה זו מקבלת צורה פשוטה במיוחד. אם α_i היא הפעולה המתאימה למצב עולם ω_i ו- λ_{ij} הוא המחיר שנשלם על הפעולה α_i במצב עולם ω_j, $\lambda_{ij} = \lambda\left(\alpha_i \mid \omega_j\right)$, אז הסיכון המותנה בבחירת הפעולה α_1 הוא

$$R(\alpha_1 \mid x) = \lambda_{11} P(\omega_1 \mid x) + \lambda_{12} P(\omega_2 \mid x)$$

והסיכון המותנה בבחירת הפעולה α_2 הוא

$$R(\alpha_2 \mid x) = \lambda_{21} P(\omega_1 \mid x) + \lambda_{22} P(\omega_2 \mid x)$$

ובגבול ההכרעה יהיו כל התצפיות שעבורן מתקיים שוויון בין הסיכונים,

$$\lambda_{21} P(\omega_1 \mid x) + \lambda_{22} P(\omega_2 \mid x) = \lambda_{11} P(\omega_1 \mid x) + \lambda_{12} P(\omega_2 \mid x)$$

$$\left(\lambda_{21} - \lambda_{11}\right) P(\omega_1 \mid x) = \left(\lambda_{12} - \lambda_{22}\right) P(\omega_2 \mid x)$$

כלומר כאשר

$$\frac{P(\omega_1 \mid x)}{P(\omega_2 \mid x)} = \frac{\lambda_{22} - \lambda_{12}}{\lambda_{11} - \lambda_{21}} \qquad (2.3)$$

נשתמש בנוסחת בייס, נעביר אגפים ונקבל

$$\frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} = \frac{P_0(\omega_2)}{P_0(\omega_1)} \cdot \frac{\lambda_{22} - \lambda_{12}}{\lambda_{11} - \lambda_{21}}$$

אגף שמאל של המשוואה נקרא **יחס הנראות** (Likelihood ratio), זהו היחס בין הנראות של התצפית (ההסתברות לראות תצפית) במצב העולם הראשון לבין הנראות במצב העולם השני. נוכל אם כן להגדיר סף Θ:

$$\Theta = \frac{P_0(\omega_2)}{P_0(\omega_1)} \cdot \frac{\lambda_{22} - \lambda_{12}}{\lambda_{11} - \lambda_{21}},$$

ולחלק באמצעותו את מרחב התצפיות לשני אזורים זרים: אזור בו יחס הנראות גדול מהסף Θ ובו נכריע ω_1 ואזור בו יחס הנראות קטן מהסף Θ ובו נכריע ω_2. הגבול בין שני אזורים אלו יהיה כל התצפיות עבורן מתקיים

$P(x|\omega_1)/P(x|\omega_2)=\Theta$, והוא נקרא **גבול ההכרעה** (Decision Boundary).
המבחן שבו נשתמש יהיה אם כן להכריע ω_1 אם ורק אם

$$\frac{P(x|\omega_1)}{P(x|\omega_2)}>\Theta \qquad (2.4)$$

מקרה פרטי: פונקצית δ כפונקצית מחיר

נטפל כעת במקרה בו פונקצית המחיר היא $\lambda(\alpha_k,\omega_i)=1-\delta_{kj}$ ועלינו להכריע בין שני מצבי העולם. הסיכון הכולל $R[\alpha]$ במקרה כזה הוא עבור משתנים מקריים בדידים

$$\begin{aligned}P_{error}&=\sum_t \min\left(P(\omega_1|x_t),P(\omega_2|x_t)\right)P(x_t)=\\&=\sum_t \min\left(\frac{P(x_t|\omega_1)P(\omega_1)}{P(x_t)},\frac{P(x_t|\omega_2)P(\omega_2)}{P(x_t)}\right)P(x_t)\\&=\sum_t \min\left(P(x_t|\omega_1)P(\omega_1),P(x_t|\omega_2)P(\omega_2)\right)\end{aligned}$$

ושוב מרחב התצפיות מתחלק לאזור בו מתקיים $P(x|\omega_1)P(\omega_1)>P(x|\omega_2)P(\omega_2)$, ובו נכריע לטובת מצב העולם ω_1, ושאר המרחב בו נכריע לטובת מצב העולם ω_2.

דוגמא

עלי הכותרת של הפרח המצוי "לבלב מצוי" ניחנים באורך מופלג המתפלג באופן אחיד בין סנטימטר אחד לבין 1.1 סנטימטר.

$$P(x|\omega_1)=\begin{cases}10 & 1\le x\le 1.1\\ 0 & otherwise\end{cases}$$

עלי הכותרת של הזן הנדיר "לבלב נדיר" (הזהה לחלוטין לאחיו) הם בעלי נטייה להיות ארוכים יותר, על פי פונקצית ההתפלגות

$$P(x|\omega_2)=\begin{cases}20(x-1) & 1\le x\le 1.1\\ 0 & otherwise\end{cases}$$

קל לוודא כי פונקציות אלו הן התפלגויות והאינטגרל עליהם הוא אחד. מהו כלל ההכרעה האופטימלי לאבחנה בין שני סוגי הלבלבים אם ידוע כי בדיוק 55 אחוזים מהלבלבים הפורחים במחוזותינו נמנים על פרח הלבלב המצוי, והשאר הם לבלבים "נדירים"?

נרצה למצוא כלל הכרעה כפונקציה של אורך העלים, כך שלכל פרח שנמצא, נוכל להכריע בין שני מצבי העולם. נרצה להכריע "לבלב מצוי" אם (ורק אם) מתקיים $P(\omega_1 \mid x) > P(\omega_2 \mid x)$. נרשום אם כן

$$P(\omega_1 \mid x) = \frac{P(x \mid \omega_1)P_0(\omega_1)}{P(x \mid \omega_1)P_0(\omega_1) + P(x \mid \omega_2)P_0(\omega_2)} =$$

$$= \frac{10 \cdot 0.55}{10 \cdot 0.55 + 20(x-1) \cdot 0.45} = \frac{5.5}{9x - 3.5}$$

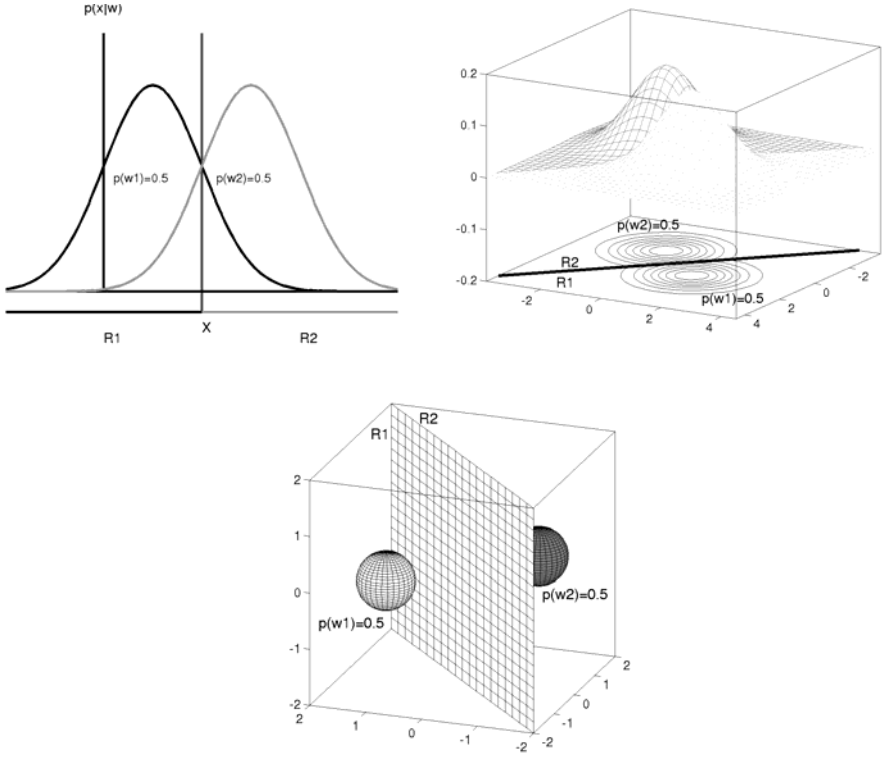
$$P(\omega_2 \mid x) = 1 - P(\omega_2 \mid x) = \frac{9x - 9}{9x - 3.5}$$

