NLP (67658) | תרגיל 1

יואב שפירא| ת"ז: 312492838 & נדב פוקס| ת"ז: 206073108

13 בנובמבר 2022

חלק תאורטי

שאלה 1

ראשית נראה שההסתברות ליצור רצף מילים $w_1, w_2....$ שלא נגמר במילה $stop$ כלומר לא נגמר אף פעם, שואפת ל־ 0. ההסתברות זו מחושבת על ידי:

$$\prod_{i=0}^{\infty} p((w_i \neq stop)|w_{i-1}) = \prod_{i=0}^{\infty}(1 - p(stop|w_{i-1})$$

משום שאוצר המילים שלנו הוא סופי, קיימת $\widehat{w}$ כך שהטרנזקציה $p(stop|\widehat{w})$ מינימלית וגם גדולה מ־0 ולכן מתקיים:

$$\prod_{i=0}^{\infty}(1 - p(stop|w_i) \leq \prod_{i=0}^{\infty}(1 - p(stop|\widehat{w}) = \prod_{i=0}^{\infty} C \longrightarrow^{*} 0$$

* משום שלכל קבוע $0 \leq C < 1$ מתקיים $\prod_{i=0}^{\infty} C \longrightarrow 0$.

לכן אם ההסתברות של כל המשפטים האינסופיים שואפת ל־0 ההסתברות המשלימה היא שההסתברות של כל המשפטים הסופיים שואפת ל־1.

שאלה 2

(א)

נגדיר $markov\ LM$ מסדר $k=1$.
לכל מילה $\omega_j \in WORDS \cup \{START, STOP\}$ (כאן $WORDS$ הוא מאגר המילים שבקורפוס), נגדיר:

$$\mathbb{P}(\omega_j) = \frac{\{\#\omega_j\}}{|WORDS|}$$

כלומר ההסתברות של כל מילה להופיע, היא לפי השכיחות שלה. במודל הזה, ההסתברות לקבל את המשפט $(\omega_1...\omega_n)$ היא:

$$\mathbb{P}(\omega_1...\omega_n) = \prod_{i=1}^{n} \mathbb{P}(\omega_i)$$

נתון שכל המילים מופיעות בהסתברות גדולה מ0 ולכן לא נבצע $smoothing$.
(זהו מודל מאוד מעפן, שבו כנראה המשפט $the\ the\ the\ the$ הוא די סביר).
במקרה שלנו, בין המילים $where$ ו$were$, תיבחר המילה שנפוצה יותר בקורפוס, ללא שום קשר להקשר שלה במשפט.
לדוגמא עבור המשפט $He\ went\ where\ there\ where\ more\ opportunities$, נקבל תיקון נכון מהאלגוריתם עבור המופע הראשון של $where$, רק במקרה שבו המילה $where$ מופיעה יותר פעמים מהמילה $were$ בקורפוס. נקבל תיקון נכון עבור המופע השני של $where$, במקרה ההפוך שבו $were$ מופיעה יותר פעמים.
תיקון נכון לשני המופעים, יוכל להתקבל רק בעזרת הגדרה הסתברותית חדשה של המודל: עבור שתי מילים ω_i, ω_j כך שמתקיים $\mathbb{P}(\omega_i) = \mathbb{P}(\omega_j)$, תבחר באקראי או את ω_i או את ω_j. כך, אם יש קורפוס שבו כמות המופעים של $where$ שווה לכמות המופעים של $were$, יכול להיות שנקבל תיקון נכון עבור שני המופעים פשוט $by\ chance$.

(ב)

בדומה, נגדיר $markov\ LM$ מסדר $k=2$.
לכל צמד מילים $\omega_j, \omega_i \in WORDS \cup \{START, STOP\}$ נגדיר:

$$\mathbb{P}(\omega_j|\omega_i) = \frac{\{\#(\omega_i\omega_j)\}}{|WORDS|^2}$$

כלומר ההסתברות של כל מילה להופיע, תלויה במילה שלפניה ולמעשה אנחנו בודקים את כל הצמדים בקורפוס, ומוצאים את השכיחות שלהם. במודל הזה, ההסתברות לקבל את המשפט $(\omega_1...\omega_n)$ היא:

$$\mathbb{P}(\omega_1...\omega_n) = \prod_{i=1}^{n} \mathbb{P}(\omega_i|\omega_{i-1})$$

המודל הזה יותר טוב מהמודל הקודם כי הוא לוקח בחשבון את ההקשר של המילים, ולכן למשל הגיוני שהוא יחזה ש$went\ were$ יהיה פחות סביר מאשר $went\ where$, ובצדק.

