

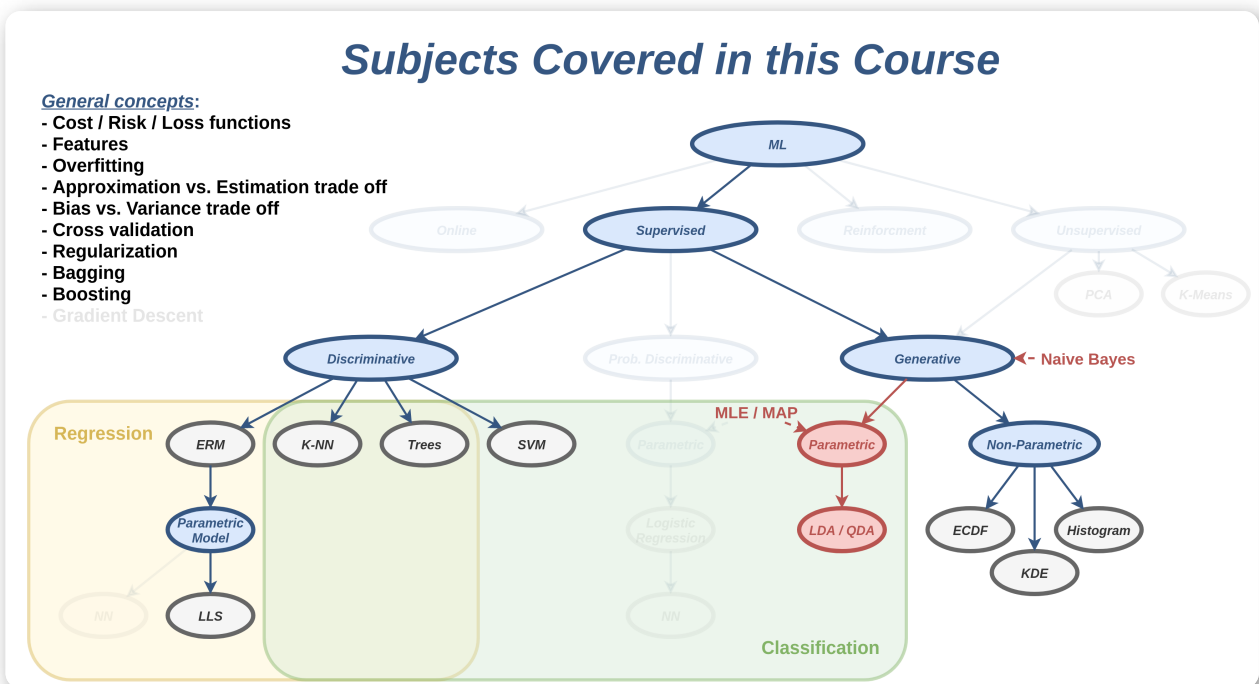
# הרצאה 8 - שיערוך פילוג

## בשיטות פרמטריות וסיווג

### גנרטיבי

Slides PDF Code

מה נלמד היום



## בעיות בגישה הלא פרמטרית

### הערה לגבי זמן הריצה בשלב בניית המודל ובשלב החיזוי

במרבית המקרים בשלב בניית המודל כמות המשאבים שיעמדו לרשותנו (זמן ריצה, כוח חישוב זיכרון וכו') תהיה מאד גדולה. לדוגמא יש כיום הרבה מודלים אשר מאומנים בחוות שרתים על מחשבים מאד חזקים, לרוב עם חומרה יעודית, במשך כמה ימים. מצד שני, בזמן החיזוי אנו לנרצה במקרים רבים לבצע את החישוב על פלטפורמה יחסית חלשה ולקבל תוצאות מאד מהירות. לדוגמא מערכת להתראה על סכנות בכביש צריכה לבצע חיזוי עבור תמונות שמצולמות בקצב של מספר תמונות בשנייה ולרוץ על מערכת פשוטה יחסית שיושבת בתוך האוטו.

דוגמא

נסתכל גם על דוגמא מספרית. נניח שאנו רוצים לחזות האם אדם מסויים שקיבל מכה בכתף פרק אותה על פי הסימפטומים שלו. לשם כך נסתכל על המדגם הבא

פריקה	כאב בכתף	נפיחות	סימנים כחולים	נימול ביד	נזלת	
+	+	+	+	+	-	1
+	+	+	+	-	-	2
+	+	+	-	+	-	3
+	+	+	+	-	+	4
-	-	-	-	-	-	5
-	+	-	-	-	+	6
-	-	-	+	-	-	7
-	+	-	-	-	-	8

נסתכל על ניסיון לחזות את הסבירות שלאדם עם כל הסימפטומים יש פריקה בכתף.

## שיערוך נאיבי - הנחת חוסר תלות בין המשתנים

נניח שאנו רוצים לשערך צפיפות הסתברות של משתנה  $D$  ממדי. שיטה מאד נאיבית (לא מתוחכמת) לפתור את הבעיה היא להתעלם מהקשר בין המשתנים השונים ולהניח שהם בלתי תלויים סטטיסטית. זאת אומרת ש:

$$p_{\mathbf{x}}(\mathbf{x}) = p_{x_1}(x_1)p_{x_2}(x_2) \dots p_{x_D}(x_D) = \prod_{d=1}^D p_{x_d}(x_d)$$

תחת הנחה זו נוכל לשערך את הפילוג של כל אחד מה  $p_{x_d}(x_d)$  בנפרד. החיסרון של שיטה זו הוא שהיא מגבילה מאד את הפילוגים שניתן ללמוד.

