Machine learning- Ex2, Osher Elhadad 318969748

1. (20 pts) Gaussian noise as a regularizer

For a set of samples $(x_1, y_1) \dots, (x_n, y_n)$ with $x_i \in R$ and $y_i \in R$, we defined in class the problem of polynomial curve fitting as an error-minimization problem. Assume that the labels $\{y_i\}_{i=1}^N$ are i.i.d. and are drawn from a Gaussian distribution with a mean that is the prediction of the model $h_{\mathbf{w}}(x)$, and a standard deviation σ. Formally, $y_i \sim \mathcal{N}(h_{\mathbf{w}}(x_i), \sigma)$.

Show that maximizing the log of the likelihood $P(y_1, \dots y_n|\mathbf{w})$ is equivalent to minimizing the squared error $Err(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n (h_{\mathbf{w}}(x_i) - y_i)^2$ used in class. Recall that the likelihood of a dataset $S_n = \{x_1, \dots, x_n\}$ is defined as $\mathcal{L} = P(S|\mathbf{w})$.

Hint: No need to compute any derivatives.

.1

בהוכחה זו נראה שהw שמביא למקסימום של הlog של הlikelihood $\log(P(y_1, \dots, y_n|w))$ מקביל לw שמביא למינימום של $Err(w) = \frac{1}{n}\sum_{i=1}^n (h_w(x_i) - y_i)^2$.

$$\log\left(P(y_1, \dots, y_n|w)\right)$$

נשים לב כי מכיוון שנתון כי $\{y_i\}_{i=1}^N$ הינם בלתי תלויים לכן $P(y_1, \dots, y_n|w) = \prod_{i=1}^n P(y_i|w)$ ומכיוון שנתון כי $y_i \sim N(h_w(x_i), \sigma)$ לכן למעשה נקבל

$$\log\left(P(y_1, \dots, y_n|w)\right) = \log\left(\prod_{i=1}^n P(y_i|w)\right) = \sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{y_i - h_w(x_i)}{\sqrt{\sigma}}\right)^2}\right)$$
$$= \sum_{i=1}^n \log\left(e^{-\frac{1}{2}\left(\frac{y_i - h_w(x_i)}{\sqrt{\sigma}}\right)^2}\right) - \log\left(\sqrt{2\pi\sigma}\right)$$
$$= \log(e) \sum_{i=1}^n \left(-\frac{1}{2}\left(\frac{y_i - h_w(x_i)}{\sqrt{\sigma}}\right)^2\right) - n\log\left(\sqrt{2\pi\sigma}\right)$$
$$= -n\log\left(\sqrt{2\pi\sigma}\right) + \frac{1}{2\sigma}\log(e) \cdot \left(-\sum_{i=1}^n (h_w(x_i) - y_i)^2\right)$$

כעת נשים לב כי על מנת למצוא את המקסימום של הביטוי שקיבלנו אזי עלינו למצוא את הw שממקסם את הביטוי ולכן מכיוון ש$-n\log\left(\sqrt{2\pi\sigma}\right)$ הינו קבוע שנוסף לפונקציה ואינו תלוי בw לכן נתעלם ממנו במציאת הw שיתן מקסימום, כמו כן נתעלם מהקבוע $\frac{1}{2\sigma}\log(e)$ זאת מכיוון שהוא קבוע שכופל את החלק המשתנה של הפונקציה $\left(-\sum_{i=1}^n (h_w(x_i) - y_i)^2\right)$.

וכעת נשים לב כי בפונקציה $Err(w)$ גם נתעלם מ$\frac{1}{n}$ במציאת המינימום מכיוון שהוא קבוע שכופל את החלק המשתנה של הפונקציה $\sum_{i=1}^n (h_w(x_i) - y_i)^2$.

ולכן סה"כ קיבלנו כי הw שממקסם את הביטוי $\left(-\sum_{i=1}^n (h_w(x_i) - y_i)^2\right)$ הינו הw שמביא למינימום את הביטוי $\sum_{i=1}^n (h_w(x_i) - y_i)^2$ זאת מכיוון שהם אותם ביטויים רק אחד הוא כפול מינוס 1, כלומר מחליף את המקסימום למינימום. ולכן סה"כ הוכחנו כי הw שמביא למקסימום של הlog של הlikelihood $\log(P(y_1, \dots, y_n|w))$ מקביל לw שמביא למינימום של $Err(w) = \frac{1}{n}\sum_{i=1}^n (h_w(x_i) - y_i)^2$ כדרוש.

