(20942) מבוא ללמידה חישובית | סיכום הרצאה 13

מנחה: ד"ר שי מימון
סמסטר: 2022א'
נכתב על ידי: מתן כהן

שערוך - Estimation

1 מוטיבציה

במידה ויש לנו וקטור **y** של מדידות אשר קשורות לוקטור **x** כלשהו ונרצה להסיק מסקנות לגבי **x**

דוגמה: נניח ש- **x** הוא וקטור של מיקומים ומהירויות של כלי טיס כלשהו ו-**y** הוא וקטור של מדידות של מכ"מים מכמה סנסורים.

היינו רוצים לאמוד את ערכי **x** על פי אותם נתונים בוקטור **y**.

ישנן 2 גישות להתמודדות עם שערוך:

(1) הוקטור שמנסים לשערך הוא וקטור של פרמטרים דטרמיניסטיים אך הוא לא ידוע.
(2) הוקטור שמנסים לשערך הוא וקטור אקראי.

2 Nonrandom Parameter Estimation

- בהרבה מצבים יותר נוח להניח שיש לנו פרמטרים דטרמיניסטיים אך לא ידועים מאשר פרמטרים אקראיים
- לדוגמה - שערוך של תדר של גל סינוסי כלשהו שמעורבב ברעשים

כאשר נרצה לשערך פרמטר או וקטור פרמטרים נסמן את השערוך ב- $\hat{x}$ בתור פונקציה של y ונגדיר את השגיאה שלנו:

$$e\left(y\right)=\hat{x}\left(y\right)-x=\hat{x}-x$$

(הרבה פעמים נעלים את התלות ב-y מתוך נוחות)

יש לזכור נקודה קריטית מאוד:

אסור ליצור תלות בין $\hat{x}$ ל- x !

כלומר - אסור שבפונקצית השערוך שלנו $\hat{x}$ הפרמטרים של x יופיעו.

2.1 נרצה למדוד את טיב המשערך שלנו בעזרת 2 מדדים

כיוון שידוע לנו ש x הוא לא אקראי אבל y כן (אלו המדידות שלנו) הרי שנוכל להשתמש במאפיינים סטטיסטיים כמו התוחלת של משתנה אקראי.

<u>הטיה - bias</u>

משערך חסר הטיה מוגדרת בתור תוחלת השגיאה:

$$b_{\hat{x}}\left(x\right)=E\left[e\left(y\right)\right]=E\left(\hat{x}\left(y\right)\right)-x$$

משמע:

- אם תוחלת השגיאה היא אפס - נגיד שהמשערך חסר הטייה
- אם המשערך חסר הטיה אז התוחלת שלו שווה ערך לפרמטר אותו אנחנו מנסים לשערך!

Error-Covariance

נרצה, בנוסף להטיה נמוכה, שגם השונות המשותפת תהיה מאוד קטנה (מה שיצביע על שערוך טוב).
לשם כך נתבונן במטריצת השונות המשותפת שלנו על פני השגיאות (שונות בין כל 2 שגיאות) שמוגדרת:

$$\Lambda_e(x) = E\left[\left(e(y) - E(e(y))\right) \cdot \left(e(y) - E(e(y))\right)^T\right]$$

כמו כן, התוחלת של כפל השגיאות מוגדרת על ידי:

$$E\left(e(y)\, e^T(y)\right) = \Lambda_e(x) + b_{\hat{x}}(x)\, b_{\hat{x}}^T(x)$$

2.2 גישת הסבירות המרבית - Maximum Likelihood Estimation

- אחת מהגישות הרווחות ביותר לשערוך פרמטרים בהנחה שהפרמטרים דטרמניסטים לא ידועים.
- גישה זו ידוע בעיקר בשל התכונות האספימטוטיות שלה (כאשר מספר המדידות מאוד גדול) אשר משליכות על כך ש**ככל שמספר הדגימות שלנו גדל המשערך שלנו יהיה יותר קרוב לפרמטר** - משמע ההטיה שלו תקטן ותשאף לאפס ביחד עם השונות.
  - במילים אחרות ניתן להגיד שככל שישנן יותר מדידות המשערך של הפרמטר ישאף לערך האמיתי של הפרמטר!
  - במצב זה המשערך ייקרא **עקבי**.

כיוון שאמרנו ש-$\mathbf{y}$ הוא וקטור של משתנים אקראיים אז יש לו פונקציית צפיפות פילוג אשר תלויה ב- $\mathbf{x}$ נגדיר את פונקצית ה-Likelihood:

$$L(\mathbf{x}) = f_{\mathbf{y}}(\mathbf{y};\mathbf{x})$$

הערה. כאשר פונקציה מוגדרת בעזרת ; - $f_{\mathbf{y}}(\mathbf{y};\mathbf{x})$מבינים שמדובר בהנחה שהפרמטרים $\mathbf{x}$ הם דטרמיניסטיים לא ידועים.

