NLP | תרגיל 1

13 בנובמבר 2022

שאלה 1

1. (10 pts) Given a bigram language model for sentences of the form START $w_1\ w_2\ w_3\ \cdots\ w_n$ STOP (where w_i for $1 \le i \le n$ is a word), show that if the transition probabilities are well-defined (i.e., sum up to 1) and each word has some non-zero probability for generating STOP ($\forall w, p(\text{STOP}|w) > 0$), then the sum of the probabilities over all finite sequences is 1.

   Hint: prove that the complement probability (i.e., the probability to never generate STOP, which is the same as the sum of all the sequences that don't have STOP) is 0.

נגדיר את המאורע - כל קבוצות המשפטים בגודל ∞ כך שכל מילה במשפט אינה $stop$:

$A = \{(w_1...)|s.t. \forall i \in [1, \infty]\ \ w_i \neq stop\}$

$A_i = \{(w_1...)|s.t\ \forall j < i \quad w_j \neq stop\}$ - קבוצות משפטים ∞ כך שכל i המילים הראשונות במשפט שונים מ $stop$

נבחים כי $A = \cap A_i$ וגם $A_i \subset A_{i+1}$ לכל i

ע"פ מה שלמדנו באינפי נקבל כי $P(A) = lim_{i\to\infty} P(A_i)$

נגיד $m = min_w P(stop|w)$ - ההסתברות הכי קטנה לקבל $stop$ אחרי מילה w כלשהי (ביטוי זה מוגדר היטב מכיוון ונתון כי $\forall w\ P(stop|w) > 0$)

כעת נחשב את $P(A_i)$

$$P(A_i) = P(w_1 \neq STOP, ..., w_i \neq STOP) =$$

$$= P(w_1 \neq STOP) \cdot P(w_2 \neq STOP|w_1 \neq STOP) \cdot ... \cdot P(w_i \neq STOP|w_{i-1} \neq STOP) \le$$

$$\overset{*1}{\le} (1-m)^i$$

$*1$ ע"פ הגדרה של m

כאשר $i \to \infty$ מתקיים כי $P(A_i) \le (1-m)^i \to 0$ והרי $P(A) = lim_{i\to\infty} P(A_i)$ לכן $P(A) = 0$

מכיוון וקיבלנו כי ההסתברות של משפט לא לצעור שווה ל 0 ההסתברות המשלימה אותה נדרשו לחשב שווה ל 1.

שאלה 2

(15 pts) We want to build a spelling corrector, focusing on the distinction between "where" and "were". Given a sentence as input, the corrector should predict the true spelling for each instance of "where" or "were" and correct the spelling in the case of mistake.

For example, given the sentence "He went where there where more opportunities", the corrector should predict "where" for the first instance and "were" for the second one. It should also correct the word in the second case.

Suppose we use a language model for this task. Given a language model $p(w_1, w_2, \cdots, w_n)$ where n is the length of the sentence, the corrector returns the spelling that gives the highest probability.

In our example, the spelling corrector will output "were" for the second instance if:

p(He went where there **were** more opportunities) $>$ p(He went where there **where** more opportunities)

(a) Describe formally a unigram language model for the spelling corrector. Assume that the probability of a word is given by its proportion in the corpus (the training set) and that the number of instances in the corpus of each word in the vocabulary is strictly bigger than 0. Given the sentence "He went where there where more opportunities", under which conditions will the spelling corrector give a right answer for the first instance of "where"? for the second instance of "where"? for both instances?

(א)

בהינתן $corpus$ (סט אימון) נגדיר את ההסתברות של כל מילה להיות מספר המופעים שהיא מופיע בקורפוס חלקי מספר המילים השונות בטקס, כלומר

נניח כי $\forall i \in [n]$ מתקיים כי $p(w_i) = proportion\ in\ the\ corpus$

בנוסף נתון כי כל מילה מופיע בקורפוס

נגדיר את מודל $unigram$ כך שההסתברות של משפט היא מכפלת ההסתבוריות של מילים בו:

$$P(w_1, ..., w_n) = \prod_{i=1}^{n} p(w_1)$$

המודל יחשב את $P(w = "where")$ ואת $P(w = "were")$

בהינתן משפט $w_1, ..., w_n$ המודל יחזיר את המשפט $w'_1, ..., w'_n$ כך לכל $i \in [n]$:

אם $w_i = "where"$ ו $P(w = "where") > P(w = "were")$ המודל יחזיר את w_i, אחרת יחזיר את $w'_i = "were"$