2. (12 pts) Bayesian decision theory

Show that for the squared cost function $\lambda(\hat{\theta}|\theta) = (\hat{\theta} - \theta)^2$, the optimal Bayes estimator is the conditional expectation.

$$\hat{\theta}_{SE} = E\left(\theta|S_n\right) = \int_{-\infty}^{\infty} \theta P(\theta|S_n) d\theta \quad .$$

Hint: The derivative operator is a linear operator, so it can switch order with the integral (you can use the Leibniz integral rule).

2. נוכיח כי עבור פונקציית ה-cost $\lambda(\hat{\theta}|\theta) = (\hat{\theta} - \theta)^2$ נקבל שהBayes estimator האופטימלי הוא בעצם הconditional expectation. נמצא נקודת קיצון מסוג מינימום ו ע״י גזירה (לפי $\hat{\theta}$) והשוואה ל0 של הנגזרת. וכך למעשה נוכל לקבל את הBayes estimator המינימלי (האופטימלי). ונוכיח כי נקבל $\hat{\theta}$ כך שיביא למינימום של הBayes estimator והוא יהיה למעשה הconditional expectation. במעבר הראשון נשתמש בכלל האינטגרל של לייבניץ.

$$\frac{d}{d\hat{\theta}} \int_{-\infty}^{\infty} (\hat{\theta} - \theta)^2 P(\theta|S_n) d\theta = \int_{-\infty}^{\infty} \frac{d}{d\hat{\theta}} (\hat{\theta} - \theta)^2 P(\theta|S_n) d\theta = \int_{-\infty}^{\infty} \frac{d}{d\hat{\theta}} (\hat{\theta}^2 - 2\theta\hat{\theta} + \theta^2) P(\theta|S_n) d\theta$$

$$= \int_{-\infty}^{\infty} \frac{d}{d\hat{\theta}} \hat{\theta}^2 P(\theta|S_n) - 2\theta\hat{\theta} P(\theta|S_n) + \theta^2 P(\theta|S_n) d\theta = \int_{-\infty}^{\infty} 2\hat{\theta} P(\theta|S_n) - 2\theta P(\theta|S_n) d\theta$$

$$\Downarrow$$

$$\int_{-\infty}^{\infty} 2\hat{\theta} P(\theta|S_n) - 2\theta P(\theta|S_n) d\theta = 0$$

$$\int_{-\infty}^{\infty} 2\hat{\theta} P(\theta|S_n) d\theta = \int_{-\infty}^{\infty} 2\theta P(\theta|S_n) d\theta$$

$$\int_{-\infty}^{\infty} \hat{\theta} P(\theta|S_n) d\theta = \int_{-\infty}^{\infty} \theta P(\theta|S_n) d\theta$$

$$\hat{\theta} \int_{-\infty}^{\infty} P(\theta|S_n) d\theta = \int_{-\infty}^{\infty} \theta P(\theta|S_n) d\theta$$

נשים לב ש-1 $\int_{-\infty}^{\infty} P(\theta|S_n) d\theta = 1$ מאחר ומדובר באינטגרל של פונקציית הצפיפות ממינוס אינסוף ועד אינסוף ולפי הגדרה אינטגרל זה שווה ל 1. ונקבל

$$\hat{\theta} = \int_{-\infty}^{\infty} \theta P(\theta|S_n) d\theta$$

כעת נגזור שוב ונבדוק האם הנגזרת השנייה חיובית וכך נוכיח למעשה שזו אכן $\hat{\theta}$ שתביא למינימום את ה conditional expectation.

$$\frac{d}{d\hat{\theta}}\int_{-\infty}^{\infty} 2\hat{\theta}P(\theta|S_n) - 2\theta P(\theta|S_n)d\theta = \int_{-\infty}^{\infty}\frac{d}{d\hat{\theta}}\big(\hat{\theta}-\theta\big)2P(\theta|S_n)d\theta = \int_{-\infty}^{\infty} 2P(\theta|S_n)d\theta = 2\int_{-\infty}^{\infty} P(\theta|S_n)d\theta$$

$$= 2 > 0$$

ואכן קיבלנו כי הנגזרת השנייה חיובית לכל $\hat{\theta}$ ובפרט עבור $\hat{\theta} = \int_{-\infty}^{\infty}\theta P(\theta|S_n)d\theta$ ונקבל $\hat{\theta}$ כך שיביא למינימום של הBayes estimator והוא למעשה הconditional expectation כדרוש.