כאן נתון לנו שכל מילה מופיעה יותר מפעם אחת בקורפוס, אבל לא נתון לנו את זה על כל הצמדים של המילים. לכן, אם נשאיר את המודל כמו שהוא, אכן יכול להיות מצב שבו ההסתברות $\mathbb{P}(\omega_1 ... \omega_n) = 0$, כי יכול להיות שקיים $i \leq n$ כך ש $\mathbb{P}(\omega_i|\omega_{i-1}) = 0$ (למשל, אם הצירוף $went\ were$ בכלל לא הופיעה בקורפוס, משפט שיכיל את הצמד הזה יקבל סבירות 0). זו אכן בעיה.
אבל כמו שראינו אפשר לעשות מניפולציות על פונקציות ההסתברות על מנת לגרום לכך שלכל i יתקיים $\mathbb{P}(\omega_i|\omega_{i-1}) > 0$, לדוגמא בעזרת $back - off\ model$.

שאלה 3

(א)

טענה:

$$\sum_{c=2}^{c_{max}} cN_c = N - N_1$$

הוכחה:
לכל c, מוגדר ש N_c הוא מספר המילים **הייחודיות** שמופיעות בקרופוס c פעמים. כדי לספור את כל מופעי המילים האלה בסה"כ, נצטרך לחשב את cN_c. לכן סכימה על כל c תתן לנו את סך כל מופעי המילים עבור כל c, כלומר את כל הקורפוס:

$$\sum_{c=1}^{c_{max}} cN_c = N$$

ומכאן ש

$$\sum_{c=2}^{c_{max}} cN_c = N - 1N_1 = N - N_1$$

אנחנו מתבקשים למצוא את $\sum_{c=1}^{c_{max}} \mathbb{P}\left(\omega_j | \{\#\omega_j\} = c\right)$. נשים לב שלא התבקשנו עבור $c = 0$, אלא התבקשנו עבור כל המילים שכן מופיעות בקורפוס.
מההגדרה ידוע שהסיכוי לבחור מילה ספציפית שמופיעה c פעמים הוא:

$$\forall \omega_j : \quad \mathbb{P}\left(\omega_j | \{\#\omega_j\} = c\right) = \frac{(c+1)N_{c+1}}{N_c N}$$

לכן, הסיכוי לבחור מילה כלשהי שמופיעה c פעמים צריך להיות מוכפל בN_c, כי יש N_c מילים שונות כאלו:

$$\forall c : \quad \mathbb{P}\left(\omega_j | \{\#\omega_j\} = c\right) = \frac{(c+1)N_{c+1}}{N_c N} \cdot N_c = \frac{(c+1)N_{c+1}}{N}$$

נסכום על כל $c \geq 1$:

$$\begin{aligned}\sum_{c=1}^{c_{max}} \frac{(c+1)N_{c+1}}{N} &= \sum_{c=1}^{c_{max}-1} \frac{(c+1)N_{c+1}}{N} + \frac{(c_{max}+1)N_{c_{max}+1}}{N} \\ &\stackrel{1}{=} \sum_{c=1}^{c_{max}-1} \frac{(c+1)N_{c+1}}{N} \stackrel{2}{=} \sum_{c=2}^{c_{max}} \frac{cN_c}{N} \stackrel{3}{=} \frac{N-N_1}{N} \\ &\stackrel{4}{=} 1 - \frac{N_1}{N} = 1 - \mathbb{P}_{unseen}\end{aligned}$$

1: נתון שלכל $c > c_{max}$ מתקיים כי $N_c = 0$ ולכן בפרט $N_{c_{max}+1} = 0$.

2: אינדוקס מחדש של הסיגמא.

3: מטענת העזר.

4: מעבר ישיר לתוצאת הדרושה.

(ב)

תהי מילה כלשהי שמופיעה c פעמים ע"י ω_c. ההסתברות לבחור את ω_c עם שיטת $add-1$ היא:

$$\mathbb{P}_{add-1}(\omega_c) = \frac{c+1}{\sum_{c=1}^{c_{max}} (c+1)}$$

מכיוון שN הוא כלל המילים, מתקיים $\sum_{c=1}^{c_{max}} c = N$, ולכן

$$\mathbb{P}_{add-1}(\omega_c) = \frac{c+1}{2N}$$

לעומת זאת, הMLE נתון ע"י השכיחות של המילה בקורפוס. כלומר:

$$MLE(\omega_c) = \frac{c}{N}$$

בהינתן שN קבוע, קל לראות שעבור $c = 1$ הMLE וה$add-1$ שווים. עבור $c = 0$ יוצא שהMLE ערכו קטן מאשר $add-1$:

$$\mathbb{P}_{add-1}(\omega_0) = \frac{1}{2N} > 0 = MLE(\omega_0)$$

עבור $c = 2$ יוצא שערך הMLE גדול יותר מה$add-1$:

$$\mathbb{P}_{add-1}(\omega_2) = \frac{3}{2N} < \frac{2}{N} = MLE(\omega_2)$$

ברור שגם הMLE וגם $add-1$ הם ליניאריים ובפרט מונוטוניים ממש ולכן קיבלנו שיש סף יחיד $\mu = 1$ שמקיים את הדרוש.

