תרגיל 1 | NLP (67658)

שם: רונאל חרדים, עומרי טויטו | ת"ז:208472761, 208917641

חלק I

תיאורטי:

שאלה 1

נוכיח את הדרוש:

תחילה נוכיח את הרמז - נוכיח כי ההסתברות המשלימה לעולם לא לקבל $stop$ שווה ל 0 :

נגדיר משפט באורך n להיות S_n, ונגדיר את ההסתברות לקבל את המשפט להיות $P(S_n)$ ואת המאורע נגדיר להיות $A = \sum_{n=0}^{\infty} P(S_n)$.

כעת נסתכל על המאורע המשלים - A^C, ההסתברות למאורע שווה ל $P\left(A^C\right) = \sum P(S_\infty)$ מכיוון שההסתברות של המאורע המשלים שווה לסכום ההסתברויות של כל המשפטים האינסופיים.

יהי משפט אינסופי $S_\infty = w_1...w_n w_{n+1}...$ נחשב את ההסתברות לקבל את המשפט.

$$P\left(S_\infty\right) = P\left(w_1 \mid START\right) \cdot P\left(w_2 \mid w_1\right) \cdots$$

כאשר $STOP \notin S_\infty$ מכיוון שהמשפט אינסופי.

מכיוון שלכל מילה w_j מתקיים כי $\sum P(w_i|w_j) = 1$ וגם $P(STOP \mid w_j) > 0$.

נוסע מכך כי לכל זוג מילים $w_i, w_j \neq STOp$ מתקיים כי $P(w_i|w_j) < 1$ לכן נקבל:

$$P\left(A_\infty\right) = \prod_{w_i, w_j \neq STOP} P\left(w_i \mid w_j\right)$$

קיבלנו מכפלה אינסופית של איברים שקטנים מ 1 ולכן המכפלה תשאף ל 0 , ולכן:

$$P(A^C) = 0 = 1 - P(A) \Rightarrow P(A) = 1$$

כנדרש.

שאלה 2

(א)

נתאר מודל $unigram$**:**
בתהליך הלמידה המודל יספור את מספר הפעמים שכל אחת מהמילים מופיעה ויחלק בסך כל המילים, כך נקבל את ההסתברות לכל מילה. כלומר, עבור $|corpus| = n$:

$$P(where) = \frac{count(where)}{n}, P(were) = \frac{count(were)}{n}$$

כעת, כאשר הוא יצטרך לחזות את המילה הבאה הוא יסתכל על ההסתברות שלה ויחזיר את המילה עם ההסתברות הגבוה ביותר. עבור משפט - המודל יחשב את מכפלת הסתברות המילים שמרכיבות את המשפט.
המודל יפעל כך:
בהינתן קלט של משפט הוא יכפול את ההסתברות לקבל את המילים שמופיעות במשפט ויחזיר את המשפט עם ההסתברות הגבוה ביותר שיכולה להתקבל.
לכן:

- המודל יחזה את המילה $where$ הראשונה רק אם מספר הפעמים שהמילה $where$ הופיעה ב $corpus$ גדולה ממספר הפעמים שהמילה $were$ הופיעה בו.
- המודל יתקן את המילה $where$ השניה להיות $were$ רק אם המילה $were$ מופיעה יותר פעמים מהמילה $where$ ב $corpus$.
- המודל יחזיר את אותה ההסתברות עבור שתי המילים אם שתיהן מופיעות אותו מספר של פעמים ב $corpus$.

(ב)

נתאר מודל $Bigram$**:**
המודל מקבל משפט ומתקן אותו בהתאם להסתברות של המילה x_i להופיע אחרי המילה x_{i-1}. לכן בהינתן משפט המודל מחזיר משפט שלדעתו עם ההסתברות הגבוה ביותר להיות נכון.