נזכור תמיד:

רוצים למצוא את סט הפרמטרים $\mathbf{x}$ אשר יסבירו את $\mathbf{y}$ בצורה הטובה ביותר

ולכן בפועל כאשר נקבל מצב בו ישנו $\mathbf{y}$ נתון ושני סטים של פרמטרים $\mathbf{x}_1$ ו- $\mathbf{x}_2$ המקיימים:

$$f_{\mathbf{y}}(\mathbf{y};\mathbf{x}_1) > f_{\mathbf{y}}(\mathbf{y};\mathbf{x}_2)$$

נרצה להשתמש ב- $\mathbf{x}_1$ כיוון שהסבירות שלו הרבה יותר גבוהה!
באופן כללי, נרצה למקסם את L:

Maximum Likelihood Estimator

$$\hat{\mathbf{x}}_{ML} = \underset{\mathbf{x}}{\operatorname{argmax}}\, L(\mathbf{x}) = \underset{\mathbf{x}}{\operatorname{argmax}}\, f_{\mathbf{y}}(\mathbf{y};\mathbf{x})$$

בהרבה מאוד מקרים נראה שיהיה לנו הרבה יותר קל למקסם פונקציה מונוטונית של הסבירות מאשר את פונקצית הסבירות עצמה, לדוגמה:

$$\log L(\mathbf{x}) = \log f_{\mathbf{y}}(\mathbf{y};\mathbf{x})$$

ולמקסם:

$$\hat{\mathbf{x}}_{ML} = \underset{\mathbf{x}}{\operatorname{argmax}}\, \log L(\mathbf{x}) = \underset{\mathbf{x}}{\operatorname{argmax}}\, \log f_{\mathbf{y}}(\mathbf{y};\mathbf{x})$$

2.2.1 מציאת המקסימום

- במקרים בהם פונקצית הסבירות תהיה גזירה נוכל להשתמש בנגזרת על מנת למצוא את המקסימום בו אנו חושקים:

$$\frac{\partial f_{\mathbf{y}}(\mathbf{y};\mathbf{x})}{\partial \mathbf{x}} = \mathbf{0}$$
$$\frac{\partial \log f_{\mathbf{y}}(\mathbf{y};\mathbf{x})}{\partial \mathbf{x}} = \mathbf{0}$$

- לא תמיד יהיו לנו פתרונות סגורים או גלובליים ונוכל להשתמש באלגוריתמי גרדיאנט על מנת לחפש נקודות מקסימום
- גישה מוכרת שנתפרה עבור בעית ה Maximum Likelihoos היא גישת ה EM - Estimate Maximize

דוגמה. התפלגות ברנולי - הטלת מטבע

נתון סט של N הטלות מטבע בת"ס (iid) עם ההסתברות θ עבור עץ (1) ו- $1-\theta$ עבור פאלי (0)

$$y_n = \begin{cases} 1 & \theta \\ 0 & 1-\theta \end{cases}, \quad n = 1,2,\ldots,N, \quad iid$$

נשתמש בגישת הסבירות המירבית על מנת לשערך את פרמטר θ - ההסתברות להופעת "1".

בשלב הראשון נרשם את פונקצית הצפיפות בעזרת θ, תחילה עבור n יחיד כלשהו:

$$f_{y_n}(y_n;\theta) = \begin{cases} \theta & y_n = 1 \\ 1-\theta & y_n = 0 \end{cases} = \theta^{y_n}\cdot(1-\theta)^{1-y_n}, \quad y_n \in \{0,1\}$$

באופן כללי, כיוון שכלל המשתנים בת"ס נוכל להשתמש במכפלה ולקבל:

$$L(\theta) = f_{\mathbf{y}}(\mathbf{y};\theta) = \prod_{n=1}^{N} f_{y_n}(y_n;\theta) = \prod_{n=1}^{N}\theta^{y_n}\cdot(1-\theta)^{1-y_n} = \theta^{\sum_{n=1}^{N} y_n}\cdot(1-\theta)^{N-\sum_{n=1}^{N} y_n}$$

קיבלנו פונקציה די "מפחידה" - ברור שבגלל שמדובר בהמון מעריכים נוכל להשתמש בפונקצית הלוגריתם ולפשט את הדברים:

$$\log f_{\mathbf{y}}(\mathbf{y};\theta) = \left(\sum_{n=1}^{N} y_n\right)\cdot\log\theta + \left(N - \sum_{n=1}^{N} y_n\right)\cdot\log(1-\theta)$$

כעת נוכל למצוא מקסימום על ידי גזירה והשוואה לאפס ונקבל שמדובר בממוצע (חייב לגזור שוב כדי לוודא שזו נק' מקס'):

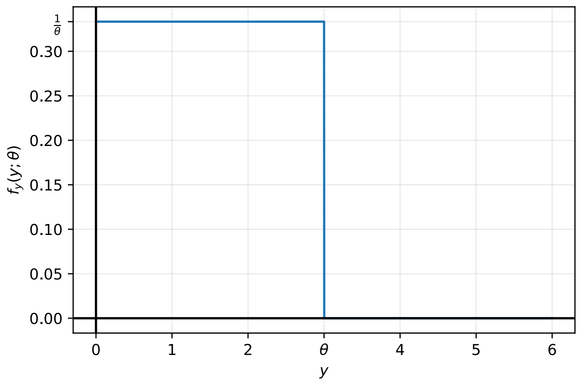
$$\frac{\partial \log f_{\mathbf{y}}(\mathbf{y};\theta)}{\partial\theta} = \left(\sum_{n=1}^{N} y_n\right)\cdot\frac{1}{\theta} + \left(N - \sum_{n=1}^{N} y_n\right)\cdot\frac{-1}{1-\theta} = 0$$
$$\Rightarrow \frac{1}{\theta(1-\theta)}\cdot\left[\left(\sum_{n=1}^{N} y_n\right)\cdot(1-\theta) - \left(N - \sum_{n=1}^{N} y_n\right)\cdot\theta\right] = 0$$
$$\Rightarrow \frac{1}{\theta(1-\theta)}\cdot\left[\left(\sum_{n=1}^{N} y_n\right) - \left(\theta\sum_{n=1}^{N} y_n\right) - N\theta + \left(\theta\sum_{n=1}^{N} y_n\right)\right] = 0$$
$$\Rightarrow \frac{1}{\theta(1-\theta)}\cdot\left[\left(\sum_{n=1}^{N} y_n\right) - N\theta\right] = \frac{N}{\theta(1-\theta)}\cdot\left[\frac{1}{N}\sum_{n=1}^{N} y_n - \theta\right] = 0$$
$$\Rightarrow \hat{\theta}_{ML} = \overline{y} = \frac{1}{N}\sum_{n=1}^{N} y_n$$

