1 esicrexE PLN

דניאל אזולאי ת"ז 311119895

אסף שול ת"ז 207042714

8 בנובמבר 2022

(1)

יהי מודל ביגרם כמוגדר בשאלה. נסמן $\Omega=\{START,w_1,...,w_n,STOP\}$

נתון $p\left(STOP|w\right)>0$, $\forall w\in\Omega \sum\limits_{u\in\Omega}P\left(u|w\right)=1$.

נניח בשלילה שקיימת סדרה אינסופית $v_0,v_1,v_2,...$ כך שלכל $i\in\mathbb{N}$ מתקיים $v_i\neq STOP$ וגם $P\left(v_0v_1v_2...\right)>0$.

כיוון שהמודל הוא $bigram$ מתקיים $P\left(v_j|v_{j-1}...v_0\right)=P\left(v_j|v_{j-1}\right)$ ולכן:

$$0<P\left(v_0v_1v_2...\right)=P\left(v_1|v_0\right)\cdot P\left(v_2|v_1\right)\cdot...$$

אזי נבחין כי הסדרה $\left\{P\left(v_j|v_{j-1}\right)\right\}_{j=1}^{\infty}$ מתכנסת ל־1, אחרת כיוון שהסדרה חסומה ע"י 1 נקבל

שניתן למצוא תת־סדרה שלה שמתכנסת למספר קטן מ־1,

כלומר תת סדרה שקטנה ממש החל ממקום מסויים ממכפלה אינסופית של מספר $p<1$, ועל כן המכפלה תהיה 0 בסתירה להנחה.

נבחין כי מספר האפשרויות של מספרים שונים ל־ $P\left(v_j|v_{j-1}\right)$ הוא מספר סופי,

כי יש מספר סופי של זוגות סדורים מעל Ω (חסום ע"י $(n+2)^2$). על כן קיים מקסימום $t=\max\limits_{j\in\mathbb{N}}P\left(v_j|v_{j-1}\right)\leq1$.

אם $t<1$ אז $\left\{P\left(v_j|v_{j-1}\right)\right\}_{j=1}^{\infty}$ לא מתכנסת ל־1 בסתירה למה שהוכחנו. לכן $t=1$ כלומר קיים $j\in\mathbb{N}$ כך שמתקיים $P\left(v_j|v_{j-1}\right)=1$.

כיוון שלקחנו סדרה שמקיימת $v_i\neq STOP$ לכל $i\in\mathbb{N}$ נקבל כי:

$$P\left(STOP|v_{j-1}\right)=1-\sum_{v\neq STOP}P\left(v|v_{j-1}\right)\leq1-P\left(v_j|v_{j-1}\right)=1-1=0$$

בסתירה להנחה $\forall w\in\Omega\ p\left(STOP|w\right)>0$.

כלומר קיבלנו כי ההסתברות לקבל סדרה אינסופית היא 0 ולכן ההסתברות של המאורע המשלים היא 1,

כלומר ההסתברות לקבל סדרה סופית הוא 1. כלומר מתקיים:

$$\sum_{n\in\mathbb{N}_{\geq 0}}\sum_{v_0,...,v_n} P\left(v_0,...,v_n\right) = P\left(\text{finite seq.}\right) = 1 - P\left(\text{infinite seq.}\right) = 1 - P\left(v_0v_1v_2...\right) = 1-0 = 1$$

כנדרש.

(2)

יהי אוצר מילים V וקורפוס C שמכיל מופע של כל מילה ב־V .

(א)

נסמן את מספרי המופעים של "where","were" ב־ $n_{wh} > 0, n_{we} > 0$ בהתאמה.

נגדיר מודל שפה UNIGRAM באופן הבא:

לכל $w \in V$, נסמן n_w להיות מספר המופעים של w ב־C .

וכעת לכל $i \in \mathbb{N}_{\geq 0}$ נגדיר מרחב הסתברות כאשר $\Omega_i = V$, המוגדר על יחידונים כך: $P_i\left(w\right) = \frac{n_w}{\sum\limits_{u\in V} n_u}$.