מודל זה יהיה מודל טוב יותר משום ש - הוא בודק את ההסתברות לקבל את המילה $where$ בהינתן המילה $went$. ואת ההסתברות לקבל את המילה $were$ בהינתן המילה $went$. כמו כן הוא יבדוק את ההסתברות לקבל כל אחת מהמילים $where, were$ בהינתן המילה $there$. לכן ודל זה יוכל לתקן באופן טוב יתר את הטעות מאשר מודל $unigram$ שמסתכל על מילה אחת בלבד.

משפט יכול לקבל הסתברות 0 לפי המודל הזה, אם המשפט הזה לא הופיע ב $corpus$. זאת יכולה להיות בעיה אם משפט זה הוא נכון אך נדיר ולכן לא נצפה קודם.

שאלה 3

(א)

אנו יודעים כי $punseen = \frac{N_1}{N}$, לכן:

$$1 - punseen = 1 - \frac{N_1}{N} = \frac{N - N_1}{N}$$

כעת, נחשב עבור כל מספר מילים c שנמצאות ב $corpus$: נשים לב כי קיבלנו טור טלסקופי וכל האיברים מצטמצמים למעט האיבר הראשון והאחרון -

$$\sum_{c=1}^{C_{max}} \frac{(c+1)N_{c+1}}{N_c \cdot N} \cdot N_c = \frac{1}{N} \cdot \left(\sum_{c=1}^{C_{max}} \frac{(c+1)N_{c+1}}{N_c} \cdot N_c \right) = \frac{1}{N} \cdot \left(\sum_{c=1}^{C_{max}} (c+1)N_{c+1} \right) =$$

$$\frac{1}{N} \cdot \left(\sum_{c=1}^{C_{max+1}} c \cdot N_{c+1} \right) = \frac{1}{N} \cdot \left(\sum_{c=2}^{C_{max+1}} c \cdot N_{c+1} - N_1 \right) = \frac{N - N_1}{N} = 1 - punseen$$

כנדרש.

(ב)

נכתוב את המודל:

$$q_{add-1}(w) = \frac{c+1}{N + \sum_{d=1}^{C_{max}} N_d}$$

כאשר $MLE = \frac{c}{N}$.
ומתקיים:

$$\frac{c+1}{N + \sum_{d=1}^{C_{max}} N_d} < MLE = \frac{c}{N} \iff N \cdot c + N < N \cdot c + c \cdot \sum_{d=1}^{C_{max}} N_d \iff N < c \cdot \sum_{d=1}^{C_{max}} N_d$$

מסעיף א, נובע כי לכל $d < c_{max}$ מתקיים כי $N_d > 0$ ולכן בפרט מתקיים כי $\sum_{d=1}^{C_{max}} N_d > 0$ ולכן:

$$\iff \frac{N}{\sum_{d=1}^{C_{max}} N_d} < c$$

כעת נסמן $\mu = \frac{N}{\sum_{d=1}^{C_{max}} N_d}$ ונובע ממה שחישבנו כי עבור מילה w שמקיימת $\mu < c(w)$ ה $add-one$ נמוך מה - MLE שלה.

מכיוון שכל הגרירות היו גרירות דו כיווניות, נקבל כי כל מילה w שמקיימת $c(w) < \mu$ הא"ש ההפוך מתקיים ולכן הגבול הנדרש בשאלה מתקיים גם הוא כנדרש.

(ג)

נראה כי מה שכתבנו בסעיף ב לא בהכרח מתקיים עבור $smoothed\ Good-Turing$ **:**
כלומר קיים $courpos-V$ כך שעבור מילה $w \in V$ מתקיים:

$$c(w) < \mu = \frac{N}{\sum_{d=1}^{C_{\max}} N_d}$$

אך

$$\frac{(C(w)+1)N_{C(w)+1}}{N_{C(w)} \cdot C} \leq \frac{C(w)}{N}$$

נגדר את V להיות $V = \{the, the, an, a, is\}$.
ואנו נקבל:

$$N = 5, \sum_{d=1}^{2} N_d = 4 \Rightarrow \mu = \frac{5}{4}$$

ועבור המילה is נקבל כי:

$$\frac{C(w)N_2}{N_1 \cdot N} = \frac{2 \cdot 1}{3 \cdot 5} = \frac{2}{15} \leq \frac{1}{5} = \frac{C(w)}{N}$$

ואכן $smoothed\ Good-Turing$ לא מקיים את סעיף ב כנדרש.