משמע $\theta \rightarrow \hat{\theta} \Leftarrow \infty \rightarrow N$, בנוסף נראה שההטיה אפסית - $E\left(\hat{\theta}_{ML}\right) = \frac{1}{N}\sum_{n=1}^{N} E(y_n) = \theta$ והשונות קטנה ככל ש-N גדל.

דוגמה. התפלגות אחידה

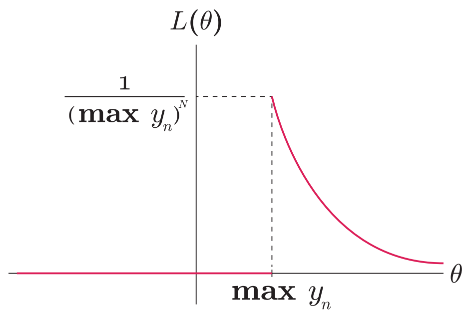
ניזכר בפונקציית צפיפות הפילוג של התפלגות האחידה עם פרמטר θ של מדידות בת"ס:

$$f_{y_n}\left(y_n;\theta\right)=\begin{cases}\frac{1}{\theta} & 0\leq y_n\leq\theta\\ 0 & \text{else}\end{cases}\qquad iid.$$



נחשב את פונקציית צפיפות הפילוג המשותפת כמו בדוגמה הקודמת:

$$L\left(\theta\right)=f_{\mathbf{y}}\left(\mathbf{y};\theta\right)=\prod_{n=1}^{N}f_{y_n}\left(y_n;\theta\right)=\begin{cases}\frac{1}{\theta^N} & \max\left(y_n\right)\leq\theta\\ 0 & \text{else}\end{cases}$$



וכעת, קל להבחין על פי הגדרת הפונקציה ש:

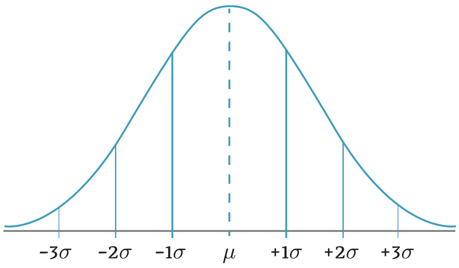
$$\hat{\theta}_{ML}=\max\left\{y_1,y_2,\ldots y_N\right\}$$

ועל פי הגדרה זו ומכך שראינו שהפונקציה מגדירה $\max\left(y_n\right)\leq\theta$ הרי שבהכרח אנחנו נתקרב אל θ **מלמטה** - משמע אנחנו תמיד נהיה עם הטיה מסוימת.

הנקודה חשובה כאן היא להבחין שככל ש-N יגדל ככה המקסימום שנקבל עבור $\hat{\theta}_{ML}$ יהיה קרוב יותר ל-θ עצמו!

דוגמה. התפלגות נורמלית (גאוסית)

$$f_{y_n}\left(y_n;\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\cdot\exp\left\{-\frac{1}{2\sigma^2}\left(y_n-\mu\right)^2\right\}\quad iid$$



פונקצית צפיפות הפילוג המשותפת:

$$\begin{aligned}L\left(\theta\right)=f_{\mathbf{y}}\left(y_n;\mu,\sigma^2\right)&=\prod_{n=1}^{N}f_{y_n}\left(y_n;\mu,\sigma^2\right)\\&=\left(2\pi\sigma^2\right)^{-N/2}\cdot\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(y_n-\mu\right)^2\right\}\end{aligned}$$

מתבקש להשתמש ב log:

$$\log\ f_{\mathbf{y}}\left(\mathbf{y};\mu,\sigma^2\right)=\frac{N}{2}\cdot\log\left(2\pi\right)-\frac{N}{2}\log\left(\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(y_n-\mu\right)^2$$

נגזור לפי μ ונשווה לאפס (פונקציה ריבועית אז נקבל מקסימום גלובלי):

$$\begin{aligned}\frac{\partial\log\ f_{\mathbf{y}}\left(y_n;\mu,\sigma^2\right)}{\partial\mu}&=\frac{1}{\sigma^2}\sum_{n=1}^{N}\left(y_n-\mu\right)=0\\\Rightarrow\hat{\mu}_{ML}&=\frac{1}{N}\sum_{n=1}^{N}y_n=\overline{y}\end{aligned}$$

באותה צורה נוכל לגזור על פי σ^2 כדי לקבל את התוצאה הרצויה:

$$\begin{aligned}\frac{\partial\log\ f_{\mathbf{y}}\left(y_n;\mu,\sigma^2\right)}{\partial\sigma^2}&=-\frac{N}{2}\cdot\frac{1}{\sigma^2}+\frac{1}{2\left(\sigma^2\right)^2}\cdot\sum_{n=1}^{N}\left(y_n-\mu\right)^2=0\\\Rightarrow\hat{\sigma^2}_{ML}&=\frac{1}{N}\sum_{n=1}^{N}\left(y_n-\underbrace{\overline{y}}_{\hat{\mu}_{ML}}\right)^2\end{aligned}$$

2.2.2 Maximum Likelihood and Least-Squares

נרצה לקשור בין הגישה לבין שיטת הריבועים הפחותים.