נבחין כי זהו אכן מרחב הסתברות: בהגדרת P_i המספרים $\left\{n_u\right\}_{u\in V}$ הם רק מספרים חיוביים מהנתון,

והמונה משתתף בסכימה במכנה, ולכן $0 < P_i\left(w\right) \leq 1 \quad \forall w$. כמו־כן הוא כמובן נסכם ל־1:

$$\sum_{w\in V} P_i\left(w\right) = \sum_{w\in V}\frac{n_w}{\sum\limits_{u\in V} n_u} = \frac{\sum\limits_{w\in V} n_w}{\sum\limits_{u\in V} n_u} = 1$$

וכיוון שזה הוגדר על יחידונים (ההסתברות לקבל מילה בודדה) זה מספיק כדי להגדיר מרחב הסתברות

(אין צורך לבדוק אדיטיביות כי מרחב ההסתברות מוגדר מהיחידונים כך שיהיה אדיטיבי).

כעת נגדיר את מרחב ההסתברות של מודל השפה באופן הבא:

$$\Omega = \prod_{i=0}^{\infty}\Omega_i$$
$$P\left(w_0,...,w_n\right) = \prod_{i=0}^{n} P_i\left(w_i\right)$$

זהו בעצם מרחב המכפלה של מרחבי ההסתברות שהגדרנו לעיל, ועל־כן הוא אכן מרחב הסתברות.

מאופן הגדרת מרחב המכפלה

$$P\left(w_n | w_0...w_{n-1}\right) = \frac{P\left(w_0...w_n\right)}{P\left(w_0...w_{n-1}\right)} = \frac{\prod\limits_{i=0}^{n} P_i\left(w_i\right)}{\prod\limits_{i=0}^{n-1} P_i\left(w_i\right)} = P_n\left(w_n\right) = P\left(w_n\right)$$

לכן זהו אכן מודל unigram (כי ההסתברות ל־w_n לא תלויה בהקשר).

בהנתן המשפט “He went **where** there **where** more opportunities” , נקבל כי:

- המופע הראשון יזוהה נכונה כ־where אמ"ם $P\left(\text{"where"}\right) > P\left(\text{"were"}\right)$ אמ"ם $\frac{n_{wh}}{\sum\limits_{u\in V} n_u} > \frac{n_{we}}{\sum\limits_{u\in V} n_u}$ אמ"ם $n_{wh} > n_{we}$, כלומר אמ"ם מספר המופעים של “where” ב־C גדול ממספר המופעים של “were” .
- המופע השני יזוהה נכונה כ־were אמ"ם $P\left(\text{"where"}\right) < P\left(\text{"were"}\right)$ אמ"ם $\frac{n_{wh}}{\sum\limits_{u\in V} n_u} < \frac{n_{we}}{\sum\limits_{u\in V} n_u}$ אמ"ם $n_{wh} < n_{we}$, כלומר אמ"ם מספר המופעים של “where” ב־C קטן ממספר המופעים של “were” .
- משני הסעיפים הקודמים נקבל כי שני המופעים יזוהו נכונה אמ"ם $n_{wh} < n_{we}$ וגם $n_{wh} > n_{we}$, כלומר זה לעולם לא יקרה.

(ב)

נגדיר מודל שפה BIGRAM באופן הבא:

לכל $w, u \in V$, נסמן n_{wu} להיות מספר המופעים של (w, u) בקורפוס (כאשר u מופיע לאחר w).

נגדיר הסתברויות מעבר באופן הבא:

$$P\left(u|w\right) = \frac{n_{wu}}{\sum\limits_{v\in V} n_{wv}}$$

נבחין שהמכנה אינו 0 כי כל מילה ב־V מופיעה בקורפוס ולכן יש גורם שאינו מתאפס בסכימה במכנה.

כמו־כן סכום הסתברויות המעבר מכל מילה הוא 1:

$$\sum_{u\in V} P\left(u|w\right) = \sum_{u\in V} \frac{n_{wu}}{\sum\limits_{v\in V} n_{wv}} = \frac{\sum\limits_{u\in V} n_{wu}}{\sum\limits_{v\in V} n_{wv}} = 1$$

כיוון שהמעברים מוגדרים היטב מכל מילה (אי־שליליים ונסכמים ל־1), נקבל כי המודל המרקובי מוגדר היטב, ומאופן הגדרתו כתלוי במילה האחרונה הוא מודל BIGRAM כנדרש.

