פרק 2: יסודות סטטיסטיים – הכרעה בייסיאנית

מקור: DHS(2001):2.1-2.7.

בהרצאה זו נדון בהכרעה בייסיאנית על גווניה השונים.

2.1. הגדרת בעיית ההכרעה

בבעיית ההכרעה (decision), אנו מתייחסים למדידות (measurement) אשר נוצאו כתוצאה ממספר מצבי עולם (states) שונים. בהינתן מדידה כזו עלינו לפעול בהתאם לה, ללא מידע מפורש על מצב העולם.

פונקציית הערכה ממפה מדידות לפעולות.
המרכיבים הבסיסיים של הבעיה הם:

Ω - אוסף סופי של מצבי עולם: $\Omega=\{1,2,\ldots,C\}$. איברים אופייניים יסומנו ע"י $\omega,\omega_i,\omega_j\in\Omega$.

X - מרחב הקלט או המדידות (input space). אלמנט במרחב זה יסומן ע"י $x\in X$, ויכונה "תבנית קלט", "קלט" או "מדידה". באופן טיפוסי מרחב הקלט הוא רב-מימדי: למשל $x=(x_1,\ldots x_d)\in\mathbb{R}^d$, כאשר $d>1$. כינוי נוסף ל- x הינו "וקטור המאפיינים".

Y - אוסף סופי של החלטות אפשריות. זהו למעשה מרחב הפלט של המסווג.

דוגמא 1: בחורף ישראלי טיפוסי מזג האוויר יכול להיות נעים (ω_1), קר (ω_2) או גשום (וקר) (ω_3). עלינו להחליט האם ללבוש חולצה בלבד (y_1) או מעיל (y_2) על סמך הצבע הדומיננטי השמיים – כחול, לבן או אפור (x), הצבע יכול לייצג רמת וסוג עננות.

דוגמא 2: עלינו להגיע לשיעור חשוב אולם לעיתים הכבישים פקוקים (ω_1) ולעיתים פנויים (ω_2). עלינו להחליט האם לצאת מוקדם (y_1), ולהסתכן שנגיע מוקדם מידי, לצאת בזמן (y_2) ולהסתכן בפקק, או לצאת מאוחר ולוותר מראש על השיעור והפקקים (y_3).

הערה: הרבה פעמים מרחב ההחלטות הינו בדיוק מרחב המצבים $\underline{Y=\Omega}$, כאשר $y=\omega$ פירושו "הקלט מסווג למחלקה ω". אולם ניתן להוסיף על כך אפשרויות שונות כגון: "לא ניתן להחליט", "דחייה (הקלט אינו באחת המחלקות הנתונות)", או "הקלט שייך למחלקה ω_1 או ω_2". במקרה זה נקרא לפונקציית ההכרעה מסווג (classifier), ולבעייה נקרא בעיית הסיווג (classification).

דוגמא 3: נדרש לסווג אנשים בוגרים ל"אישה" (ω_1) או "גבר" (ω_2) לפי גובהם (x) בלבד.

<u>דוגמא 4 :</u> במקלט של מערכת תקשורת, נקלט אות $\{s(t),\, 0 \le t \le T\}$. יש להחליט לפי האות הנקלט אם שודר "0" (ω_0) או "1" (ω_1).

<u>דוגמא 5 :</u> במערכת זיהוי מיקרוביולוגית יש להפריד בין "סטפילוקוקוס", "שמרים", "פטריות", ו- "חידקים לא ידועים". נשים לב שבמקרה זה יש ענין בתשובה כמו "שמרים או פטריות" או בתשובה כגון "לא ידוע".

פונקציית ההכרעה, אם כן, הינה פונקציה $f: X \to Y$ ממרחב הקלט למרחב ההחלטות :

מסווג

$$x \in X \longrightarrow \boxed{f} \longrightarrow y = f(x) \in Y$$

המטרה היא כמובן לבחור פעולה מיטבית ביחס למצב העולם (שאיננו יודעים) מתוך מדידות. בהמרכה של בעיית הסווג, מטרה סבירה היא להקטין ככל האפשר את מספר טעויות הסיווג. בהמשך נציג מטרה זו באופן מתמטי.

