תרגיל 1 - NLP

27 בינואר 2024

Avraham Asraf : 315774570
Schnaidman Elchanan: 316092436

חלק 1

סעיף a

לפי ההדרכה, נוכיח כי המשלים, כלומר הסיכוי למשפטים שלא נגמרים הוא 0.
נסמן

$$\Omega = \{s = (w_1, ...w_n) \,|n \in \mathbb{N} \cup \infty\}$$

מתקיים

$$\Omega = \Omega_{\mathbb{N}} \cup \Omega_{\infty}$$

נוכיח כי

$$\mathbb{P}\left(\Omega_{\infty}\right) = 0$$

מתקיים

$$\mathbb{P}\left(\Omega_{\infty}\right) = \sum_{s \in \Omega_{\infty}} \mathbb{P}\left(s\right)$$

כיון שמהאורעות זרים,
קיים $\varepsilon > 0$ המקיים

$$\varepsilon = \min_{w \in W} \{\mathbb{P}\left(\text{stop}|w\right)\}$$

אזי לכל $w_i, w_j \in W$ מתקיים

$$\mathbb{P}\left(w_i|w_1..., w_j\right) \leq 1 - \varepsilon$$

כיון שהסיכוי ל$\mathbb{P}\left(\text{stop}|w_1..., w_j\right) = \mathbb{P}\left(\text{stop}|w_j\right) \geq \varepsilon$
א"כ לכל $s \in \Omega_\infty$ מתקיים

$$\mathbb{P}\left(s\right) \leq \prod_{i=1}^{\infty} 1 - \varepsilon = \lim_{n \to \infty} \left(1 - \varepsilon\right)^n = 0$$

לכן

$$\mathbb{P}\left(\Omega_\infty\right) = \sum_{s \in \Omega_\infty} \mathbb{P}\left(s\right) = \sum_{s \in \Omega_\infty} 0 = 0$$

כנדרש.

b1.

במודל לא מרקובי המקיים את ההנחות לעיל יתכן שיהיה משפט אינסופי עם הסתברות גדולה מ0.

דוגמה:

יהי P מודל שפה על אוצר המילים $(\text{start}, \text{w}, \text{end})$ מוגדר בצורה הבאה:

$$P\left(\text{end}|s\right) = \begin{cases} 0 & s = (\text{start}) \\ \frac{1}{2^n} & s = (\text{start}, \text{w}..., \text{w}_n) \end{cases}, P\left(\text{w}|s\right) = \begin{cases} 1 & s = (\text{start}) \\ 1 - \frac{1}{2^n} & s = (\text{start}, \text{w}..., \text{w}_n) \end{cases}$$

זהו מודל מוגדר היטב:

$$\sum_{n\to\infty} P\left(s_n\right) = \frac{1}{2} + \frac{1}{4}.... + \frac{1}{2^n} = \frac{\frac{1}{2}}{1 - \frac{1}{2}} = 1$$

כאשר s_n מסמן את מספר הw במילה.

בכל רצף מילים קיים סיכוי חיובי להגיע לend

קיים סיכוי חיובי לטור אינסופי

$$\lim_{n\to\infty}\left(1 - \frac{1}{2^n}\right)^n > \lim_{n\to\infty}\left(1 - \frac{1}{n}\right)^n = \frac{1}{e} > 0$$

כנדרש.

a2.

במודל unigram קיימות שלוש אפשרויות:

$\mathbb{P}(\text{where}) > \mathbb{P}(\text{were})$ או $\mathbb{P}(\text{where}) < \mathbb{P}(\text{were})$ או $\mathbb{P}(\text{where}) = \mathbb{P}(\text{were})$ ואז ההחלטה בן where לwere תהיה לא דטרמניסטית.

אם $\mathbb{P}(\text{where}) > \mathbb{P}(\text{were})$ אזי מתקן השגיאות לא יחליף את ההופעה של were במקום הלא נכון.

אם $\mathbb{P}(\text{where}) < \mathbb{P}(\text{were})$ מתקן השגיאות יחליף את שתי ההופעות של where. במקרה בו הסיכוי שווה לא תהיה החלטה קבועה ובכל מקרה לא נקבל פתרון נכון בוודאות.

b2.