- מודל זה עשוי להיות טוב יותר ממודל ה־UNIGRAM בסעיף הקודם, כי הוא שומר במידה מסויימת על ההקשר: יש מילים שאמנם פחות נפוצות לבדן, אך מאוד נפוצות כחלק מצירוף של שתי מילים - הזכרנו בכיתה למשל את הדוגמה של “San”-”Francisco” , אכן נצפה שאחרי “San” נקבל את “Fransisco” ולא את המילה “the” שהיא כנראה המילה הנפוצה ביותר בפני עצמה.
- משפט עשוי לקבל במודל זה הסתברות 0: אם קיים במשפט כלשהו צמד מילים wu שאינו מופיע בקורפוס, אזי $n_{wu} = 0$, כלומר גם $P\left(u|w\right) = 0$ ואז כיוון שמודל BIGRAM יתן את מכפלת ההסתברויות המותנות של כל צמד, המכפלה תתאפס ונקבל הסתברות 0.

- הנקודה הקודמת עשויה להיות בעיה של המודל, כי מספר הזוגות של מילים הוא משמעותית גדול יותר ממספר המילים, ועל כן למרות שנתון היה לנו שכל מילה ב־V מופיעה בקורפוס, ממש לא בהכרח כל צמד יופיע, ובמקרה זה עלולים להיות צמדים הגיוניים שהיינו עשויים לפספס.

(3)

(א) לכל $w \in V$ נסמן ב־$c(w)$ את מספר ההופעות של w. נראה כי מתקיים:

$$\sum_{w\in V}\frac{(c(w)+1)\,N_{c(w)+1}}{N_{c(w)}\cdot N}=1-p_{unseen}$$

ראשית נשנה סדר סכימה בסכום הנתון לעיל כך שנסכום לפי השכיחות של מילים קודם:

$$\begin{aligned}\sum_{w\in V}\tfrac{(c(w)+1)N_{c(w)+1}}{N_{c(w)}\cdot N}&=\sum_{c=1}^{c_{max}}\sum_{\substack{w\in V\\c(w)=c}}\tfrac{(c(w)+1)N_{c(w)+1}}{N_{c(w)}\cdot N}=\sum_{c=1}^{c_{max}}\sum_{\substack{w\in V\\c(w)=c}}\tfrac{(c+1)N_{c+1}}{N_c\cdot N}=\sum_{c=1}^{c_{max}}\left(\tfrac{(c+1)N_{c+1}}{N_c\cdot N}\cdot\sum_{\substack{w\in V\\c(w)=c}}1\right)=\\&=\sum_{c=1}^{c_{max}}\tfrac{(c+1)N_{c+1}}{N_c\cdot N}\cdot\left|\{w\in V\mid c(w)=c\}\right|=\sum_{c=1}^{c_{max}}\tfrac{(c+1)N_{c+1}}{N_c\cdot N}\cdot N_c=\sum_{c=1}^{c_{max}}\tfrac{(c+1)N_{c+1}}{N}\end{aligned}$$

נבחין כי משיקולי ספירה מתקיים $\sum_{c=1}^{c_{max}}cN_c=N$.

כמו־כן מתקיים $N_{c_{max}+1}=0$ ועל כן ניתן לבצע הזזת אינדקסים כזו:

$$\sum_{c=1}^{c_{max}}(c+1)\,N_{c+1}=\sum_{c=2}^{c_{max}+1}cN_c=\left(\sum_{c=2}^{c_{max}}cN_c\right)+(c_{max}+1)\underbrace{N_{c_{max}+1}}_{=0}=\sum_{c=2}^{c_{max}}cN_c$$