<u>הערה</u> : בעיית הסיווג (classification) ידועה גם בשם זיהוי תבניות (pattern recognition), במיוחד בהקשר של ראיה ממוחשבת. הגישה הבייסיאנית מציעה פתרון אנליטי לבעיית הסיווג, אשר מסתמך על ניסוח הסתברותי (סטטיסטי) של הבעיה. בתחום הסטטיסטיקה בעייה זו ידועה גם כ"בחינת השערות" (Hypothesis Testing).

<u>2.2. סיווג בייסיאני: המבנה ההסתברותי</u>

בניסוח הבייסאני של בעיית הסיווג, אנו מניחים כי לבעייה מבנה הסתברותי הידוע לנו. במילים אחרות, אנו מניחים קיום מידע סטטיסטי מלא (מידע א-פריורי) לגבי פילוג ההסתברות של המחלקה ω והקלט x. מידע זה מאפשר לנו לסווג באופן "אופטימאלי" כל קלט נצפה.

אנו מניחים אם כן כי נתונים פילוגי ההסתברות הבאים :

1. פילוג הסתברות $P(\omega)$ על המרחב Ω של המחלקות האפשריות. פילוג זה מתאר את שכיחות ההופעה של המחלקות השונות. $P(\omega)$ נקרא <u>פילוג א-פריורי</u> של ω.

2. לכל מחלקה $\omega\in\Omega$, נתון פילוג הסתברות מותנה $p(x|\omega)$ על מרחב הקלט X. פילוג זה מתאר כיצד נראה "איבר אופייני" מכל מחלקה. הפילוג המותנה $p(x|\omega)$ נקרא לעיתים <u>פונקצית הסבירות</u> (likelihood function).

<u>הערה לגבי הסימון ההסתברותי</u>: Ω הוא מרחב סופי (בדיד), ולכן $P(\omega)$ מציין את הסתברות המחלקה ω. לעומת זאת, מרחב הקלט X עשוי להיות רציף או בדיד. במקרה הראשון, $p(x|\omega)$ מציין את פונקציית **צפיפות ההסתברות** (pdf, probability density function) המתאימה על X. כאשר X בדיד, $p(x|\omega)$ הינו ההסתברות עצמה (pmf, probability mass function). לעיתים נרשום $P(x|\omega)$ במקרה הבדיד כאשר נדרש להבדיל בין שני המקרים.

<u>הערה נוספת לגבי הסימון</u>: לעיתים נשתמש בסימון דומה לפונקציות הסתברות שונות, למשל $P(\omega)$ ו- $P(x)$, כאשר הראשונה מציינת פילוג הסתברות על X, והשנייה על Ω. מבחינה מתמטית, נדרש סימון שונה להסתברויות אילו, דהיינו $P_\Omega(\omega)$ ו- $P_X(x)$. על מנת למנוע סרבול, אנו נסתפק באבחנה שמספק הארגומנט של P או p.

נציין כי הפילוגים $P(\omega)$ ו- $p(x|\omega)$ יחדיו מגדירים למעשה באופן מלא את הפילוג המשותף $p(x,\omega)$, לפי $p(x,\omega)=p(x|\omega)P(\omega)$. הסיבה שנוקבים דווקא בפילוגים אלה קשורה לייצוג הטבעי של פילוגי ההסתברות בבעיות אופייניות.

<u>דוגמא 1 (המשך)</u>: בחורף ישראלי טיפוסי מזג האוויר יכול להיות נעים (ω_1), קר (ω_2) או גשום (וקר) (ω_3). עלינו להחליט האם ללבוש חולצה בלבד (y_1) או מעיל (y_2) על סמך מראה העננים מבעד לחלון (x). בהתאם לשכיחות מזג האוויר בחורף ישראלי טיפוסי נבחר $P(\omega_3)=0.10$ $P(\omega_2)=0.27$ $P(\omega_1)=0.63$. כמו כן, עלינו לבחור את הפילוג של התצפית בהנתן כל אחד ממצבי העולם. נגדיר את הפילוגים $p(white|\omega_1)$, $p(gray|\omega_1)$ וכמובן $p(blue|\omega_1)$ עבור מצב העולם הראשון, ובאופן דומה פילוגים בהנתן כל אחד ממצבי העולם. בסך הכל עלינו להגדיר 3 משתנים תלויים (כי סכומם אחד, כדי להיות הסתברות מלאה, או שני משתנים בלתי תלויים) לכל אחד ממצבי העולם, וסה"כ נקבל שישה משתנים בלתי תלויים עבור התצפיות בהנתן מצב העולם.

