

תרגול 8 - שיערוך פילוג בשיטות פרמטריות וסיווג גנרטיבי

הבעיה בגישה הלא פרמטרית

• **Curse of dimensionality**:

○ שיטות לא פרמטריות לומדות את הפילוג בכל איזור על פי הדגימות שנמצאות באותו איזור באופן בלתי תלוי באיזורים האחרים במרחב.

○ לכן, נדרש מספר רב של דוגמאות לכסות את מרחב הדגימות האפשריות.

○ הגודל האפקטיבי של מרחב הדגימות גדל מעריכית עם המימד של הדגימות (האורך של הוקטור x)

• המודלים המתקבלים אינם פונקציות שנוח לעבוד איתן.

○ לדוגמא: חישוב הצפיפות בנקודה מסוימת ב- KDE נעשה באמצעות סכימה על כל הנקודות שנמצאות ב

הגישה הפרמטרית

- נעשה שימוש במודלים פרמטרים בדומה לאופן שבו הדבר נעשה בגישה הדיסקרימינטיבית:
- נחפש פילוג בתוך משפחה פרמטרית מסוימת, על ידי מציאת הפרמטרים האופטימליים.
 - לרוב ננסה למדל את צפיפות הפילוג (ה PDF).
- באופן כללי ישנן 2 דרכים להתייחס לפרמטרי המודל: הגישה הבייסיאנית, והגישה התדירותית (לא בייסיאנית).
 - בגישה הבייסיאנית אנו מתייחסים לפרמטרים כאל משתנים אקראיים.
 - בגישה התדירותית אנו מתייחסים לפרמטרים כקבועים.

הגישה הלא-בייסיאנית (קלאסית או תדירותית) (**Frequentist**)

- בגישה זו אנו נתייחס לפרמטרים באופן דומה לשיטות הדיסקרימינטיביות.
- תחת גישה זו אין כל העדפה של ערך מסויים של הפרמטרים על פני ערך אחר. את המודל הפרמטרי להסתברות / צפיפות הסתברות של משתנה אקראי x נסמן ב:

$$p_x(x; \theta)$$

משערך Maximum Likelihood Estimator (MLE)

- הדרך הנפוצה ביותר לבחור את הערך של θ תחת הגישה הלא בייסאנית היא בעזרת MLE.
- נחפש את הערך של θ אשר מסביר בצורה הכי טובה את המדגם הנתון.
- נסמן ב $p_{\mathcal{D}}(\mathcal{D}; \theta)$ את ההסתברות לקבלת מדגם $\mathcal{D} = \{x^{(i)}\}$.
- גודל זה מכונה **הסבירות (likelihood)** של המדגם כפונקציה של θ .

- כדי להדגיש שהמדגם הוא "גודל" ידוע ואילו הגודל הלא ידוע שאותו נרצה לבדוק הינו θ , מקובל לסמן את פונקציית ה likelihood באופן הבא:

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

- משערך ה MLE של θ הוא הערך שממקסם את ה-likelihood:

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

- כאשר הדגימות במדגם הן i.i.d נוכל להסיק כי:

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

ולכן:

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

• נוכל להחליף את המכפלה על כל הדגימות בסכום
:(Maximum Log-Likelihood Estimator)

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

הגישה הבייסיאנית

- וקטור הפרמטרים θ הינו ריאליזציה של וקטור אקראי בעל פילוג כלשהוא $p_\theta(\theta)$.
- פילוג זה מכונה **הא-פריורי (a priori distribution)**, הפילוג של θ לפני שראינו את המדגם.
- תחת גישה זו, המודל שלנו יהיה הפילוג של x **בהינתן** θ :

$$p_{x|\theta}(x|\theta)$$

Maximum

A-posteriori

משערך

(Probability (MAP

- בגישה זו נבחר את הערך של θ ע"י משערך MAP.

- בשיטה זו נחפש את הערך הכי סביר של θ בהינתן המדגם
 $p_{\theta|\mathcal{D}}(\theta|\mathcal{D})$.

- פילוג זה מכונה הפילוג **א-פוסטריורי (a posteriori distribution)** (או הפילוג בדיעבד) - הפילוג אחרי שראינו את המדגם.

אם כן, משערך ה MAP הוא וקטור הפרמטרים אשר ממקסמים את ההסתברות ה א-פוסטרירית:

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

על פי חוק בייס, נוכל לכתוב זאת כ:

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

כאשר הדגימות במדגם **בהינתן** θ הן i.i.d מתקיים כי:

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

ולכן:

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

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

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

(Linear Discriminant Analysis (LDA

LDA הינו אלגוריתם לפתרון בעיות סיווג בגישה גנרטיבית פרמטרית (לא בייסיאנית).
המודל הפרמטרי:

1. את הפילוג של $p_y(y)$ נשערך ישירות מתוך התווית (זה פילוג דיסקרטי).

2. את הפילוג של $p_{x|y}(x|y)$ נמדל כפילוג נורמאלי.

3. אנו נניח כי מטריצת ה covariance של הפילוג הנורמאלי אינה תלויה בערך של y .

- נסמן את מטריצת הקווריאנס של הפילוגים הנורמאליים (אותה נרצה לשערך) ב- Σ .

- בנוסף, בעבור כל מחלקה c של y נסמן:

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

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

- μ_c - וקטורי התוחלת של הפילוג הנורמאלי $p_{\mathbf{x}|y}(\mathbf{x}|c)$.

