

## פרק 2: יסודות סטטיסטיים – מבוא לשיערוך וסיווג

בפרק זה נתבונן בצמד משתנים אקראיים,  $(X, Y)$ . על סמך מדידת ערכו של  $X$ , עלינו להעריך (לאמוד) את ערכו של  $Y$ .

לשם כך נדרש כמובן מודל המתאר את הקשר בין המשתנים. אנו מבחינים בין שני סוגי מודלים סטטיסטיים:

1. מודל בייסיאני: נתון פילוג המשותף  $p_{X,Y}(x, y)$ .

באופן שקול: נתונים הפילוגים  $p_Y(y)$ ,  $p_{X|Y}(x|y)$ .

2. מודל לא-בייסיאני: נתונים הפילוגים המותנים  $p_{X|Y}(x|y)$  בלבד.

למעשה, במקרה זה  $Y$  אינו חייב להיות משתנה מקרי, וניתן להתייחס אליו כפרמטר לא ידוע.

### הערות לסימון:

- הסימון  $p_X(x)$  מתייחס לצפיפות ההסתברות במקרה של משתנה מקרי רציף, ולערך ההסתברות במקרה של משתנה בדיד.
- למשתנה מקרי (מ"מ)  $X$ , ערך אפשרי יסומן  $x$ , ואוסף הערכים האפשריים (הטווח) יסומן  $\underline{X}$ . לפיכך,  $x, X \in \underline{X}$ .
- הסימון  $p_X(x)$  (ללא ציון ערכו של  $x$ ) משמש תחליף נוח לסימון המלא של פונקציית הצפיפות:  $p_X(\cdot) = \{p_X(x) : x \in \underline{X}\}$ .

**2.1 אמידת ערכו של משתנה מקרי בעל פילוג נתון**

כהקדמה, נתבונן בבעיה הבאה: נתון משתנה מקרי (מ"מ) ממשי  $Y$  בעל פילוג  $p_Y(y)$ . אנו

מתבקשים לבחור ערך  $\hat{y}$  שמתאר בצורה מיטבית את ערכו של  $Y$ . במה נבחר?

אפשרויות נפוצות:

א. תוחלת:  $\hat{y} = E(Y) = \int y p_Y(y) dy$

ב. הערך הסביר ביותר:  $\hat{y} = \arg \max_y p_Y(y)$

ג. חציון (מדיאן):  $\hat{y} = \text{median}(Y) : \text{prob}(Y \leq \hat{y}) = 0.5$

ניתן לראות כי אמד התוחלת ממזער את קריטריון השגיאה הריבועית:

$$\min_{\hat{y}} E(Y - \hat{y})^2$$

בעוד אמד החציון ממזער את קריטריון השגיאה המוחלטת:

$$\min_{\hat{y}} E |Y - \hat{y}|$$

**מדדי אי-וודאות:** בנוסף לערך שבחרנו, אנו נדרשים לרוב להעריך גם את מידת הדיוק (או חוסר

הדיוק) בהערכה זו. המדדים המקובלים הם:

א. הסטייה הריבועית סביב האמד הנבחר:  $E(Y - \hat{y})^2$

ב. השונות (סטיה ריבועית מסביב לתוחלת):  $\text{Var}(Y) = E(Y - E(Y))^2$

ג. סטיית התקן:  $\sigma_Y = \sqrt{\text{var}(Y)}$

ד. רווח סמך (Confidence Interval): מקטע  $[a, b]$  כך שמתקיים

$$\text{prob}(Y \in [a, b]) \geq 1 - \alpha$$

כאשר  $\alpha$  ערך קטן (מקובל:  $\alpha = 0.05$ ).

רווח סמך עבור פילוג גאוס ו- $\alpha = 0.05$  הינו

$$E(Y) \pm 3\sigma_Y \equiv [E(Y) - 3\sigma_Y, E(Y) + 3\sigma_Y]$$

## 2.2 שערך בייסיאני

נחזור לבעיית השערך (אמידה) שבה פתחנו :

- נתון הפילוג המשותף  $p_{X,Y}(x, y)$  של המשתנים  $(X, Y)$ .
- נמדד הערך  $X = x$ .
- יש להעריך את ערכו של  $Y$ .

כפי שצוין לעיל, בפועל הפילוג המשותף מוגדר לרוב באופן מופרד על ידי שני הגדלים הבאים :

- הפילוג הראשוני (אפריורי, prior distribution),  $p_Y(y)$ .
- פונקציית הסבירות,  $p_{X|Y}(x | y)$ .