<u>דוגמא 3 (המשך)</u>: נדרש לסווג אנשים בוגרים ל"אישה" (ω_1) או "גבר" (ω_2) לפי גובהם (x) בלבד . בהתאם לשכיחות המינים באוכלוסיה הרלוונטית, נבחר (למשל) $P(\omega_1)=0.52$, $P(\omega_2)=0.48$. כמו כן, נבחר את $p(x|\omega_1)$ ו- $p(x|\omega_2)$ כפילוגים נורמליים (גאוסיים) עם ממוצע ושונות מתאימים (בהתאם לנתונים סטטיסטיים לגבי פילוג הגבהים באוכלוסיה).

נשים לב כי הפילוג הגאוסי אינו מייצג פילוג "אמיתי" של הגבהים – למשל, הוא מעניק הסתברות חיובית לגבהים שליליים. אולם אנו מניחים כי הוא מהווה קרוב מספק לצורך התכן.

דוגמא 6 : נדרש לסווג אנשים בוגרים ל"אישה" (ω_1) או "גבר" (ω_2) לפי תמונתם. במקרה זה x הינו תמונת פורטרט בייצוג נתון, למשל 128*128 פיקסלים ברמות אפור 0:255. $P(\omega_i)$ יוגדרו כמו קודם. את $p(x|\omega_i)$, לעומת זאת, קשה יותר להגדיר בדוגמא זו!

2.3. סיווג בייסיאני: מדד הביצועים

נתמקד כעת במקרה שבו הפעולה הנדרשת היא זיהוי המצב, דהיינו $\underline{Y=\Omega}$

עלינו להגדיר מהו קריטריון הטיב של מסווג. נניח כי התקבל קלט x מתוך מחלקה $\omega\in\Omega$.

מסווג אידאלי יקיים כמובן $f(x)=\omega$. מסווג טוב אמור להיות "קרוב" למסווג האידיאלי.

מדד השגיאה: ההגדרה הפשוטה ביותר של מדד ביצועים היא תוחלת מספר השגיאות (אותה נרצה להקטין ככל האפשר). עבור מסווג נתון $f(x)$, נגדיר שגיאת סיווג כמאורע $f(x)\neq\omega$.

נגדיר עתה את הסתברות השגיאה המותנית:

$$P(error\,|\,x)=P\{f(x)\neq\omega\,|\,x\}$$

וכן את הסתברות השגיאה הממוצעת:

$$P(error)=P\{f(x)\neq\omega\}$$

הערה - מדד השגיאה המשוקללת: ביישומים מסוימים, ייתכן כי לשגיאות שונות תהיה משמעות שונה, ולכן מחיר שונה. במקרה זה ניתן להגדיר **פונקציית הפסד** (loss) $\ell(y,\omega)$, אשר תקיים (בה"כ) את התנאים הבאים:

א. $\ell(y,\omega)\geq 0$

ב. $\ell(y,\omega)=0$ אם $y=\omega$.

ניתן עתה להגדיר את מדד הביצועים המותנה:

$$L(x)=E(\ell(f(x),\omega)\,|\,x)$$

ואת מדד הביצועים הממוצע:

$$L=E(\ell(f(x),\omega)).$$

נשים לב כי מדד השגיאה ה"רגיל" מתקבל כמקרה פרטי, כאשר

$$\ell(y,\omega)=\begin{cases}1\ :\ f(x)\neq\omega\\ 0\ :\ f(x)=\omega\end{cases}.$$

בדוגמת הזיהוי המקרוביולוגי ברור ששגיאות מסוימות יקרות הרבה יותר משגיאות אחרות.