אם $w_i = "were"$ ו $P(w = "where") > P(w = "were")$ המודל יחזיר את $w'_i = "where"$, אחרת יחזיר את w_i

מקרה זה לא טופל ולכן נטפל בו כעת:

אם יש שוויון בין מס' המופעים של $where$ ו $were$ בקורפוס נגדיר כי המודל יטיל מטבע בלתי מוטה ובהסתברות חצי יבחר בין $"where", "were"$

ולכל w_i ששונה מ $"where"$ ו $"were"$ יחזיר את w_i המקורי

מקרה ראשון:

במקרה בו $P(w = "where") > P(w = "were")$ ה $spelling\ corector$ יחזיר את התשובה הנכונה

מקרה שני:

במקרה בו במקרה בו $P(w = "where") < P(w = "were")$ ה $spelling\ corector$ יחזיר את התשובה הנכונה

מקרה שלישי:

המודל יצדק במקרה בו $P(w = "where") = P(w = "were")$ ובהטלת מטבע הראשונה הוא יבחר ב $"where"$ ובהטלת מטבע השניה הוא יבחר ב $"were'$.

(b) Describe formally a bigram language model for the spelling corrector. Assume again that we estimate the parameters of the model using relative frequency and that the number of instances in the corpus of each word in the vocabulary is strictly bigger than 0. Why might this model be better that the model in (a)? Can a sentence in this model get a zero-probability? Would it be a problem for the model?

(ב)

1. מודל זה יכול להיות יותר טוב מהמודל בסעיף א' מכיוון תופס קשר בין מילים ולא מתייחס רק לשכיחות שלהם בקורפוס
כמו כן המודל מורכב יותר מכיוון והוא משתמש ב (V^2) פרמרטרים - קומבינציות של זוגות של מילים
2. למרות שבמודל זה ההסתברות של כל מילה להופיע בקורפוס גדולה מאפס, הסתברות של צמד מילים להופיע בקורפוס יכולה להיות אפס במקרה וצמד מילים זה לא הופיעה ולכן ההסתברות למשפט במודל זה עלולה להיות שווה לאפס.
3. במודל זה יתכן ונקבל משפט בעל היגיון שההסתברות שלו גדולה מאפס אך בגלל שבמשפט קיימות צמד מילים שלא הופיעו ברצף בקורפוס המודל ישערך כי ההסתברות של המשפט כאפס.

(ג)

שאלה 3

(15 pts) Consider the advanced smoothing method called Good-Turing smoothing. Let N_c be the number of word types (unique words) which appeared exactly c times in the training corpus (e.g., N_1 is the number of unique words that appeared one time in the training corpus). N denotes the total number of word instances in the training corpus. An estimate of the total probability of all unseen words (i.e., words that do not appear in the training corpus) is given by $p_{unseen} = \frac{N_1}{N}$.

The smoothed Good-Turing estimate of a frequency of a word that appears c times in the training corpus is $\frac{(c+1)N_{c+1}}{N_c \cdot N}$.

Note: Assume that $N_c > 0$ for all values of c up to a certain maximum value c_{max} and $N_c = 0$ for all $c > c_{max}$.

(a) Show that the sum of smoothed Good-Turing frequency estimates over all word types in the training corpus is $1 - p_{unseen}$

(b) Write down the equation for the smoothed Add-One estimate of a frequency of a word that appears c times in the training corpus. Show that there is a threshold μ, such that for all words of frequency less than μ, their smoothed estimate is higher than the MLE, and for all words of frequency more than μ, their smoothed estimate is lower than the MLE.

(c) Show that the property in (b) does not necessarily hold for the smoothed Good-Turing estimate.

(א)

נתון כי $N = \Sigma_{c=1}^{cmax} c \cdot N_c$

$$\boxed{\sum_{c=1}^{cmax} \cancel{N_c} \cdot \underbrace{\frac{(c+1)N_{c+1}}{\cancel{N_c} \cdot N}}_{frequency\ estimate}} = \frac{1}{N} \sum_{c=1}^{cmax} (c+1)N_{c+1} =$$

$$\overset{c-1=i}{=} \frac{1}{N}\left(\sum_{i=2}^{cmax} iN_i\right) = \frac{1}{N}\left(\underbrace{\left(\sum_{i=1}^{cmax} iN_i\right)}_{=N} - N_1\right) = \frac{1}{N}\cdot(N-N_1) =$$

$$= 1 - \frac{N_1}{N} = \boxed{1 - P_{unseen}}$$

קיבלנו כי ההסתברות אותה נדרשנו לחשב שווה ל $1 - P_{unseen}$

(ב)