4. (28 pts) Maximum likelihood estimation

Let $S_n = \{x_1, \dots, x_n\}$ be a random sample from the distributions below. Find the maximum likelihood estimate of the parameter θ.

(a) Poisson: $P(k|\theta) = \frac{\theta^k}{k!}\exp(-k)$ for $k = 0, 1, \dots$.

(b) Exponential: $P(x|\theta) = \frac{1}{\theta}\exp(-x/\theta)$ for $x \geq 0$.

(c) $P(x|\theta) = \frac{1}{2}\exp(-|x-\theta|)$.

(d) $P(x|\theta) = \theta x^{\theta-1}$.

4. a. נמצא את המקסימום של ההערכת ה$likelihood$ של הפרמטר θ עבור התפלגות פואסיונית של $S_n = \{x_1, \dots, x_n\}$.
כלומר נמצא את המקסימום של ההסתברות הבאה (נשים לב כי מכיוון שהדוגמאות הינם שרירותיות לכן גם בלתי תלויות ולכן נקבל כפל בין ההסתברויות שלהם):

$$P(S_n|\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!} e^{-x_i}$$

נוכל לבצע ln על שני הצדדים ומכיוון שln הינה מונוטונית עולה לכן הפרמטר θ שעבורו נקבל את המקסימום של הln של שני צידי המשוואה יתן גם את המקסימום של $P(k|\theta)$ אם נציבו.

$$\ln\big(P(S_n|\theta)\big) = \ln\left(\prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!} e^{-x_i}\right) = \sum_{i=1}^{n} \left(\ln(\theta^{x_i}) + \ln(e^{-x_i}) - \ln(x_i!)\right) = \sum_{i=1}^{n} \left(x_i \cdot ln(\theta) - x_i - \ln(x_i!)\right)$$

כעת נגזור את הפונקציה הזו על מנת למצוא מקסימום:

$$\frac{d}{d\theta}\left(\sum_{i=1}^{n} \left(x_i \cdot ln(\theta) - x_i - \ln(x_i!)\right)\right) = \sum_{i=1}^{n} x_i$$

ומכיוון שנתון כי בהתפלגות זו $x_i \geq 0$ לכל $1 \leq i \leq n$ לכן למעשה קיבלנו שהנגזרת הינה חיובית לכל θ ולכן נקבל כי הפונקציה $\sum_{i=1}^{n}\left(x_i \cdot ln(\theta) - x_i - \ln(x_i!)\right)$ הינה מונוטונית לא יורדת ולכן למעשה נקבל את המקסימום של פונקציה זו כאשר $\theta \to \infty$ וכפי שהסברנו למעלה נקבל גם את המקסימום של ה$likelihood$ שאותו חיפשנו. כלומר (נתון כי $x_i \geq 0$ בהתפלגות זו):

$$\max\big(P(S_n|\theta)\big) = \lim_{\theta\to\infty} P(S_n|\theta) = \lim_{\theta\to\infty} \prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!} e^{-x_i} = \infty$$

b. נמצא את המקסימום של ההערכת ה$likelihood$ של הפרמטר θ עבור התפלגות אקפוננציאלית של $S_n = \{x_1, \dots, x_n\}$.
כלומר נמצא את המקסימום של ההסתברות הבאה (נשים לב כי מכיוון שהדוגמאות הינם שרירותיות לכן גם בלתי תלויות ולכן נקבל כפל בין ההסתברויות שלהם):

$$P(S_n|\theta) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-\frac{x_i}{\theta}}$$

נוכל לבצע ln על שני הצדדים ומכיוון שln הינה מונוטונית עולה לכן הפרמטר θ שעבורו נקבל את המקסימום של הln של שני צידי המשוואה יתן גם את המקסימום של $P(S_n|\theta)$ אם נציבו.

$$\ln\big(P(S_n|\theta)\big) = \ln\left(\prod_{i=1}^{n} \frac{1}{\theta} e^{-\frac{x_i}{\theta}}\right) = \sum_{i=1}^{n} \left(\ln\left(e^{-\frac{x_i}{\theta}}\right) - \ln(\theta)\right) = \sum_{i=1}^{n} \left(-\frac{x_i}{\theta}\right) - n \cdot \ln(\theta)$$