נניח שוקטור המדידות שלנו $\mathbf{y} \in \mathbb{R}^N$ מוגדר על ידי **המדידות** X, וקטור $\mathbf{w}$ ווקטור **אקראי** $\mathbf{v} \in \mathbb{R}^N$ כך ש $v_n \sim N\left(0, \sigma^2\right)$ בת"ס:

$$\mathbf{y} = X\mathbf{w} + \mathbf{v}, \quad \mathbf{w} \in \mathbb{R}^{d+1}, \ \mathbf{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_N \end{bmatrix} \sim N\left(\mathbf{o}, \sigma^2 \cdot I\right)$$

נוכל לרשום את הפילוג של $\mathbf{y}$ (כיוון שהזזה של משתנה אקראי פשוט מסיטה את הממוצע, נבחין כי השונות לא תשתנה אך הממוצע יזוז על פי אותו היסט - במקרה שלנו $X\mathbf{w}$):

$$\mathbf{y} \sim N\left(X \cdot \mathbf{w}, \sigma^2 \cdot I\right) \tag{2.1}$$

נכליל את המקרה הסקלרי של ההתפלגות הנורמלית: עבור משתנה אקראי $\mathbf{y} \sim N\left(\boldsymbol{\mu}, C\right)$ פונקצית צפיפות הפילוג שלו היא:

$$f_{\mathbf{y}}\left(\mathbf{y}; \boldsymbol{\mu}, C\right) = \frac{1}{\sqrt{\det\left(2\pi C\right)}} \cdot \exp\left\{-\frac{1}{2} \cdot \left(\mathbf{y} - \boldsymbol{\mu}\right)^T \cdot C^{-1} \cdot \left(\mathbf{y} - \boldsymbol{\mu}\right)\right\}$$

כעת נציב את 2.1:

$$\begin{aligned} f_{\mathbf{y}}\left(\mathbf{y}; X \cdot \mathbf{w}, \sigma^2 \cdot I\right) &= \frac{1}{\sqrt{\det\left(2\pi \cdot \sigma^2 I\right)}} \cdot \exp\left\{-\frac{1}{2} \cdot \left(\mathbf{y} - X \cdot \mathbf{w}\right)^T \cdot \underbrace{C^{-1}}_{\frac{1}{\sigma^2} I} \cdot \left(\mathbf{y} - X \cdot \mathbf{w}\right)\right\} \\ &= \frac{1}{\sqrt{\left(2\pi \cdot \sigma^2\right)^N \cdot \underbrace{\det\left(I\right)}_{1}}} \cdot \exp\left\{-\frac{1}{2\sigma^2} \cdot \left\|\mathbf{y} - X \cdot \mathbf{w}\right\|^2\right\} \end{aligned}$$

נשתמש ב- log על מנת לפשט את החישובים שלנו:

$$\log \ f_{\mathbf{y}}\left(\mathbf{y}; X \cdot \mathbf{w}, \sigma^2 \cdot I\right) = -\frac{N}{2} \cdot \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left\|\mathbf{y} - X \cdot \mathbf{w}\right\|^2$$

ונקבל:

$$\hat{\mathbf{w}}_{ML} = \underset{\mathbf{w}}{\text{argmax}} \ \log \ L\left(\mathbf{w}\right)$$

כיוון שהחלק הראשון של הביטוי לא תלוי כלל ב $\mathbf{w}$ ונוכל להתעלם מקבועים - נוכל פשוט למצוא ערך $\mathbf{w}$ אשר יביא למינימום (החלפת הסימן) את הביטוי ונקבל בדיוק את גישת ה- Least Squares:

$$\hat{\mathbf{w}}_{ML} = \underset{\mathbf{w}}{\text{argmin}} \ \left\|\mathbf{y} - X \cdot \mathbf{w}\right\|^2 = \hat{\mathbf{w}}_{LS}$$

2.2.3 Maximum Likelihood and Logistic Regression

נזכיר שרגרסיה לוגיסטית קיבלנו פונקציה הממודלת על ידי $\theta(s)=\dfrac{1}{1+e^{-s}}$:

$$P\left(y_n=1\mid \mathbf{x}_n\right)=\theta\left(\mathbf{w}^T\mathbf{x}_n\right)$$
$$P\left(y_n=-1\mid \mathbf{x}_n\right)=1-\theta\left(\mathbf{w}^T\mathbf{x}_n\right)=\theta\left(-\mathbf{w}^T\mathbf{x}_n\right)$$

כעת נשתמש ב-Maximum Likelihood על מנת למדל את אותה בעיה.
תחילה ננסה לרשום את הפונקציה בצורה פשוטה:

$$P\left(y_n\mid \mathbf{x}_n\right)=\theta\left(y_n\cdot\mathbf{w}^T\mathbf{x}_n\right),\quad y_n\in\{-1,1\}$$

ונוכל לרשום את המכפלה עבור N המדידות בת"ס:

$$\prod_{n=1}^{N}P\left(y_n\mid \mathbf{x}_n\right)=\prod_{n=1}^{N}\theta\left(y_n\cdot\mathbf{w}^T\mathbf{x}_n\right)$$