נמצא את סף מספר המילים עבורו $Add-one$ נותן הסתברות גבוהה מה MLE

$$\frac{c+1}{N+|V|} > \frac{c}{N} \iff N\,(c+1) > c\,(N+|V|)$$

$$\iff \cancel{Nc} + N > \cancel{cN} + c\,|V| \iff N > c\,|V|$$

$$\iff c < \frac{N}{|V|}$$

באותו אופן נקבל כי סף מספר המילים עבורו $Add-one$ נותן הסתברות נמוכה מה MLE יהיה:

$$c > \frac{N}{|V|}$$

לכן הסף אותו התבקשנו לחשב הינו $\boxed{\frac{N}{|V|}}$ - ממוצע ההפועות של מילים $unique$ בקורפוס

(ג)

נראה כי התכונה מסעיף b לא תמיד מתקיימת עבור $smoothed\ good\ turing$:

$$\frac{(c+1)\,N_{c+1}}{N_c \cdot N} > \underbrace{\frac{c}{N}}_{MLE} \iff (c+1)\,N_{c+1} > \frac{N_c \cdot \cancel{N} \cdot c}{\cancel{N}}$$

$$\iff cN_{c+1} + N_{c+1} > N_c \cdot c \iff c(N_{c+1} - N_c) > -N_{c+1}$$

קיבלנו כי $c(N_{c+1} - N_c) > -N_{c+1}$
למשל עבור $N_{c+1} = 2$ ו $c = N_c = 4$ נקבל כי

$$2(2-4) > -2 \iff -8 > -2$$

סתירה.

שאלה 4

(15 pts)

(a) Write down the equation for a trigram language model (without detailing the probability estimations). Which (conditional) independence assumption is made in the model?

(b) Give an example of an English sentence and a Hebrew sentence where the phenomenon of verb-subject agreement (see below) is captured by the model in (a). That is, give an example where the model in (a) is likely to predict the correct inflection of the verb, given the subject.

(c) Give an example of an English sentence and a Hebrew sentence where subject-verb agreement is not captured. Which n (for an n-gram model) is necessary for capturing this phenomenon in your example?

(א)

נגדיר את המודל בצורה הבאה:
בהינתן קורפוס נחשב את ההסתברויות הבאות $P(w_n|w_{n-2}w_{n-1}) = \frac{count(w_{n-2},w_{n-1},w_n)}{count(w_{n-2},w_{n-1}w_j)}$
בהינתן משפט המודל ישערך את הסתברות המשפט בצורה הבאה:

$$P(w_1, ..., w_n) = \prod_{i=1}^{n} p(w_i|w_{i-2}w_{i-1})$$

עבור $i = 1$ ו נגדיר $p(w_i|w_{i-2}w_{i-1}) = p(w_1|start, start) = p(w_1)$
עבור $i = 2$ נגדיר את $p(w_i|w_{i-2}w_{i-1}) = p(w_2|start, w_1) = p(w_2|w_1)$
במודל זה מתקיימת הנחה של אי תלות בין ההסתברות למילה w_i למילים $w_{i-2,}w_{i-3}....$
כלמור w_i תלויה רק בשתי המילים שקדמו לה.

(ב)

$"a\ girl\ cries\ in\ the\ park"$
"הילד אכל תפוח"

(ג)

$"girls\ who\ have\ fun\ are\ nice"$ - המודל לא יחזה את המילה are בהינתן $have\ fun$ מכיוון ו are תלויה ב $girls$
"הכלב שנראה צולע רץ" - המודל לא יחזה את המילה "רץ" בהינתן "שנראה צולע" מכיוון ו "רץ" תלוי "הכלה"

שאלה 5

משפט שכל שני מילים בו הגיוניות אך הוא לא הגיוני:
"היום אוכל מחר"
משפט שכל שלוש מילים בו הגיוניות אך הוא לא הגיוני:
"היום אצא לריצה מחר"
משפט שכל ארבע מילים בו הגיוניות אך הוא לא הגיוני:
"היום נצא לשתות בבר מחרתיים"

תשובות לחלק המעשי:

```
### TASK 2 ###
Predicted word: the

### TASK 3 ###
Probability of first sentence: -inf
Probability of second sentence: -29.707484205602142
Perplexity of both sentences: inf

inf
### TASK 4 ###
Linear interpolation smoothing for first sentence: -36.20200520715635
Linear interpolation smoothing for second sentence: -30.99123945115062
Perplexity of both sentences: 270.2742150360732
```