שיערוך של הפרמטריים בעזרת משעריך MLE נותן את הפתרון הבא:

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

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

הפרדה לינארית

בעבור המקרה של סיווג בינארי (סיווג לשתי מחלקות) ופונקציית מחיר misclassification rate מתקבל החזאי הבא:

$$h(x) = \begin{cases} 1 & \mathbf{a}^T \mathbf{x} + b > 0 \\ 0 & \text{otherwise} \end{cases}$$

כאשר:

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

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

• נשים לב כי תנאי ההחלטה שבין שני התחומים הינו לינארי, ומכאן מקבל האלגוריתם את שמו.

תרגיל 8.1 - שיערוך MLE

נתון מדגם $D = \{x^{(i)}\}_{i=1}^N$ של דגימות בלתי תלויות של משתנה אקראי x . מצאו את משערוך ה MLE של המודלים הבאים:

- (1)** פילוג נורמלי: $x \sim N(\mu, \sigma^2)$ עם פרמטרים μ ו σ^2 לא ידועים.
- (2)** פילוג אחיד: $x \sim U[0, \theta]$, עם פרמטר θ לא ידוע.
- (3)** פילוג אקספוננציאלי (**לקריאה עצמית**): $x \sim \exp(\theta)$. עם פרמטר θ לא ידוע.

פיתרון 8.1

(1)

המודל של פונקציית ה PDF יהיה:

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

נסמן את וקטור הפרמטרים: $\theta = [\mu, \sigma^2]^T$. המשערך הינו:

$$\begin{aligned}\hat{\theta}_{\text{MLE}} &= \arg \min_{\theta} - \sum_{i=1}^N \log\left(p(x^{(i)}; \theta)\right) \\ &= \arg \min_{\theta} - \sum_{i=1}^N \log\left(\frac{1}{\sqrt{2\pi\theta_2}} \exp\left(-\frac{1}{2\theta_2} \left(x^{(i)} - \theta_1\right)^2\right)\right) \\ &= \arg \min_{\theta} \frac{N}{2} \log(2\pi\theta_2) + \sum_{i=1}^N \frac{1}{2\theta_2} \left(x^{(i)} - \theta_1\right)^2\end{aligned}$$

נפתור על ידי גזירה והשוואה ל 0 (נסמן ב $f(\theta)$ את פונקציית המטרה אותה יש למזער):

$$\begin{aligned} & \begin{cases} \frac{\partial}{\partial \theta_1} f(\theta) = 0 \\ \frac{\partial}{\partial \theta_2} f(\theta) = 0 \end{cases} \\ \Leftrightarrow & \begin{cases} \sum_{i=1}^N \frac{1}{\theta_2} (x^{(i)} - \theta_1) = 0 \\ \frac{N}{2\theta_2} - \sum_{i=1}^N \frac{1}{2\theta_2^2} (x^{(i)} - \theta_1)^2 = 0 \end{cases} \\ \Leftrightarrow & \begin{cases} \theta_1 = \frac{1}{N} \sum_{i=1}^N x^{(i)} \\ \theta_2 = \frac{1}{N} \sum_{i=1}^N (x^{(i)} - \theta_1)^2 \end{cases} \end{aligned}$$

מכאן ש:

$$\hat{\mu}_{\text{MLE}} = \hat{\theta}_1 = \frac{1}{N} \sum_{i=1}^N x^{(i)}$$

$$\hat{\sigma}_{\text{MLE}}^2 = \hat{\theta}_2 = \frac{1}{N} \sum_{i=1}^N \left(x^{(i)} - \hat{\mu}_{\text{MLE}} \right)^2$$

המודל של פונקציית ה PDF יהיה:

$$p(x; \theta) = \begin{cases} \frac{1}{\theta} & \theta \geq x_i \geq 0 \\ 0 & \text{else} \end{cases}$$

ולכן:

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} \prod_{i=1}^N p(x^{(i)}; \theta) = \begin{cases} \frac{1}{\theta^N} & \theta \geq x^{(i)} \quad \forall i \\ 0 & \text{else} \end{cases}$$

התנאי $\theta \geq x^{(i)}$ לכל i שקול ל $\theta \geq \max_i \{x^{(i)}\}$. מצד אחד נרצה לקיים תנאי זה בכדי שה likelihood לא יתאפס, מצד שני נרצה ש θ יהיה כמה שיותר קטן בכדי למקסם את $1/\theta^N$. לכן,

$$\hat{\theta}_{\text{MLE}} = \max_i \{x^{(i)}\}$$

המודל של פונקציית ה PDF יהיה:

$$p(x; \theta) = \theta \exp(-\theta x)$$

משערך ה MLE נתון על ידי:

$$\begin{aligned} \hat{\theta}_{\text{MLE}} &= \arg \min_{\theta} - \sum_{i=1}^N \log \left(p(x^{(i)}; \theta) \right) \\ &= \arg \min_{\theta} - N \log(\theta) + \theta \sum_{i=1}^N x^{(i)} \end{aligned}$$

נפתור על ידי גזירה והשוואה ל 0 (נסמן ב $f(\theta)$ את פונקציית המטרה אותה יש למזער):

$$\begin{aligned}\frac{\partial}{\partial \theta} f(\theta) &= 0 \\ \Leftrightarrow -\frac{N}{\theta} + \sum_{i=1}^N x^{(i)} &= 0 \\ \Leftrightarrow \theta &= \frac{1}{\frac{1}{N} \sum_{i=1}^N x^{(i)}}\end{aligned}$$

מכאן ש:

$$\hat{\theta}_{\text{MLE}} = \frac{1}{\frac{1}{N} \sum_{i=1}^N x^{(i)}}$$

תרגיל 8.2 - MAP

ביום טוב, עומרי כספי קולע בהסתברות p מהקו. ביום רע, הוא קולע בהסתברות q מהקו. α מהימיהם הם ימים טובים עבור עומרי.