כעת נגזור את הביטוי ונחפש את θ שיביא למקסימום של פונקציה זו.

$$\frac{d}{d\theta}\left(\sum_{i=1}^{n}\left(-\frac{x_i}{\theta}\right)-n\cdot\ln(\theta)\right)=\sum_{i=1}^{n}\left(\frac{x_i}{\theta^2}\right)-\frac{n}{\theta}$$

נשווה ל0 את הנגזרת הנ"ל:

$$\sum_{i=1}^{n}\left(\frac{x_i}{\theta^2}\right)-\frac{n}{\theta}=0\rightarrow\sum_{i=1}^{n}(x_i)-n\cdot\theta=0\rightarrow\theta=\sum_{i=1}^{n}\frac{x_i}{n}$$

כעת נשים לב להתנהגות הפונקציה $\prod_{i=1}^{n}\frac{1}{\theta}e^{-\frac{x_i}{\theta}}$ וכי עבור θ מאוד גדולים (גדולים מ$\sum_{i=1}^{n}\frac{x_i}{n}$) אזי נקבל כי הביטוי $\frac{1}{\theta}e^{-\frac{x_i}{\theta}}$ ילך ויקטן ולכן הפונקציה תרד עבור θ מאוד גדולים כאשר הם הולכים וגדלים. כמו כן עבור ערכי θ קטנים (קטנים מ$\sum_{i=1}^{n}\frac{x_i}{n}$ וגדולים מ0) אשר גדלים לאט ניתן לשים לב כי הביטוי $\frac{1}{\theta}e^{-\frac{x_i}{\theta}}$ ילך ויגדל ולכן הפונקציה תעלה עבור θ מאוד קטנים כאשר הם הולכים וגדלים.

וסה"כ קיבלנו כי לפונקציה זו ישנה נקודת קיצון מסוג מקסימום וכעת נמצא את הערך המקסימלי שמתקבל עבור θ זו שמצאנו.

$$\max\left(P(S_n|\theta)\right)=\prod_{i=1}^{n}\frac{1}{\sum_{j=1}^{n}\frac{x_j}{n}}e^{-\frac{x_i}{\sum_{j=1}^{n}\frac{x_j}{n}}}$$

c. נמצא את המקסימום של ההערכת ה$likelihood$ של הפרמטר θ עבור ההתפלגות הנ"ל של $S_n=\{x_1,\dots,x_n\}$. כלומר נמצא את המקסימום של ההסתברות הבאה (נשים לב כי מכיוון שהדוגמאות הינם שרירותיות לכן גם בלתי תלויות ולכן נקבל כפל בין ההסתברויות שלהם):

$$P(S_n|\theta)=\prod_{i=1}^{n}\frac{1}{2}e^{-|x_i-\theta|}$$

נוכל לבצע ln על שני הצדדים ומכיוון שln הינה מונוטונית עולה לכן הפרמטר θ שעבורו נקבל את המקסימום של הln של שני צידי המשוואה יתן גם את המקסימום של $P(k|\theta)$ אם נציבו.

$$\ln\left(P(S_n|\theta)\right)=\ln\left(\prod_{i=1}^{n}\frac{1}{2}e^{-|x_i-\theta|}\right)=\sum_{i=1}^{n}\left(\ln(1)-\ln(2)-\ln\left(e^{-|x_i-\theta|}\right)\right)=\sum_{i=1}^{n}(|x_i-\theta|-\ln(2))$$

נסדר את ערכי הx כך שיהיו מסודרים בסדר עולה ($x_i<x_j$ לכל $1\le i<j\le n$) ויהי $k\in\mathbb{N}$ $1\le k\le n$. נשים לב כי הפונקציה $\sum_{i=1}^{n}(|x_i-\theta|-\ln(2))$ היא בעצם בעלת ערך מקסימום מכיוון שנשים לב כי עבור θ קטנים, כלומר אם קיימים יותר x_i כך ש$\theta>x_i$ מאשר כאלו ש$\theta\le x_i$ אזי נקבל כי הנגזרת הינה מספר חיובי קבוע כלשהו, כלומר הנגזרת חיובית ולכן הפונקציה עולה וזה עבור θ קטנים. ועבור θ גדולים, כלומר אם קיימים פחות x_i כך ש$\theta>x_i$ מאשר כאלו ש$\theta\le x_i$ אזי נקבל כי הנגזרת הינה מספר שלילי קבוע כלשהו, כלומר הנגזרת שלילית ולכן הפונקציה יורדת וזה עבור θ גדולים. וסה"כ קיבלנו כי קיים ערך מקסימום לפונקציה זו ולכן אנו נמצא את ערך הθ עבורו ההערכת ה$likelihood$ תהיה מקסימלית ע"י גזירה והשוואה ל0. כעת נחלק למקרים (לא נתייחס למקרים שבהם θ גדול או שווה מכל ערכי x_i ובמקרה שבו θ קטן או שווה מכל ערכי x_i , מכיוון שבערכים אלו בהכרח לא נקבל נגזרת שווה ל0 בהכרח מכיוון שערך הנגזרות יהיה n או $-n$ בהתאמה וכפי שידוע אנו מתעסקים עם $n>0$)-