נשתמש בכל השוויונות הנ"ל כדי לקבל שמתקיים:

$$\begin{aligned}\sum_{w\in V}\tfrac{(c(w)+1)N_{c(w)+1}}{N}+\tfrac{N_1}{N}&=\left(\sum_{c=1}^{c_{max}}\tfrac{(c+1)N_{c+1}}{N}\right)+\tfrac{N_1}{N}=\tfrac{1}{N}\left(\left(\sum_{c=1}^{c_{max}}(c+1)\,N_{c+1}\right)+N_1\right)=\\&=\tfrac{1}{N}\left(\left(\sum_{c=2}^{c_{max}}cN_c\right)+N_1\right)=\tfrac{1}{N}\cdot\sum_{c=1}^{c_{max}}cN_c=1\end{aligned}$$

נחסר $\frac{N_1}{N}$ משני האגפים ונקבל:

$$\sum_{w\in V}\frac{(c(w)+1)\,N_{c(w)+1}}{N_{c(w)}\cdot N}=1-p_{unseen}$$

(ב) נזכור כי בהרצאה הגדרנו עבור מילה t את $c(t)$ להיות מספר המופעים של t בקורפוס.

בהתאם להוראות הסעיף בהינתן w מסויימת נסמן $c(w) := c$. אז כעת הנוסחה היא:

$$q_{add-One}(w) = \frac{c(w)+1}{\sum_{w'}(c(w')+1)} = \frac{c+1}{\sum_{w'}(c(w')+1)} = \frac{c+1}{N+\sum_{i=1}^{c_{max}} N_i}$$

כעת ניזכר כי $MLE(w) = \frac{c(w)}{\sum_{w'} c(w')}$ ובכתיב שלנו נקבל שזה בדיוק $MLE(w) = \frac{c}{N}$.

נבחר $\mu = \frac{N}{\sum_{i=1}^{c_{max}} N_i}$. נניח $c > \mu$ אז נבחין כי אנחנו עוסקים במספרים חיוביים ולכן כל המעברים הבאים תקינים:

$$\mu = \frac{N}{\sum_{i=1}^{c_{max}} N_i} < c \implies N < c \cdot \sum_{i=1}^{c_{max}} N_i \implies cN + N < cN + c \cdot \sum_{i=1}^{c_{max}} N_i \implies$$
$$\implies N(c+1) < c\left(N + \sum_{i=1}^{c_{max}} N_i\right) \implies \frac{c+1}{N+\sum_{i=1}^{c_{max}} N_i} < \frac{c}{N}$$

ובאופן סימטרי ניתן להפוך את אי השוויון בכל משוואה ולקבל את הגרירה עבור אי שוויון הפוך.

(ג) הראינו קודם שקיים μ כך שלכל $c > \mu$ מתקיים שהsmoothing הוא קטן מה־MLE ולכל $c < \mu$ הוא גדול מה־MLE .

השלילה של המשפט עבור גוד־טיורינג:

לכל μ קיים $c > \mu$ כך ש־ $\frac{(c+1)N_{c+1}}{N_c \cdot N} > \frac{c}{N}$ או קיים $c < \mu$ כך ש־ $\frac{(c+1)N_{c+1}}{N \cdot N_c} < \frac{c}{N}$

נכפול את שני האגפים ב־$N_c \cdot N$ בכל אחד משני האי שוויונות (הכל אי־שלילי) ונקבל שצריך להוכיח את המשפט:

לכל μ קיים $c > \mu$ כך ש־ $(c+1)N_{c+1} > cN_c$ או קיים $c < \mu$ כך ש־$(c+1)N_{c+1} < cN_c$.

נבחר קורפוס שמקיים: $N_1 = 6, N_2 = 2, N_3 = 3$ וגם $c_{max} = 3$ (כלומר $N_4 = N_5 = ... = 0$).

למשל אפשר לבחור את הקורפוס שכולל את שלושת המשפטים הבאים:

None of these words appear twice

two two times times

Three Three Three Hello Hello Hello World World World

כעת נוכיח את הטענה השקולה שהראינו למעלה. יהי $\mu \in \mathbb{R}$.

