למידה חישובית - שיעור 6

MLE – Maximum Likelihood Estimate

זוהי גישה ישירה לשערך פרמטרים. בשיטה זו אנו מתאימים מודל לנתונים, כלומר אם נתון לי instance שאני יודעת שההתפלגות שלו היא נורמלית, אני יכולה למצוא אומדן לתוחלת של ה-instance הזה. נרצה למצוא את הפרמטר שיסביר בצורה הטובה ביותר את המדגם.

מודל זה מביא למקסימום את הסבירות של ה-data.

נתון:

- ה-data. נקבל קבוצת נקודות: $D = \{x_1, x_2, \dots, x_n\}$
- סוג המודל איתו אנו עובדים (לדוגמה: נורמלי, פואסוני וכו'),
- וקטור θ. אוסף וקטור פרמטרים שמתאר את ה-instance הספציפי של המודל שאותו אני רוצה למצוא. מתאר את סוג המודל. θ בעצם יכיל את התוחלת והשונות של ה-instance, כתלות במיימד, כלומר אם אני במימד 1 תהיה תוחלת ושונות אחת, אם אני במימד 5 יהיו 5 תוחלות ושונויות וזה יהיה סוג של מטריצה.

נחשב:

- Likelihood של המודל עבור ה-data יהיה $L(D|\theta) = P(D|\theta)$
- log-Likelihood של המודל בהינתן ה-data יהיה $L(\theta) = \log P(D|\theta)$

נחפש את:

- $\theta_{ML} = \underset{\theta\in\Omega}{\operatorname{argmax}} L(\theta)$

בשימוש בMLE אנו מניחים כי ה-instances בלתי תלויים זה בזה, לכן ההסתברות של כל ה-data בהינתן המודל הוא מכפלת ההסתברויות:

$$\begin{aligned}
\theta_{ML} &= \underset{\theta\in\Omega}{\arg\max}\ L(\Theta) \\
&= \underset{\theta\in\Omega}{\arg\max}\ \log P(D\,|\,\Theta) \\
&= \underset{\theta\in\Omega}{\arg\max}\ \log P(x_1,\dots,x_n\,|\,\Theta) \\
&= \underset{\theta\in\Omega}{\arg\max}\ \log \prod_i P(x_i\,|\,\Theta) \\
&= \underset{\theta\in\Omega}{\arg\max}\ \sum_i \log P(x_i\,|\,\Theta)
\end{aligned}$$

log המכפלה שווה
לסכום הלוגים

דוגמת הטלת המטבע (מודל בינומי):

הנחות:

- מטבע יקבל ערך T בהסתברות p, וערך H בהסתברות $q = 1 - p$
- נזרוק את המטבע N פעמים.

במקרה הנוכחי, θ היא רק p, כי זה הפרמטר היחיד שמשפיע על ההסתברות. נחשב:

$$L(\Theta) = \log P(D \mid \Theta) = \log p^m (1-p)^{N-m}$$
$$= m \log p + (N-m) \log(1-p)$$

הערה: השמטנו את המקדם הבינומי, כי הוא זהה בכל ההסתברויות.

$$\frac{dL(\Theta)}{dp} = \frac{d(m \log p + (N-m)\log(1-p))}{dp} = \frac{m}{p} - \frac{N-m}{1-p} = 0$$

$$p = \frac{m}{N}$$

דוגמת מכונה (מודל פואסוני):

$$L(\theta = \lambda) = P(D|\lambda)$$
$$L(\theta = \lambda) = e^{-\lambda n} \prod_{i=1}^{n} \frac{\lambda^{k_i}}{k_i!}$$

נרצה להימנע מגזירת הביטוי (לפי λ), אז נוציא לוג:

$$\log L = -\lambda n + \sum_{i=1}^{n} k_i \, log\lambda - \cancel{\sum_{i=1}^{n} \log(k_i!)}$$

גוזרים לפי λ לכן הביטוי הזה הוא קבוע.

$$\log L = -\lambda n + log\lambda \cdot \sum_{i=1}^{n} k_i$$
$$\frac{d}{d\lambda} L(\lambda) = -n + \frac{1}{\lambda} \cdot \sum_{i=1}^{n} k_i$$

לכן קיבלנו:

$$\lambda = \frac{\sum k_i}{n}$$

כלומר ה-MLE עבור מודל פואסוני הוא $\lambda = \frac{\sum k_i}{n}$.