עבור n אי זוגי:

- **אם $x_1,\dots,x_k<\theta<x_{k+1},\dots,x_n$:**

$$\sum_{i=1}^{n}(|x_i-\theta|-\ln(2))=\left(\sum_{i=1}^{k}(\theta-x_i)+\sum_{i=k+1}^{n}(x_i-\theta)-n\cdot ln(2)\right)$$

נגזור לפי θ ונשווה ל-0:

$$\frac{d}{d\theta}\left(\sum_{i=1}^{k}(\theta - x_i) + \sum_{i=k+1}^{n}(x_i - \theta) - n \cdot ln(2)\right) = k - (n - (k+1) + 1) = 2k - n$$

$$2k - n = 0$$

$$k = \frac{n}{2}$$

וקיבלנו סתירה לכך שk טבעי כפי שהגדרנו מכיוון שn הוא אי זוגי ולכן $\frac{n}{2}$ אינו טבעי, כלומר לא קיבלנו מקסימום במקרה זה.

- **אם $x_1, \dots, x_{k-1} < (x_k = \theta) < x_{k+1}, \dots, x_n$ (במידה וקיימים גם עוד x_j עבורם $x_j = \theta$ אזי נוריד כפילויות אלה ונשאיר רק את $x_k = \theta$ ונעבור למקרה האי זוגי או הזוגי עם n חדש לאחר הורדת הכפילויות):**

$$\sum_{i=1}^{n}(|x_i - \theta| - \ln(2)) = \left(\sum_{i=1}^{k-1}(\theta - x_i) + \sum_{i=k+1}^{n}(x_i - \theta) - n \cdot ln(2)\right)$$

$(\theta = x_k)$

נגזור לפי θ ונשווה ל-0:

$$\frac{d}{d\theta}\left(\sum_{i=1}^{k-1}(\theta - x_i) + \sum_{i=k+1}^{n}(x_i - \theta) - n \cdot ln(2)\right) = k - 1 - (n - (k+1) + 1) = 2k - n - 1$$

$$2k - n - 1 = 0$$

$$2k = n + 1$$

$$k = \frac{n+1}{2}$$

ובמקרה זה קיבלנו כי עבור $k = \frac{n+1}{2}$ אזי הנגזרת שווה ל0 וכמו שהסברנו למעלה עבור $\theta = x_k = x_{\frac{n+1}{2}}$ (החציון עבור n אי זוגי במקרה שלנו) אזי ערך ה$likelihood$ המקסימלי הוא $\max(P(S_n|\theta)) = \prod_{i=1}^{n}\frac{1}{2}e^{-|x_i - x_{\frac{n+1}{2}}|}$.

עבור n זוגי:

- **אם $x_1, \dots, x_{k-1} < (x_k = \theta) < x_{k+1}, \dots, x_n$ (במידה וקיימים גם עוד x_j עבורם $x_j = \theta$ אזי נוריד כפילויות אלה ונשאיר רק את $x_k = \theta$ ונעבור למקרה האי זוגי או הזוגי עם n חדש לאחר הורדת הכפילויות):**

$$\sum_{i=1}^{n}(|x_i - \theta| - \ln(2)) = \left(\sum_{i=1}^{k-1}(\theta - x_i) + \sum_{i=k+1}^{n}(x_i - \theta) - n \cdot ln(2)\right)$$

$(\theta = x_k)$

נגזור לפי θ ונשווה ל-0:

$$\frac{d}{d\theta}\left(\sum_{i=1}^{k-1}(\theta - x_i) + \sum_{i=k+1}^{n}(x_i - \theta) - n \cdot ln(2)\right) = k - 1 - (n - (k+1) + 1) = 2k - n - 1$$

$$2k - n - 1 = 0$$

$$2k = n + 1$$

$$k = \frac{n+1}{2}$$

וקיבלנו סתירה לכך שk טבעי כפי שהגדרנו מכיוון ש$n+1$ הוא אי זוגי ולכן $\frac{n+1}{2}$ אינו טבעי, כלומר לא קיבלנו מקסימום במקרה זה.

