Amit Sindani 206902694

Raz Olewsky 315341396

Natural Language Processing – Assignment #2 - Solution

Task #1: the Viterbi algorithm for sequence decoding

1. Does Viterbi guarantee to find the optimal solution? That is, does it find a sequence with the highest probability, given a sequence of observations (a sentence)? Explain shortly (in your own words), referring to structures and the algorithm we saw in the class.

   Answer:

תלוי, אלגוריתם ויטרבי שואף למצוא את הפתרון האופטימלי (רצף המצבים - במקרה שלנו רצף חלקי הדיבר) כאשר אנו מניחים הנחות מסוימות כמו שהנחת מרקוב מתקיימת (מה שלא בהכרח קורה בעולם האמיתי).

במידה ובאמת הנחות אלה מתקיימות, האלגוריתם כן מבטיח פתרון אופטימלי.

האלגוריתם בנוי בצורת תכנות דינמי המבטיח שבכל שלב נקבל את הפתרון האופטימלי (תת רצף של חלקי דיבר) יחסית לאותו שלב הסביר ביותר בהינתן תת הרצף של הפלטים (תת רצף של מילים – משפט חלקי) ולכן בסוף התהליך נקבל פתרון אופטימלי כולל.

2. During forward pass, the algorithm assigns each cell in the matrix C a value according to the following computation:

   (a) $C(i, j) = \max_k C(k, j-1) * A(k, i) * B(i, ind(w_j))$

   Dropping the first factor in this computation will result in the below assignment:

   (b) $C(i, j) = \max_k A(k, i) * B(i, ind(w_j))$

   Give a small example for transition and emission matrices, where the alternative (b) is insufficient for computation of the most probable sequence, but the original alternative (a) would get it right. Example with two tags and two words should suffice for this purpose.

   Answer:

transition matrix

$$A = \begin{array}{c|cc} & \text{noun} & \text{verb} \\ \hline \text{noun} & 0.6 & 0.4 \\ \text{verb} & 0.9 & 0.1 \end{array}$$

emission matrix

$$B = \begin{array}{c|cc} & \text{pet} & \text{store} \\ \hline \text{noun} & 0.5 & 0.5 \\ \text{verb} & 0.5 & 0.5 \end{array}$$

initial probability distribution

$$\pi = \begin{array}{cc} \text{noun} & \text{verb} \\ 0.9 & 0.1 \end{array}$$

נניח שרצף המילים שלנו הוא: "pet store" ואנחנו רוצים לתייג את רצף תלקי הדיבר המתאימים לו -

(רצף תלקי הדיבר הנכון לרצף מלים זה הוא חעם חעם "חנות חיות מחמד").

נריץ את אלגוריתם ויטרבי:

$$C = \begin{array}{c|c|c} & w_1 = \text{pet} & w_2 = \text{store} \\ \hline \text{noun} & 0.45 & 0.135/0.45 \\ \hline \text{verb} & 0.05 & 0.09/0.2 \end{array} \qquad D = \begin{array}{c|c|c} & w_1 = \text{pet} & w_2 = \text{store} \\ \hline \text{noun} & 0 & \text{noun}/\text{verb} \\ \hline \text{verb} & 0 & \text{noun}/\text{noun} \end{array}$$

(a) $C(i, j) = \max_k C(k, j-1) * A(k, i) * B(i, ind(w_j))$ שימוש בנוסחה המקורית

(b) $C(i, j) = \max_k A(k, i) * B(i, ind(w_j))$ שימוש בנוסחה האלטרנטיבית

$$C(noun, \overset{w_1}{pet}) = \pi(noun) \cdot B(noun, pet) = 0.9 \cdot 0.5 = 0.45$$

$$C(verb, \overset{w_1}{pet}) = \pi(verb) \cdot B(verb, pet) = 0.1 \cdot 0.5 = 0.05$$

מקורי

$$C(noun, store) = \max\{\overbrace{C(noun,pet) \cdot A(noun,noun) \cdot B(noun,store)}^{0.135}, \overbrace{C(verb,pet) \cdot A(verb,noun) \cdot B(noun,store)}^{0.0225}\} = 0.135$$

$$\begin{array}{ccc} 0.45 & 0.6 & 0.5 \end{array} \qquad \begin{array}{ccc} 0.05 & 0.9 & 0.5 \end{array}$$

noun

מקורי

$$C(verb, store) = \max\{\overbrace{C(noun,pet) \cdot A(noun,verb) \cdot B(verb,store)}^{0.09}, \overbrace{C(verb,pet) \cdot A(verb,verb) \cdot B(verb,store)}^{0.0025}\} = 0.09$$

$$\begin{array}{ccc} 0.45 & 0.4 & 0.5 \end{array} \qquad \begin{array}{ccc} 0.05 & 0.1 & 0.5 \end{array}$$

noun

אלטרנטיבי

$$C(noun, store) = \max\{\overbrace{A(noun,noun) \cdot B(noun,store)}^{0.3}, \overbrace{A(verb,noun) \cdot B(noun,store)}^{0.45}\} = 0.45$$

$$\begin{array}{cc} 0.6 & 0.5 \end{array} \qquad \begin{array}{cc} 0.9 & 0.5 \end{array}$$

verb

אלטרנטיבי

$$C(verb, store) = \max\{\overbrace{A(noun,verb) \cdot B(verb,store)}^{0.2}, \overbrace{A(verb,verb) \cdot B(verb,store)}^{0.05}\} = 0.2$$

$$\begin{array}{cc} 0.4 & 0.5 \end{array} \qquad \begin{array}{cc} 0.1 & 0.5 \end{array}$$

noun

בדרך המקורית נחצה את הלקי הדבר חשמל חשמל

ובדרך האלטרנטיבית נחצה חשמל verb ← פחות סבר לרצף המילים שלנו

Task #2: text classification – Amazon products reviews

1. Print your confusion matrix; which classes share the highest confusion? Why? Write your short interpretation of the confusion matrix in the report you're submitting.