נעבור שוב ל-log ונשתמש בהגדרת θ:

$$\begin{aligned}\log\ L=\prod_{n=1}^{N}P\left(y_n\mid \mathbf{x}_n\right)&=\sum_{n=1}^{N}\log\theta\left(y_n\cdot\mathbf{w}^T\mathbf{x}_n\right)\\&=-\sum_{n=1}^{N}\log\left(1+e^{-y_n\mathbf{w}^T\mathbf{x}_n}\right)\end{aligned}$$

ועל פי הצבה פשוטה נבחין כי זו פונקצית המחיר שפיתחנו עבור cross entropy כאשר דיברנו על Logistic Regression:

$$\hat{\mathbf{w}}_{ML}=\underset{\mathbf{w}}{\operatorname{argmax}}\ \log L\left(\mathbf{w}\right)=\underset{\mathbf{w}}{\operatorname{argmin}}\ \sum_{n=1}^{N}\log\left(1+e^{-y_n\mathbf{w}^T\mathbf{x}_n}\right)$$

זוהי פונקציה קונבקסית של **w** ולכן ההתכנסות שלנו עם אלגוריתם איטרטיבי תביא אותנו על מינימום גלובלי.

3 שערוך פרמטר אקראי

בגישה זו אנחנו מתייחסים ל $\mathbf{x}$ (הפרמטר שנרצה לשערך) בתור משתנה אקראי כלשהו שיש לנו מידע מקדים עליו (נניח אי שלילי) - משמע ישנו פילוג $p_{\mathbf{x}}(\mathbf{x})$ אשר מתאר את הפילוג **הא-פריורי (A-priori)** של $\mathbf{x}$ עוד לפני שצפינו בתצפיות!
בצורה זו, במידה ובידינו הפילוג $p_{\mathbf{x}}(\mathbf{x})$ והפילוג $p_{\mathbf{y}|\mathbf{x}}(\mathbf{y} \mid \mathbf{x})$ נוכל לקבל את הפילוג המשותף (הסטטיסטיקה המלאה):

$$p_{\mathbf{y},\mathbf{x}}(\mathbf{y},\mathbf{x}) = p_{\mathbf{y}|\mathbf{x}}(\mathbf{y} \mid \mathbf{x})\, p_{\mathbf{x}}(\mathbf{x})$$

וכמובן שנוכל להשתמש בנוסחאת בייס ולקבל את הפילוג **הא-פוסטריורי (A-posteriori)** אשר משמעו "איך הפילוג של $\mathbf{x}$ מושפע לאחר שצפינו במדידות $\mathbf{y}$:

Bayes' theorem

$$p_{\mathbf{x}|\mathbf{y}}(\mathbf{x} \mid \mathbf{y}) = \frac{p_{\mathbf{y}|\mathbf{x}}(\mathbf{y} \mid \mathbf{x})\, p_{\mathbf{x}}(\mathbf{x})}{p_{\mathbf{y}}(\mathbf{y})}$$

לגישה זו קוראים **Baysian Estimation**

3.1 שערוך בייסיאני - Bayesian Estimation

בשיטה זו נרצה להגדיר פונקצית מחיר בין פרמטר $\mathbf{x}$ לבין המשערך שלנו שתלוי ב-$\mathbf{y}$: $\mathbf{f}(\mathbf{y})$ ועל מנת למצוא את המשערך הטוב ביותר נרצה לפתור בעית אופטימיזציה מהצורה הבאה:

$$\hat{\mathbf{x}}(\cdot) = \underset{\mathbf{f}(\cdot)}{\operatorname{argmin}}\, E\left(C\left(\mathbf{x},\mathbf{f}(\mathbf{y})\right)\right)$$

נתבונן בתוחלת שבבעית האופטימיזציה:

$$\overbrace{E\left(C\left(\mathbf{x},\mathbf{f}(\mathbf{y})\right)\right)}^{\iint C(\mathbf{x},\mathbf{f}(\mathbf{y}))\cdot p_{\mathbf{xy}}(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y}} \overset{\text{Bayes}}{=} \int \left[\underbrace{\int C\left(\mathbf{x},\mathbf{f}(\mathbf{y})\right) p_{\mathbf{x}|\mathbf{y}}(\mathbf{x} \mid \mathbf{y})\, d\mathbf{x}}_{(1)}\right] p_{\mathbf{y}}(\mathbf{y})\, d\mathbf{y}$$

ונוכל להתבונן ב- (1) ולרשום את $\mathbf{f}(\mathbf{y})$ בתור $\mathbf{a}$ ולקבל:

$$\hat{\mathbf{x}}(\mathbf{y}) = \underset{\mathbf{a}}{\operatorname{argmin}} \int C(\mathbf{x},\mathbf{a})\, p_{\mathbf{x}|\mathbf{y}}(\mathbf{x} \mid \mathbf{y})\, d\mathbf{x} \tag{3.1}$$

נשתמש שוב בנוסחאת בייס על הביטוי: $p_{\mathbf{x}|\mathbf{y}}(\mathbf{x} \mid \mathbf{y})$, נשמיט את המכנה $p_{\mathbf{y}}$ (כיוון שלא קשור לבעיה) ונקבל את בעית האופטימיזציה:

$$\hat{\mathbf{x}}(\mathbf{y}) = \underset{\mathbf{a}}{\operatorname{argmin}} \int C(\mathbf{x},\mathbf{a})\, p_{\mathbf{y}|\mathbf{x}}(\mathbf{y} \mid \mathbf{x})\, p_{\mathbf{x}}(\mathbf{x})\, d\mathbf{x}$$

בדוגמה הבאה נדבר על מקרה בו המשתנה האקראי הוא סקלרי, קל לעבור למצב וקטורי בצורה של component wise.