2.4. חוק בייס, ההסתברות בדיעבד

מתוך פונקציות ההסתברות הבסיסיות $P(\omega)$ ו- $p(x|\omega)$, נחשב עתה את ההסתברות $P(\omega|x)$. הסתברות זאת נקראת ההסתברות בדיעבד (*a-posteriori*), וניתנת לחשבה בעזרת חוק בייס (Bayes):

$$P(\omega|x)=\frac{p(x|\omega)P(\omega)}{p(x)}$$

כאשר

$$p(x)=\sum_{\omega\in\Omega}p(x|\omega)P(\omega)$$

$P(\omega|x)$ מציין את הסתברות המחלקה ω, <u>לאחר</u> שראינו את תבנית הקלט x. חוק סיווג הגיוני, לאור הבחנה זו, הינו לבחור עבור כל קלט x את המחלקה ω שהיא בעלת הסבירות הגבוהה ביותר לפי $P(\omega|x)$, דהיינו:

$$f(x)=\arg\max_{\omega\in\Omega}P(\omega|x)$$

מסווג זה נקרא מסווג בייס, וגם מסווג ההסתברות-בדיעבד המירבית, או MAP classifier (Maximum a-posteriori classifier), ויסומן $f_{MAP}(x)$. כפי שנראה מייד, מסווג זה מביא למינימום את תוחלת הסתברות השגיאה.

<u>נוסחה חליפית:</u> ע"י הצבת הנוסחה עבור $P(\omega|x)$ נקבל כי

$$f_{MAP}(x)=\arg\max_{\omega\in\Omega}\left\{\frac{p(x|\omega)P(\omega)}{p(x)}\right\}$$

אולם מכיוון ש- $p(x)$ אינו תלוי ב- ω, נקבל כי

$$f_{MAP}(x)=\arg\max_{\omega\in\Omega}\left\{p(x|\omega)P(\omega)\right\}$$

נוסחה זו חוסכת את החישוב (המיותר) של $p(x)$.

<u>המקרה של שתי מחלקות:</u> כאשר Ω כולל שתי מחלקות בלבד, כלומר $\Omega=\{\omega_1,\omega_2\}$, ניתן לבטא את $f_{MAP}(x)$ באופן הבא:

$$p(x|\omega_1)P(\omega_1)>p(x|\omega_2)P(\omega_2)\quad\Rightarrow\quad f(x)=\omega_1$$
$$p(x|\omega_1)P(\omega_1)<p(x|\omega_2)P(\omega_2)\quad\Rightarrow\quad f(x)=\omega_2$$

באופן שקול,

$$\frac{p(x|\omega_1)}{p(x|\omega_2)}\frac{P(\omega_1)}{P(\omega_2)} \gtrless 1 \quad \Rightarrow \log\left(\frac{p(x|\omega_1)}{p(x|\omega_2)}\right)+\log\left(\frac{P(\omega_1)}{P(\omega_2)}\right) \gtrless 0 \Rightarrow f(x)={}^{\omega_1}\!/_{\omega_2}$$

היחס בצד שמאל נקרא <u>יחס הסבירות</u>.

<u>2.5. המסווג הבייסיאני האופטימלי</u>

נזכור את הגדרת הסתברות השגיאה המותנית :

$$P(error\,|\,x)=P\{f(x)\neq\omega\,|\,x\}$$

וכן את הגדרת הסתברות השגיאה הממוצעת :

$$P(error)=P\{f(x)\neq\omega\}$$

<u>משפט 1</u> : המסווג $f_{MAP}(x)$ מביא למינימום הן את הסתברות השגיאה המותנית, והן את הסתברות השגיאה הממוצעת.

<u>הוכחה</u> : נתבונן ראשית בהסתברות השגיאה המותנית :

$$P(error\,|\,x)=P\{f(x)\neq\omega\,|\,x\}=1-P\{\omega=f(x)\,|\,x\}$$

מכאן שמזעור $P(error\,|\,x)$ שקול לבחירת $f(x)$ שמביא למכסימום את $p(\omega=f(x)\,|\,x)$

. אבל זו בדיוק הגדרת $f_{MAP}(x)$.