הנקודות על גבול ההכרעה מקיימות $P(\omega_1 \mid x) = P(\omega_2 \mid x)$, דהיינו

$$\frac{5.5}{9x - 3.5} = \frac{9x - 9}{9x - 3.5} \qquad \Rightarrow \qquad x = 1.611$$

ולכן נכריע לטובת הלבלב הנדיר אם ורק אם אורך עלי הכותרת יהיה גדול מ-1.611 , כלומר אף פעם.

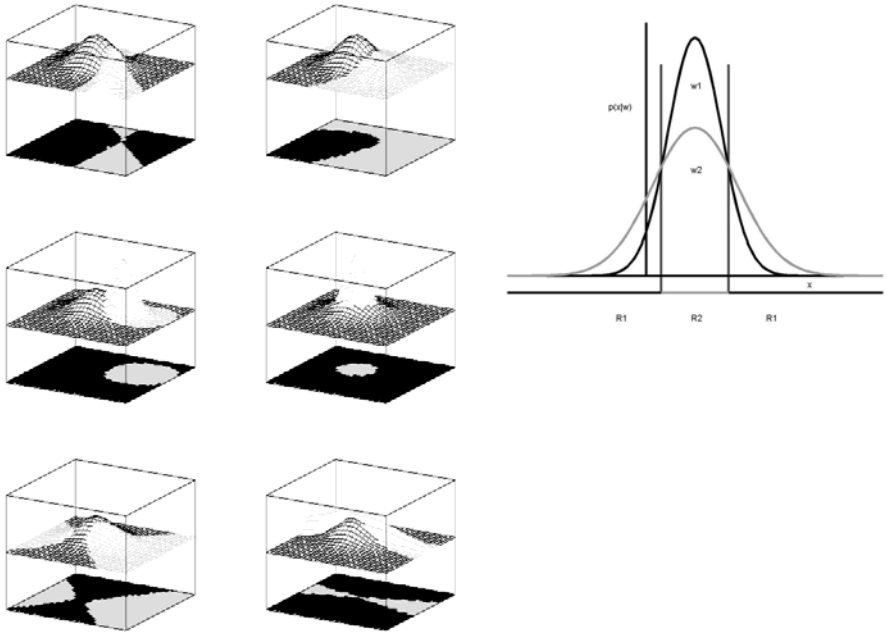
דוגמא: גבול הכרעה עבור שני מצבי עולם והתפלגויות נורמליות שוות שונות



איור 2.1

גבול ההכרעה בין שתי התפלגויות נורמליות בעלות שונויות שוות הוא מפריד לינארי. הדגמה עבור התפלגויות חד מימדיות, דו מימדיות ותלת מימדיות.

דוגמא: גבול הכרעה עבור שני מצבי עולם והתפלגויות נורמליות דו ממדיות



איור 2.2

גבולות ההכרעה בין שתי התפלגויות נורמליות בעלות שונויות שונות. במקרה החד ממדי מתקבלים תחום שאינו רצוף. במקרה הדו ממדי גבולות ההכרעה הן פונקציות ממעלה שניה (אליפסות, היפרבולות).

2.3 בדיקת השערות פשוטות ומבחן סף

בסעיף הקודם תיארנו את הגישה הבייסיאנית לקבלת החלטות בתנאי אי ודאות. על מנת להשלים את התמונה, נתאר כעת בקצרה גישה סטטיסטית שונה לבדיקת השערות.

2.3.1 מושגים בבדיקת השערות

הגדרות

נניח כי אוסף מצבי העולם מתחלק לשתי קבוצות זרות אותן נסמן Ω_0 ו- Ω_1. נסמן ב- H_0 את ההשערה כי מצב העולם הוא בקבוצה Ω_0 וכן נסמן ב- H_1 את ההשערה כי מצב העולם הוא בקבוצה Ω_1. כאשר Ω_0 מכילה רק מצב עולם יחיד, אזי ההשערה H_0 מכונה **השערה פשוטה** (Simple hypothesis), בעוד שבמקרה בו הקבוצה מכילה מספר מצבי עולם אפשריים היא מכונה **השערה מורכבת** (composite hypothesis). באופן דומה מגדירים עבור H_1 ו- Ω_1.

עד כה התייחסנו להשערות H_0 ו- H_1 באופן סימטרי, אך בבעיות רבות נהוג להתייחס אליהן באופן שונה. נהוג ש- H_0 מסמלת את המצב השכיח (ברירת המחדל) ומכונה **השערת האפס** (The null hypothesis), בעוד H_1 מסמלת את המצב הנדיר או המסוכן ומכונה **ההשערה האלטרנטיבית** (The alternative hypothesis).

דוגמא:

נאמר שאנחנו רוצים לזהות האם בבדיקת משטח גרון ישנו זיהום חיידקי. ידוע כי תוצאת ספירת החיידקים באדם בריא מתפלג נורמלית עם ממוצע 10 ושונות 20, ואילו באדם חולה הספירה מתפלגת נורמלית עם ממוצע בין 15 ל- 20 ושונות 25. במקרה זה השערת האפס תהיה כי האדם בריא, והיא השערה פשוטה, בעוד שההשערה האלטרנטיבית H_1 היא ההשערה שהאדם חולה והיא השערה מורכבת היות והקבוצה Ω_1 מכילה קבוצה שלמה של מצבי עולם אפשריים, לכל אחד מהם תוחלת אחרת.

שני סוגי שגיאות

כאשר קיימות שתי קבוצות של מצבי עולם יש גם שני סוגים של שגיאות אפשריות. שגיאה ראשונה (false positive) היא המקרה בו נקבל בטעות את H_1 למרות שמצב העולם הוא ב- Ω_0. במקרה של השגיאה השניה (false negative) נקבל בטעות את H_0.

