תרגיל 2 - עיבוד שפה טבעית

6:21 PM Tuesday, December 8, 2020

רועי כהן-אופנהיים
אלה בר-יעקב

שאלה 1

נתון לנו מודל גנרטיבי שמניח ש:

- סדרת התגיות מתנהגת באופן מרקובי (מסדר k כלשהו)
- ייצור המילה x תלוי אך ורק בתגית שלו.

נסמן ב-Y את אוסף כל התגיות, אליהם נוסיף גם START ו-STOP, וב-V את אוסף כל המילים.
מנתונים אלו אנחנו יכולים ללמוד שקיימות שתי פונקציות מוגדרות היטב שמחשבות לנו:

$$\forall y_1, \dots, y_k \in Y \; p'(y_1|y_2, \dots, y_k) \; is\ well\ defined$$

$$\forall x \in V, y \in T \; p''(x|y) \; is\ well\ defined$$

כיוון שזה מודל גנרטיבי, מוגדר בסופו של דבר לכל xi במשפט:

$$p(x_i, y_i, \dots, y_{i-k}) = \; p''(x_i|y_i)p'(y_i|y_{i-\mathrm{i}}, \dots, y_{i-k})$$

ומהנחת המרקוביות ומחוסר התלות בין פונקציות הייצור יתקיים:

$$p(x_1, \dots, x_n, y_1, \dots, y_n, y_{n+1}) = \Pi_{\mathrm{i}}^{\mathrm{n}+1} p(x_i, y_i, \dots, y_{i-k}) =$$
$$\Pi_{\mathrm{i}}^{\mathrm{n}+1} \; p''(x_i|y_i)p'(y_i|y_{i-\mathrm{i}}, \dots, y_{i-k}) = \Pi_{\mathrm{i}=1}^{\mathrm{n}+1} \mathrm{p}'(\mathrm{y_i}|\mathrm{y_{i-1}}, \dots, \mathrm{y_{i-k}}) \Pi_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}''(\mathrm{x_i}|\mathrm{y_i})$$

כאשר המעבר האחרון נובע משינוי סדר הכפילה, וכן מכך שעבור
y_n+1=STOP
לא נפלטת אף מילה.

נזכור כי הנוסחה למודל HMM מסדר מרקובי k מגדירה התפלגות משותפת באופן הבא, בהינתן פונקציית מעברים q ופונקציית ייצור e:

$$p(x_1, \dots, x_n, y_1, \dots, y_n) = \Pi_{\mathrm{i}=1}^{\mathrm{n}+1} \mathrm{q}(\mathrm{y_i}|\mathrm{y_{i-1}}, \dots, \mathrm{y_{i-k}}) \Pi_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{e}(\mathrm{x_i}|\mathrm{y_i})$$

לכן, נשים לב שנוכל לבחור q=p', e=p'' ולקבל מודל של HMM כנדרש.

שאלה 2

Transition table (T):

from\to	H	L
H	0.5	0.5
L	0.4	0.6

Emission Table (E):

	A	C	G	T
H	0.2	0.3	0.3	0.2
L	0.3	0.2	0.2	0.3

Dynamic table (D):

	A	C	C	G	T	G	C	A
H	0.1 H	0.018 L	0.0027 H	0.000405 H	0.0000 405 H	0.00000 729 L	0.00000 10935 H	0.000000 10935 H
L	0.15 H	0.018 | L	0.00216 L	0.00027 H	0.0000 6075 H	0.00000 729 L	0.00000 08748 L	0.000000 164025 H

בכל עמודה מילאנו את הערך למעלה, ולמטה את השורה שממנה אנו הגענו מתוך העמודה הקודמת.

עבור מקרי הבסיס (העמודה השמאלית) מילאנו:

D[HA] = T[HH]*E[HA]
D[HB] = T[HL]*E[LA]

מילאנו רקורסיבית עבור כל עמודה על פי הנוסחה של ויטרבי.

נבחר את הערך המקסימלי מבין העמודה האחרונה: L.
נשחזר את המסלול ע"י הליכה אחורה בהתאם לפוינטרים, ונקבל:
HLHHHLHHL

שאלה 3

תחילה, נחשב עבור כל תג את המילה הכי נפוצה שהוא יכול לפלוט:

$$\forall y \in K, x_y = argmax_{x \in V} P(x|y)$$

כעת ניצור טבלה D כך שכל שורה תואמת לצירוף אפשרי של שלושה תגים מ-K (עם חשיבות לסדר).

$$(y_1, y_2, y_3) \in K$$

מספר העמודות יהיה בהתאם ל-n שקיבלנו כקלט.
כלומר, ב-D יש |K|^3 שורות ו-n עמודות.

נמלא כל עמודה בנפרד, החל מהעמודה הראשונה, בהתאם לחוקיות הנ"ל:

מקרה בסיס:

$$D[(** \, y_j), 1] = p(y_j | ***) \times p(x_{y_j})$$

For all other rows, 0.

צעד:

$$\forall 2 \le t \le n, D[(y_{i-2}, y_{i-1}, y_i), t] = \max_{(y_{i-3}, y_{i-2}, y_{i-1}) \in K \cup \{*\}} D[(y_{i-3}, y_{i-2}, y_{i-1}), t-1] \times q(y_i | y_{i-3}, y_{i-2}, y_{i-1}) \times p(x_{y_i})$$

<u>החזרה:</u>
בכל עמודה ושורה, נשמור את מספר השורה שנתנה את המקסימום מתוך העמודה הקודמת.
ניקח בעמודה האחרונה את השורה שמקיימת:

$$argmax_{(y_{i-3},y_{i-2},y_{i-1})\in K\cup\{*\}} D[(y_{i-3},y_{i-2},y_{i-1}),n] \times q(STOP|y_{i-3},y_{i-2},y_{i-1})$$

משם, נתחקה אחר המסלול שהשארנו בפוינטרים כדי לקבל את התגיות y1,...,yn.
כדי למצוא את המילים המתאימות, ניקח את

$$x_{y_1},\dots,x_{y_n}$$

ובכך נסיים. :)

<u>שאלה 4</u>

Method/ Error Rate	Known	Unknown	Total
MLE	0.070446	0.789336	0.152556
HMM, no smoothing	0.692628	1	0.727736
HMM, add-one smoothing	0.146754	0.719406	0.212161
HMM, pseudowords, no smoothing	0.1465284	0.545455	0.192093
HMM, pseudowords, add-one smoothing	0.140555	0.540210	0.186202

ניתוח ה-confusion matrix:

להלן חמשת הטעויות הכי נפוצות:

Tag	Prediction	Count
NNS	NN	236
NN	NP	102
JJ	NN	71
NN	JJ	71
NP	NN	54

רוב הטעויות של המודל שלנו נסובו סביב NN, שהיא כנראה קטגוריה מבלבלת עבור המודל.