ביום מסויים זרק עומרי N זריקות וקלע m מתוכם. מאמנו של עומרי צריך לזהות האם מדובר ביום טוב או רע של השחקן (ולהשאיר אותו או להחליף אותו בהתאמה).

מהו חוק ההחלטה אשר ממקסם את סיכויי המאמן לצדוק?

- הניחו כי בהינתן המידע של האם יום מסויים הוא טוב או לא, ההסתברות לקלוע זריקות שונות הינה הסתברות בלתי תלויה.

פתרון 8.2

- $x^{(i)}$ - משתנה אקראי בינארי של האם עומרי קלע בזריקה ה- i . (0-1 קלע)
- y - משתנה אקראי בינארי של האם היום הינו יום טוב. (0-1 יום לא טוב, 1-יום טוב).

על פי הנתונים בשאלה:

$$p_{x|y}(x|0) = \begin{cases} 1 - q & x = 0 \\ q & x = 1 \end{cases}$$

$$p_{x|y}(x|1) = \begin{cases} 1 - p & x = 0 \\ p & x = 1 \end{cases}$$

- בכדי למקסם את הסיכוי לחזות האם היום הוא יום טוב בהינתן המדגם נרצה למצוא איזה ערך יותר סביר בהינתן המדגם (יום טוב או רע).

- במילים אחרות אנו רוצים את ה y הכי סביר בהינתן $\mathcal{D} = \{x^{(i)}\}$

$$\hat{y} = \arg \max_y p_{y|\mathcal{D}}(y|\mathcal{D})$$

- זוהי למעשה בעיית MAP קלאסית, כאשר y משמש למעשה כפרמטר בפילוג של $x|y$.

- בכדי לשמור על אחידות עם הסימונים שהגדרנו קודם לבעיות שיערוך נסמן את y ב θ .

$$\hat{\theta} = \arg \max_{\theta} p_{\theta|\mathcal{D}}(\theta|\mathcal{D}) = \arg \max_{\theta} p_{\mathcal{D}|\theta}(\mathcal{D}|\theta)p_{\theta}(\theta) = \arg \max_{\theta} p_{\theta}(\theta) \prod_i p_{x^{(i)}|\theta}(x^{(i)}|\theta)$$

מכיוון ש θ יכול לקבל רק שני ערכים נוכל לבדוק את שניהם ולקבוע מי מהם סביר יותר.
 בעבור $\theta = 0$ נקבל:

$$p_{\theta}(0) \prod_i p_{x|\theta}(x^{(i)}|0) = (1 - \alpha)q^m (1 - q)^{N-m}$$

בעבור $\theta = 1$ נקבל:

$$p_{\theta}(1) \prod_i p_{x|\theta}(x^{(i)}|1) = \alpha p^m (1 - p)^{N-m}$$

לכן החיזוי האופטימאלי יהיה:

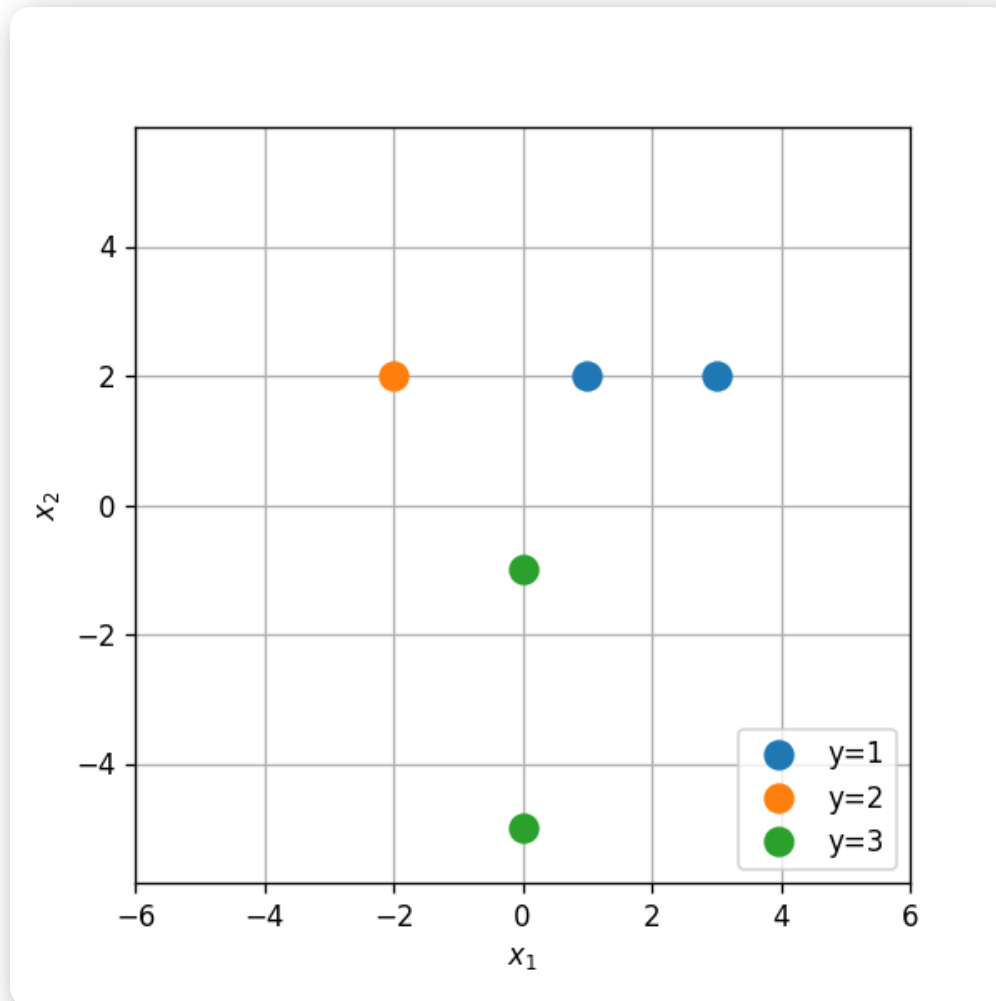
$$\hat{\theta} = \begin{cases} 0 & (1 - \alpha)q^m (1 - q)^{N-m} > \alpha p^m (1 - p)^{N-m} \\ 1 & \text{otherwise} \end{cases}$$