2.3.2 פרוצדורות הכרעה אופטימליות

פרוצדורת הכרעה להשערות פשוטות: משפט ניימן-פירסון

נתאר כעת פרוצדורת הכרעה אופטימלית כאשר שתי ההשערות הן פשוטות. תהי δ פרוצדורת הכרעה כלשהי, אז נהוג לסמן את הסתברויות השגיאה באופן הבא:

$$\begin{aligned} \alpha(\delta) &= \Pr(\text{Rejecting } H_0 | \Omega_0) \\ \beta(\delta) &= \Pr(\text{Accepting } H_0 | \Omega_1) \end{aligned} \tag{2.5}$$

השגיאה α נקראת גם המובהקות של המבחן, ו- $(1-\beta)$ נקראת עוצמת המבחן. בבואנו להגדיר פרוצדורה להכרעה בין שתי השערות נרצה להביא למינימום את השגיאות α ו- β. נוכל כמובן לקבוע מבחן שמכריע תמיד H_0, ובכך להביא את השגיאה α לאפס, אך במקרה כזה השגיאה β תהיה אחת. קריטריון שנראה סביר הוא לנסות ולהביא למינימום קומבינציה לינארית של השגיאות מהצורה $a\alpha(\delta)+b\beta(\delta)$. עבור קריטריון כזה קיימת פרוצדורת הכרעה שהיא אופטימלית במובן הבא: לכל בחירת ערך של α הפרוצדורה מביאה למינימום את β.

הפרוצדורה המבוקשת מתוארת על ידי **הלמה של ניימן-פירסון** (1933):

יהי $\Theta>0$ ו- δ^* פרוצדורת הכרעה בעלת המבנה הבא: ההשערה H_0 מתקבלת אם $f_0(x)>\Theta f_1(x)$ ואילו ההשערה H_1 מתקבלת אם $f_0(x)<\Theta f_1(x)$ (כאשר f_i היא ההסתברות לקבל התצפית x בהנחת H_i). אז לכל פרוצדורת הכרעה אחרת δ, המקיימת $\alpha(\delta)\le\alpha(\delta^*)$ מתקיים $\beta(\delta)\ge\beta(\delta^*)$, ואם $\alpha(\delta)<\alpha(\delta^*)$ אז $\beta(\delta)>\beta(\delta^*)$.

למרות פשטות ההוכחה לא נוכיח את הלמה כאן מטעמי קיצור. המסקנה ממשפט זה היא שלכל רמת מובהקות α, מבחן יחס נראות מהצורה

$$\frac{f_1(x)}{f_0(x)}>\Theta \tag{2.6}$$

משיג עוצמה מקסימלית (דהיינו שגיאת β מינימלית). בסעיף הקודם הגענו למסקנה דומה לגבי מבחן יחס נראות כאשר נקטנו בגישה בייסיאנית, אבל כאן לא נדרשנו להניח כי ידועות לנו ההתפלגויות האפריוריות של מצבי העולם, אלא קיבלנו כי מבחן יחס נראות הוא אופטימלי במקרה של הכרעה בין שתי השערות פשוטות.

השערות מורכבות

כאשר עוברים לטפל בהשערות מורכבות, דהיינו להכריע בין קבוצות אפשריות של מצבי עולם, הסתברויות השגיאה α ו- β, אינן מוגדרות היטב ויש להגדירן כראוי. פתרון בגישה הבייסיאנית יהיה להביט על השגיאות הממוצעות מסוג α ו- β, (למשל α תהיה הסיכוי לדחות את H_0 באופן ממוצע על פני מצבי העולם ב- Ω_0),

אך גישה זו דורשת לדעת את ההסתברויות האפריוריות לכל אחד ממצבי העולם ב- Ω_0 . הגישה הסטטיסטית המקובלת נמנעת מלהגדיר הסתברויות אפריוריות כאלו, ובמקום זה מגדירה

$$\alpha = \sup_{\omega \in \Omega_0} \left(\Pr(\text{Reject } H_0 \mid \omega)\right), \tag{2.7}$$

דהיינו ניקח את המקרה הגרוע ביותר מבין כל מצבי העולם בקבוצה Ω_0. במקרה זה לא קיים משפט מקביל ללמה של ניימן-פירסון ולא קיים מבחן שמבטיח עוצמה מקסימלית לכל מצב עולם ω_0;. ניתן עם זאת להגדיר פרוצדורה דומה של יחס נראות המקיימת תכונות מועילות אחרות שלא נכנס אליהן כאן.

2.4 תצפיות מרובות ומבחן סדרתי

2.4.1 שימוש בתצפיות מרובות

עד כה התמקדנו במקרה בו נתונה לנו תצפית בודדת x, וראינו כלל הכרעה אופטימלי מהצורה

$$\frac{P(x_1 \mid \omega_1)}{P(x_1 \mid \omega_2)} > \Theta .$$

אך למעשה כל הניתוח שלנו מתאים גם למקרה בו נתונות לנו תצפיות מרובות, שאז נשתמש בכלל הכרעה מהצורה

$$\frac{P(x_1,\ldots,x_n \mid \omega_1)}{P(x_1,\ldots,x_n \mid \omega_2)} > \Theta ,$$

וכפי שראינו, עבור בחירה נכונה של הסף, כלל הכרעה זו הוא אופטימלי במובן של מינימום סיכון. פעמים רבות, התצפיות שלנו נאספות על ידי חזרות מרובות על אותו ניסוי. במקרה כזה (ואם הניסוי נערך כהלכה), לכל התצפיות ישנה אותה התפלגות והן בלתי תלויות. במקרה זה המשתנים המקריים המתאימים הם שווי התפלגות ו**בלתי תלויים** זה בזה **בהנתן מצב העולם**, כך שכלל ההכרעה עבור תצפיות מרובות מקבל את הצורה

$$\prod_{i=1}^{n} \frac{P(x_i \mid \omega_1)}{P(x_i \mid \omega_2)} > \Theta .$$

2.4.2 בחינת תצפיות מרובות באופן סדרתי

הניתוח לעיל מתאים למקרה בו כל התצפיות ניתנות "בבת אחת". קיימים מקרים רבים בהם התצפיות נאספות בזו אחר זו ויש לנו אפשרות לנסות ולהכריע במהלך איסוף התצפיות. בעיות מסוג זה נקראות בעיות למידת on-line (בניגוד למקרה בו כל הדגימות נתונות מראש הנקרא למידת batch). נפנה כעת לנתח את התפתחות הציונים שתיארנו עבור תצפיות הניתנות בזו אחר זו. כפי שראינו, עבור תצפית בודדת מתקיים

$$P(\omega \mid x_1) = P_0(\omega) \cdot \frac{P(x_1 \mid \omega)}{P(x_1)} = P_0(\omega) \cdot \frac{P(\omega, x_1)}{P_0(\omega)P(x_1)}$$

הסתכלות אפשרית על נוסחה זו היא כי ההסתברות האפריורית למצב העולם $P_0(\omega)$ מוכפלת ב- "גורם תיקון" $P(\omega, x_1)/[P_0(\omega)P(x_1)]$, וכאשר גורם תיקון זה שונה מאחד, כלומר כאשר $P(x_1, \omega) \neq P_0(\omega)P(x_1)$, המדידה מספקת אינפורמציה על מצב העולם.

אם יש לנו שתי מדידות, x_1, x_2 אזי ההסתברות למצב העולם לאור שתי התצפיות תהיה

$$\begin{aligned} P(\omega \mid x_1, x_2) &= \\ &= \frac{P_0(\omega)P(x_1, x_2 \mid \omega)}{P_0(x_1, x_2)} \\ &= P_0(\omega)\frac{P(x_1 \mid \omega)}{P(x_1)} \cdot \frac{P(x_2 \mid \omega, x_1)}{P(x_2 \mid x_1)} \end{aligned}$$

וגורמי התיקון כאן הולכים ומסתבכים.