## מסווג בייס נאיבי

נוכל להשתמש בשערוך הנאיבי לפתרון בעיות מסווג. לשם כך אנו נשתמש בהנחת החוסר תלות לשיערוך של  $p_{\mathbf{x}|y}(\mathbf{x}|y)$  ונניח כי ניתן לרשום את הפונקציית ההסתברות / צפיפות הסתברות המונתית באופן הבא:

$$p_{\mathbf{x}|y}(\mathbf{x}|y) = p_{x_1|y}(x_1|y)p_{x_2|y}(x_2|y) \dots p_{x_D|y}(x_D|y) = \prod_{d=1}^D p_{x_d|y}(x_d|y)$$

זאת אומרת ש**שבהינתן**  $y$  הרכיבים של  $\mathbf{x}$  בלתי תלויים סטטיסטית. אנו לא נרצה להניח חוסר תלות בין  $\mathbf{x}$  ל  $y$  מיכוון שזוהי בדיוק התלות שאנו מחפשים על מנת לבנות עלפיה את החיזוי של  $y$  בהינתן  $\mathbf{x}$ .

בעזרת הנחה זו נוכל לבנות חזאים אשר מבוססים על הפילוג של כל אחד מהמשתנים בנפרד. לדוגמא, אם נסתכל על החזאי אשר ממזער את ה misclassification rate:

$$\hat{y} = h(\mathbf{x}) = \arg \max_y p_{y|\mathbf{x}}(y|\mathbf{x})$$

נוכל תחת הנחת החוסר תלות (ובעזרת חוק בייס) לרשום את החזאי כ:

$$\hat{y} = \arg \max_y p_{\mathbf{x}|y}(\mathbf{x}|y)p_y(y) = \arg \max_y p_y(y) \prod_{d=1}^D p_{x_d|y}(x_d|y)$$

שיטה זו לחיזוי בעזרת הנחת החוסר תלות מכונה **סיווג בייס נאיבי (naïve Bayes classification)**.

## דוגמא 1 - זיהוי פריקה של הכתף

נסתכל שוב על הדוגמא של הפריקה של הכתף. על מנת לבנות חזאי, נשערך את  $p_y(y)$  ואז נשתמש בהנחת החוסר תלות ונשערך בנפרד את ההסתברות המותנית של כל אחד מהרכיבים  $p_{x_d|y}(x_d|y)$ . (לשם הפשטות נשתמש ב-1 ו-0 במקום ה-+ וה- (בהתאמה) אשר מופיעים בטבלה):

$$p_y(y) = \begin{cases} \frac{4}{8} = 0.5 & 1 \\ \frac{4}{8} = 0.5 & 0 \end{cases}$$

$$p_{x_{\text{pain}}|y}(x_{\text{pain}}|1) = \begin{cases} \frac{4}{4} = 1 & 1 \\ \frac{0}{4} = 0 & 0 \end{cases}$$

$$p_{x_{\text{pain}}|y}(x_{\text{pain}}|0) = \begin{cases} \frac{2}{4} = 0.5 & 1 \\ \frac{2}{4} = 0.5 & 0 \end{cases}$$

$$p_{x_{\text{swelling}}|y}(x_{\text{swelling}}|1) = \begin{cases} \frac{4}{4} = 1 & 1 \\ \frac{0}{4} = 0 & 0 \end{cases}$$

$$p_{x_{\text{swelling}}|y}(x_{\text{swelling}}|0) = \begin{cases} \frac{0}{4} = 0 & 1 \\ \frac{1}{4} = 0.25 & 0 \end{cases}$$

נמשיך כך גם לשלושת העמודות של "סימנים כחולים", "נימול" ו"נזלת".

החיזוי עבור המקרה שבו מופיעים כל שלושת הסימטומים יהיה:

$$\hat{y} = \arg \max_y p_y(y) \prod_{d=1}^D p_{x_d|y}(1|y)$$

זאת אומרת שהחיזוי יהיה 1 (שאכן יש פריקה) כאשר:

$$p_y(1) \prod_{d=1}^D p_{x_d|y}(1|1) \stackrel{?}{>} p_y(0) \prod_{d=1}^D p_{x_d|y}(1|0)$$

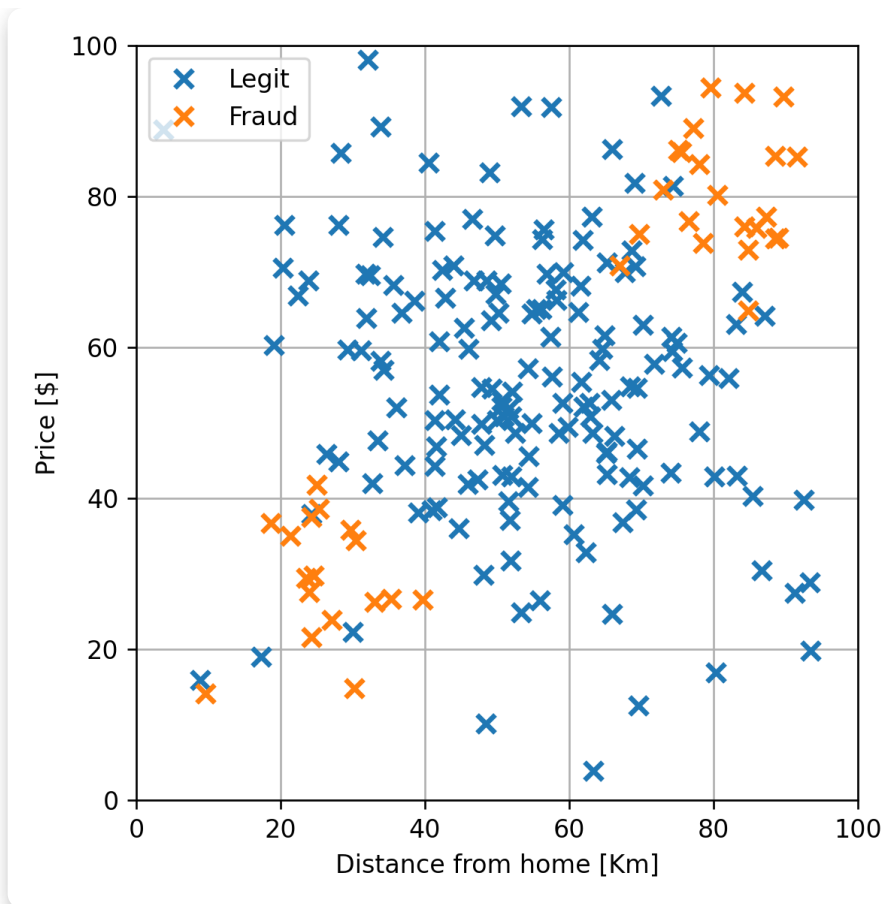
$$0.5 \times 1 \times 1 \times 0.75 \times 0.5 \times 0.25 \stackrel{?}{>} 0.5 \times 0.5 \times 0 \times 0.25 \times 0 \times 0.25$$

$$0.09375 \stackrel{?}{>} 0$$

מכיוון שתנאי זה מתקיים, החיזוי במקרה שבו מופיעים כל הסימפטומים יהיה שאכן ישנה פריקה של כתף.