נעבור להסתברות השגיאה הממוצעת :

$$P(error)=E(P(error\,|\,x))=\int_x P(error\,|\,x)\,p(x)dx$$

אולם ראינו כי $f_{MAP}(x)$ ממזער את $P(error\,|\,x)$ לכל x, ומכאן שהוא ממזער את $P(error)$. □

<u>תרגיל</u>: עבור פונקציית הפסד כללית $\ell(y,\omega)$, נגדיר :

$$L_\ell(x)=E(\ell(f(x),\omega)\,|\,x),\quad L_\ell=E(\ell(f(x),\omega))$$

(כאמור לעיל, $L_\ell(x)$ מכונה *הסיכון המותנה*, ואילו L_ℓ הינו *הסיכון הממוצע*, Risk). הניחו עתה כי פונקציית ההפסד $\ell(y,\omega)$ הינה מהצורה :

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

כאשר $\ell(\omega)>0$ לכל $\omega\in\Omega$. מצאו את המשערך $f_\ell(x)$ אשר ממזער את הסיכון המותנה (לכל x) ואת הסיכון הממוצע.

רמז : הראו ראשית כי $L_\ell(x)=\sum_\omega P(\omega\,|\,x)\ell(\omega)I\{f(x)\neq\omega\}$.

2.6. סיווג בייסיאני – המקרה הגאוסי

נדגים עתה את חישוב המסווג הבייסיאני האופטימלי במקרה הפשוט שבו פונקציות הסבירות $p(x\,|\,\omega)$ הן בעלות פילוג גאוסי.

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

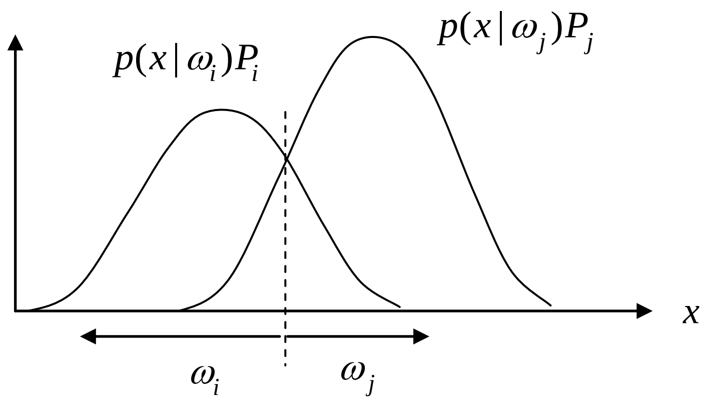
$$p(x\,|\,\omega_i)=\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right),\quad \omega_i\in\Omega$$

כזכור, $f_{MAP}(x)=\underset{\omega\in\Omega}{\arg\max}\left\{p(x\,|\,\omega)P(\omega)\right\}$. בפרט, הסיווג של קלט $x\in\mathbb{R}$ למחלקה ω_i עדיף על סיווגו למחלקה ω_j אם

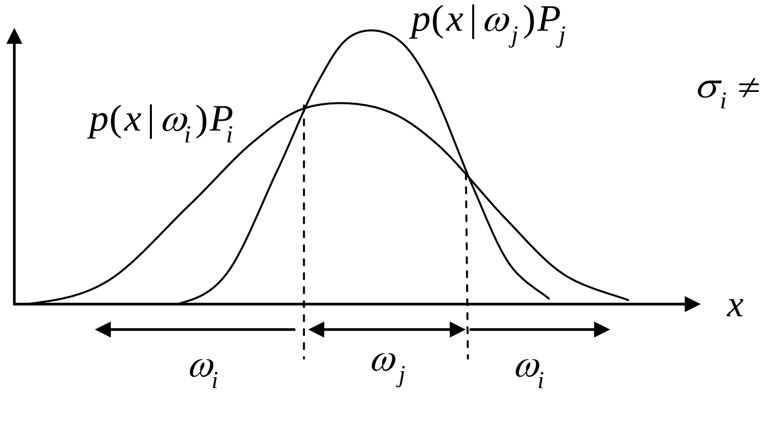
$$p(x\,|\,\omega_i)P(\omega_i)\geq p(x\,|\,\omega_j)P(\omega_j) .$$

ניתן לתאר זאת באופן גרפי :

- שונות זהה : $\sigma_i=\sigma_j$



- שונות בלתי-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 נקראים Bayes decision boundary.
נקודות השפה ניתנות לחישוב ע״י פתרון משוואה ריבועית (לאחר הוצאת לוגריתם).