- **אם $x_1, \dots, x_k < \theta < x_{k+1}, \dots, x_n$:**

$$\sum_{i=1}^{n} (|x_i - \theta| - \ln(2)) = \left(\sum_{i=1}^{k} (\theta - x_i) + \sum_{i=k+1}^{n} (x_i - \theta) - n \cdot ln(2) \right)$$

נגזור לפי θ ונשווה ל-0 :

$$\frac{d}{d\theta} \left(\sum_{i=1}^{k} (\theta - x_i) + \sum_{i=k+1}^{n} (x_i - \theta) - n \cdot ln(2) \right) = k - (n - (k+1) + 1) = 2k - n$$

$$2k - n = 0$$

$$k = \frac{n}{2}$$

ובמקרה זה קיבלנו כי עבור $k = \frac{n}{2}$ אזי הנגזרת שווה ל0 וכמו שהסברנו למעלה עבור $x_k = x_{\frac{n}{2}} < \theta < x_{k+1} = x_{\frac{n}{2}+1}$ (החציון עבור n זוגי במקרה שלנו, נבחר את ערכו להיות ממש החציון- $\frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$, אך כל ערך בין $x_{\frac{n}{2}}$ ל $x_{\frac{n}{2}+1}$ יהיה אופטימלי) אזי ערך ה$likelihood$ המקסימלי הוא $\max\left(P(S_n|\theta)\right) = \prod_{i=1}^{n} \frac{1}{2} e^{-|x_i - \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}|}$.

וסה״כ גם למקרה שn זוגי וגם למקרה שהוא אי זוגי קיבלנו כי החציון הוא אכן הפתרון האופטימלי $\theta = median\ of\ \{x_1, \dots, x_n\}$ להביא את ה$likelihood$ למקסימום.

d. נמצא את המקסימום של ההערכת ה$likelihood$ של הפרמטר θ עבור ההתפלגות הנ״ל של $S_n = \{x_1, \dots, x_n\}$. כלומר נמצא את המקסימום של ההסתברות הבאה (נשים לב כי מכיוון שהדוגמאות הינם שרירותיות לכן גם בלתי תלויות ולכן נקבל כפל בין ההסתברויות שלהם) :

$$P(S_n|\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1}$$

נוכל לבצע ln על שני הצדדים ומכיוון שln הינה מונוטונית עולה לכן הפרמטר θ שעבורו נקבל את המקסימום של הln של שני צידי המשוואה יתן גם את המקסימום של $P(S_n|\theta)$ אם נציבו.

$$\ln\left(P(S_n|\theta)\right) = \ln\left(\prod_{i=1}^{n} \theta x_i^{\theta-1}\right) = \sum_{i=1}^{n} \left(\ln(\theta) + \ln\left(x_i^{\theta-1}\right)\right) = n\ln(\theta) + (\theta - 1)\sum_{i=1}^{n} \ln(x_i)$$

כעת נגזור את הביטוי ונחפש את θ שיביא למקסימום של פונקציה זו.

$$\frac{d}{d\theta}\left(n\ln(\theta) + (\theta - 1)\sum_{i=1}^{n} \ln(x_i)\right) = \frac{n}{\theta} + \sum_{i=1}^{n} \ln(x_i)$$

נשווה ל0 את הנגזרת הנ״ל:

$$\frac{n}{\theta} + \sum_{i=1}^{n} \ln(x_i) = 0 \to \theta = -\sum_{i=1}^{n} \frac{n}{\ln(x_i)}$$

כעת אם נגזור שוב את הנגזרת נקבל:

$$\frac{d}{d\theta}\left(\frac{n}{\theta}+\sum_{i=1}^{n}\ln(x_i)\right)=-\frac{n}{\theta^2}$$

כעת נשים לב כי עבור כל θ ובפרט עבור הθ שמצאנו שעבורה הנגזרת שווה ל0 אזי הנגזרת השנייה שלילית ולכן סה"כ קיבלנו כי לפונקציה זו ישנה נקודת קיצון מסוג מקסימום וכעת נמצא את הערך המקסימלי שמתקבל עבור θ זו שמצאנו.