## דוגמא 2 - זיהוי הונאות אשראי

ננסה להשתמש בשיטה זו לבעיית חיזוי הונאות האשראי



## שיטות פרמטריות

דרך נוספת לשיערוך פילוגים היא על ידי שימוש במודל פרמטרי. הדבר מאד דומה לשימוש שעשינו במודלים פרמטריים כאשר עסקנו בגישה הדטרמיניסטית. בשיטה זו אנו נגביל את הצורה של הפונקציה שאותה אנו רוצים לשערך (לרוב פונקציית צפיפות ההסתברות) למשפחה מומצמת של פונקציות על ידי שימוש במודל פרמטרי. גם כאן אנו נסמן את וקטור הפרמטרים של המודל ב  $\theta$ .

חשוב לשים לב שבניגוד לשימוש במודלים פרמטרים בגישה הדטרמיניסטית, שם לא הייתה שום מגבלה על המודל הפרמטרי, כאן המודל חייב לייצר פילוג חוקי עבור כל בחירה של פרמטרים (במקרה של PDF זה אומר פונקציה חיובית שאינטגרל עליה נותן 1). מגבלה זו הינה מגבלה קשה אשר מצמצמת מאד את המודלים הפרמטריים שאיתם ניתן לעבוד. מגבלה זו למשל מונעת מאיתנו מלהשתמש אפילו במודל לינארי פשוט. המודלים שאיתם נעבוד יהיו לרוב פונקציות פילוג ידועות כגון פילוג אחיד, נורמלי, אקפוננציאלי וכו'.

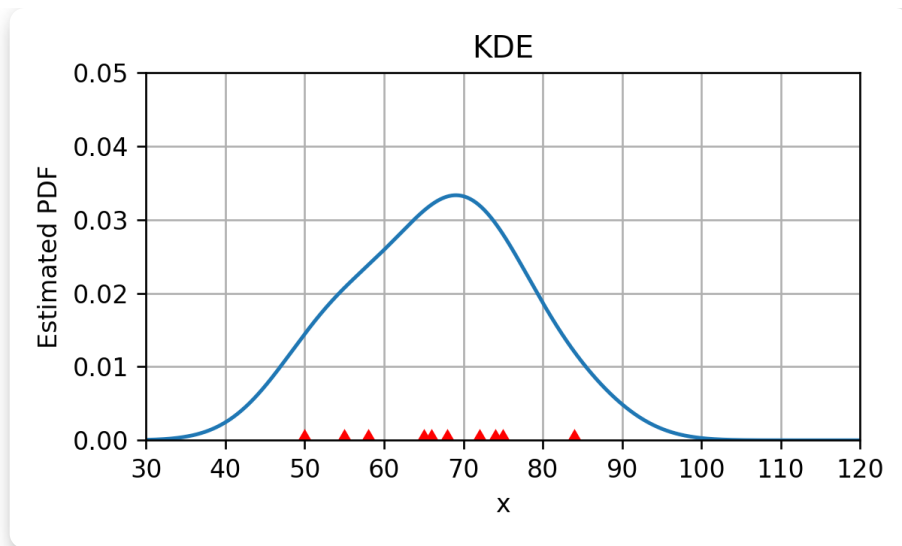
כדי למצוא את הפרמטרים של המודל נרצה גם כאן למצוא דרך לתת "ציון" לכל בחירה של פרמטרים ולחפש את הפרמטרים אשר מניבים את הציון הטוב ביותר. נציג כעת שתי דרכים שונות להתייחס לפרמטרים של המודל. שני דרכים אלו מגיעות משתי גישות הקיימות בתחום של תורת השיערוך. כל גישה מובילה לדרך מעט שונה לבחירה של הפרמטרים האופטימאליים.

## דוגמא: שיערוך הפילוג של זמן הנסיעה בכביש החוף

בהרצאה הקודמת הסתכלנו על שיערוך של הפילוג של זמן הנסיעה בכביש החוף מתוך הדגם הבא:

$$\mathcal{D} = \{x^{(i)}\} = \{55, 68, 75, 50, 72, 84, 65, 58, 74, 66\}$$

משערך ה KDE (עם גרעין גאוסיאני עם הרוחב אשר נקבע על פי כלל האצבע) של הפילוג (לא דנו הסמסטר) הינו:



(הנקודות האדומות על ציר ה- $x$  מסמנות את המיקומים של הדגימות מהמדגם)

בהרצאה זו ננסה לשערך פרמטרים של פילוג נורמלי שיתאר בצורה טובה את הדגימות במדגם.

## הגישה הלא-בייסיאנית (המכונה גם: קלאסית או תדירותית (Frequentist))

תחת גישה זו אנו נתייחס לפרמטרים בצורה פשוטה כאל מספרים שאותם יש לקבוע על מנת שהמודל יתאר בצורה טובה את המדגם הנתון. ההנחה היא כי יש ערך לא ידוע של הפרמטר שהוא ה"טוב" ביותר. את המודל הפרמטרי להסתברות / צפיפות הסתברות של משתנה אקראי  $x$  נסמן ב:

$$p_x(\mathbf{x}; \theta)$$

ונרצה לבחון עד כמה טוב מתאר מודל עם פרמטרים מסוימים את הפילוג של הדגימות במדגם. אחת הדרכים הנפוצות ביותר לעשות זאת הינה בעזרת פונקציית הסבירות.