מנתונים אלה, ניתן לחשב (לפחות להלכה) את הפילוג המותנה של  $Y$  בהינתן המדידה  $X = x$  :

$$p_{Y|X}(y | x) = \frac{p_{X,Y}(x, y)}{p_X(x)} = \frac{p_{X,Y}(x, y)}{\int p_{X,Y}(x, y) dy}$$

או (נוסחת בייס, Bayes) :

$$p_{Y|X}(y | x) = \frac{p_{X|Y}(x | y) p_Y(y)}{\int p_{X|Y}(x | y) p_Y(y) dy}$$

הערה : נשים לב שאנו מחשבים פה פילוג על  $y$ , עבור ערך נתון של  $x$ . כדי להדגיש זאת נסמן

לעיתים  $p_{Y|X}(\cdot | x)$ .

פילוג זה נקרא הפילוג בדיעבד (posterior distribution) של  $Y$ .

לאחר שחישבנו את הפילוג בדיעבד של  $Y$ , ניתן למצוא אמד נקודה של  $Y$  מתוך פילוג זה, לפי אחת מהאפשרויות שסקרנו בתת-הפרק הקודם. בפרט :

א. משערך התוחלת המותנית (MMSE) :

$$\hat{y} = E(Y | X = x) = \int y p_{Y|X}(y | x) dy \triangleq \hat{y}_{MMSE}$$

כפי שראינו, משערך זה ממזער את תוחלת השגיאה הריבועית, ומכאן השם : משערך שגיאה ריבועית ממוצעת מינימלית : MMSE=Minimal Mean Square Error.

ב. משערך ההסתברות המירבית בדיעבד (MAP) :

$$\hat{y} = \max_y p_{Y|X}(y | x) \triangleq \hat{y}_{MAP}$$

פה MAP=Maximum A-posteriori Probability

מדדי אי-הוודאות לאומדן יתקבלו פה מתוך הפילוג המותנה  $p_{Y|X}(\cdot|x)$ . הגודל הנפוץ ביותר הוא השונות המותנית:

$$\begin{aligned} \text{Var}(Y | X = x) &= E((Y - \hat{y}_{MMSE})^2 | X = x) \\ &= \int (y - \hat{y}_{MMSE})^2 p_{Y|X}(y|x) dy \end{aligned}$$

גודל זה (ומדדי אי-הוודאות אחרים כגון רווח סמך, ואף המשערך עצמו) קשים לחישוב אנליטי, ופרט למקרים מיוחדים יש לחשבם באופן נומרי. מקרים מיוחדים אלה כוללים את המקרה הגאוסי (מ"מ גאוסיים במשותף), בו יש לנו נוסחאות סגורות לתוחלת המותנית ולשונות המותנית.

### 2.3 שיעור לא-בייסיאני

במקרה זה, נתון אוסף הפילוגים המותנים (הסבירות) בלבד:  $\forall y, p_{X|Y}(\cdot|y)$ .

הפילוג  $p_Y(y)$  אינו נתון, כך שלא ניתן לחשב הסתברויות של  $Y$ .

המשערך המקובל פה הוא משערך הסבירות המירבית –

MLE, Maximum Likelihood Estimator

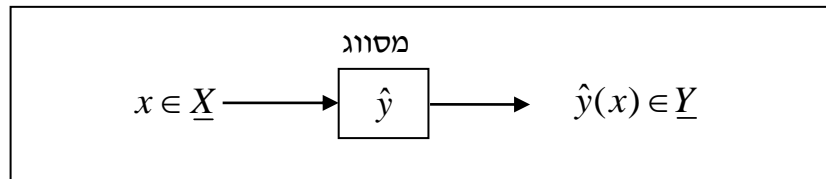
$$\hat{y}_{MLE} = \arg \max_y p_{X|Y}(x|y)$$

על משערך זה נרחיב בפרק הבא, הדן בשיעור מודל.

## 2.4 סיווג בייסיאני

### א. הגדרת הבעיה

בעיית הסיווג הבייסיאני (הקרויה גם בדיקת השערות בייסיאנית) היא בעיית שערך מסוג מסוים. כרגיל, נתונים הפילוגים  $p_Y(y)$ ,  $p_{X|Y}(x|y)$ , ויש להעריך את ערכו של  $Y$  על סמך הערך הנצפה  $x$  של  $X$ . המשערך (אמד) הוא פונקציה  $\hat{y}: X \rightarrow Y$ , בעל ערך  $\hat{y}(x)$ .