$$\max\left(P(S_n|\theta)\right)=\prod_{i=1}^{n}\left(-\sum_{j=1}^{n}\frac{n}{\ln(x_j)}\right)x_i^{-\sum_{j=1}^{n}\frac{n}{\ln(x_j)}-1}$$

5. (20 pts) Bayesian decision boundary

Let X be an R.V. with a Gaussian distribution, where $P(x|\omega = 1) = \mathcal{N}(\mu_1, \sigma_1^2)$ and $P(x|\omega = 2) = \mathcal{N}(\mu_2, \sigma_2^2)$. σ_1 and σ_2 may not be identical. Assuming the prior is uniform $P(\omega = 1) = P(\omega = 2)$, find the decision boundary as a function of x, and the parameters μ_1, μ_2, σ_1, σ_2 .

5. תחילה נתון כי X מתפלג נורמלית ומתקיים $P(x|w=1)=N(\mu_1,\sigma_1^2)$ ו$P(x|w=2)=N(\mu_2,\sigma_2^2)$. כמו כן נתון כי σ_1,σ_2 לא בהכרח זהים וכי התפלגות ה*prior* הינה יוניפורמית, כלומר $P(w=1)=P(w=2)$. כמו כן נניח שפונקציה העלות הינה $zero-one-lost$ כלומר מתקיים $\frac{\lambda_{22}-\lambda_{12}}{\lambda_{11}-\lambda_{21}}=\frac{0-1}{0-1}=1$ ונמצא את ה*decision boundary* כפונקציה של $\mu_1,\mu_2,\sigma_1,\sigma_2$, כלומר עבור ערכי ה$x$ שמקיימים את האי השוויון הבא, נבחר במצב העולם $w=1$ ועבור שאר ערכי הx נבחר במצב העולם $w=2$:

$$\frac{P(x|w=1)}{P(x|w=2)} > \frac{P(w=2)\lambda_{22}-\lambda_{12}}{P(w=1)\lambda_{11}-\lambda_{21}}$$

$$\frac{P(x|w=1)}{P(x|w=2)} > 1$$

$$\frac{\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2}}{\frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2}} > 1$$

$$\frac{\sigma_2}{\sigma_1}e^{\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} > 1$$

$$e^{\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} > \frac{\sigma_1}{\sigma_2}$$

$$\ln\left(e^{\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2}\right) > \ln\left(\frac{\sigma_1}{\sigma_2}\right)$$

$$\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2 > \ln\left(\frac{\sigma_1}{\sigma_2}\right)$$

$$\sigma_1^2x^2-2\sigma_1^2x\mu_2+\sigma_1^2\mu_2^2-\sigma_2^2x^2+2\sigma_2^2x\mu_1-\sigma_2^2\mu_1^2 > 2\sigma_1^2\sigma_2^2\ln\left(\frac{\sigma_1}{\sigma_2}\right)$$

$$(\sigma_1^2-\sigma_2^2)x^2+(2\sigma_2^2\mu_1-2\sigma_1^2\mu_2)x+\left(\sigma_1^2\mu_2^2-\sigma_2^2\mu_1^2-2\sigma_1^2\sigma_2^2\ln\left(\frac{\sigma_1}{\sigma_2}\right)\right) > 0$$

כעת נבצע נוסחת שורשים-

$$x_{1,2}=\frac{-(2\sigma_2^2\mu_1-2\sigma_1^2\mu_2)\pm\sqrt{(2\sigma_2^2\mu_1-2\sigma_1^2\mu_2)^2-4(\sigma_1^2-\sigma_2^2)\left(\sigma_1^2\mu_2^2-\sigma_2^2\mu_1^2-2\sigma_1^2\sigma_2^2\ln\left(\frac{\sigma_1}{\sigma_2}\right)\right)}}{2(\sigma_1^2-\sigma_2^2)}$$

$$= \frac{-(2\sigma_2^2\mu_1 - 2\sigma_1^2\mu_2) \pm \sqrt{-4\sigma_2^2\mu_1\sigma_1^2\mu_2 + 4\sigma_1^2\sigma_2^2\mu_1^2 + 8\sigma_1^4\sigma_2^2 \ln\left(\frac{\sigma_1}{\sigma_2}\right) + 4\sigma_1^2\sigma_2^2\mu_2^2 - 8\sigma_1^2\sigma_2^4 \ln\left(\frac{\sigma_1}{\sigma_2}\right)}}{2(\sigma_1^2 - \sigma_2^2)}$$