$$= \begin{cases} 0 & \frac{1-\alpha}{\alpha} \left(\frac{q}{p}\right)^m \left(\frac{1-q}{1-p}\right)^{N-m} > 1 \\ 1 & \text{otherwise} \end{cases}$$

- בסוואנה חיים שלושה זני פילים אשר נמצאים בסכנת הכחדה. ידוע כי כל אחד משלושת הזנים ניזון מצמחיה מעט שונה, ועל מנת לשמר את אוכלוסיית הפילים מעוניינים לפזר להם אוכל ברחבי הסוואנה.
- בכדי למקסם את האפקטיביות של פעולה זו מעוניינים לשערך בכל נקודת חלוקה מהו הזן שהכי סביר להמצא באותה נקודה על מנת להתאים את סוג המזון לזן זה.

-
- הפילוג של זני הפילים על פני הסוואנה אינו ידוע אך נתונות לנו התצפית הבאה של הקואורדינטות בהן נצפו הפילים:

Type	x_1	x_2
1	1	2
1	3	2
2	-2	2
3	0	-1
3	0	-5



השתמשו במסוג LDA על מנת לבנות חזאי אשר ישערך את הזן הנפוץ ביותר בכל קואורדינטה.

פתרון 8.3

נחשב את הפרמטרים של המודל הפרמטרי של LDA.
נסמן ב \mathcal{I}_c את אוסף כל התצפיות שבהם הזן הוא c :

$$\mathcal{I}_1 = \{1, 2\}$$

$$\mathcal{I}_2 = \{3\}$$

$$\mathcal{I}_3 = \{4, 5\}$$

נשערך את $p_y(y)$:

$$p_y(y) = \begin{cases} \frac{|\mathcal{I}_1|}{N} = \frac{2}{5} & 1 \\ \frac{|\mathcal{I}_2|}{N} = \frac{1}{5} & 2 \\ \frac{|\mathcal{I}_3|}{N} = \frac{2}{5} & 3 \end{cases}$$

נחשב את התוחלות של כל אחת משלושת הפילוגים $p_{\mathbf{x}|y}(\mathbf{x}|c)$:

$$\boldsymbol{\mu}_1 = \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} \mathbf{x}^{(i)} = \frac{1}{2} \left(\begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 3 \\ 2 \end{pmatrix} \right) = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

$$\boldsymbol{\mu}_2 = \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \mathbf{x}^{(i)} = \begin{pmatrix} -2 \\ 2 \end{pmatrix}$$

$$\boldsymbol{\mu}_3 = \frac{1}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} \mathbf{x}^{(i)} = \frac{1}{2} \left(\begin{pmatrix} 0 \\ -1 \end{pmatrix} + \begin{pmatrix} 0 \\ -5 \end{pmatrix} \right) = \begin{pmatrix} 0 \\ -3 \end{pmatrix}$$

נחשב את מטריצת covariance המשותפת של הפילוגים:

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

דרך נוחה לחשב את הסכום בביטוי זה הינה באופן הבא: נגדיר את המטריצה של התצפיות לאחר חיסור של התוחלת המתאימה לכל זן:

$$\tilde{X} = \begin{pmatrix} -x_1 \\ -x_2 \\ -x_3 \\ -x_4 \\ -x_5 \end{pmatrix} - \begin{pmatrix} -\mu_{y_1} \\ -\mu_{y_2} \\ -\mu_{y_3} \\ -\mu_{y_4} \\ -\mu_{y_5} \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ -2 & 2 \\ 0 & -1 \\ 0 & -5 \end{pmatrix} - \begin{pmatrix} 2 & 2 \\ 2 & 2 \\ -2 & 2 \\ 0 & -3 \\ 0 & -3 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 2 \\ 0 & -2 \end{pmatrix}$$

ניתן להראות כי ניתן לכתוב את הסכום בביטוי ל Σ באופן הבא:

• נשתמש כעת בפילוגים שאותם שיערכנו על מנת לבנות את החזאי.

• האיזור שבו זן 1 הינו הזן הסביר ביותר הינו האיזור שבו מתקיים:

$$\begin{cases} p_{y|x}(1|x) > p_{y|x}(2|x) \\ p_{y|x}(1|x) > p_{y|x}(3|x) \end{cases}$$

נחשב את התנאי הראשון

$$\begin{aligned} p_{y|\mathbf{x}}(1|\mathbf{x}) &> p_{y|\mathbf{x}}(2|\mathbf{x}) \\ \Leftrightarrow p_{\mathbf{x}|y}(\mathbf{x}|1)p_y(1) &> p_{\mathbf{x}|y}(\mathbf{x}|2)p_y(2) \\ \Leftrightarrow \frac{1}{\sqrt{4\pi^2|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)} p_y(1) &> \frac{1}{\sqrt{4\pi^2|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)} p_y(2) \\ \Leftrightarrow -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \log(p_y(1)) &> -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) + \log(p_y(2)) \\ \Leftrightarrow \mathbf{x}^T \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2}(\boldsymbol{\mu}_2^T \Sigma^{-1} \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1) + \log\left(\frac{p_y(1)}{p_y(2)}\right) &> 0 \end{aligned}$$

זוהי למעשה הפרדה לשני תחומים על ידי הקו הבא:

$$\mathbf{a}^T \mathbf{x} + b = 0$$

כאשר:

$$\mathbf{a} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \begin{pmatrix} 10 \\ 0 \end{pmatrix}$$

$$b = \frac{1}{2}(\boldsymbol{\mu}_2^T \Sigma^{-1} \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1) + \log\left(\frac{p_y(1)}{p_y(2)}\right) = \log(2)$$

- זוהי כמובן התוצאה עבור מסוג LDA בינארי בין שני הזנים של $y = 1$ ו $y = 2$.