בעיית הסיווג מאופיינת על ידי התכונות הבאות:

1. המשתנה  $Y$  הוא משתנה קטגורי, המקבל מספר סופי של ערכים:

$$Y \in \Omega \triangleq \{\omega_1, \omega_2, \dots, \omega_K\}$$

בהתאם לכך, המשערך  $\hat{y}$  נבחר מתוך אותו סט  $\Omega$ .

2. קריטריון הביצועים הוא הסתברות השגיאה: בהינתן הערך  $X = x$ , יש לבחור ערך  $\hat{y}(x) \in \Omega$  אשר ממזער את הסתברות השגיאה:

$$P_e(\hat{y}|x) \triangleq \text{prob}(\hat{y}(x) \neq Y | X = x)$$

המשמעות: מתקבל קלט  $x$  מתוך אחת המחלקות  $\{\omega_1, \omega_2, \dots, \omega_K\}$ . יש לסווג את  $x$  לאחת ממחלקות אלו, במטרה לזהות את המחלקה הנכונה.

דוגמאות:

1. נתון גובה  $x$  של אדם אלמוני. יש להחליט את מדובר באישה ( $\omega_1$ ) או גבר ( $\omega_2$ ).

2. במקלט של מערכת תקשורת, נקלט אות  $\{s(t), 0 \leq t \leq T\}$ . יש להחליט אם שודר "0" ( $\omega_0$ ) או "1" ( $\omega_1$ ).

3. הקלט  $x$  הינו תמונה. יש להחליט אם בתמונה מופיע אדם מסוים.

הערות :

- הערך הנבחר  $\hat{y}(x)$  נקרא גם ההחלטה (המתאימה לקלט  $x$ ).
- בתיאור לעיל הנחנו כי  $\hat{y}(x) \in \Omega$ , כלומר: מרחב ההחלטה הוא  $\Omega$ , אוסף המחלקות. לעיתים נרצה להוסיף אפשרויות נוספות, כגון "לא ניתן להחליט", או "הקלט שייך למחלקה  $\omega_1$  או  $\omega_3$ ". למעשה ניתן להכליל את בעיית הסיווג למרחב החלטות כללי, נזכיר זאת בהמשך.

### ב. המסווג הבייסיאני האופטימאלי

נזכור את הגדרת הסתברות השגיאה:

$$P_e(\hat{y} | x) \triangleq \text{prob}(\hat{y}(x) \neq Y | X = x)$$

זו הסתברות השגיאה המותנית. נגדיר גם את הסתברות השגיאה הממוצעת:

$$P_e(\hat{y}) \triangleq \text{prob}(\hat{y}(X) \neq Y)$$

נגדיר עתה את **מסווג MAP** (מסווג ההסתברות המירבית בדיעבד) לבעיה הנדונה:

$$\hat{y}_{MAP}(x) = \arg \max_{\omega \in \Omega} p_{Y|X}(\omega | x)$$

מסווג זה נקרא גם מסווג בייס (Bayes Classifier).

נבטא עתה מסווג זה באמצעות הגדלים הנתונים:  $p_{X|Y}, p_Y$ . בהצבת נוסחת בייס נקבל:

$$\hat{y}_{MAP}(x) = \arg \max_{\omega \in \Omega} \left\{ \frac{p_{X|Y}(x | \omega) p_Y(\omega)}{p_X(x)} \right\}$$

אולם מכיוון ש-  $p_X(x)$  אינו תלוי ב-  $\omega$ , נקבל כי

$$\hat{y}_{MAP}(x) = \arg \max_{\omega \in \Omega} \left\{ p_{X|Y}(x | \omega) p_Y(\omega) \right\}$$

ביטוי זה חוסך את החישוב (המיותר) של  $p(x)$ .

**משפט** (מסווג בייס)

המסווג  $\hat{y}_{MAP}(x)$  הינו המסווג האופטימאלי, המביא למינימום את הסתברות השגיאה המותנית עבור כל קלט  $x$ , וכן את הסתברות השגיאה הממוצעת.