אם נשתמש במודל bigram

אם $\mathbb{P}(\text{were}|\text{went}) < \mathbb{P}(\text{where}|\text{went})$ וגם $\mathbb{P}(\text{were}|\text{there}) < \mathbb{P}(\text{where}|\text{there})$

כלומר יש יותר מקרים כאלה בקרפוס, אזי המודל יתקן נכון.

יתכן שמודל כזה יתן סיכוי 0 לצמד מילים שלא נמצאים בקורפוס, גם אם הוא הגיוני.

כמו כן מילה אחת לפני לא מספיקה בשביל לדעת האם המשפט נכון או לא. ויתכן שצימדי מילים שנמצאות הרבה יקבלו סיכוי גבוה ויחזו כנכונים גם במקרים לא הגיוניים.

עיבוד שפה טבעית

תרגיל 1

3. נענה לפי הסעיפים

א. עבור מילה שהופיע c פעמים, יש N_c מילים כאלה בקורפוס, לכן סכום התדירויות של סוג מילה כזה (כל המילים) הוא $\frac{(c+1)*N_{c+1}}{N_c*N} * N_c = \frac{(c+1)*N_{c+1}}{N}$.

כדי לקבל את סכום התדירויות של כלל המילים בקורפוס, נסכום על כל האפשרויות של c.

$$\sum_{c\in[c_{max}]} \frac{(c+1)*N_{c+1}}{N} = \frac{1}{N} * \sum_{c\in[c_{max}]} \frac{(c+1)*N_{c+1}}{N} \overset{*}{=}$$
$$= \frac{1}{N} * \sum_{c\in[c_{max}-1]} \frac{(c+1)*N_{c+1}}{N} = \overset{**}{=} \frac{1}{N} * \sum_{c\in[2,\ c_{max}]} \frac{c*N_c}{N} \overset{***}{=}$$
$$= \frac{1}{N} * (N - 1 * N_1) = \frac{N}{N} - \frac{N_1}{N} = 1 - p_{unseen}$$

מעבר *: מכיוון ש $N_{c+1} = 0$ ניתן להשמיט את האיבר האחרון

מעבר **: במקום להתחיל לספור מ-1 ולסיים ב- $c_{max} - 1$, מתחילים ומסיים במספר העוקב ומפחיתים איד בכל מופע של c.

מעבר ***: מתבצעת שם סכימה של כל המילים שמופיעים לפחות פעמיים, חוץ מהמילים שמופיעות פעם אחת.

ב. הנוסחה ל-Add-One הינה $q_{Add-One} = \frac{c+1}{N+|V|}$. הנוסחה ל-MLE הינה $q_{MLE} = \frac{c}{N}$.

נראה מתי מתקיים $q_{MLE} > q_{Add-One}$.

$$q_{MLE} > q_{Add-One} \leftrightarrow \frac{c}{N} > \frac{c+1}{N+|V|} \leftrightarrow cN + c|V| > cN + N \leftrightarrow c|V| > N \leftrightarrow \frac{c}{N} > \frac{1}{|V|}$$
$$\leftrightarrow q_{MLE} > \frac{1}{|V|}$$

כלומר, קיבלנו שכאשר $q_{MLE} > \frac{1}{|V|}$ אזי מתקיים $q_{MLE} > q_{Add-One}$.

נראה מתי מתקיים $q_{MLE} < q_{Add-One}$.

$$q_{MLE} < q_{Add-One} \leftrightarrow \frac{c}{N} < \frac{c+1}{N+|V|} \leftrightarrow cN + c|V| < cN + N \leftrightarrow c|V| < N \leftrightarrow \frac{c}{N} > \frac{1}{|V|}$$
$$\leftrightarrow q_{MLE} < \frac{1}{|V|}$$

כלומר, קיבלנו שכאשר $q_{MLE} < \frac{1}{|V|}$ אזי מתקיים $q_{MLE} < q_{Add-One}$.

מחיבור שני החלקים, מצאנו שיש *threshold* $\mu = \frac{1}{|V|}$ שמקיים את מה שהיה צריך להוכיח.

ג. הנוסחה ל-GTE הינה $q_{GTE} = \frac{(c+1)*N_{c+1}}{N_c*N}$. הנוסחה ל-MLE הינה $q_{MLE} = \frac{c}{N}$.