(ג)

נראה שני מקרים בהם התכונה מהסעיף הקודמת לא מתקיימת לגבי $good-smoothing$:

1. מקרה בו לכל c מתקיים שערך ה$MLE(\omega_c)$ גבוה יותר: אם כל המילים בקורפוס הופיעו בדיוק פעם אחת. במקרה כזה, $c_{max}=1$, ולכן $N_{c+1}=0$. מכאן ש$\mathbb{P}_{good-smoothing}(\omega_c)=0$ בעוד ש $MLE(\omega_c)=\frac{1}{N}$.
2. מקרה הפוך: בקורפוס שבו עבור כל $c<c_{max}$ מתקיים $N_c=1$ כלומר כל $type$ של מילים מכיל מילה יחידה בלבד. במקרה כזה מתקיים $\forall_{i,j\leq c_{max}}: \quad N_i=N_j$, ולכן:

$$\forall c: \quad \mathbb{P}_{good-smoothing}(\omega_c)=\frac{(c+1)N_{c+1}}{N_cN}=\frac{c+1}{N}>\frac{c}{N}=MLE(\omega_c)$$

שאלה 4

א)

מודל $trigram$ נתון ע"י נוסחת ההסתברות הבאה:

$$\mathbb{P}\left(\omega_j|\omega_{j-1},\omega_{j-2}\right)=\prod_{j=1}^{n}\mathbb{P}\left(\omega_j|\omega_{j-1},\omega_{j-2}\right)$$

המודל הזה מניח שכל מילה איננה תלויה במילה ה3 שמופיעה לפניה, אלא רק בשתיים שלפניה.

ב)

דוגמא בעברית: צורת רבים או מול צורת יחיד: המילה "כלבים" מופיעה באותה השלשה עם הפועל "נובח" ולכן המודל יפרש את הצורה נכונה:
"יש **כלבים** שלא **נובחים**".
דוגמא באנגלית: המילה dog מופיעה בצורת יחיד בתוך אותה שלשה עם הפועל $bark$, ולכן המודל יפרש את הצורה נכונה:

In israel, a **dog** only bark**s** at list one time a day.

ג)

דוגמא בעברית: המילה "כלבים" **לא** מופיעה באותה השלשה עם הפועל "נובח" ולכן המודל יפרש את הצורה לא נכונה:
"יש **כלבים** גדולים שלא **נובחים**. נצטרך כאן מודל 4־גרמי.
דוגמא באנגלית: המילה dog **לא** מופיעה באותה שלשה עם הפועל $bark$, ולכן המודל לא יפרש את הצורה נכונה (כאן נצטרך מודל ־6גרמי)

In israel, a **dog** that have vocal cords, bark**s** atlist one time a day.

שאלה 5

האישה שישבה על כיסא עץ שקד נפל ממנו.
האישה שישבה על כיסא עץ שקד גדול נפל ממנו.
האישה שישבה על כיסא עץ שקד גדול כמעט נפל ממנו.

דוגמאות אלו מראות שלא משנה כמה תגדיל את מודל מרקוב הוא לא יצליח לתפוס את כל השפה משום שתמיד יהיה קיים משפט שנצטרך יותר תלויות כדי לעבד את כולו.

חלק פרקטי

(א)

מצורף קוד

(ב)

לאחר אימון המודל הביגראמי, הוא השלים את המשפט כך:

$$I\ have\ a\ house\ in\ the$$

(ג)

.1

(א)

$$\mathbb{P}\left(Brad\ Pitt\ was\ born\ in\ Oklahoma\right) = -\infty$$

(ב)

$$\mathbb{P}\left(The\ actor\ was\ born\ in\ USA\right) = -29.74$$

2. סהכ בשני המשפטים יש 12 מילים. אז ה$perplexity$:

$$e^{-l}\ ,\ l=\frac{1}{12}\left(-\infty-29.74\right)=-\infty$$

ולכן

$$perplexity=e^{-(-\infty)}=\infty$$

(ד)

1.

(א)

$$\mathbb{P}\left(Brad\ Pitt\ was\ born\ in\ Oklahoma\right)=-36.18$$

(ב)

$$\mathbb{P}\left(The\ actor\ was\ born\ in\ USA\right)=-31.04$$

2. $perplexity$:

$$perplexity=e^{\frac{-(-36.18-31.04)}{12}}=271.11$$