במקרה בו התצפיות בלתי-תלויות בהינתן מצב העולם, דהיינו $P(x_1, x_2 \mid \omega) = P(x_1 \mid \omega)P(x_2 \mid \omega)$, אז ניתן לרשום

$$\begin{aligned} P(\omega_i \mid x_1, \ldots, x_n) &= P_0(\omega_i)\frac{P(x_1, \ldots, x_n \mid \omega_i)}{P(x_1, \ldots, x_n)} \\ &= P_0(\omega_i)\frac{P(x_1, \ldots, x_n \mid \omega_i)}{\sum_{j=1}^{m} P(x_1, \ldots, x_n \mid \omega_j)P_0(\omega_j)} \\ &= \frac{1}{\sum_{j=1}^{m} \frac{P(x_1, \ldots, x_n \mid \omega_j)}{P(x_1, \ldots, x_n \mid \omega_i)} \frac{P_0(\omega_j)}{P_0(\omega_i)}} \\ &= \frac{1}{1 + \sum_{j \neq i}^{m} \frac{P(x_1, \ldots, x_n \mid \omega_j)}{P(x_1, \ldots, x_n \mid \omega_i)} \frac{P_0(\omega_j)}{P_0(\omega_i)}} \end{aligned}$$

ובמקרה שקיימים רק שני מצבי עולם נקבל

$$
\begin{aligned}
P(\omega_1 \mid x_1,\ldots,x_n) &= \frac{1}{1+\dfrac{P(x_1,\ldots,x_n \mid \omega_2)}{P(x_1,\ldots,x_n \mid \omega_1)}\dfrac{P_0(\omega_2)}{P_0(\omega_1)}} \\
&= \frac{1}{1+\dfrac{P_0(\omega_2)}{P_0(\omega_1)}\prod_{i=1}^{n}\dfrac{P(x_i \mid \omega_2)}{P(x_i \mid \omega_1)}} \qquad (2.8) \\
&= \frac{1}{1+\exp\left[\sum_{i=1}^{n}\log\left(\dfrac{P(x_i \mid \omega_2)}{P(x_i \mid \omega_1)}\right)+\log\left(\dfrac{P_0(\omega_2)}{P_0(\omega_1)}\right)\right]}
\end{aligned}
$$

וקיבלנו פונקציה סיגמואידית שהשיפוע שלה גדל עם n , כלומר, היכולת להבחין בין שני מצבי העולם גדלה וההסתברויות נעשות חדות עם הגידול במספר התצפיות. כאשר n גדול, ההסתברות למצב עולם בהנתן התצפיות היא או אפס, או אחת.

2.4.3 מבחן סדרתי להכרעה - Sequential Probability Ratio Test (SPRT)

נשוב לבעיית ההכרעה הבייסיאנית. בפרק הקודם תיארנו פרוצדורה להכרעה בין שני מצבי עולם בה השווינו את יחס הנראות לסף. כעת, כאשר אנחנו פועלים בתרחיש של למידת on-line, יש לפנינו שלוש אפשרויות במקום שתיים: בנוסף לשתי ההכרעות (לקבל מצב עולם 1 או לקבל מצב עולם 0) אנחנו יכולים ל"החליט שלא להחליט", ולדרוש נתונים נוספים לצורך הכרעה. מסתבר כי בדומה למבחן ההשוואה לסף אותו תיארנו בפרק הקודם, ניתן לבחור ספים עבור פרוצדורה מסוג זה כך שיובטחו הסתברויות השגיאה הנדרשות. נעבור אם כן לתיאור פורמלי של פרוצדורת ההכרעה מסוג זה.

משפט: (Wald 1942) בהנתן $1 \ge \alpha,\beta \ge 0$ נגדיר מבחן "סדרתי" המשתמש בשני ספים

$$
(2.9) \qquad decision = \begin{cases}
\dfrac{1-\alpha}{\beta} < \dfrac{L\left(x^{(n)} \mid \omega_0\right)}{L\left(x^{(n)} \mid \omega_1\right)} & \omega_1 \\[2ex]
\dfrac{\alpha}{1-\beta} < \dfrac{L\left(x^{(n)} \mid \omega_0\right)}{L\left(x^{(n)} \mid \omega_1\right)} < \dfrac{1-\alpha}{\beta} & continue \\[2ex]
\dfrac{L\left(x^{(n)} \mid \omega_0\right)}{L\left(x^{(n)} \mid \omega_1\right)} < \dfrac{\alpha}{1-\beta} & \omega_0
\end{cases}
$$

אם נסמן ב- α' את הסתברות השגיאה מסוג ראשון של מבחן זה, וב- β' את הסתברות השגיאה מסוג שני, אזי מובטח כי $\alpha' \le \frac{\alpha}{1-\beta}$ ו- $\beta' \le \frac{\beta}{1-\alpha}$.

מבחן זה מכריע מצב עולם 1 אם חוצים את הסף העליון, מצב עולם 0 אם יורדים מתחת לסף התחתון, ובמקרה שערכו של יחס הנראות הוא בין שני הספים, יש לחכות לתצפיות נוספות. בפועל, פרט לאי דיוק הנובע מכך דגימות הן אלמנטים בדידים, מתקיים $\alpha' \le \alpha$ ו- $\beta' \le \beta$.

הוכחה:
יהיו A ו- B שני ספים (מאוחר יותר נגדיר את הערכים שלהם במפורש, ולעת עתה יהיו מספרים כלשהם), ובאמצעותם נגדיר את קבוצת סדרות התצפיות באורך n שעבורן אנחנו מכריעים ω_1 בדיוק כשהגענו לתצפית ה- n- ית

$$C_n = \left\{ x^{(n)} \text{ such that decide } \omega_1 \text{ exactly after } n \text{ observations} \right\}$$

$$= \left\{ x^{(n)} \text{ such that } B \le \frac{L\left(x^{(l)} \mid \omega_0\right)}{L\left(x^{(l)} \mid \omega_1\right)} \le A \text{ for } l = 1..n-1 \text{ and } \frac{L\left(x^{(n)} \mid \omega_0\right)}{L\left(x^{(n)} \mid \omega_1\right)} < B \right\} \tag{2.10}$$

ובאופן דומה את קבוצת הסדרות באורך n עבורן נכריע ω_0 בתצפית ה- n-ית

$$D_n \equiv \left\{ x^{(n)} \text{ such that decide } \omega_0 \text{ exactly after } n \text{ observations} \right\}$$

$$= \left\{ x^{(n)} \text{ such that } B \le \frac{L\left(x^{(l)} \mid \omega_0\right)}{L\left(x^{(l)} \mid \omega_1\right)} \le A \text{ for } l = 1..n-1 \text{ and } A < \frac{L\left(x^{(n)} \mid \omega_0\right)}{L\left(x^{(n)} \mid \omega_1\right)} \right\} \tag{2.11}$$

המאורע בו נכריע ω_1 הוא איחוד המאורעות $\{C_n\}$, שהם מאורעות זרים וממצים. לכן, הסיכוי הכולל שנכריע ω_1 הוא פשוט סכום הסיכויים שנכריע ω_1 בכל צעד n, כלומר

$$P\left(\text{decide } \omega_1\right) = \sum_{n=1}^{\infty} P\left(C_n\right)$$

והסתברות השגיאה מסוג ראשון α נתונה על ידי:

$$\alpha = \sum_n P\left(C_n \mid \omega_0\right) = \sum_n \int_{Cn} \Pr\left(X^{(n)} \mid \omega_0\right) \tag{2.12}$$

וכן

$$\begin{aligned} 1-\beta &= 1-\left(\text{probability to decide } \omega_0 \text{ while } \omega_1\right) = \\ &= \left(\text{probability to decide } \omega_1 \text{ while } \omega_1\right) = \\ &= \sum_n \int_{Cn} \Pr\left(X^{(n)} \mid \omega_1\right) \end{aligned}$$

ובאופן דומה

$$\begin{aligned} 1-\alpha &= \sum_n \int_{Dn} \Pr\left(X^{(n)} \mid \omega_0\right) \\ \beta &= \sum_n \int_{Dn} \Pr\left(X^{(n)} \mid \omega_1\right) \end{aligned} \tag{2.13}$$

.