דוגמה של מודל נורמלי:

$$\theta = \{\mu, \sigma\}$$
$$D = \{x_1, x_2, \dots, x_n\}$$

נרצה למצוא את μ, σ. נרצה למצוא את פונקציית הנראות (Likelihood). לשם כך נכתוב את פונקציית הצפיפות:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \ e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
$$L(\theta) = P(D|\theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \cdot \left(\frac{1}{\sigma}\right)^n \cdot \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

כעת נוציא $\log L$:

$$\log L(\theta) = -nlog\left(\sqrt{2\pi}\right) - n\log\sigma - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$$

נרצה להשוות את הגרדיאנט ל-0, לכן נרצה לגזור לפי μ ולפי σ

נגזור לפי μ:

$$\frac{d}{d\mu}\log L(\theta) = -\sum_{i=1}^{n}\frac{2\cdot(x_i-\mu)}{2\sigma^2} \cdot (-1)$$
$$\frac{1}{\sigma^2}\cdot\sum_{i=1}^{n}(x_i-\mu) = 0 \quad /\cdot\sigma^2$$
$$\sum_{i=1}^{n}(x_i-\mu) = 0 \quad \Rightarrow \quad n\cdot\mu = \sum_{i=1}^{n}x_i$$
$$\mu = \frac{1}{n}\sum_{i=1}^{n}x_i$$

נגזור לפי σ (לא הראנו בכיתה את הגזירה) ונקבל:

$$\sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

זוהי השונות האימפירית

EM – Expectation Maximization

במקרים רבים של למידה חישובית, יהיו ברשותנו רק חלק מה-features הרלוונטיים לנו לחישוב. אם יש משתנים שאנו לפעמים לומדים ולפעמים לא, נרצה להשתמש במקרים בהם כן למדנו את המשתנים הללו על מנת לתת פרדיקציה לערכם כאשר אנו לא לומדים אותם.

נוכל להשתמש באלגוריתם במקרים בהם יש משתנים שמעולם לא למדנו ולבצע פרדיקציה טובה באופן יחסי.

זאת שיטה איטרטיבית למצוא את הוקטור θ_{ML}.

יש 3 שכבות של data- : C = (X,Z) כאשר:

- **C**- המידע בשלמותו.
- **X**- המידע שאנו רואים.
- **Z**- המידע המוסתר (המידע החסר).

אנו מעוניינים להסיק את C מתוך X.

האלגוריתם מתחלק ל-2:

שלב E- Expectation

בשלב זה אנו יוצרים פונקציית expectation של הlog likelihood משוערך על ידי הפרמטרים הקודמים שמצאתי.

שלב M- Maximization

חישוב מחדש של היפותזת maximum likelihood באמצעות הערכים החזויים של המשתנים החבויים.

<u>האלגוריתם:</u>

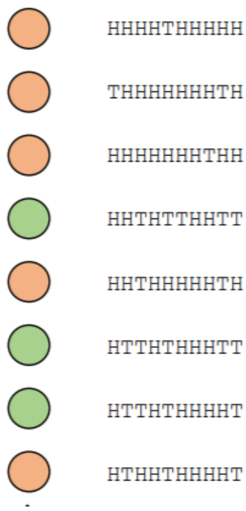
1. ננחש את הנקודה ההתחלתית.
2. נשערך את ערכי הdata החסר, כלומר נחשב את הסיכוי לכל data שהוא הגיע מהתפלגות A או התפלגות B. (חישוב responsibilities)
3. נעדכן את הפרמטרים בהתאם לתוצאות שקיבלנו
4. נעדכן מהיכן ה-data הגיע
5. נחזור על שלבים (2)-(4) כאשר הנקודה ש

<u>דוגמא</u>

יש 2 מטבעות עם הסתברויות P_A, P_B.

נבחר מטבע רנדומלי מבין ה-2 בהסתברות w_A, w_B ונטיל את המטבע 10 פעמים. נחזור על תהליך הבחירה 8 פעמים. אנו לא יודעים איזה מטבע מבין ה-2 הוטל, ולכן נרצה להשתמש באלגוריתם EM (אם היינו יודעים איזה מטבע הוטל היינו משתמשים בMLE)

תוצאות ההטלות:



W_A הוא ה-prior שלי, ובדוגמה זו $W_A = P(A)$, וכמובן $W_B = P(B)$.

ראשית, נאתחל את המשתנים הלא ידועים בערכים רנדומליים: $P_A = 0.6, P_B = 0.5, W_B = 0.5$.

נרצה כעת לחשב את ה-responsibilities, כלומר הaposterior probabilities של Bayes Classifier.