דוגמה. Minimum Uniform Cost (MUC)

נניח שפונקציית המחיר שלנו היא:

$$C(a,\hat{a}) = \begin{cases} 1 & |a-\hat{a}| > \varepsilon \\ 0 & \text{otherwise} \end{cases}$$

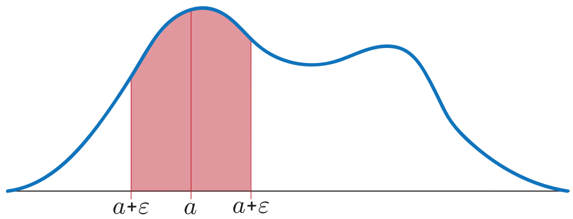
נציב את C בבעית האופטימיזציה שהגדרנו ב- 3.1 ונשתמש בהסתברות המשלימה:

$$\hat{x}_{MUC}(\mathbf{y}) = \underset{a}{\operatorname{argmin}} \left[\int\limits_{\{\mathbf{x}|C=1\}=|\mathbf{x}-a|>\varepsilon} p_{\mathbf{x}|\mathbf{y}}(\mathbf{x} \mid \mathbf{y})\, d\mathbf{x} \right] = \boxed{\underset{a}{\operatorname{argmin}} \left[1 - \int\limits_{a-\varepsilon}^{a+\varepsilon} p_{\mathbf{x}|\mathbf{y}}(\mathbf{x} \mid \mathbf{y})\, d\mathbf{x} \right]}$$

כיוון שאנחנו רוצים להביא למינימום את הביטוי, אנחנו בעצם רוצים להביא למקסימום את הביטוי:

$$\underset{a}{\operatorname{argmax}} \int\limits_{a-\varepsilon}^{a+\varepsilon} p_{\mathbf{x}|\mathbf{y}}(\mathbf{x} \mid \mathbf{y})\, d\mathbf{x}$$

בעצם מה שעשינו כאן זה למצוא נקודת מקסימום בתחום מסוים של פונקצית ההתפלגות שלנו על מנת להחזיר לנו **את ההסתברות הגבוהה ביותר**:



אם נשאיף את ε לאפס נמצא את הנקודה המקסימלית של ה a-posteriori distribution ונקבל המשערך הא-פוסטריורי המקסימלי:

Maximum A posteriori Estimation (MAP)

$$\hat{x}_{MAP}(\mathbf{y}) = \underset{a}{\operatorname{argmax}}\, p_{x|\mathbf{y}}(a \mid \mathbf{y}) = \lim_{\varepsilon \to 0} \hat{x}_{MUC}(\mathbf{y})$$

ואם נתבונן בהבדלים בין Maximum Likelihood לבין Maximuim A-posteriori Estimation נראה שהעיקר הוא המידע המקדים על **x**:

$$\hat{x}_{ML}(y) = \underset{a}{\operatorname{argmax}}\, p(y;a) = \underset{a}{\operatorname{argmax}}\, \log p(y;a)$$

$$\hat{x}_{MAP}(y) \overset{\text{Bayes}}{=} \underset{a}{\operatorname{argmax}} \frac{p_{y|x}(y \mid a) \cdot p_x(a)}{p_y(y)} \overset{\log \text{ is monotonic}}{=} \underset{a}{\operatorname{argmax}} \log p_{y|x}(y \mid a) + \log p_x(a)$$

דוגמה. Minimum Absolute Error (MAE)

ניתן להשתמש בפונקצית המחיר של הערך המוחלט:

$$C\left(a,\hat{a}\right)=\left|a-\hat{a}\right|$$

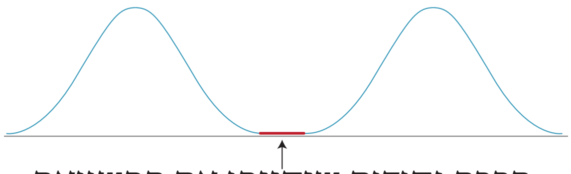
נקבל בעזרת הצבה ב- 3.1 את:

$$\hat{\mathbf{x}}\left(\mathbf{y}\right)=\underset{a}{\operatorname{argmin}}\int\limits_{-\infty}^{\infty}\left|x-a\right|p_{\mathbf{x}|\mathbf{y}}\left(\mathbf{x}\mid\mathbf{y}\right)d\mathbf{x}$$

ובעצם את החציון של פונקצית ההתפלגות - הערך שבו ההסתברות של ההתפלגות האפוסטריורית היא בדיוק חצי:

$$\int\limits_{-\infty}^{\hat{x}_{MAE}(\mathbf{y})}p_{\mathbf{x}|\mathbf{y}}\left(\mathbf{x}\mid\mathbf{y}\right)d\mathbf{x}=\int\limits_{\hat{x}_{MAE}(\mathbf{y})}^{\infty}p_{\mathbf{x}|\mathbf{y}}\left(\mathbf{x}\mid\mathbf{y}\right)d\mathbf{x}=\frac{1}{2}$$

הערה. יכולים להיות כמה משערכים עם אותו ערך (החציון הוא לא בהכרח משערך יחיד) - חשוב לזכור!