## משערך (Maximum Likelihood Estimator (MLE

נסמן ב-  $p_{\mathcal{D}}(\mathcal{D}; \theta)$  את ההסתברות לקבלת המדגם הנתון  $\mathcal{D} = \{\mathbf{x}^{(i)}\}$  על פי המודל שבידינו. גודל זה מכונה **הסבירות (likelihood)** של המדגם. אנו מעוניינים למצוא את הפרמטרים  $\theta$  אשר מניבים את הסבירות הכי גבוהה. על מנת להדגיש את העובדה שהמדגם הוא למעשה גודל ידוע ואילו הגודל הלא ידוע, שאותו נרצה לבדוק, הינו  $\theta$ , מקובל לסמן את פונקציית likelihood באופן הבא:

$$\mathcal{L}(\theta; \mathcal{D}) \triangleq p_{\mathcal{D}}(\mathcal{D}; \theta)$$

משערך ה- MLE של  $\theta$  הוא הערך אשר ממקסם את פונקציית ה-likelihood:

$$\hat{\theta}_{MLE} = \arg \max_{\theta} \mathcal{L}(\theta; \mathcal{D})$$

מקובל לרשום בעיות אופטימיזציה כבעיות מינימיזציה, לכן במקרים רבים נרשום אותה כבעיה של מציאת הפרמטרים אשר ממזערים את המינוס של פונקציית הסבירות:

$$\hat{\theta}_{MLE} = \arg \min_{\theta} -\mathcal{L}(\theta; \mathcal{D})$$

כאשר הדגימות במדגם הם i.i.d (בעלות פילוג זהה ובלתי תלויות, כפי שנניח תמיד שמתקיים בבעיות supervised learning) נוכל להסיק כי:

$$p_{\mathcal{D}}(\mathcal{D}; \theta) = \prod_i p_x(\mathbf{x}^{(i)}; \theta)$$

ולכן:

$$\hat{\theta}_{\text{MLE}} = \arg \min_{\theta} -\mathcal{L}(\theta; \mathcal{D}) = \arg \min_{\theta} -\prod_i p_{\mathbf{x}}(\mathbf{x}^{(i)}; \theta)$$

במקרים רבים נוכל להחליף את המכפלה על כל הדגימות בסכום, על ידי החלפת פונקציית ה-likelihood ב- log-likelihood (בזכות המונוטוניות העולה של פונקציית ה- log מובטח לנו שקבל את אותם פרמטרים אופטימאליים בשתי הבעיות):

$$\hat{\theta}_{\text{MLE}} = \arg \min_{\theta} -\log \mathcal{L}(\theta; \mathcal{D}) = \arg \min_{\theta} -\sum_i \log \left( p_{\mathbf{x}}(\mathbf{x}^{(i)}; \theta) \right)$$

**הערה:** בקורסים "עיבוד אותות אקראיים" ו-"הסקה סטטיסטית" מרחיבים הרבה בנושא תכונות משערך זה ואחרים.

## דוגמא

נסתכל על הדוגמא של התאמת פילוג נורמלי לנסיעות בכביש החוף. הפרמטרים של מודל זה הינם התוחלת  $\mu$  וסטיית התקן  $\sigma$ . נסמן את וקטור הפרמטרים ב-  $\theta = [\mu, \sigma]^T$ . המודל שלנו יהיה:

$$p_{\mathbf{x}}(x; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

נרשום את בעיית האופטימיזציה של מציאת משערך ה- MLE:

$$\begin{aligned} \hat{\theta}_{\text{MLE}} &= \arg \min_{\theta} -\sum_i \log \left( p_{\mathbf{x}}(x^{(i)}; \theta) \right) \\ &= \arg \min_{\theta} -\sum_i \log \left( \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x^{(i)}-\mu)^2}{2\sigma^2}\right) \right) \\ &= \arg \min_{\theta} \sum_i \log(\sigma) + \frac{1}{2} \log(2\pi) + \frac{(x^{(i)}-\mu)^2}{2\sigma^2} \\ &= \arg \min_{\theta} N \log(\sigma) + \frac{1}{2\sigma^2} \sum_i (x^{(i)}-\mu)^2 \end{aligned}$$

את בעיית האופטימיזציה הזו ניתן לפתור על ידי גזירה והשוואה ל-0. נסמן את הפונקציה שאותה יש למזער (ה- objective)  $f$ :  
פונקציית המטרה) ב-

$$f(\theta; x) = N \log(\sigma) + \frac{1}{2\sigma^2} \sum_i (x^{(i)}-\mu)^2$$

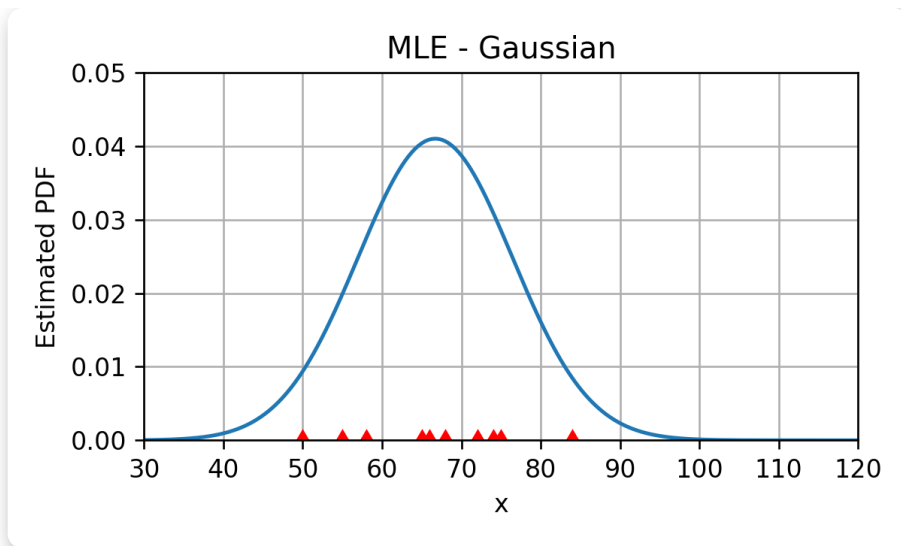
$$\begin{aligned} &\begin{cases} \frac{\partial f(\theta; x)}{\partial \mu} = 0 \\ \frac{\partial f(\theta; x)}{\partial \sigma} = 0 \end{cases} \\ \Leftrightarrow &\begin{cases} -\frac{1}{\sigma^2} \sum_i (x^{(i)}-\mu) = 0 \\ \frac{N}{\sigma} - \frac{1}{\sigma^3} \sum_i (x^{(i)}-\mu)^2 = 0 \end{cases} \\ \Leftrightarrow &\begin{cases} N\mu - \sum_i x^{(i)} = 0 \\ N\sigma^2 - \sum_i (x^{(i)}-\mu)^2 = 0 \end{cases} \\ \Leftrightarrow &\begin{cases} \mu = \frac{1}{N} \sum_i x^{(i)} \\ \sigma = \sqrt{\frac{1}{N} \sum_i (x^{(i)}-\mu)^2} \end{cases} \end{aligned}$$