שאלה 4

(א)

נוסחת מודל ה $trigram$ היא הנוסחה הבאה:

$$p\left(x_i \mid x_{i-1} \ldots x_0\right) = \prod_{i=1}^{n} p\left(x_i \mid x_{i-1}, x_{i-2}\right)$$

המודל מניח כי מתקיימת אי תלות בין שתי המילים הקודמות x_{i-1}, x_{i-2} לבין שאר המילים שבמודל $x_0 ... x_{i-3}$.
המודל חוזה בדרך הבאה - הוא סופר את מפר ההופעות של שלשת המילים יחד חלקי ההופעות של כל המילים ב $corpus$.

(ב)

משפט באנגלית: "$The\ dogs\ is$" VS "$The\ dogs\ are$", המודל יצליח לזהות כי המילה הנכונה העוקבת היא are ולא is צשום שהוא מסתכל על שתי המילים האחרונות ומסתכל על ההסתברות הגבוה ביותר לקבל את המילה השלישית.

משפט בעברית: "מה את אוכלת" VS "מה את אוכל". המודל יצליח לחזות כי המילה הנכונה היא "אוכלת", ובהתאם לכך המשפט הראשון הוא הנכון. בהינתן ההסתברות שלה להופיע אחרי שתי המילים "מה את".

(ג)

משפט באנגלית:

$"They\ didn't\ know\ how\ much\ \boldsymbol{their}\ help\ was\ needed"\ \ VS\ \ "They\ didn't\ know\ how\ much\ \boldsymbol{his}\ help\ was\ needed"$

מכיוון שהמודל מסתכל רק על שתי המילים האחרונות $how\ much$ הוא לא יידע אם המילה שוא צריך להשלים (או המילה הנכונה) היא $thier$ או his ולכן הוא ייכשל.

משפט בעברית: "אתה לא רוצה לבצע את המוטל **עליך**" VS "אתה לא רוצה לבצע את המוטל **עלייך**", במקרה זה המודל לא ידע לחזות את המילה הנכונה משום שהוא מסתכל על שתי המילים האחרונות "את המוטל" והוא לא יודע אם צריך להשלים צורת זכר או צורת נקבה.

שאלה 5

הדוגמאות הדרושות:

- **עבור 2 מילים:** חם מאד קשה
- **עבור 3 מילים:** את לא יותר חם
- **עבור 4 מילים:** היום לא ניתן לסדר מחר

המסקנה היא כי עבור מודל מרקוב מסדר n, ככל ש n גבוה יותר אנו נצליח לחזות בסיכוי גבוה יותר את המילה הנכונה, משום שכפי שנוכחנו לראות ככל שמסתכלים אחורה יותר, כך יותר קשה למצוא משפט לא הגיוני. לכן המודל אולי לא יצליח בחלק מהמשפטים, אך אלו יהיו משפטים נדירים.

חלק II

פרקטי:

שאלה 2

(א)

המילה הבאה שהמודל חזה היא: the

שאלה 3

(א)

1. ההסתברות לקבל את משפט 1 היא: $-inf$
2. ההסתברות לקבל את משפט 2 היא: -29.714080217781863

(ב)

ה $perplexity$ של המשפטים = inf.

שאלה 4

$linear\ interpolation\ smoothing$

1. ההסתברות לקבל את המשפט הראשון היא: -36.197090487266074
2. ההסבתרות לקבל את המשפט השני היא: -31.03218010933515

ה $perplexity$ של המשפטים = 271.0868410925168