נחשב:

$$P_A(x_1) = w_A \binom{10}{9} 0.6^9 \cdot 0.4^1 = 0.04$$

$$P_B(x_1) = w_B \binom{10}{9} 0.5^9 \cdot 0.5^1 = 0.01$$

נשים לב ש: $P_A(x_1) > P\_B(x_1)$ לכן נבחר ב-$P_A(x_1)$.

נחשב $r(x_1, A) = \frac{p_A(x_i)}{p_A(x_i)+p_B(x_i)}$:

$$r(x_1, A) = \frac{0.04}{0.05} = 0.8$$

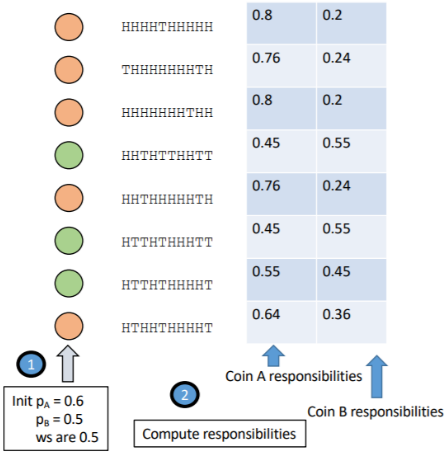
$$r(x_1, B) = \frac{0.01}{0.05} = 0.2$$

<u>הערה:</u> נשים לב שהם נסכמים ל-1 (לכן חילקנו ב-0.05).

נחזור על התהליך עבור כל אחד מ-10 ההטלות עד שנמלא את כל הטבלה.

המשך תרגיל המטבעות:

הטבלה שתתקבל אחרי שנבצע את 10 החישובים:



נעדכן את $new\ w_A, new\ w_B$ באמצעות הנוסחאות:

$$new\ w_A = \frac{1}{N}\sum_{i=1}^{N} r(x_i, A)\,, \qquad new\ w_B = \frac{1}{N}\sum_{i=1}^{N} r(x_i, B)$$

ונקבל:

$$New\ w_A = \frac{1}{8}\sum_{i=1}^{8} r(x_i, A) = 0.65$$

$$New\ w_B = \frac{1}{8}\sum_{i=1}^{8} r(x_i, B) = 0.35$$

נעדכן את הטבלה:

נעדכן את $new\ p_A, new\ p_B$ באמצעות הנוסחאות:

$$new\ p_A = \frac{1}{new\ w_A \cdot N}\sum_{i=1}^{N} r(x_i, A)v(i)\,, \qquad new\ p_B = \frac{1}{new\ w_B \cdot N}\sum_{i=1}^{N} r(x_i, B)v(i)$$

כאשר

- $\boldsymbol{v(i)}$– המספר שכתוב בעמודה האדומה. ה"דעה" של ההטלה לגבי מה המטבע הנוכחי. ערך $v(i)$ נובע מכמה פעמים ראיתי H. הערך אנלוגי לMLE.

נחזור על התהליך עד שנגיע להתכנסות או לנקודת עצירה כלשהי עליה נחליט.

תערובת גאוסיאנית – Guassian Mixture

נאמר שמשתנה מקרי x מתפלג Guassian Mixture אם פונקציית הצפיפות שלו היא:

$$f(x) = \sum_{i=1}^{k} w_i f_i(x)$$

כך שפונקציית הצפיפות של כל f_i היא פונקציית צפיפות גאוסיאנית:

$$f_i(x) = \frac{1}{\sigma_i\sqrt{2\pi}}\, e^{-\frac{(x-\mu_i)^2}{2\sigma^2}}$$

ומתקיים:

$$\sum_{i=1}^{n} w_i = 1$$

כאשר w_i הן משקולות המייצגות את הסיכוי להגריל גאוסייאן כלשהו.

נגדיר, N הוא הצפיפות הנורמלית:

$$N(x|\mu,\sigma\,) = \frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

נחשב את הresponsibilities:

$$r(x,k) = \frac{w_k N(x|\mu_k\,,\sigma_k)}{\sum_{j=1}^{K} w_j N(x|\mu_j\,,\sigma_j)}$$

ולבסוף נעדכן:

$$New\ w_j = \frac{1}{N}\sum_{i=1}^{N} r(x_i,j)$$

$$New\ \mu_k = \frac{1}{(New\ w_k)N}\sum_{i=1}^{N} r(x_i,k)x_i$$

$$(New\ \sigma_k)^2 = \frac{1}{(New\ w_k)N}\sum_{i=1}^{N} r(x_i,k)(x_i - New\ \mu_k)^2$$