אם $\mu > 3$ אז נבחר $c = 3 < \mu$ ואז עבורו אכן מתקיים:

$$(c+1)N_{c+1} = 4 \cdot N_4 = 4 \cdot 0 = 0 < 3 \cdot 3 = 3 \cdot N_3 = cN_c$$

אם $1 < \mu \leq 3$ אז נבחר $c = 1 < \mu$ ואז עבורו אכן מתקיים:

$$(c+1)\,N_{c+1} = 2 \cdot N_2 = 2 \cdot 2 < 1 \cdot 6 = 1 \cdot N_1 = cN_c$$

אם $\mu \leq 1$ אז נבחר $c = 2 > \mu$ ומתקיים:

$$(c+1)\,N_{c+1} = 3N_3 = 3 \cdot 3 > 2 \cdot 2 = 2N_2 = cN_c$$

כיסינו את כל המקרים עבור $\mu \in \mathbb{R}$, ובכל אחד מהמקרים קיבלנו סתירה,

ולכן עבור קורפוס זה לא קיים μ שהוא threshold כמתואר בסעיף ב' עבור משערך smooth Good-Turing .

(4)

(א) המשוואה עבור מודל $TRIGRAM$:

$$P\left(w_n|w_0...w_{n-1}\right) = P\left(w_n|w_{n-2}w_{n-1}\right)$$

ואז נקבל כי:

$$P\left(w_0...w_n\right) = P\left(w_n|w_0...w_{n-1}\right) \cdot P\left(w_{n-1}|w_0...w_{n-2}\right) \cdot ... \cdot P\left(w_0\right) = P\left(w_0\right) \cdot P\left(w_1|w_0\right) \cdot \prod_{i=2}^{n} P\left(w_i|w_{i-2}w_{i-1}\right)$$

ההנחה במודל זה היא שההסתברות לקבל מילה תלויה אך ורק בשתי המילים שקדמו לה.

או במילים אחרות, שאין תלות בין מילה למילים שקדמו ל־2 המילים שקדמו לה.

(ב) משפט בעברית: "משחק הילדים **מאתגר** אותי".

במשפט זה בגלל שאנחנו במודל TRIGRAM נבין מההקשר שאחרי שתי המילים "משחק הילדים", המילה הבאה מתייחסת למשחק,

בשונה ממודל BIGRAM למשל, שסביר שהיה נוטה לתת פועל עם ריבוי, כי המילה האחרונה הייתה ילדים.

משפט באנגלית: "A cat usually **drinks** milk".

באותו אופן, מודל ה־TRIGRAM שלנו יתן עדיפות להטיה הנכונה של הפועל drink כך שיתאים ל"חתול" (יחיד),

על סמך המילה חתול שהופיעה כחלק מצמד המילים שקדמו למילה.

(ג) משפט בעברית: "המשחק של הילדים האלו **מאתגר** אותי"

עבור משפט זה צריך מודל 5-GRAM כי הנושא הוא המילה "המשחק", אשר מופיע 4 מילים לפני הפועל "מאתגר",

ושתי המילים האחרונות לפני הפועל אינן מסגירות האם הנושא הוא יחיד או רבים - להפך.

משפט באנגלית: "The cats from across the street **drink** milk"

עבור משפט זה צריך מודל 6-GRAM כי הנושא הוא החתולים והוא מופיע 5 מילים לפני הפועל,

ושתי המילים האחרונות לפני הפועל אינן מסגירות האם הנושא הוא יחיד או רבים - להפך.

(5)

נציג שלושה צירופי מילים שהם בעצמם לא חלק מאף משפט תקני אך כל k מילים צמודות שלהן ניתן לכתוב בתוך משפט תקני,

עבור $k = 2, 3, 4$ (לפי סדר עולה). לכל צירוף מילים, נכתוב משפט תקני לכל k-יה רצופה בצירוף.

שתה אכל שתה:

כל מה שהאיש שתה אכל האיש השני

כל מה שהאיש אכל שתה האיש השני

שתה לא אכל שתה:

כל מה שהאיש שתה לא אכל האיש השני

כל מה שהוא לא אכל שתה האיש השני

שתה לא אכל לו שתה:

כל מה שהאיש שתה לא אכל לו האיש השני

כל מה שהוא לא אכל לו שתה האיש השני