הוכחה : נתבונן ראשית בהסתברות השגיאה המותנית :

$$P_e(\hat{y} | x) \triangleq \text{prob}(\hat{y}(x) \neq Y | X = x) \\ = 1 - \text{prob}(\hat{y}(x) = Y | X = x)$$

מכאן שמזעור  $P_e(\hat{y} | x)$  שקול לבחירת  $\hat{y}(x)$  שמביא למכסימום את הגורם האחרון. אבל זו בדיוק הגדרת  $\hat{y}_{MAP}(x)$ .

נעבור להסתברות השגיאה הממוצעת :

$$P_e(\hat{y}) = \text{prob}(\hat{y}(X) \neq Y) \\ = \int_x \text{prob}(\hat{y}(x) \neq Y | X = x) p_X(x) dx$$

אולם ראינו כי  $\hat{y}_{MAP}$  מביא למינימום את הסתברות השגיאה לכל  $x$  בנפרד, ומכאן נובע מיידית שהוא מביא למינימום את האינטגרל האחרון.

□

**ג. מסווג בייס במקרה הגאומטרי**

מקרה פרטי חשוב מתקבל כאשר פונקציות הסבירות  $p_{X|Y}(\cdot | \omega_i)$  הן בעלות פילוג גאומטרי. במקרה זה חוק החלטה מבוטא באמצעות ביטוי ריבועי, או אף לינארי, במשתנים  $x$ .

(1) המקרה החד-מימדי :  $x \in \mathbb{R}$ . נניח כי

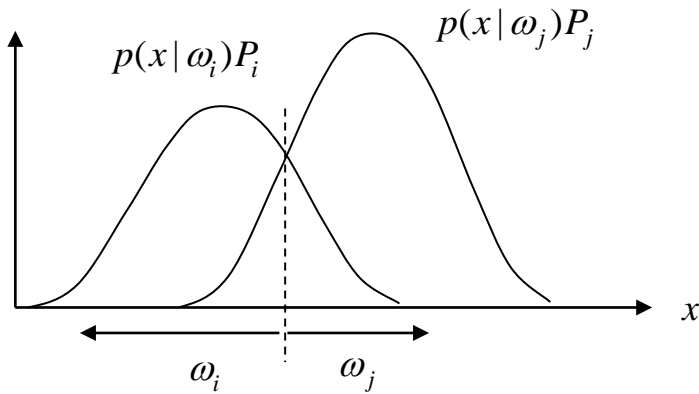
$$p_{X|Y}(x | \omega_i) = \frac{1}{(2\pi\sigma_i^2)^{1/2}} \exp\{-(x - \mu_i)^2 / 2\sigma_i^2\}, \quad \forall \omega_i \in \Omega$$

$$\hat{y}_{MAP}(x) = \arg \max_{\omega \in \Omega} \{p_{X|Y}(x | \omega) p_X(\omega)\} \quad \text{, כזכור}$$

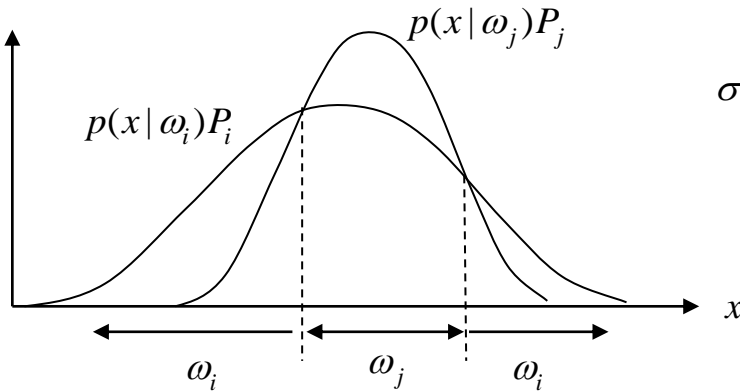
לפיכך, הסיווג  $\omega_i$  עדיף על  $\omega_j$  אם מתקיים

$$p_{X|Y}(x | \omega_i) p_Y(\omega_i) \geq p_{X|Y}(x | \omega_j) p_Y(\omega_j)$$

באופן גרפי, נקבל את התמונה הבאה :



• שונות זהה:  $\sigma_i = \sigma_j$



• שונות לא-שווה:  $\sigma_i \neq \sigma_j$

נשים לב כי שפות התחומים בהם עדיפה מחלקה אחת על השנייה מוגדרות על ידי השוויון  $p(x|\omega_i)p(\omega_i) = p(x|\omega_j)p(\omega_j)$ . שפות אלו נקראות Bayes decision boundary. נקודות השפה ניתנות לחישוב ע"י פתרון משוואה ריבועית (לאחר הוצאת לוגריתם).