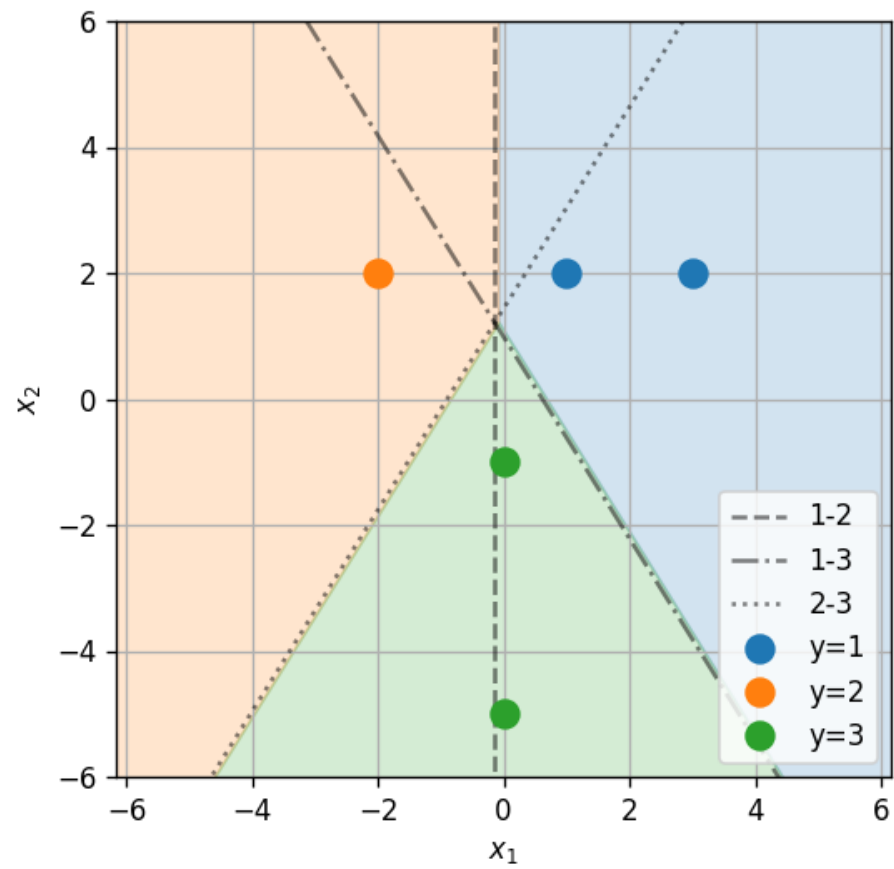
- מכאן שהקו המפריד בין זן 1 ל זן 2 נתון על ידי:

$$1 - 2 : \quad 10x_1 + \log(2) = 0$$

- באופן דומה ניתן לחשב גם את שני קווי ההפרדה האחרים (בין 1 ל 3 ובין 2 ל 3):

$$1 - 3 : \quad 5x_1 + \frac{25}{8}x_2 + \frac{55}{16} = 0$$

$$2 - 3 : \quad -5x_1 + \frac{25}{8}x_2 + \frac{55}{16} - \log(2) = 0$$



תרגיל מעשי - שיערוך הפילוג של זמני נסיעה בניו יורק

Code

נחזור לבעיה מהתרגול הקודם של שיערוך הפילוג של זמן הנסיעה של מונית מתוך המדגם הבא:

id	day of week	duration	dropoff northing	dropoff easting	pickup northing	pickup easting	tip amount	fare amount	payment type	trip distance	passenger count	passenger count
3	3	11.5167	4515.18	588.155	4512.98	586.997	0	9.5	2	2.76806	2	0
6	6	12.6667	4512.63	584.85	4512.92	587.152	0	10	2	3.21868	1	1
1	0	5.51667	4513.17	585.434	4513.36	587.005	2.49	7	1	2.57494	1	2
5	1	9.88333	4512.55	586.672	4511.73	586.649	1.65	7.5	1	0.965604	1	3
5	2	8.68333	4511.76	585.262	4511.89	586.967	1.66	7.5	1	2.46229	1	4
0	3	9.43333	4511.54	585.169	4512.88	585.926	2.2	7.5	1	1.56106	5	5
8	5	7.95	4514.21	588.71	4515.08	586.731	1	8	1	2.57494	1	6
9	5	4.95	4509.55	585.844	4509.71	585.345	0	5	2	0.80467	1	7
8	5	11.0667	4507.74	583.671	4509.48	585.422	1.1	10	1	3.6532	1	8

ניסיון 1: פילוג גאوسي

- נשתמש במודל של פילוג נורמלי לתיאור הפילוג של משך הנסיעה. למודל זה שני פרמטרים, התוחלת μ והשונות σ .

סימונים והנחות:

- N - מספר הדגמים במדגם.