הרבה נקודות שיקיימו את המשוואה

דוגמה. Minimum Mean Square Error (MMSE)

המשערך הפופולרי ביותר, קובע את פונקצית המחיר הריבועית המוכרת (במקרה הוקטורי):

$$C\left(\mathbf{a},\hat{\mathbf{a}}\right)=\left\|\mathbf{a}-\hat{\mathbf{a}}\right\|^{2}=\left(\mathbf{a}-\hat{\mathbf{a}}\right)^{T}\left(\mathbf{a}-\hat{\mathbf{a}}\right)=\sum_{i=1}^{N}\left(a_{i}-\hat{a}_{i}\right)^{2}$$

נציב ב- 3.1 ונקבל:

$$\hat{\mathbf{x}}_{MMSE}\left(\mathbf{y}\right)=\underset{a}{\operatorname{argmin}}\int\left(\mathbf{a}-\mathbf{x}\right)^{T}\left(\mathbf{a}-\mathbf{x}\right)p_{\mathbf{x}|\mathbf{y}}\left(\mathbf{x}\mid\mathbf{y}\right)d\mathbf{x}$$

במקרה הסקלרי:

$$\hat{\mathbf{x}}_{MMSE}\left(\mathbf{y}\right)=\underset{a}{\operatorname{argmin}}\int\limits_{-\infty}^{\infty}\left(a-x\right)^{2}p_{x|\mathbf{y}}\left(x\mid\mathbf{y}\right)dx$$

נמצא נקודת מינימום על ידי גזירה על פי a ונקבל את **משערך התוחלת המותנית**:

$$\hat{\mathbf{x}}_{MMSE}\left(\mathbf{y}\right)=\int\limits_{-\infty}^{\infty}xp\left(x\mid\mathbf{y}\right)dx=E\left(x\mid\mathbf{y}\right)$$

משערך התוחלת המותנית

Minimum Mean Square Error (MMSE)

$$\hat{\mathbf{x}}_{MMSE}\left(\mathbf{y}\right)=E\left(x\mid\mathbf{y}\right)$$

תכונות ה- MMSE

הטיה

המשערך חסר הטיה:

$$b_{MMSE} = E\left(\mathbf{e}\left(\mathbf{x},\mathbf{y}\right)\right) = E\left(\hat{\mathbf{x}}_{MMSE}\left(\mathbf{y}\right) - \mathbf{x}\right) = E_{\mathbf{y}}\left(E_{\mathbf{x}|\mathbf{y}}\left(\mathbf{x} \mid \mathbf{y}\right)\right) - E\left(\mathbf{x}\right) = 0$$

שונות משותפת - covariance

$$\begin{aligned}\Lambda_{MMSE} &= E\left(\mathbf{e}\mathbf{e}^T\right) = E\left[\left(\mathbf{x} - E\left(\mathbf{x} \mid \mathbf{y}\right)\right)\left(\mathbf{x} - E\left(\mathbf{x} \mid \mathbf{y}\right)\right)^T\right] \\ &= E_{\mathbf{y}}\left\{E_{\mathbf{x}|\mathbf{y}}\left[\left(\mathbf{x} - E\left(\mathbf{x} \mid \mathbf{y}\right)\right)\left(\mathbf{x} - E\left(\mathbf{x} \mid \mathbf{y}\right)\right) \mid \mathbf{y}\right]\right\} \\ &= E\left(\Lambda_{\mathbf{x}|\mathbf{y}}\left(\mathbf{y}\right)\right)\end{aligned}$$

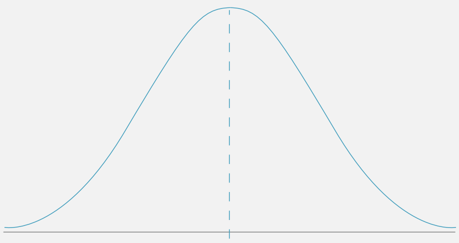
שגיאת השערוך אורתוגונלית לכל פונקציה (לינארית או לא לינארית) של המדידות שלנו

$$E\left[\left(\hat{\mathbf{x}}\left(\mathbf{y}\right) - \mathbf{x}\right)\mathbf{g}^T\left(\mathbf{y}\right)\right] = 0$$

סיכום הדוגמאות:

- MAP מחפש את נקודת המקסימום של הפילוג
- MAE מחפש את נקודת (נקודות) החציון של הפילוג
- MMSE מחפש את נקודת התוחלת של הפילוג

במידה ו-**x** ו-**y** מפולגים גאוסית במשותף - שלושת המשערכים יתכנסו לאותה נקודה



$E(x|y)$=Median=Max $p(x|y)$

4 דוגמה לשערוך

נניח שיש בידינו סט מדידות y_n המוגדר על ידי המודל:

$$y_n = w_0 + \mathbf{x}_n^T \cdot \mathbf{w} + V_n, \quad n = 1, 2, \ldots, N$$

כאשר:

- w_0 - פרמטר
- $\mathbf{x}_n$ - מדידות ידועות
- $\mathbf{w}$ - סט משתנים אקראיים בת"ס אחד בשני וב- V_n המתפלגים: $w_j \sim N\left(0, \tau^2\right),\ j = 1, 2, \ldots, d$
- V_n - רעשים המתפלגים: $V_n \sim N\left(0, \sigma^2\right),\ iid$

MAP

נכתוב את פונקציית צפיפות הפילוג של $\mathbf{w}$, תוך שימוש בנתון שהמשתנים הם בת"ס:

$$f_{\mathbf{w}}\left(\mathbf{w}\right) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\tau^2}} \cdot \exp\left\{-\frac{1}{2\tau^2} \cdot \left(w_j\right)^2\right\}$$