$$= \frac{-(2\sigma_2^2\mu_1 - 2\sigma_1^2\mu_2) \pm \sqrt{4\sigma_1^2\sigma_2^2\left(-\mu_1\mu_2 + \mu_1^2 + 2\sigma_1^2 \ln\left(\frac{\sigma_1}{\sigma_2}\right) + \mu_2^2 - 2\sigma_2^2 \ln\left(\frac{\sigma_1}{\sigma_2}\right)\right)}}{2(\sigma_1^2 - \sigma_2^2)}$$

$$= \frac{-(2\sigma_2^2\mu_1 - 2\sigma_1^2\mu_2) \pm 2\sigma_1\sigma_2\sqrt{-\mu_1\mu_2 + \mu_1^2 + 2\sigma_1^2 \ln\left(\frac{\sigma_1}{\sigma_2}\right) + \mu_2^2 - 2\sigma_2^2 \ln\left(\frac{\sigma_1}{\sigma_2}\right)}}{2(\sigma_1^2 - \sigma_2^2)}$$

כעת נפצל למקרים-

- אם $(\sigma_1^2 - \sigma_2^2)$ חיובי אזי נקבל פרבולה "מחייכת" ולכן $(\sigma_1^2 - \sigma_2^2)x^2 + (2\sigma_2^2\mu_1 - 2\sigma_1^2\mu_2)x + \left(\sigma_1^2\mu_2^2 - \sigma_2^2\mu_1^2 - 2\sigma_1^2\sigma_2^2 \ln\left(\frac{\sigma_1}{\sigma_2}\right)\right) > 0$ מתקיים עבור:
  - אם נקבל כי באמת קיימים $x_{1,2}$ שמקיימים את משוואת השורשים ונניח בלי הגבלת הכלליות כי $x_1 < x_2$ אזי נקבל כי עבור ערכי x שמקיימים $x < x_1$ או $x > x_2$ הפונקציה הנ"ל חיובית ומקיימת את אי השוויון ואלו בעצם ערכי הx עבורם נבחר במצב העולם $w = 1$.
  - אם נקבל כי קיים x_1 שמקיים את משוואת השורשים אזי נקבל כי עבור ערכי x שמקיימים $x \neq x_1$ הפונקציה הנ"ל חיובית ומקיימת את אי השוויון ואלו בעצם ערכי הx עבורם נבחר במצב העולם $w = 1$.
  - אם נקבל כי לא קיים x שמקיימים את משוואת השורשים אזי נקבל כי עבור כל ערכי x הפונקציה הנ"ל חיובית ומקיימת את אי השוויון ואלו בעצם ערכי הx עבורם נבחר במצב העולם $w = 1$.
- אם $(\sigma_1^2 - \sigma_2^2)$ שלילי אזי נקבל פרבולה "עצובה" ולכן $(\sigma_1^2 - \sigma_2^2)x^2 + (2\sigma_2^2\mu_1 - 2\sigma_1^2\mu_2)x + \left(\sigma_1^2\mu_2^2 - \sigma_2^2\mu_1^2 - 2\sigma_1^2\sigma_2^2 \ln\left(\frac{\sigma_1}{\sigma_2}\right)\right) > 0$ מתקיים עבור:
  - אם נקבל כי באמת קיימים $x_{1,2}$ שמקיימים את משוואת השורשים ונניח בלי הגבלת הכלליות כי $x_1 < x_2$ אזי נקבל כי עבור ערכי x שמקיימים $x_1 < x < x_2$ הפונקציה הנ"ל חיובית ומקיימת את אי השוויון ואלו בעצם ערכי הx עבורם נבחר במצב העולם $w = 1$.
  - אם נקבל כי קיים x_1 שמקיים את משוואת השורשים אזי נקבל כי לא קיים x עבורו הפונקציה הנ"ל חיובית ומקיימת את אי השוויון ולכן לא קיימים ערכי x עבורם נבחר במצב העולם $w = 1$.
  - אם נקבל כי לא קיים x שמקיימים את משוואת השורשים אזי נקבל כי לא קיים x עבורו הפונקציה הנ"ל חיובית ומקיימת את אי השוויון ולכן לא קיימים ערכי x עבורם נבחר במצב העולם $w = 1$.