- $\theta = [\mu, \sigma]^T$ - וקטור הפרמטרים של המודל

- המודל - $p_{\text{normal}}(x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x^{(i)} - \mu)^2}{2\sigma^2}\right), i = 1, \dots, N$

ראינו כי בעבור המודל הנורמלי, ניתן למצוא את הפרמטרים של משערך ה-MLE באופן מפורש (אנליטית), והפתרון נתון על ידי:

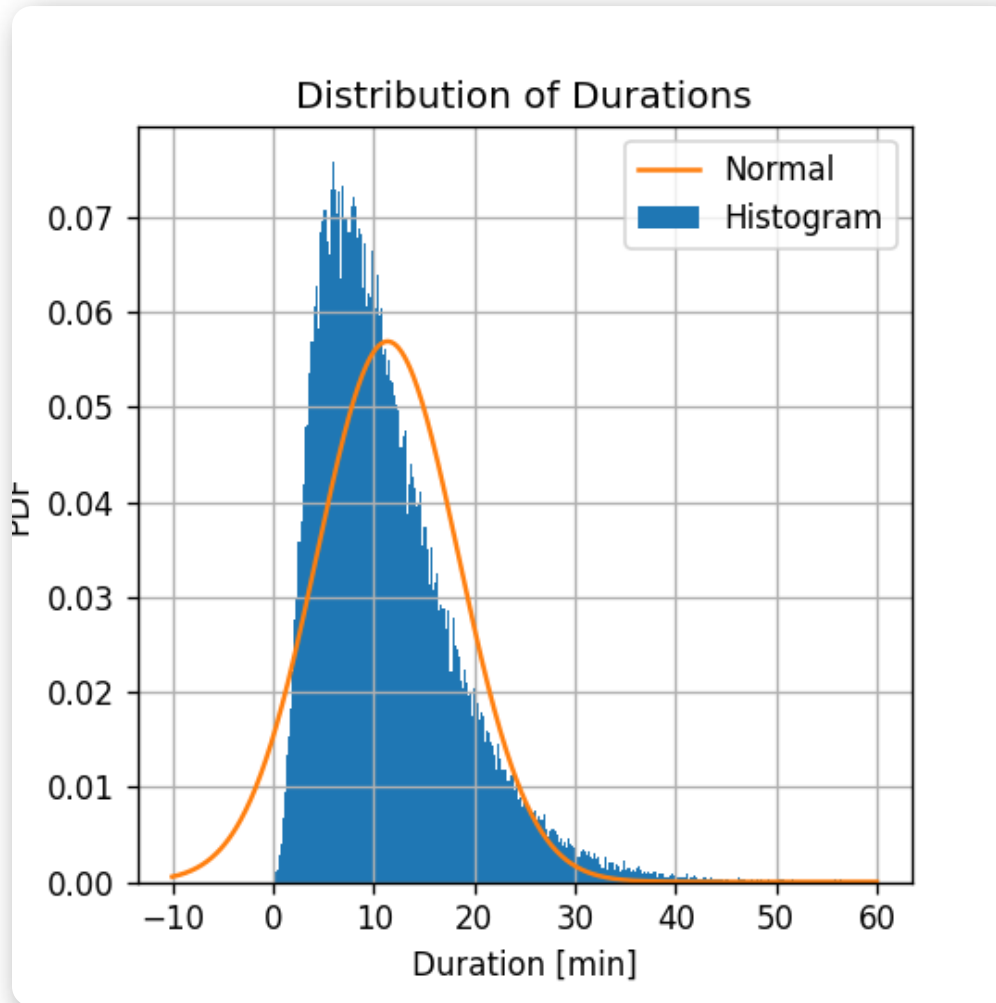
$$\mu = \frac{1}{N} \sum_i x^{(i)}$$
$$\sigma = \sqrt{\frac{1}{N} \sum_i \left(x^{(i)} - \mu\right)^2}$$

בעבור המדגם הנתון נקבל:

$$\hat{\mu} = 11.4 \text{ min}$$

$$\hat{\sigma} = 7.0 \text{ min}$$

ההיסטוגרמה של של משכי הנסיעה יחד עם הפילוג המשוער:



- הפילוג הנורמלי נותן קירוב מאד גס לפילוג האמיתי.

- עובדה אחת שמאד מטרידה לגבי הפילוג שקיבלנו הינה שישנו סיכוי לא אפסי לקבל נסיעות עם משך נסיעה שלילי.

נסיון 2: פילוג Rayleigh

- פילוג Rayleigh מתאר את הפילוג של האורך האוקלידי (l_2 norm) של וקטור גאוסי דו מימדי עם תוחלת 0 וחוסר קורלציה ופילוג זה לשני רכיבי הוקטור.
- במלים אחרות, עבור וקטור בעל הפילוג הבא:

$$\mathbf{Z} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} \right)$$

- פילוג Rayleigh מתאר את הגודל $\|\mathbf{Z}\|_2 = \sqrt{Z_x^2 + Z_y^2}$.

פונקציית צפיפות ההסתברות של פילוג Rayleigh נתונה על ידי:

פונקציית צפיפות ההסתברות של פילוג Rayleigh נתונה על ידי:

$$p_{\text{Rayleigh}}(z; \sigma) = \frac{z}{\sigma^2} \exp\left(-\frac{z^2}{2\sigma^2}\right), \quad z \geq 0$$

• שימו לב: הפילוג מוגדר רק בעבור ערכים חיוביים.

• לפילוג זה פרמטר יחיד σ שנקרא פרמטר סקאלה (scale parameter). בניגוד לפילוג הנורמלי, פה σ אינה שווה לסטיית התקן של הפילוג.