כעת לכל סדרה $\left(x_1,\ldots,x_n\right) \in C_n$ (הכרענו ω_1 אחרי n תצפיות בדיוק) מתקיים $\Pr\left(x^{(n)} \mid \omega_0\right) \le B \cdot \Pr\left(X^{(n)} \mid \omega_1\right)$, ולכן

$$\begin{aligned} \alpha &= \sum_n \int_{Cn} \Pr\left(x^{(n)} \mid \omega_0\right) \le \sum_n \int_{Cn} B \cdot \Pr\left(X^{(n)} \mid \omega_1\right) \\ &= B\left(1-\beta\right) \end{aligned}$$

ולכל סדרה ב- D_n מתקיים

$$\begin{aligned} 1-\alpha &= \sum_n \int_{Dn} \Pr\left(X^{(n)} \mid \omega_0\right) \ge \sum_n \int_{Dn} A \cdot \Pr\left(X^{(n)} \mid \omega_1\right) \\ &= A \cdot \beta \end{aligned}$$

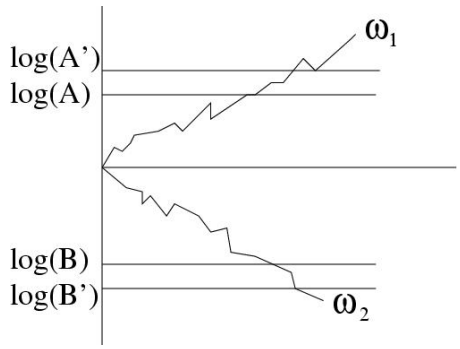
כלומר מצאנו חסמים על הסתברויות השגיאה במונחים של ערכי הסף A ו-B.

המסקנה מניתוח זה היא שבהנתן ערכי שגיאות α_0 ו- β_0 רצויים נוכל לקבוע ספים A' ו- B' , התלויים רק ב- α_0 ו- β_0

$$\begin{aligned} B' &\equiv \frac{\alpha_0}{1-\beta_0} \\ A' &\equiv \frac{1-\alpha_0}{\beta_0} \end{aligned} \tag{2.14}$$

וספים אלו מקיימים $B' \le B$ ו- $A' \ge A$. ולכן מבטיחים כי אם נעבור אותם – נעבור גם את A ו-B והסתברויות השגיאה בפועל של המבחן יקיימו $\alpha' \le \alpha_0 / (1-\beta_0)$ ו-

$\beta' \le \beta_0/(1-\alpha_0)$. האיור המצורף להלן מבהיר את היחס בין ספים אלו. כאמור ניתן גם להראות כי בפועל, פרט לאי דיוק הנובע מכך שהדגימות הן אלמנטים בדידים, מתקיים $\alpha' \le \alpha$ ו- $\beta' \le \beta$. כפי שכבר ציינו, בדרך כלל נשתמש במבחן לוג יחס הנראות במקום ביחס הנראות, ולכן גם בלוג של הספים.



דוגמא

בתכנון ערכה לזיהוי תאים סרטניים נרצה הסתברות גילוי של 99.99% ($\alpha = 10^{-4}$) והסתברות התראות שווא של 0.1% ($\beta = 10^{-3}$), ונקבל את הספים הבאים המבטיחים שלא נחרוג מהסתברויות השגיאה הנדרשות

$$A = \frac{1-\alpha}{\beta} = \frac{0.9999}{10^{-3}} \cong 1000 \Rightarrow \log_{10}(A) = 3$$

$$B = \frac{\alpha}{1-\beta} = \frac{10^{-4}}{0.999} \cong 10^{-4} \Rightarrow \log_{10}(B) = -4$$

כמות התצפיות הדרושות לקבלת הכרעה

המבחן הסדרתי מאפשר לנו להגיע להכרעה עם מספר תצפיות שמשתנה באופן גמיש: אם בשל מזל טוב במיוחד התצפיות הראשונות שקיבלנו הן כאלו שעבורן קל להכריע, הרי שנסתפק בהן. אם לעומת זאת נקבל תצפיות שאינו מאפשרות הכרעה, הרי שנצטרך להשתמש ביותר תצפיות. מסתבר, שבאופן ממוצע פרוצדורת המבחן הסדרתי דורשת שימוש בפחות תצפיות מאשר כמות התצפיות הדרושה במבחן יחס נראות שאיננו סדרתי. לכן פרוצדורה כזו היא שימושית במיוחד במקרה שיש עלות גבוהה לייצר דגימה (למשל כאשר כל אחת מהתצפיות דורשת לבצע ניסוי ארוך/יקר/מסוכן על נבדקים מתנדבים). למעשה, התיאוריה אותה אנו מתארים פותחה לראשונה על ידי Wald לצורך בדיקת איכות של סדרות פגזים במלחמת העולם השניה: בהינתן סדרת ייצור של פגזים, היה צורך לבצע ניסויי ירי ולהכריע האם הסדרה תקינה או פגומה. השאיפה להכריע לגבי תקינות הסדרה על ידי שימוש בכמה שפחות פגזים, הביאה את הצי האמריקאי לפנות לסטטיסטיקאים שיפתחו פרוצדורות יעילות לבחינת הפגזים.

כדי להעריך כמה תצפיות בממוצע דרושות על מנת לקבל הכרעה, נתבונן כיצד מתנהג $\log\frac{L\left(X^{(n)},\omega_0\right)}{L\left(X^{(n)},\omega_1\right)}$ כפונקציה של n. כאשר הדגימות הן בלתי תלויות בהנתן מצב העולם, אז $P\left(X^{(n)}\mid\omega\right)=\prod_{i=1}^{n}P\left(X_i\mid\omega\right)$, ונרשום

$$
\begin{aligned}
y &\equiv \log\left(\frac{P_0\left(\omega_0\right)}{P_0\left(\omega_1\right)}\prod_{i=1}^{n}\frac{P\left(X_i\mid\omega_0\right)}{P\left(X_i\mid\omega_1\right)}\right)= \\
&= \log\left(P_0\left(\omega_0\right)\right)+\sum_{i=1}^{n}\log\left(P\left(x_i\mid\omega_0\right)\right) \\
&\qquad -\log\left(P_0\left(\omega_1\right)\right)-\sum_{i=1}^{n}\log\left(P\left(x_i\mid\omega_1\right)\right)= \\
&= \log\left(\frac{P_0\left(\omega_0\right)}{P_0\left(\omega_1\right)}\right)+\sum_{i=1}^{n}\log\left(\frac{P\left(x_i\mid\omega_0\right)}{P\left(x_i\mid\omega_1\right)}\right)= \\
&= \log\left(\frac{P_0\left(\omega_0\right)}{P_0\left(\omega_1\right)}\right)+n\left(\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{P\left(x_i\mid\omega_0\right)}{P\left(x_i\mid\omega_1\right)}\right)\right)\equiv \\
&\equiv a\cdot n+b
\end{aligned}
$$