נראה מתי מתקיים $q_{MLE} > q_{GTE}$.

$$q_{MLE} > q_{GTE} \leftrightarrow \frac{c}{N} > \frac{(c+1)*N_{c+1}}{N_c*N} \leftrightarrow \frac{c}{c+1} > \frac{N_{c+1}}{N_c}$$

נניח שקיים *threshold* μ שמקיים שאם $q_{MLE} > \mu$ אזי המשפט מתקיים.

כלומר, $q_{MLE} > \mu \leftrightarrow \frac{c}{N} > \mu \leftrightarrow c > N*\mu$.

נראה כי ניתן לבחור c כרצוננו בתחום הנ"ל שפעם האי שוויון $\frac{c}{c+1} > \frac{N_{c+1}}{N_c}$ יתקיים ופעם לא. (בעקבותיו גם $q_{MLE} > q_{GTE}$)

c	**c+1**	$\frac{c}{c+1}$	N_{c+1}	N_c	$\frac{N_{c+1}}{N_c}$	**האי שוויון מתקיים?**
$N*\mu+1$	$N*\mu+2$	$\frac{N*\mu+1}{N*\mu+2}$	$N*\mu+1$	$N*\mu+10$	$\frac{N*\mu+1}{N*\mu+10}$	כן
$N*\mu+7$	$N*\mu+8$	$\frac{N*\mu+7}{N*\mu+8}$	$N*\mu$	$N*\mu$	$\frac{N*\mu}{N*\mu}=1$	לא

כלומר, הראינו כי לא קיים *threshold* שבו בהכרח יתקיים תמיד $q_{MLE} > q_{GTE}$.

4. לפי סעיפים

א. הנוסחה למודל *trigram* הינה: $p(w_1, w_2, \dots, w_n) = \prod_{i=1}^{n} p(w_i | w_{i-2},\ w_{i-1})$.

ההנחה שההסתברות להופעת כל מילה מושפעת רק <u>משתי המילים הקודמות</u> ולא יותר.

ב. **משפט בעברית**: הכלב אוהב לרוץ.
משפט באנגלית: *The dog likes to run*.

הפועל (*likes*) הוא צמוד למילת הנושא (*dog*), בגלל שהמודל trigram מניח שמילה מושפעת משתי המילים הקודמות, וכאן הנושא הוא פחות משתי מילים לפני הפועל, לכן כנראה המודל ינבא היטב את ההטיה הנכונה.

ג. **משפט בעברית**: הכלב עם השיניים הגדולות אוהב לרוץ.
משפט באנגלית: *The dog with the big teeth likes to run*.

הפועל (*likes*) הוא לא צמוד מספיק למילת הנושא (*dog*), בגלל שהמודל trigram מניח שמילה מושפעת **משתי** המילים הקודמות, וכאן הנושא הוא יותר משתי מילים לפני הפועל, לכן כנראה המודל לא ינבא היטב את ההטיה הנכונה.

בגלל שבמשפט שבחרנו הנושא הוא 5 מילים לפני הפועל, נדרש מודל *6-gram* כדי לחזות כהלכה את ההטיה. (במשפט בעברית זה יהיה מודל *5-gram*).

5. עבור צמדים: **הילד** אכל ואז **הלכה** לבית הספר
עבור שלישיות: **הילד** אכל ולאחר מכן **הלכה** לבית הספר
עבור רביעיות: **הילד** עם העיניים הכחולות **הלכה** לבית הספר

בצמדים – כל צמד הוא חלק ממשפט תקף ורק המשפט המלא לא נכון תחבירית כי יש הטיה של נקבה עבור הילד (זכר).
אותו הדבר עבור שלישיות ורביעיות.

ככל שהמודל המרקובי מניח תלות ביותר מילים קודמות, ככה נלקח בחשבון גם יותר מידע הקשרי. לכן, המודל יהיה יותר מדוייק ויתפוס יותר הקשרים בתוך המשפט, כאשר הוא ייקח בחשבון יותר מילים לפני המילה הנוכחית. לעשות משפט שהוא לא תקף בכללותו אבל במקטעיו הוא כן, ככל המודל הוא בסדר יותר גבוה, ככה יותר קשה כי ההקשרים במשפט נוטים להיות יחסית סמוכים.