מוטיבציה לשימוש בפילוג Rayleigh

- נניח שוקטור המחבר את נקודת תחילת הנסיעה עם נקודת סיום הנסיעה הינו וקטור דו מימדי אשר מפולג נרמלית, ולשם הפשטות נניח כי רכיביו מפולגים עם פילוג זהה וחסר קורלציה.
- נניח כי המונית נוסעת בקירוב בקו ישר בין נקודת ההתחלה והסיום. לכן, המרחק אותו נוסעת המכונית יהיה מפולג על פי פילוג Reyleigh.
- נניח בנוסף כי מהירות הנסיעה קבועה ולכן משך הנסיעה פורפורציוני למרחק ולכן גם הוא יהיה מפולג על פי פילוג Reyleigh.

לשם השלמות נסמן את וקטור הפרמטרים ב: $\theta = [\sigma]$
במקרה זה המודל נתון על ידי:

$$p_{\text{rayleigh}}(\mathbf{x}; \theta) = \prod_{i=1}^N \frac{x^{(i)}}{\theta^2} \exp\left(-\frac{(x^{(i)})^2}{2\theta^2}\right)$$

ופונקציית ה **log likelihood** תהיה:

$$\begin{aligned} l_{\text{rayleigh}}(\theta) &= \sum_i \log\left(p_{\text{rayleigh}}\left(x^{(i)}; \theta\right)\right) \\ &= \sum_i \log\left(x^{(i)}\right) - 2N \log(\theta) - \frac{1}{2\theta^2} \sum_i (x^{(i)})^2 \end{aligned}$$

בעיית האופטימיזציה הינה:

$$\hat{\theta} = \arg \min_{\theta} - \sum_i \log(x^{(i)}) + 2N \log(\theta) + \frac{1}{2\theta^2} \sum_i (x^{(i)})^2$$

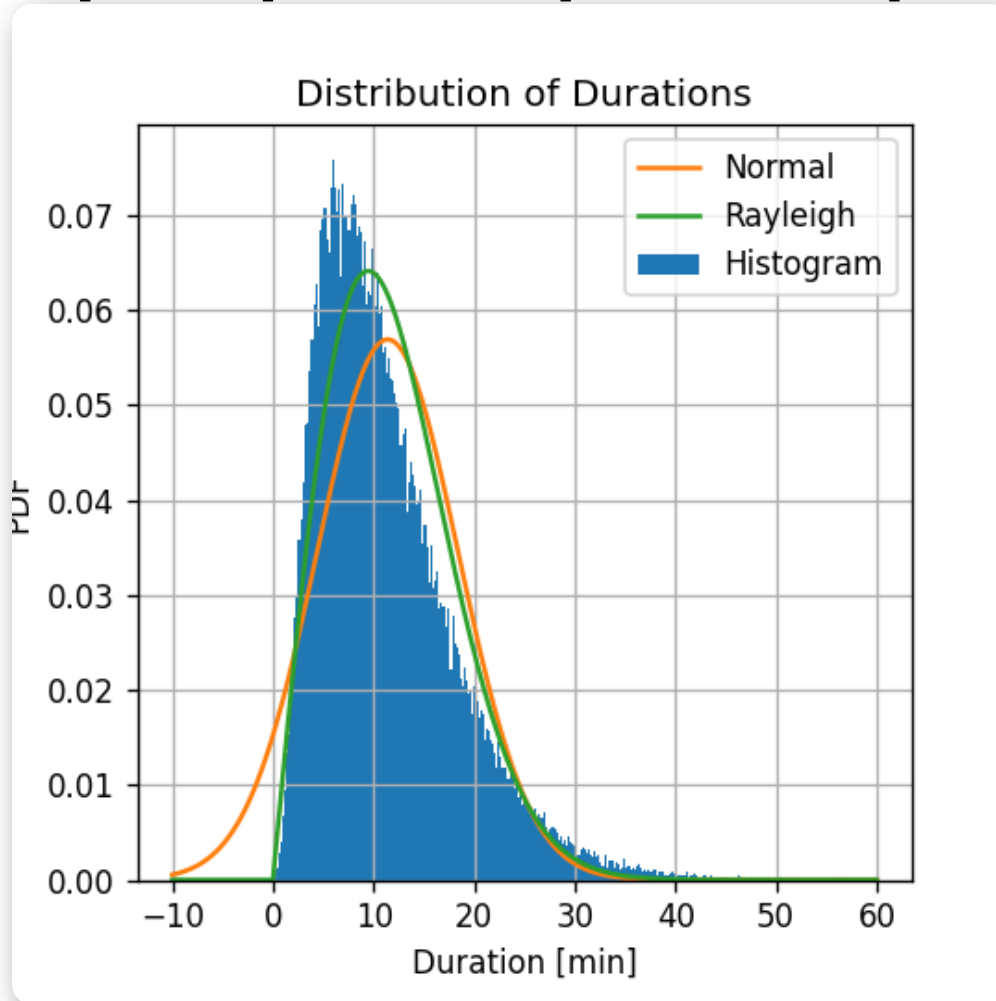
גם בעבור המקרה הזה נוכל לפתור את בעיית האופטימיזציה באופן אנליטי על ידי גזירה והשוואה לאפס:

$$\begin{aligned} \frac{\partial l_{\text{rayleigh}}(\theta)}{\partial \theta} &= 0 \\ \Leftrightarrow -\frac{2N}{\theta} + \frac{\sum_i (x^{(i)})^2}{\theta^3} &= 0 \\ \Leftrightarrow \hat{\sigma} = \theta &= \sqrt{\frac{1}{2N} \sum_i (x^{(i)})^2} \end{aligned}$$

בעבור המדגם הנתון נקבל:

$$\hat{\sigma} = 9.5$$

נוסיף את השיערוך החדש שקיבלנו לגרף ממקודם:



- המודל של פילוג Rayleigh טוב יותר מהמודל הנורמלי.
- אין הסתברות שונה מ0 לקבל משך נסיעה שלילי.