כלומר קבלנו משוואה לינארית מהצורה y = a·n + b כשהשיפוע

$$
a=\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{P\left(x_i\mid\omega_0\right)}{P\left(x_i\mid\omega_1\right)}\right) \tag{2.15}
$$

הוא הממוצע האמפירי של לוג יחס הנראות. על-פי החוק החלש של המספרים הגדולים, ממוצע של n משתנים מקריים המתפלגים i.i.d. שואף לתוחלת $\frac{1}{n}\sum_{i=1}^{n}x_i\to\sum_{\{x\}}x\cdot p(x)$. ההסתברות לפיה נחשב את התוחלת תלויה במצב העולם האמיתי, ולכן נקבל במצב ω_0

$$
\begin{aligned}
&\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{P\left(x_i\mid\omega_0\right)}{P\left(x_i\mid\omega_1\right)}\right) \\
&\to\sum_{x'}P\left(x'\mid\omega_0\right)\log\left(\frac{P\left(x'\mid\omega_0\right)}{P\left(x'\mid\omega_1\right)}\right)
\end{aligned} \tag{2.16}
$$

ובמצב ω_1, שוב על פי חוק המספרים הגדולים

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{P(x_i\,|\,\omega_1)}{P(x_i\,|\,\omega_{12})}\right)$$

$$(2.17)\quad \to\sum_{x'}P(x'\,|\,\omega_1)\log\left(\frac{P(x'\,|\,\omega_0)}{P(x'\,|\,\omega_1)}\right)=$$

$$=-\sum_{x'}P(x'\,|\,\omega_1)\log\left(\frac{P(x'\,|\,\omega_1)}{P(x'\,|\,\omega_0)}\right)$$

הביטויים שקיבלנו מכילים תלויות במדד חשוב לדמיון בין התפלגויות שאותו נתאר בסעיף הבא.

2.4.4 מדד לדמיון בין התפלגויות – The Kullback Leibler Divergence

הגדרה: המרחק הסטטיסטי

עבור X מ"מ בדיד ו- P, ו- Q שתי התפלגויות, הגודל

$$(2.18)\quad D[p\,\|\,q]=\sum_{x}p(x)\log\left(\frac{p(x)}{q(x)}\right)$$

מהווה מדד למידת הדמיון הסטטיסטי בין ההתפלגויות. לגודל מספר רב של שמות: Kullback Leibler Divergence, Relative Entropy, Cross Entropy , וחשיבותו רבה בתורת האינפורמציה, בלמידה חישובית ובפיסיקה סטטיסטית. למרבה הבלבול, תחומי מדע שונים נוהגים לבחור בסיס שונה לפונקצית הלוג במשוואה: בפיסיקה נהוג השימוש בלוגריתם הטבעי ובמדעי המחשב בלוג בבסיס 2. אנחנו נשתמש בבסיסים שונים לפי הצורך, ונשים לב כי שינוי בסיס הלוגריתם מתבטא בהכפלת המרחק בקבוע.
מדד זה אינו עונה לקריטריונים של מרחק היות והוא אינו סימטרי ואינו מקיים את אי שוויון המשולש. קל להבין את הסיבה לחוסר הסימטריה אם נזכר כי הראנו ש- D מודד עד כמה קל להבחין בין שתי התפלגויות הנובעות משני מצבי עולם. היות והתצפיות שאנו רואים בפועל תלויות במצב העולם, אז יתכן שאחד ממצבי העולם יספק תצפיות שיקלו על ההכרעה.
למרות שאינו עונה על הקריטריונים של מרחק, המדד D מקיים תכונות חשובות ההופכות אותו לשימושי להשוואת התפלגויות. נראה כעת שלוש תכונות כאלו: נראה כי הוא מדד חיובי, וכן את הקשר שלו לשני מדדי מרחק אחרים.

טענה: D[p||q] הוא אי שלילי, ומקבל ערך אפס אם ורק אם p=q כמעט בכל מקום

הוכחה

נסמן ב- A את קבוצת המאורעות שעבורם p(x)>0 , $A=\{x:p(x)>0\}$. נשתמש באי השוויון $\log(x)\le x-1$ (עבור הבסיס הטבעי), ונרשום

$$
\begin{aligned}
-D\left[p\,\|\,q\right] &= -\sum_{x\in A} p(x)\log\left(\frac{p(x)}{q(x)}\right) = \\
&= \sum_{x\in A} p(x)\log\left(\frac{q(x)}{p(x)}\right) \le \\
&= \sum_{x\in A} p(x)\left(\frac{q(x)}{p(x)}-1\right) = \\
&= \sum_{x\in A} q(x) - \sum_{x\in A} p(x) \le \\
&= \sum_{x\in \Omega} q(x) - \sum_{x\in A} p(x) = 1-1=0
\end{aligned}
\tag{2.19}
$$

נשים לב כי על מנת שיתקיים שוויון , דרוש כי לכל x ב- A מתקיים $\log\left(q(x)/p(x)\right)=1-\left(q(x)/p(x)\right)$, ושוויון זה מתקיים אם ורק אם *p(x)=q(x)* . קיבלנו כי $D\left[p\,\|\,q\right]=0$ אם ורק אם *p(x)=q(x)* לכל x שעבורו *p(x)>0*.

טענה: D(p||q) מקיים

$$
D\left[p\,\|\,q\right] \ge \frac{1}{2\ln 2}\left(\sum_x \left|p(x)-q(x)\right|\right)^2 \tag{2.20}
$$

כאשר D מחושב עם לוג בבסיס 2. הוכחה בתרגיל.

טענה: D[p||q] חסום על ידי

$$
\frac{1}{2}\sum_x \frac{(p(x)-q(x))^2}{\max(p(x),q(x))} \le D[p\,\|\,q] \le \frac{1}{2}\sum_x \frac{(p(x)-q(x))^2}{\min(p(x),q(x))} \tag{2.21}
$$

טענה:

כאשר $p \approx q$, ניתן לקרב את D[p||q] על ידי

$$D[p \| q] \approx \frac{1}{2\ln 2}\sum_{i=1}^{n}\frac{(p_i - q_i)^2}{p_i} = \frac{1}{2\ln 2}\chi_{p,q}^{2} \qquad (2.22)$$

ומכאן שניתן לקרב את D על ידי מדד χ^2 , שהוא מדד נפוץ בסטטיסטיקה קלאסית להשוואה בין התפלגויות. ההוכחה בתרגיל.

טענה: D(p||q) מקיים את כלל השרשרת הבא:

$$\begin{aligned} &D[p(x,y) \| q(x,y)] \\ &= D[p(x) \| q(x)] + D[p(y|x) \| q(y|x)] \end{aligned} \qquad (2.23)$$

הוכחה

$$\begin{aligned} D[p(x,y) \| q(x,y)] &= \sum_x \sum_y p(x,y)\log\left(\frac{p(x,y)}{q(x,y)}\right) = \\ &= \sum_x \sum_y p(x,y)\log\left(\frac{p(x)p(y|x)}{q(x)q(y|x)}\right) = \\ &= \sum_x \sum_y p(x,y)\log\left(\frac{p(x)}{q(x)}\right) + \sum_x \sum_y p(x,y)\log\left(\frac{p(y|x)}{q(y|x)}\right) = \\ &= D[p(x) \| q(x)] + D[p(y|x) \| q(y|x)] \end{aligned}$$

שימוש במרחק סטטיסטי להערכת סבירות של תוצאות

נניח שאנחנו מבצעים n ניסויי ברנולי שלכל אחד הסתברות p להצלחה. מהו הסיכוי לקבל m הצלחות?