נשתמש בלוגריתם על מנת לפשט את הדברים:

$$\begin{aligned} \log f_{\mathbf{w}}\left(\mathbf{w}\right) &= -\frac{d}{2}\log\left(2\pi\tau^2\right) - \frac{1}{2\tau^2}\sum_{j=1}^{N} w_j^2 \\ &= -\frac{d}{2}\log\left(2\pi\tau^2\right) - \frac{1}{2\tau^2}\left\|\mathbf{w}\right\|^2 \end{aligned}$$

כעת נבחין כי הביטוי $\mathbf{w} \cdot \mathbf{x}_n^T + w_0$ משמש לנו מעין קבוע, ולכן נוכל להבחין כי ההתפלגות של $y_n \mid \mathbf{w}$ היא נורמלית כהתפלגות V_n אשר מוסטת בקבוע $w_0 + \mathbf{x}_n^T \cdot \mathbf{w}$:

$$y_n \mid \mathbf{w} \sim N\left(w_0 + \mathbf{x}_n^T \cdot \mathbf{w}, \sigma^2\right)$$

ונוכל למצוא את פונקציית ההתפלגות:

$$f_{\mathbf{y}|\mathbf{w}} = \prod_{n=1}^{N} f_{y_n|\mathbf{w}}\left(y_n \mid \mathbf{w}\right)$$

כל אחד מהביטויים $f_{y_n|\mathbf{w}}\left(y_n \mid \mathbf{w}\right)$ מוגדר:

$$f_{y_n|\mathbf{w}}\left(y_n \mid \mathbf{w}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{1}{2\sigma^2} \cdot \left(y_n - w_0 - \mathbf{x}_n^T \cdot \mathbf{w}\right)^2\right\}$$

נשתמש שוב ב-log:

$$\log f_{y_n|\mathbf{w}}\left(y_n \mid \mathbf{w}\right) = -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \cdot \left(y_n - w_0 - \mathbf{x}_n^T \cdot \mathbf{w}\right)^2$$

ולכן:

$$\log f_{\mathbf{y}|\mathbf{w}} = -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(y_n - w_0 - \mathbf{x}_n^T \cdot \mathbf{w}\right)^2$$

מצאנו את שני המרכיבים של MAP וכעת צריך למקסם:

$$\hat{\mathbf{w}}_{MAP} = \underset{\mathbf{w}}{\text{argmax}} \log f_{\mathbf{y}|\mathbf{w}}\left(\mathbf{y} \mid \mathbf{w}\right) + \log f_{\mathbf{w}}\left(\mathbf{w}\right)$$

קל לראות ש:

- ב-$\log f_{\mathbf{y}|\mathbf{w}}\left(\mathbf{y} \mid \mathbf{w}\right)$ הביטוי התלוי ב- $\mathbf{w}$ הוא:

$$-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(y_n - w_0 - \mathbf{x}_n^T \cdot \mathbf{w}\right)^2$$

  - ביטוי זה מוכפל ב- $-\frac{1}{2}$, לכן נוכל להשמיט אותו
- ב- $\log f_{\mathbf{w}}\left(\mathbf{w}\right)$ הביטוי התלוי ב- $\mathbf{w}$ הוא:

$$-\frac{1}{2\tau^2}\left\|\mathbf{w}\right\|^2$$

  - ביטוי זה גם מוכפל ב- $-\frac{1}{2}$ ונוכל להשמיט את המכפלה גם כאן

נקבל את בעית האופטימיזציה הפשוטה יותר (לאחר מכפלה ב- σ^2 על מנת לקבל ביטוי נוח יותר):

$$\hat{\mathbf{w}}_{MAP} = \underset{\mathbf{w}}{\text{argmin}}\left\{\sum_{n=1}^{N}\left(y_n - w_0 - \mathbf{x}_n^T \cdot \mathbf{w}\right)^2 + \frac{\sigma^2}{\tau^2}\cdot\left\|\mathbf{w}\right\|^2\right\}$$

ונבחין כי זה ביטוי זהה לביטוי שקיבלנו ב- Ridge Regression (אם נסמן $\lambda = \frac{\sigma^2}{\tau^2}$), רק שכאן λ נקבע על פי שונות הרעש V_n והשונות של סט המשתנים האקראיים $\mathbf{w}$.

- השונויות σ^2 ו-τ^2 יהוו מדד לעד כמה אנו מאמינים במידע המקדים ($\mathbf{w}$) לעומת המדידות (y_n)
- אם τ^2 מאוד קטן - נסיק שיש לנו וודאות מאוד גדולה שהערכים מתפלגים סביב 0 (פעמון מאוד צר)
  - משמע נקבל פרמטר רגולריזציה λ מאוד גדול - מה שיגרור אילוץ של פרמטרים מאוד קטנים
- אם σ^2 מאוד קטנה (ביחס ל τ^2 כמובן) - מדובר במדידות שהן כמעט ונטולות ברעש
  - משמעש נקבל פרמטר רגולריזציה מאוד קטן ונסתמך על ערכי $\mathbf{w}$ בשביל לאפטם את הבעיה שלנו

הערה. במידה ונשנה את הפלגות w_j כך ש:

$$w_j \sim Laplace\left(0, \tau^2\right), \quad j = 1, 2, \ldots, d$$

עם אותן הנחות של חוסר תלות - נקבל את Lasso Regression.