(2) המקרה הרב-מימדי:  $x = (x_1, \dots, x_d)^T \in \mathbb{R}^d$ .

נניח כי פונקציות הסבירות נתונות ע"י הפילוגים הגאוסיים

$$p_{X|Y}(x|\omega_i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i)\}, \quad \omega_i \in \Omega$$

גם במקרה זה, ההעדפה בין שתי מחלקות תתבצע בהתאם לאי השוויון  $p(x|\omega_i)p(\omega_i) \gtrless p(x|\omega_j)p(\omega_j)$ . לאחר הוצאת לוגריתם נקבל את אי השוויון

השקול:

$$g_{ij}(x) \triangleq g_i(x) - g_j(x) \gtrless 0$$

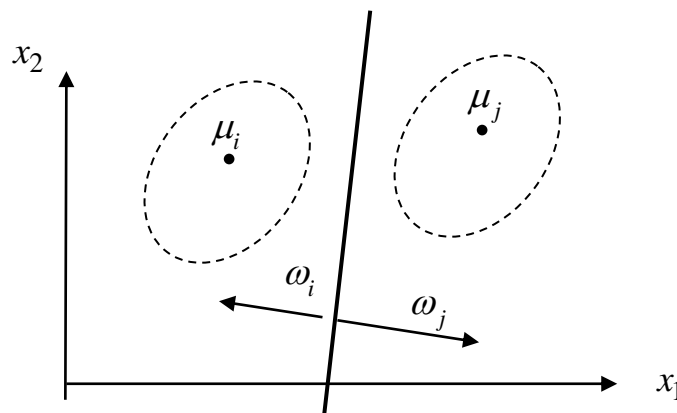
כאשר

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i) + \alpha_i, \quad \forall \omega_i \in \Omega$$

$$\alpha_i = \log\{p(\omega_i) / (2\pi)^{d/2} |\Sigma_i|^{1/2}\}$$



- הפונקציה  $g_{ij}(x)$  קרויה פונקציית האבחנה (discriminant function) בין המחלקות  $\omega_i$ ,  $\omega_j$ . במקרה הגאוסי זו פונקציה ריבועית באיברי וקטור הקלט  $x$ .
- שפות תחומי ההחלטה נקבעים על ידי השוויון  $g_{ij}(x) = 0$ . במקרה הדו-מימדי  $(d = 2, x = (x_1, x_2))$  נקבל אליפסה, היפרבולה, או שתי היפרבולות.
- במקרה המיוחד שבו  $\Sigma_i = \Sigma_j$  (מטריצות קווריאנס זהות) האיברים הריבועיים מתבטלים, ומתקבל משטח ישר (על-מישור) המפריד בין התחומים :



**ד. סיווג בינארי**

נניח כי  $\Omega = \{\omega_1, \omega_2\}$  אזי

$$\hat{y}_{MAP}(x) = \arg \max_{\omega \in \{\omega_1, \omega_2\}} \{p_{X|Y}(x|\omega)p_Y(\omega)\}$$

במקרה זה ניתן לבטא את  $\hat{y}_{MAP}(x)$  כך :

$$p_{X|Y}(x|\omega_1)p_Y(\omega_1) > p_{X|Y}(x|\omega_2)p_Y(\omega_2) \Rightarrow \hat{y}(x) = \omega_1$$

$$p_{X|Y}(x|\omega_1)p_Y(\omega_1) < p_{X|Y}(x|\omega_2)p_Y(\omega_2) \Rightarrow \hat{y}(x) = \omega_2$$

(במקרה של שוויון ניתן לבחור שרירותית). באופן שקול,

$$\frac{p_{X|Y}(x|\omega_1)}{p_{X|Y}(x|\omega_2)} \begin{matrix} > \\ < \end{matrix} \frac{p_Y(\omega_2)}{p_Y(\omega_1)} \Rightarrow \hat{y}(x) = \omega_1 / \omega_2$$

היחס בצד שמאל נקרא יחס הסבירות, שנסמנו  $\Lambda_{1,2}(x)$ . היחס בצד ימין הוא קבוע,  $C$ . חוק

ההחלטה ניתן לפיכך לכתיבה מקוצרת כך :

$$\Lambda_{1,2}(x) \gtrless C \Rightarrow \omega_1 / \omega_2$$

חוק החלטה זה נקרא **מבחן יחס הסבירות** (לסיווג בינארי). בהקשר מעט שונה (בחינת השערות לא בייסיאנית) הוא ידוע גם כ"מבחן ניימן-פירסון".