במקרה של הנסיעות בכביש החוף נקבל:

$$\mu = 66.7 \text{ [min]}$$

$$\sigma = 9.7 \text{ [min]}$$

או PDF המתקבל יראה כך:



## הגישה הבייסיאנית

בגישה זו אנו נניח כי בדומה למדגם, גם הפרמטרים של המודל הם ריאליזציות (הגרלות) של משתנה אקראי בעל פילוג כל שהוא. גישה זו למעשה מניחה שיש בידינו מודל לפילוג המשותף של הפרמטרים והמדגם. לרוב הפילוג משותף יהיה נתון בצורה של:

$$p_{\mathcal{D},\theta}(\mathcal{D}, \theta) = p_{\mathcal{D}|\theta}(\mathcal{D}|\theta)p_{\theta}(\theta)$$

תחת ההנחה שבהינתן הפרמטרים הדגימות במדגם הם i.i.d, הפילוג המשותף יהיה:

$$p_{\mathcal{D},\theta}(\mathcal{D}, \theta) = p_{\theta}(\theta) \prod_i p_{x|\theta}(\mathbf{x}^{(i)}|\theta)$$

תחת גישה זו עלינו לקבוע בנוסף לפילוג של הדגימות בהינתן הפרמטרים  $p_{x|\theta}(\mathbf{x}|\theta)$  וגם את הפילוג השולי של הפרמטרים  $p_{\theta}(\theta)$ .

## A Priori Distribution

הפילוג השולי של הפרמטרים  $p_{\theta}(\theta)$ , מכונה לרוב **הפילוג הפריורי (prior distribution)** או **הא-פריורי (a priori distribution)**, זאת אומרת הפילוג של  $\theta$  לפני שראינו את המדגם.

## A Posteriori Distribution

פילוג חשוב נוסף שנרצה להתייחס אליו הוא הפילוג של הפרמטרים בהינתן המדגם  $p_{\theta|\mathcal{D}}(\theta|\mathcal{D})$ . פילוג זה מכונה **הפילוג הפוסטריורי (posterior distribution)** או **א-פוסטריורי (a posteriori distribution)** (או הפילוג בדיעבד), זאת אומרת, הפילוג אחרי שראינו את המדגם.

## משעך (Maximum A-posteriori Probability (MAP

הדרך הנפוצה ביותר לשעך את הפרמטרים  $\theta$  היא למצוא את הערך אשר ממקסם את הפילוג הא-פוסטריורי (הערך הכי סביר של  $\theta$  בהינתן המדגם  $p_{\theta|\mathcal{D}}(\theta|\mathcal{D})$ ):

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} p_{\theta|\mathcal{D}}(\theta|\mathcal{D}) = \arg \min_{\theta} -p_{\theta|\mathcal{D}}(\theta|\mathcal{D})$$

על פי חוק ביס, נוכל לכתוב זאת כ:

$$\hat{\theta}_{\text{MAP}} = \arg \min_{\theta} - \frac{p_{\mathcal{D}|\theta}(\mathcal{D}|\theta)p_{\theta}(\theta)}{p_{\mathcal{D}}(\mathcal{D})} = \arg \min_{\theta} - p_{\mathcal{D}|\theta}(\mathcal{D}|\theta)p_{\theta}(\theta)$$

כאשר הדגימות במדגם **בהינתן  $\theta$**  הן i.i.d, מתקיים:

$$\hat{\theta}_{\text{MAP}} = \arg \min_{\theta} -p_{\theta}(\theta) \prod_i p_{x|\theta}(\mathbf{x}^{(i)}|\theta)$$

גם כאן נוכל להפוך את המכפלה לסכום על ידי מזעור מינוס הלוג של הפונקציה:

$$\hat{\theta}_{\text{MAP}} = \arg \min_{\theta} -\log(p_{\theta}(\theta)) - \sum_i \log(p_{x|\theta}(\mathbf{x}^{(i)}|\theta))$$

## ההבדל בין MLE ל MAP

ההבדל בין משעריך ה MLE למשעריך ה MAP הינו התוספת של האיבר  $-\log(p_{\theta}(\theta))$ . איבר זה, ששווה ללוג של הפילוג הא-פריורי, מוסיף למעשה את הידע שיש לנו לגבי איזה ערכים של  $\theta$  יותר סבירים. ראינו תוספת שכזו כאשר דיברנו על רגולריזציה, שם הוספנו איבר לבעיית האופטימיזציה במטרה למשוך את הפתרון לאיזורים שהנחנו שהם יותר סבירים. לכן, ניתן למעשה לחשוב על בעיית ה MAP כעל בעיית MLE עם רגולריזציה. בתרגיל הבית אתם תראו את השקילות שבין בעיית MAP לבין לבעיית MLE עם רגולריזציה.

## בגישה בייסיאנית בעיית השיערוך היא בעיית חיזוי

כפי שצינו, בגישה הבייסיאנית אנו מתייחסים גם למדגם וגם לפרמטרים בריאליזציות של משתנים אקראיים, בנוסף, אנו מניחים שאנו יודעים את הפילוג המשותף שלהם ואנו מנסים למצוא את הערך של הפרמטרים בהינתן המדגם. זוהי בדיוק בעיית חיזוי קלאסית של משתנה אקראי אחד בהינתן משתנה אקראי אחר על סמך הפילוג המשותף (חיזוי של הפרמטרים בהינתן המדגם).