(2) המקרה הדו-מימדי ($x\in\mathbb{R}^2$) : נניח כי פונקציות הסבירות נתונות ע״י הפילוגים הגאוסיים :

$$p(x\,|\,\omega_i)=\frac{1}{(2\pi)^{d/2}\,|\Sigma_i|^{1/2}}\exp\left(-\tfrac{1}{2}(x-\mu_i)\Sigma_i^{-1}(x-\mu_i)\right),\quad \omega_i\in\Omega$$

כאשר $d=2$. גם במקרה זה, ההעדפה בין שתי מחלקות תתבצע בהתאם לאי השיוויון $p(x\,|\,\omega_i)P(\omega_i) \gtrless p(x\,|\,\omega_j)P(\omega_j)$. לאחר הוצאת לוגריתם נקבל את אי השיוויון השקול:

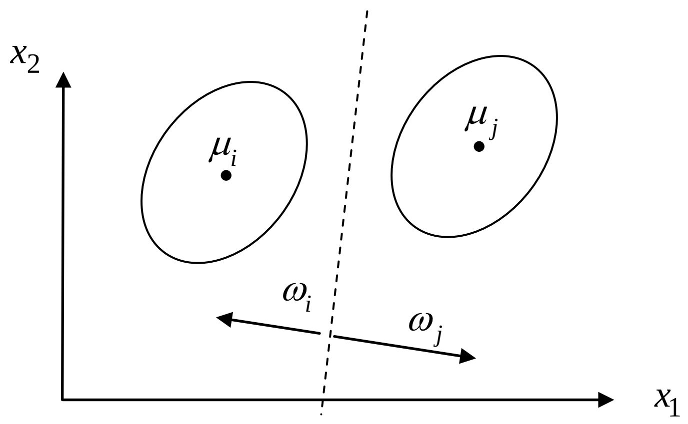
$$g_i(x)\gtrless g_j(x)$$

כאשר

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

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

שפות תחומי ההחלטה הבייסאנית נקבעים על ידי השיוויון $g_i(x)=g_j(x)$, והם <u>עקומים ריבועיים</u> : אליפסה, היפרבולה, או שתי היפרבולות. במקרה המיוחד שבו $\Sigma_i=\Sigma_j$ האיברים הריבועיים מתבטלים, ומתקבל קו ישר :



<u>בעיית האסירים</u>

לילה בבית הכלא, ישנם שלושה אסירים שלמחרת אחד ישאר כלוא ושניים אחרים ישוחררו. אחד האסירים פונה לסוהר ומבקש ממנו בדחילו ורחימו שיאמר לו את שמו של אחד האסירים שאינו הוא שישוחרר. הסוהר מסרב, בתואנה שבכך ישנה את סיכוייו להשאר. האם הסוהר צודק?

נרשום את הבעיה כבעיית הכרעה בייסינת. מצבי העולם הם A,B ו- C. הפעולות המתאימות להם הם a,b,c. נניח פריור אחיד $p(a)=p(b)=p(c)=1/3$

נניח כי האסיר המדובר היא A. מהם הסתברויות התצפית (מה שהסוהר אמר) בהנתן מצבי העולם, ידוע כי

P(told B | C) = P(told C|B) = 1

אנו נגדיר אם כן,

P(told B | A) = p

P(told C | A) = 1-p

עלינו לחשב את ההסתברות כי A ישאר בהנתן כל אחת מן התצפיות, בה"כ B

$$P(A|\text{told B})=\frac{P(\text{told B}|A)}{P(\text{told B})}P(A)=\frac{p}{p\,1/3+1\,1/3}\left(\frac{1}{3}\right)=\frac{p}{p+1}$$

ולכן, למעט המקרה שבו p=0.5, שאז p=1/2 , נקבל כי העובדה שנמסר שמו של B שאכן מוסיפה מידע לאסיר שלנו!