מספר ההצלחות מתפלג בינומית

$$P_n(m) = \binom{n}{m} p^m (1-p)^{n-m} = \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m}$$

נשתמש בנוסחת סטירלינג לקירוב העצרת

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n} \Rightarrow \log(n!) \approx n\log(n) - n + \frac{1}{2}\log(2\pi n)$$

ואם נזניח את האיבר השלישי, נוכל לרשום

$$\begin{aligned}\log\binom{n}{m} &= \log\left(\frac{n!}{m!(n-m)!}\right)\\ &\approx \left[n\log n - n\right] - \left[m\log m - m\right] - \left[(n-m)\log(n-m) - (n-m)\right]\\ &= n\left(\log(n) - \frac{m}{n}\log(m) - \frac{n-m}{n}\log(n-m)\right)\\ &= n\left(-\frac{m}{n}\log\left(\frac{m}{n}\right) - \frac{n-m}{n}\log\left(\frac{n-m}{n}\right)\right)\end{aligned}$$

ואם נסמן ב- $q_{m/n}$ את התפלגות ברנולי עם סיכוי $\frac{m}{n}$ להצלחה, אז קיבלנו

$$\begin{aligned}\log\left(P_n(m)\right) &\approx n\left[-\frac{m}{n}\log\left(\frac{m}{n}\right) - \frac{n-m}{n}\log\left(\frac{n-m}{n}\right) + \right.\\ &\qquad \left. + \frac{m}{n}\log(p) + \frac{(n-m)}{n}\log(1-p)\right]\\ &= -nD\left[q_{m/n} \,\|\, p\right]\end{aligned}$$

או

$$P_n(m) \approx \exp\left(-nD[q_{m/n} \,\|\, p]\right)$$

(הערה: אם נחשב את D לפי בסיס שתיים אז נקבל "שתיים בחזקת..." במקום האקספוננט).

הקירוב שקיבלנו יכול לשמש אותנו לא רק להערכת ההסתברות לקבל תוצאה מסוימת (m הצלחות) אלא אף לצורך הערכת הסתברות הזנב כולו (m הצלחות או יותר), בדומה לחסם צ'רנוף. לא נוכיח טענה זו כאן בפירוט, אך ניתן סקיצה של ההוכחה. על מנת לחסום את הסתברות הזנב (דהיינו סכום של (n-m) איברים אקספוננציאלים), נשים לב כי האיבר הגדול ביותר בסכום הוא האיבר הראשון $P_n(m)$, וישנם n-m איברים בסכום. לכן הסכום כולו קטן מביטוי מהצורה

$$n\exp\left(-nD\left[q_{m/n} \,\|\, p\right]\right) = \exp\left(-nD\left[q_{m/n} \,\|\, p\right] + \ln(n)\right) \ .$$

וקיבלנו חסם שעבור n גדול יורד אקספוננציאלית עם גודל המדגם n בדומה לחסם צ'רנוף.

לצורך ההמחשה, נציג דוגמא מספרית. נחסום את ההסתברות לקבל 70 פעמים "עץ" מתוך 100 הטלות של מטבע מאוזנת. נציב p=0.5 ונקבל

$$D\left[0.7,0.3\,\|\,0.5,0.5\right]=0.7\log\left(\frac{0.7}{0.5}\right)+0.3\log\left(\frac{0.3}{0.5}\right)=0.083\ .$$

מכאן שההסתברות לקבל 70 פעמים "עץ" מתוך 100 הטלות כאשר המטבע מאוזנת, חסומה על ידי

$$P_{100}\left(70\,|\,.5,.5\right)\le\exp\left[-100*.083\right]=\exp(-8.3)=0.00025$$

דוגמא: מרחק בין התפלגויות נורמליות

נניח שיש לנו שני מצבי עולם, אחד בו התצפיות מגיעות מהתפלגות נורמלית המאופיינת ע"י תוחלת μ_1 וסטית תקן σ_1, והשני בו התצפיות מפולגות נורמלית עם תוחלת μ_2 וסטית תקן σ_2

$$f_1\left(x\right)=\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^{\ 2}}\right)$$
$$f_2\left(x\right)=\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^{\ 2}}\right),$$

נחשב את המרחק הסטטיסטי בין ההתפלגויות. באפן טבעי, מרחק הסטטיסטי עבור משתנים רציפים מוגדר כאינטגרל על פונקצית צפיפות ההתפלגות במקום סכום על פונקצית ההסתברות. נשתמש בביטויים עבור תוחלת ושונות של משתנים נורמליים: $E\left[x\right]=\mu_1$, $E\left[(x-\mu_1)^2\right]=\sigma_1^{\ 2}$, ונרשום

$$\begin{aligned}
&D[P_1\,\|\,P_2]\\
&=\int_{-\infty}^{\infty}P_1(x)\log\left(\frac{P_1(x)}{P_2(x)}\right)dx=\\
&=\int_{-\infty}^{\infty}\left(\log\left(\frac{\sigma_2}{\sigma_1}\right)-\frac{(x-\mu_1)^2}{2\sigma_1^2}+\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)\frac{\exp^{\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)}}{\sqrt{2\pi}\sigma_1}dx=\\
&*=\log\left(\frac{\sigma_2}{\sigma_1}\right)-\frac{\sigma_1^2}{2\sigma_1^2}+\int_{-\infty}^{\infty}\frac{\exp^{\left(\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right)}}{\sqrt{2\pi}\sigma_1}\cdot\frac{(x-\mu_2+\mu_1-\mu_1)^2}{2\sigma_2^2}dx\\
&=\log\left(\frac{\sigma_2}{\sigma_1}\right)-\frac{1}{2}+\frac{\mathrm{E}\left[(x-\mu_1)^2+2(x-\mu_1)(\mu_1-\mu_2)+(\mu_2-\mu_1)^2\right]}{2\sigma_2^2}\\
&*=\log\left(\frac{\sigma_2}{\sigma_1}\right)-\frac{1}{2}+\frac{1}{2}\frac{\sigma_1^2}{\sigma_2^2}+0+\frac{(\mu_2-\mu_1)^2}{2\sigma_2^2}
\end{aligned}$$

בשוויונים המסומנים בכוכבית השתמשנו בעובדה שתחת ההתפלגות P_1, ל-$(x-\mu_1)$ יש שונות σ_1^2 ולכן

$$(2\pi\sigma_1^2)^{-1/2}\int (x-\mu_1)^2 e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} = \sigma_1^2 \ .$$

במקרה הכללי, המרחק שקיבלנו איננו סימטרי כמובן, אך במקרה המיוחד בו השונויות זהות $\sigma_1 = \sigma_2$ נקבל כי האיברים הראשונים מתבטלים ונשאר עם

$$D[P_1 \| P_2] = \frac{(\mu_2-\mu_1)^2}{2\sigma^2} \qquad (2.24)$$

כלומר D מבטא במקרה זה את ריבוע המרחק בין התוחלות ביחידות של סטית תקן ("סיגמאות"). מרחק זה נקרא גם "מרחק מהאלאנוביס" (Mahalanobis), והשורש הריבועי שלו ידוע גם בתור "יחס אות לרעש" (Signal-to-Noise Ratio), והוא מדד נפוץ למדידת יכולת ההבחנה בין הערכים אפשריים של משתנה מקרי רציף המקבל שני ערכים שעליהם נוסף רעש.