במרבית המקרים המקרים פונקציית הצפיפות המשותפת יהיו מסובכות יכללו מכפלה של הרבה מאד איברים. לכן, חיזויים אחרים כגון התחלת המותנית או החציון יהיו לרוב מסובכים מידי לחישוב.

## דוגמא - הוספת prior

נחזור לדוגמא של התאמת מודל של פילוג נורמלי לפילוג של זמן הנסיעה בכביש החוף. לשם הפשטות בדוגמא זו נקבע את סטיית התקן של המודל ל  $\sigma = 10$  כך שיהיה לנו מודל פרמטרי בעל פרמטר יחיד  $\mu$ :

$$p_{x|\mu}(x|\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

נניח כעת שיש לנו ידע קודם על פילוג הצפוי של  $\mu$ . ידע כזה יכול לדוגמא להגיע מתוך סטטיסטיקה שאספנו על מהירויות הנסיעה בכבישים אחרים בארץ. הפילוג הצפוי של  $\mu$  יהיה הפילוג הא-פריורי של פרמטר זה. נניח אם כן שהפילוג הא-פריורי של  $\mu$  הוא גם פילוג נורמלי עם תוחלת  $\mu_{\mu} = 60$  וסטיית תקן של  $\sigma_{\mu} = 5$ :

$$p_{\mu}(\mu) = \frac{1}{\sqrt{2\pi}\sigma_{\mu}} \exp\left(-\frac{(\mu-\mu_{\mu})^2}{2\cdot\sigma_{\mu}^2}\right)$$

התיאור פה מעט מבלבל משום שאנו מניחים שגם הפילוג של זמן הנסיעה הוא נורמלי וגם הפילוג הא-פריורי של  $\mu$  נורמלי. אלא שני פילוגים שונים שבמקרה נבחרו להיות בעלי אותו מבנה. ניתן לחשוב על התהליך של ייצור זמני הנסיעה באופן הבא. בתחילה באופן חד פעמי (לצורך העניין עם הבניה של כביש החוף) מוגרל הפרמטר  $\mu$  מתוך הפילוג  $p_{\mu}$  אשר מאפיין את הנסיעות בכביש החוף. אחרי שפרמטר זה נקבע, עבור כל נסיעה מחדש מגרילים את זמן נסיעה מתוך  $p_{x|\mu}$  תוך שימוש בערך של  $\mu$  אשר הגרלנו.

נרשום את משעריך ה MAP של  $\mu$ :

$$\hat{\mu}_{\text{MAP}} = \arg \min_{\mu} -\log(p_{\mu}(\mu)) - \sum_i \log(p_{x|\mu}(\mathbf{x}^{(i)}|\mu))$$

גזירה והשוואה ל-0 נותנת את התוצאה הבאה:



$$\begin{aligned} \frac{\partial f(\boldsymbol{\theta}; \mathbf{x})}{\partial \mu} &= 0 \\ \Leftrightarrow \frac{1}{\sigma_\mu^2}(\mu - \mu_\mu) - \frac{1}{\sigma^2} \sum_i (x^{(i)} - \mu) &= 0 \\ \Leftrightarrow \left( \frac{1}{\sigma_\mu^2} + \frac{N}{\sigma^2} \right) \mu &= \frac{\mu_\mu}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_i x^{(i)} \\ \Leftrightarrow \mu &= \frac{\frac{\sigma^2}{N\sigma_\mu^2} \mu_\mu + \frac{1}{N} \sum_i x^{(i)}}{\frac{\sigma^2}{N\sigma_\mu^2} + 1} \end{aligned}$$

זו למעשה ממוצע ממושקל בין הערך הממוצע של  $x$  במדגם לבין  $\mu_\mu$ . עבור הדוגמא שלנו נקבל:

$$\mu = 64.8 \text{ [min]}$$

ערך זה מעט יותר קרוב ל-60 משאר התוצאה שקיבלנו בשיעור ה- MLE. זאת משום ה prior ש"מושך" את הפרמטרים לאיזורים הסבירים יותר ולכן הוא מקרב אותו ל  $\mu_\mu = 60$ .

## שימוש בשיעור פרמטרי לפתרון בעיות supervised learning

נראה עכשיו איך להשתמש בשיעור הצפיפות שתארנו צורך פתרון בעיות סיווג בלמידה ממוקחת. נציג שיטה אשר משתמשת במודל של פילוג נורמלי וב MLE לפתרון בעיות סיווג.

### (Quadratic Discriminant Analysis (QDA

בשיטה זו אנו נשתמש במודל של פילוג נורמלי וב MLE על מנת לשערך את הפילוג המותנה של  $\mathbf{x}$  בהינתן  $y$ . אנו למעשה צריכים לשערך מודל נורמלי אחד עבור כל אחת מ  $C$  המחלקות של  $y$  יכול לקבל. זאת אומרת שאנו נרצה לשערך את הפרמטרים הבאים:

- וקטור תוחלת  $\boldsymbol{\mu}_c$  עבור כל אחד מהמחלקות ( $c \in \{1, 2, \dots, C\}$ ).
- מטריצת קווריאנס  $\Sigma_c$  עבור כל אחד מהמחלקות.

כאשר המודל של הפילוג של  $\mathbf{x}$  בהינתן  $y = c$  יהיה:

$$p_{\mathbf{x}|y}(\mathbf{x}|c; \boldsymbol{\mu}_c, \Sigma_c) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_c|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_c)^\top \Sigma_c^{-1}(\mathbf{x} - \boldsymbol{\mu}_c)\right)$$

(כאשר  $D$  הוא המימד של  $\mathbf{x}$ ).

בנוסף יהיה עלינו לשערך את  $p_y(y)$ . (זה פילוג דיסקרטי אותו ניתן לשערך בקלות על פי השכיחות היחסית במדגם).