**סוגי שגיאות** : עד כה התייחסנו להסתברות השגיאה הכוללת :

$$P_e(\hat{y}) \triangleq \text{prob}(\hat{y}(X) \neq Y)$$

ניתן לרשום ביטוי זה כך :

$$P_e(\hat{y}) = P_e(2|1)p_Y(\omega_1) + P_e(1|2)p_Y(\omega_2)$$

כאשר

$$P_{2|1}(\hat{y}) = \text{prob}(\hat{y}(X) = \omega_2 | Y = \omega_1)$$

$$P_{1|2}(\hat{y}) = \text{prob}(\hat{y}(X) = \omega_1 | Y = \omega_2)$$

אלו ההסתברויות המותנות של שגיאה מסוג 2 ושגיאה מסוג 1, בהתאמה.

הערות :

- ניתן לראות כי הסתברויות השגיאה המותנית אינו תלויות בהסתברות הראשוניות של ההשערות  $p_Y(\omega_i)$ , אלא רק בפונקציות הסבירות  $p_{X|Y}$  ובחוק ההחלטה  $\hat{y}$ .
- כאשר ההסתברויות הראשוניות  $p_Y(\omega_i)$  אינו נתונות (או חסרות משמעות), אנו דנים בבעיית סיווג לא-בייסיאנית.
- במקרה זה, ניתן להגדיר חוק החלטה מיטבי ככזה שמביא למינימום הסתברות שגיאה מסוג אחד, כפוף לאילוץ גודל מסוים על השגיאה השניה. בפרט :

$$\min_{\hat{y}} P_{2|1}(\hat{y}), \quad \text{subject to } P_{1|2}(\hat{y}) \leq \alpha$$

כאשר  $\alpha$  קבוע ( $\alpha \ll 1$ ).

- ניתן להראות כי מבחן יחס הסבירות (מבחן ניימן פירסון) הוא המבחן האופטימאלי במובן זה, עבור בחירה מתאימה של הקבוע  $C$  (מהי בחירה זו?).

**ה. הכללה למדד סיכון כללי (\*)**

ביישומים מסוימים, ייתכן כי לשגיאות שונות תהיה משמעות שונה, ולכן מחיר שונה. כמו כן, ייתכן כי מרחב ההחלטות  $\underline{Y}$  כולל אפשרויות שונות פרט למחלקות  $\Omega$ .

במקרה זה נגדיר פונקציית הפסד  $\ell(y, \omega) : Y \times \Omega \rightarrow \mathbb{R}$ , אשר מקיימת את התנאים הבאים:

א.  $\ell(y, \omega) \geq 0$

ב.  $\ell(y, \omega) = 0$  אם  $y = \omega$ .

ניתן עתה להגדיר כמדד הביצועים את הסיכון המותנה:

$$L(\hat{y} | x) = E(\ell(\hat{y}(X), Y) | X = x)$$

ואת הסיכון הממוצע (Expected Risk):

$$L(\hat{y}) = E(\ell(\hat{y}(X), Y))$$

נשים לב כי מדד השגיאה הבסיסי אליו התייחסנו עד כה מתקבל כמקרה פרטי, כאשר

$$\ell(y, \omega) = I\{y = \omega\} \triangleq \begin{cases} 1 & : y \neq \omega \\ 0 & : y = \omega \end{cases}$$

תרגיל: הניחו כי פונקציית הפסד  $\ell(y, \omega)$  הינה מהצורה:

$$\ell(y, \omega) = \ell(\omega) I\{y = \omega\} \triangleq \begin{cases} 0 & : y = \omega \\ \ell(\omega) & : y \neq \omega \end{cases}$$

כאשר  $\ell(\omega) > 0$  לכל  $\omega \in \Omega$ . מצאו את המסווג  $\hat{y}(x)$  אשר ממזער את הסיכון המותנה (לכל  $x$ ) ואת הסיכון הממוצע.

רמז: הראו ראשית כי  $L(\hat{y} | x) = \sum_{\omega} p_{Y|X}(\omega | x) \ell(\omega) I\{\hat{y}(x) \neq \omega\}$