תרגילים

1. נניח כי אנו מחליפים את פונקצית ההכרעה הדטרמניסטית, $\alpha(x)$ בכלל הכרעה אקראי: בהינתן התצפית x אנו מבצעים את הפעולה α_i בהסתברות $P(\alpha_i \mid x_i)$

א. הראו כי הסיכון הכולל נתון כעת על-ידי $R=\int\left[\sum_i R(\alpha_i \mid x)\cdot P(\alpha_i \mid x)\right]P(x)dx$ (במקרה בדיד מופיע סכום במקום האינטגרל).

ב. הראו כי R הינו מינימאלי אם אנו בוחרים $P(\alpha_i \mid x_i)=1$ עבור הפעולה α_i המביאה למינימום את הסיכון המותנה, $R(\alpha_i \mid x)$, ולכן הכרעה דטרמיניסטית היא אופטימלית.

2. נניח שמציעים לכם להשתתף במשחק הבא: מטילים זוג קוביות הוגנות עד שיוצא "1" לפחות באחת מהקוביות. לפני כל הטלה אתם יכולים להחליט אם אתם ממשיכים להשתתף במשחק או יוצאים ממנו. אם אתם ממשיכים להשתתף במשחק אתם זוכים בשקלים עפ"י תוצאת ההטלה (סכום התוצאות בשתי הקוביות) למעט המקרה שבו יוצא "1" לפחות באחת מהקוביות שבו אתם מפסידים את כל מה שהרווחתם. אם הספקתם לצאת מהמשחק לפני שיצאה התוצאה "1" אתם נשארים עם מה שהרווחתם עד כה.

א. נסחו את הבעיה כבעיית הכרעה בייסיאנית.

ב. מהי האסטרטגיה הבייסיאנית האופטימאלית לקבלת החלטה אם להמשיך לשחק או לצאת מהמשחק?

ג. מהו הסכום המרבי שתהיו מוכנים לשלם כדי להשתתף במשחק? נמקו.

3. יהיו s_1 ו- s_2 שני "מקורות" פואסוניים, עם λ_1 ו- λ_2 בהתאמה.

א. בהנתן סדרת דגימות מאחד המקורות, כמה דגימות נחוצות על מנת להכריע מהו מקור הסדרה בוודאות של 99 אחוזים (לכל כיוון).

ב. תאר גישה בייסיאנית לטיפול במקרה בו נוסף מקור שלישי עם λ_3. מה יהיה כלל ההכרעה במקרה זה?

4. גבול הכרעה בין התפלגויות נורמליות.

א. נתונה בעיית ההכרעה הבאה: X מתפלג נורמלית (חד ממדית) עם $P(x \mid w_1)=N(\mu_1,\sigma_1^2)$, $P(x \mid w_2)=N(\mu_2,\sigma_2^2)$. מהו גבול ההכרעה בהנחה כי ההסתברויות האפריוריות לשני מצבי העולם שוות $(P(w_1)=P(w_2))$?

ב. מהו גבול ההכרעה אם $\underline{X}$ מתפלג דו-נורמלית עם $P(\underline{X} \mid w_1)=N_{\underline{X}}(\underline{\mu}_1,\Sigma_1^2)$ $P(\underline{X} \mid w_2)=N_{\underline{X}}(\underline{\mu}_2,\Sigma_2^2)$.

ג. מצא את הגבול במקרה הפרטי בו מטריצות הקוואריאנס הן אלכסוניות ושוות וכן ההסתברויות האפריוריות שוות ומטריצת המחירים מקיימת $\lambda_{11} = \lambda_{22} = 0$ ו- $\lambda_{12} = \lambda_{21}$.

5. יהיו $x_1, x_2, ..., x_n$ משתנים-מקריים המתפלגים באופן אחיד בקטע $[0,1]$ נגדיר $V_n = \prod_{i=1}^{n} x_i$ (כלומר נפח התיבה ה- n ממדית ש- $x_1, ..., x_n$ הן צלעותיה).

א. מהו $\lim_{n\to\infty} V_n^{\frac{1}{n}}$?

ב. השוו גודל זה לשורש ה- n-י של הנפח "הנאיבי", המתקבל ממכפלת האורכים הממוצעים של הצלעות, (כלומר $\left(\left(\frac{1}{2}\right)^n\right)^{1/n} = \frac{1}{2}$).

6. הוכיחו כי ה"מרחק" D בין שתי התפלגויות ברנולי עם סיכויי הצלחה p ו- q מקיים

$$D[p \| q] \geq \frac{2}{\ln 2}(p-q)^2$$

הדרכה: הגדירו פונקציה $g(p,q)$ שהיא ההפרש בין שני האגפים

$$g(p,q) = D[p \| q] - \frac{2}{\ln 2}(p-q)^2$$

הראו כי הנגזרת של פונקציה זאת קטנה או שווה לאפס כאשר $q \leq p$ והסיקו מכך כי $g(p,q) \geq 0$ עבור $q \leq p$.

7. הוכח כי המרחק D חסום על ידי

$$\frac{1}{2}\sum_x \frac{(p(x)-q(x))^2}{\max(p(x),q(x))} \leq D[p \| q] \leq \frac{1}{2}\sum_x \frac{(p(x)-q(x))^2}{\min(p(x),q(x))}$$

8. הוכח כי כאשר $p \approx q$, ניתן לקרב את D[p||q] על ידי

$$D[p \| q] \approx \frac{1}{2\ln 2}\sum_{i=1}^{n}\frac{(p_i - q_i)^2}{p_i} = \frac{1}{2\ln 2}\chi_{p,q}^{\ 2}$$

9. חשב את המרחק הסטטיסטי בין שתי התפלגויות פואסוניות.

10. חשב את המרחק הסטטיסטי בין שתי התפלגויות אקספוננציאליות.

תרגיל מחשב

כתבו תכנית להכרעה סדרתית בין טקסט הכתוב באנגלית לטקסט כתוב בצרפתית, על סמך פילוגי האותיות הבודדות בשתי השפות (כולל רווח). הקלט לתכנית יהיה הפילוגים, טקסט ארוך והסתברויות השגיאה מסוג ראשון ושני (α,β).

א. צייר גרף של הציון המצטבר (לוג הנראות) כפונקציה של אורך הטקסט. סמן את החסמים (A', B').

ב. מהו אורך הטקסט הנדרש להכרעה ומהן תוצאות המבחן.

ג. צור גרפים של אורך הטקסט כפונקציה של α עבור β קבוע ולהיפך.

ד. השוו את התוצאה המתקבלת לאורך הצפוי על-פי המרחק הסטטיסטי בין הפילוגים.

ה. מצאו טקסט בשפה (לטינית) שלישית, וחזרו על החישובים מהסעיפים הקודמים עבור השפה החדשה עם אחת משתי השפות הקודמות.