הפילוג המשותף של  $\mathbf{x}$  ו  $y$  יהיה:

$$p_{\mathbf{x},y}(\mathbf{x}, y; \{\boldsymbol{\mu}_c\}, \{\Sigma_c\}) = p_{\mathbf{x}|y}(\mathbf{x}|y; \boldsymbol{\mu}_y, \Sigma_y) p_y(y)$$

את הפרמטרים של המודל מוצאים בעזרת MLE. בעיית האופטימיזציה תהיה:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\text{MLE}} &= \arg \min_{\boldsymbol{\theta}} -\log \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) \\ &= \arg \min_{\boldsymbol{\theta}} -\sum_i \log \left( p_{\mathbf{x}|y}(\mathbf{x}^{(i)}|y^{(i)}; \boldsymbol{\mu}_y, \Sigma_y) p_y(y^{(i)}) \right) \\ &= \arg \min_{\boldsymbol{\theta}} -\sum_i \log \left( p_{\mathbf{x}|y}(\mathbf{x}^{(i)}|y^{(i)}; \boldsymbol{\mu}_y, \Sigma_y) \right) + \log \left( p_y(y^{(i)}) \right) \\ &= \arg \min_{\boldsymbol{\theta}} -\sum_i \log \left( p_{\mathbf{x}|y}(\mathbf{x}^{(i)}|y^{(i)}; \boldsymbol{\mu}_y, \Sigma_y) \right) \end{aligned}$$

בהנחה ש- $P_y$  ידוע.

נפשט מעט את הביטוי על ידי חלוקה של הסכימה לסכימות נפרדות על כל אחת מהמחלקות. לשם הפשטות נגדיר את הסימונים הבאים:

$$\bullet \mathcal{I}_c = \{i : y^{(i)} = c\} \text{ - זאת אומרת, אוסף האינדקסים של הדגמים במדגם שמקיימים } y^{(i)} = c.$$

$$\bullet |\mathcal{I}_c| \text{ - מספר האינדקסים ב } \mathcal{I}_c$$

נוכל כעת לרשום את בעיית האופטימיזציה באופן הבא:

$$\hat{\theta}_{\text{MLE}} = \arg \min_{\theta} - \sum_{i \in \mathcal{I}_1} \log(p_{\mathbf{x}|y}(\mathbf{x}^{(i)}|1; \boldsymbol{\mu}_1, \Sigma_1)) - \sum_{i \in \mathcal{I}_2} \log(p_{\mathbf{x}|y}(\mathbf{x}^{(i)}|2; \boldsymbol{\mu}_2, \Sigma_2)) - \dots$$

בשביל למצוא את הערכים האופטימאליים של  $\boldsymbol{\mu}_1$  ו  $\Sigma_1$  מספיק להסתכל על האיבר הראשון. לכן ניתן למעשה לפרק את הבעיה ל  $C$  בעיות נפרדות שבהם משערכים בנפרד את הפרמטרים של כל מחלקה.

עבור המחלקה ה  $c$  נקבל את בעיית האופטימיזציה הבאה:

$$\hat{\boldsymbol{\mu}}_{c,\text{MLE}}, \hat{\Sigma}_{c,\text{MLE}} = \arg \min_{\boldsymbol{\mu}_c, \Sigma_c} - \sum_{i \in \mathcal{I}_c} \log(p_{\mathbf{x}|y}(\mathbf{x}^{(i)}|c; \boldsymbol{\mu}_c, \Sigma_c))$$

$$= \arg \min_{\boldsymbol{\mu}_c, \Sigma_c} \sum_{i \in \mathcal{I}_c} \log(\sqrt{|\Sigma_c|}) + \frac{1}{2}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_c)^\top \Sigma_c^{-1}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_c)$$

ניתן לפתור את הבעיה הזו על ידי גזירה והשוואה ל-0. הפיתוח עבור  $\Sigma_c$  הוא מעט מורכב ואנו לא נראה אותו בקורס זה ונקפוץ לפתרון. הפיתוח מודגם בקורס "עיבוד וניתוח מידע". נראה אבל את החישוב של  $\boldsymbol{\mu}_c$

את בעיית האופטימיזציה הזו ניתן לפתור על ידי גזירה והשוואה ל-0. נסמן את ה objective ב  $f$ :

$$f(\boldsymbol{\theta}; \mathbf{x}) = \sum_{i \in \mathcal{I}_c} \log(\sqrt{|\Sigma_c|}) + \frac{1}{2}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_c)^\top \Sigma_c^{-1}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_c)$$

$$\frac{\partial f}{\partial \boldsymbol{\mu}_c} = 0$$

$$\Leftrightarrow - \sum_{i \in \mathcal{I}_c} \Sigma_c^{-1}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_c) = 0$$

$$\Leftrightarrow |\mathcal{I}_c| \boldsymbol{\mu}_c - \sum_{i \in \mathcal{I}_c} \mathbf{x}^{(i)} = 0$$

$$\Leftrightarrow \boldsymbol{\mu}_c = \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \mathbf{x}^{(i)}$$

שאר הפרמטרים של המודל יהיו:

$$p_y(c) = \frac{|\mathcal{I}_c|}{N}$$

$$\Sigma_c = \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_c) (\mathbf{x}^{(i)} - \boldsymbol{\mu}_c)^\top$$

הצגנו כאן משערך אינטואיטיבי להתסברות השיור לכל מחלקה. אפשר להראות שזו התוצאה המתקבלת גם מהנחת מודל פילוג מולטינומיאלי למשתנה הקטגורי של המחלקה.