   Answer:

להלן ה- confusion matrix שקיבלנו עבור כל אחד מהדאטה סטים:

דאטה סט Sports_and_Outdoors –

```
[[285  80  17  10   8]
 [ 90 212  68  22   8]
 [ 29  86 206  63  16]
 [ 11  25  59 245  60]
 [  9  12  16  63 300]]
```

דאטה סט Pet_Supplies –

```
[[278  76  31   5  10]
 [101 183  73  32  11]
 [ 44  72 199  58  27]
 [  9  21  58 250  62]
 [ 12   5  11  89 283]]
```

דאטה סט Automotive –

```
[[280  81  27   7   5]
 [101 208  67  16   8]
 [ 31  69 225  60  15]
 [ 18  37  53 229  63]
 [ 19  16  15  73 277]]
```

השורות מייצגות את התוויות האמיתיות (השורה הראשונה מייצגת את ציון הביקורת "1", השורה השנייה מייצגת את ציון הביקורת "2" וכן הלאה..).

העמודות מייצגות את התוויות החזויות (העמודה הראשונה מייצגת את ציון הביקורת "1", העמודה השנייה מייצגת את ציון הביקורת "2" וכן הלאה..).

מחלקות הסיווג שחולקות את ה"בלבול הגדול ביותר" הן מחלקות קרובות בערך הסיווג (בציון הביקורת) שלהן. למשל – מחלקת סיווג 2 תחווה "בלבול גדול" עם מחלקות סיווג 1 ו-3.

לדעתנו זה קורה כי ביקורות של סיווגים סמוכים תהיינה קרובות בייצוג הוקטורי שלהן ולכן סביר שלמודל יהיה נטייה יותר גדולה "להתבלבל" ולסווגם לא נכון (בפועל ביקורות עם סיווגים סמוכים - בסיכוי סביר שתתוארנה ע"י מילים זהות/דומות).

המטריצות המתוארות לעיל, אכן משקפות את התוצאות שציפינו לקבל. ניתן לראות שערכי האלכסון הראשי (המייצגים את ערכי ה-TP) גבוהים משאר הערכים באותה שורה/עמודה, כלומר הצלחנו לחזות נכון את ציון רוב הביקורות.

2. Python's scikit-learn allows extraction of K features (words or word n-grams in our case) that have the highest discriminative power, i.e., are the best for the classification task at hand. One such function is `SelectKBest` provided by scikit-learn. Use this function to extract 15 features most effective for classification and report them at the document you're submitting.

   Answer:

דאטה סט Sports_and_Outdoors –

```
The 15 words that have the highest discriminative power are: ['five' 'five stars' 'four' 'four stars' 'good four' 'great' 'great five'
 'one' 'one star' 'star' 'stars' 'three' 'three stars' 'two' 'two stars']
```

דאטה סט Pet_Supplies –

```
The 15 words that have the highest discriminative power are: ['five' 'five stars' 'four' 'four stars' 'great' 'great five' 'it five'
 'love' 'one star' 'star' 'stars' 'three' 'three stars' 'two' 'two stars']
```

דאטה סט Automotive –

```
The 15 words that have the highest discriminative power are: ['five' 'five stars' 'four' 'four stars' 'good four' 'great' 'great five'
 'ok three' 'one star' 'star' 'stars' 'three' 'three stars' 'two'
```

כפי שניתן להתרשם, המילים/הביגרמים שחולצו אכן בעלות משמעות רבה ומהסתכלות עליהן באמת ניתן להבין כי יש להן כוח תיאורי. למשל המילה "five" שמשקפת בצורה נחרצת וחד משמעית שמדובר בביקורת שציונה 5 או הביגרם "one star" שמשקף באופן נחרץ תיאור של ביקורת שציונה 1.

הערה לבודקת: השארנו את הקוד שביצע את חילוץ 15 המילים בעלות הכוח התיאורי הגדול ביותר כהערה בקובץ ה-main.py שהגשנו (עקרונית, אלה אמרה שלא צריך, אבל החלטנו להוסיף בכל זאת כדי שתוכלי לראות איך מימשנו).

3. Perform cross-domain classification: train on one domain (e.g., sports training data) and test on another (e.g., pets test data). How do your results compare to in-domain classification? Interpret your results shortly in the document.

   Answer:

נבצע את cross-domain classification הבא:

Train set - Sports_and_Outdoors

Test set - Pet_Supplies

הדיוק שקיבלנו הוא: 0.572

להשוואה נריץ in-domain classification שה-Train and Test set הוא Sports_and_Outdoors.

הדיוק שקיבלנו הוא: 0.624

ניתן לראות שקיבלנו דיוק גבוה יחסית ב-cross-domain (למרות שחשבנו שנקבל דיוק נמוך הרבה יותר), אך קטן יותר מהדיוק ב-in-domain .

הדיוק הגבוה התקבל בעקבות מגוון מילים וביגרמים חופפים בביקורות של הנושאים השונים שבחנו (מילים וביגרמים כמו "great", "great five", "three stars" ועוד).

אולם יש הבדל בין הדיוקים ביחס לin-domain, זאת כיוון שעדיין קיים הבדל בין קבוצת המילים שהביקורות בנושא הראשון עושות בהן שימוש לקבוצת המילים שהביקורות בנושא השני עושות בהן שימוש. למשל, המילה uncomfortable ("לא נוח") מתארת מילה שלילית שמשומשת יותר לתיאור בנושא sports and outdoors מאשר בנושא pet supplies.