נסיון 3: Generalized Gamma Distribution

- פילוג Rayleigh הינו מקרה פרטי של משפחה כללית יותר של פונקציות פילוג המכונה Generalized Gamma Distribution.

- פונקציית צפיפות ההסתברות של משפחה זו נתונה על ידי:

$$p_{\text{gengamma}}(z; \sigma, a, c) = \frac{cz^{ca-1} \exp(-(z/\sigma)^c)}{\sigma^{ca-1} \Gamma(a)}, \quad z \geq 0$$

- כש- Γ היא פונקציה המוכנה פונקציית גמא (gamma function)

- למודל זה 3 פרמטרים: $\theta = [\sigma, a, c]^T$.

- בעבור $c = 2$ ו $a = 1$ נקבל את פילוג Rayleigh כאשר

- $\sigma_{\text{gamma}} = 2\sigma_{\text{rayleigh}}$

- בשונה מהמקרים של פילוג נורמלי ופילוג Rayleigh, לא נוכל למצוא בקלות את הפרמטרים האופטימאליים של המשערך באופן אנליטי.

- לכן, לשם מציאת הפרמטרים נאלץ להעזר בפתרון נומרי.

- נעשה שימוש באחת החבילה של Python הנקראת SciPy. חבילה זו מכילה מודלים הסברותיים רבים ומכילה מספר רב של כלים הקשורים למודלים אלו, כגון מציאת הפרמטרים האופטימאליים בשיטת MLE על סמך מדגם נתון. את הפונקציות הקשורות למודל Generalized Gamma Distribution ניתן למצוא כאן. אתם תעשו שימוש בפונקציות אלו בתרגיל הבית הרטוב.

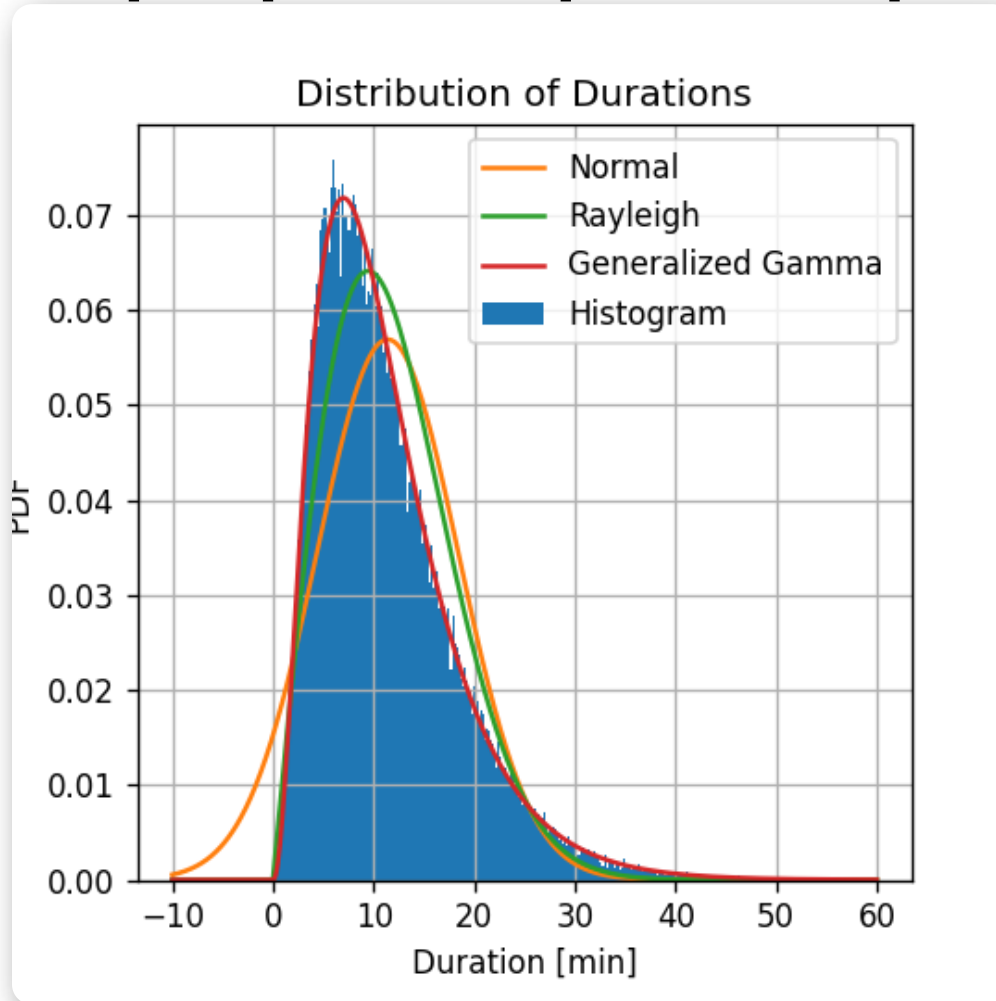
- שימוש בפונקציה הנ"ל, מניב את התוצאות הבאות:

$$\hat{a} = 4.4$$

$$\hat{c} = 0.8$$

$$\hat{\sigma} = 1.6$$

נוסיף את השיערוך החדש שקיבלנו לגרף הקודם:



- המודל של Generalized Gamma Distribution אכן מניב תוצאה אשר דומה מאד לצורת ההיסטוגרמה.