כעת נוכל למצוא את החזאי האופטימאלי לבעיה בהינתן הפילוג שמצאנו. עבור פונקציית מחיר של סיכוי הטעות, החזאי האופטימאלי יהיה:

$$\hat{y} = h(\mathbf{x}) = \arg \max_y p_{\mathbf{x}|y}(\mathbf{x}|y; \boldsymbol{\mu}_y, \Sigma_y) p_y(y)$$

$$= \arg \max_y - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_y)^\top \Sigma_y^{-1}(\mathbf{x} - \boldsymbol{\mu}_y) + \log\left(\frac{p_y(y)}{\sqrt{|\Sigma_y|}}\right)$$

## המקרה הבינארי - משטח הפרדה ריבועי

עבור המקרה של סיווג בינארי (סיווג לשתי מחלקות) מתקבל החזאי הבא:

$$h(\mathbf{x}) = \begin{cases} 1 & -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + \log\left(\frac{p_y(1)}{\sqrt{|\Sigma_1|}}\right) > -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_0)^\top \Sigma_0^{-1}(\mathbf{x} - \boldsymbol{\mu}_0) + \log\left(\frac{p_y(0)}{\sqrt{|\Sigma_0|}}\right) \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} 1 & \mathbf{x}^\top C \mathbf{x} + \mathbf{a}^\top \mathbf{x} + b > 0 \\ 0 & \text{otherwise} \end{cases}$$

כאשר:

$$C = \frac{1}{2}(\Sigma_0^{-1} - \Sigma_1^{-1})$$

$$\mathbf{a} = \Sigma^{-1} \boldsymbol{\mu}_1 - \Sigma_0^{-1} \boldsymbol{\mu}_0$$

$$b = \frac{1}{2} (\boldsymbol{\mu}_0^\top \Sigma_0^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^\top \Sigma_1^{-1} \boldsymbol{\mu}_1) + \log\left(\frac{\sqrt{|\Sigma_0|} p_y(1)}{\sqrt{|\Sigma_1|} p_y(0)}\right)$$

התנאי שקיבלנו  $\mathbf{x}^\top C \mathbf{x} + \mathbf{a}^\top \mathbf{x} + b > 0$  הוא ריבועי ב  $\mathbf{x}$  ומכאן מקבל האלגוריתם את שמו.

## (Linear Discriminant Analysis (LDA

LDA שונה מ QDA בשינוי קטן בהנחות על מודל. על מנת להקטין את כמות הפרמטרים של המודל LDA מניח שלפונקציות הפילוג של המחלקות השונות יש את אותה מטריצת הקווריאנס. זאת אומרת שיש  $\Sigma$  יחידה אשר משותפת לכולם.

הפרמטרים של המודל יהיו כעת:

- וקטור תוחלת עבור כל אחד מהמחלקות  $\{\boldsymbol{\mu}_c \mid c \in \{1, 2, \dots, C\}\}$ .
- מטריצת covariance אחת  $\Sigma$  אשר משותפת לפילוגים של כל המחלקות.

המודל של הפילוג של  $\mathbf{x}$  בהינתן  $y = c$  הינו:

$$p_{\mathbf{x}|y}(\mathbf{x}|c; \boldsymbol{\mu}_c, \Sigma) = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_c)^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_c)\right)$$

גם כאן ניתן למצוא את הפתרון של בעיית האופטימיזציה על ידי גזירה והשוואה לאפס. במקרה זה בחיפוש אחר ה  $\Sigma$  האידיאלי לא ניתן להתייחס רק לחלק מהמדגם משום מהטריצה משותפת לכל המחלקות. הפתרון המתקבל הינו:

$$p_y(c) = \frac{|\mathcal{I}_c|}{N}$$

$$\boldsymbol{\mu}_c = \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \mathbf{x}^{(i)}$$

$$\Sigma = \frac{1}{N} \sum_i (\mathbf{x}^{(i)} - \boldsymbol{\mu}_{y^{(i)}}) (\mathbf{x}^{(i)} - \boldsymbol{\mu}_{y^{(i)}})^\top$$

עבור פונקציית מחיר של misclassification rate, החזאי האופטימאלי המתקבל ממודל זה הינו:

$$\begin{aligned} \hat{y} = h(\mathbf{x}) &= \arg \max_y p_{\mathbf{x}|y}(\mathbf{x}|y; \boldsymbol{\mu}_c, \Sigma) p_y(y) \\ &= \arg \max_y -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_y)^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_y) + \log(p_y(y)) \\ &= \arg \min_y \mathbf{x}^\top \Sigma^{-1} \boldsymbol{\mu}_y - \frac{1}{2} \boldsymbol{\mu}_y^\top \Sigma^{-1} \boldsymbol{\mu}_y - \log(p_y(y)) \end{aligned}$$

## המקרה הבינארי

עבור המקרה של סיווג בינארי (סיווג לשני מחלקות) מתקבל החזאי הבא:

$$h(x) = \begin{cases} 1 & \mathbf{x}^\top \Sigma^{-1} \boldsymbol{\mu}_1 - \frac{1}{2} \boldsymbol{\mu}_1^\top \Sigma^{-1} \boldsymbol{\mu}_1 - \log(p_y(1)) > \mathbf{x}^\top \Sigma^{-1} \boldsymbol{\mu}_0 - \frac{1}{2} \boldsymbol{\mu}_0^\top \Sigma^{-1} \boldsymbol{\mu}_0 - \log(p_y(0)) \\ 0 & \text{otherwise} \end{cases}$$
$$= \begin{cases} 1 & \mathbf{a}^\top \mathbf{x} + b > 0 \\ 0 & \text{otherwise} \end{cases}$$

כאשר:

$$\mathbf{a} = \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$$

$$b = \frac{1}{2} (\boldsymbol{\mu}_0^\top \Sigma^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^\top \Sigma^{-1} \boldsymbol{\mu}_1) + \log\left(\frac{p_y(1)}{p_y(0)}\right)$$

התנאי שקיבלנו  $\mathbf{a}^\top \mathbf{x} + b > 0$  הוא לינארי ב  $\mathbf{x}$  ומכאן מקבל האלגוריתם את שמו. תנאי זה מחלק את המרחב של  $\mathbf{x}$  לשני חלקים המופרדים על ידי המישור  $\mathbf{a}^\top \mathbf{x} + b = 0$ .

## המקרה הכללי (לא בינארי)

במקרה הכללי המרחב יהיה מחולק ל  $C$  איזורים שהשפות שלהם יהיו מורכבות מהמישורים המתקבלים מהשפות שבין כל זוג מחלקות. דוגמא למקרה עם 3 מחלקות תופיע בתרגול.