תאריך: 10.2.2019

מבחן מועד א' בקורס עיבוד שפה טבעית (67658)

2018/2019 שנת לימודים התשע"ח

משך המבחן: שעתיים

מרצה הקורס: ד"ר עמרי אבנד

השימוש בכל חומר עזר אסור, נא כתבו על הכריכה על אילו שאלות עניתם.

חלק א' (80 נקודות): ענו על <u>בדיוק שתי שאלות</u> מבין שאלות 1-3

<u>שאלה 1 (40 נקודות):</u>

א. (10 נק') הגדירו מודל Naïve Bayes עבור סיווג טקסטים (Text Classification) לקטגוריה מתוך קבוצה נתונה L. הניחו שהתכוניות (features) הם מסוג Bag of Words בינריות (כלומר: "מופיעה" או "לא מופיעה"). כמה פרמטרים יש למודל?

ב. (10 נק') נמשיך לעסוק במודל מסעיף א'. נניח כעת שב-training data המילה w מופיעה רק פעם אחת, עם $y_0 \in L$. כמו כן, נניח שאנו משערכים את הפרמטרים בעזרת אמידת נראות מרבית (Maximum Likelihood Estimation). הראו שעבור ההסתברות שמגדיר המודל המאומן P, ובהינתן מסמך D שבו מופיעה המילה w, מתקיים שלכל תווית $y \in L$: $P(y_0|D) \geq P(y|D)$.

ג. (10 נק') הסבירו מדוע התופעה שאת קיומה הראיתם בסעיף ב' איננה רצויה, והסבירו כיצד ניתן להתמודד עם הבעיה.

ד. (10 נק') כעת נניח שאנו משתמשים במודל לוג-לינארי (log-linear) עבור אותה המטלה מסעיף א' (סיווג טקסטים לקטגוריה מתוך קבוצה נתונה L). כמו כן נניח שהתכוניות הן כמו בסעיף א', ושקיימת מילה w המופיעה פעם אחת בלבד ב-training data, במסמך עם התווית y_0. כאשר במקרה זה משתמשים באמידת נראות מרבית, שוב עשויה להיווצר הטיה לכיוון y_0 במסמכים בהם w מופיעה. (אין צורך להוכיח זאת) הסבירו כיצד ניתן לשנות את שערוך הפרמטרים ע"מ להתמודד עם בעיה זאת במקרה של מודל לוג ליניארי.

<u>שאלה 2 (40 נקודות):</u>

בטקסטים הכתובים בשפה הסינית נהוג שלא לכתוב רווחים בין התווים. על כן, אחת הבעיות בניתוח טקסט בסינית היא לסמן היכן נגמרת מילה ומתחילה אחרת, כלומר להוסיף רווחים בין המילים.

לדוגמא: בהינתן הקלט 日文章魚怎麼說?, הפלט הנכון הוא

日文 章魚 怎麼 說?

(משמעות המילים היא (יפנית, תמנון, איך, אומר), ומשמעות המשפט כולו היא "כיצד אומרים תמנון ביפנית?")

א. (10 נק') הגדירו באופן פורמלי מודל Bigram Maximum Entropy Markov Model (Bigram MEMM) עבור הבעיה. מהו מרחב התוויות (labels) שהמודל חוזה?

<u>הערה</u>: אין צורך להסביר כיצד מחושבת פונקציית התכוניות (feature function), או אילו תכוניות רלוונטיות לפתרון הבעיה, אלא רק להגדיר מהו התחום והטווח שלה.

ב. (10 נק') בהינתן מודל מאומן (כלומר ערך לכל פרמטר), כתבו פסאודו-קוד עבור אלגוריתם יעיל המקבל רצף תווים בסינית ומחזיר את ההסתברות שהמודל נותן לחלוקה הסבירה ביותר של הרצף למילים. (אין צורך להוכיח את נכונות האלגוריתם)

ג. (10 נק') איזו הנחת אי-תלות צריכה להתקיים עבור טקסטים בסינית על מנת שמודל MEMM מסדר ראשון (כלומר Bigram) יהיה אופטימלי עבור הבעיה? (הכוונה: בהשוואה למודל MEMM מסדר גבוה יותר)

ד. (10 נק') בהינתן training data עבור הבעיה, כלומר אוסף של N קלטים ופלטים נכונים עבורם, כתבו את פונקציית ה-log-likelihood.

<u>שאלה 3 (40 נקודות):</u>

א. (10 נק') הגדירו פורמלית מהו מודל שפה (language model).

כעת נגדיר מודל שפה באופן הבא: בהינתן משפט $x_1,\dots,x_n$, ובהינתן סדרת חלקי הדיבר (Parts of Speech) עבור כל מילה בו $y_1,\dots,y_n$, ההסתברות למשפט תוגדר להיות:

$$P(x_1 \dots x_n | y_1 \dots y_n) = \prod_{i=1..n} P(x_i | y_i)$$

כאשר לא נתונה סדרת חלקי הדיבר, נגדיר את ההסתברות למשפט בעזרת נוסחת ההסתברות השלמה:

$$P(x_1, \dots, x_n) = \sum_{(y_1..y_n) \in Y^n} P(y_1 \dots y_n) P(x_1 \dots x_n | y_1 \dots y_n)$$

כאשר Y הוא אוסף חלקי הדיבר האפשריים לכל מילה.

כמו כן, נניח ששרשרת חלקי הדיבר $y_1 \dots y_n$ מהווה מודל מרקוב מסדר ראשון (bigram).

ב. (10 נק') מהם הפרמטרים של מודל השפה תחת הנחות אלו? כמה פרמטרים יש לו בסה"כ?

ג. (10 נק') בהינתן מודל מאומן (כלומר ערך לכל פרמטר), כתבו פסאודו-קוד לאלגוריתם יעיל המקבל משפט ומחזיר את ההסתברות שלו על פי המודל.

ד. (10 נק') מהם הנחות אי-התלות המותנה שמניח מודל שפה זה (שתי הנחות עיקריות). האם הנחות אלו מתקיימות בשפה טבעית? הסבירו את תשובתכם באמצעות דוגמאות באנגלית.

הערה: בתשובתכם אתם יכולים לעשות שימוש ברשימת חלקי הדיבר הנפוצים באנגלית המופיעה בנספח.

חלק ב' (20 נקודות): ענו על <u>בדיוק שאלה אחת</u> מבין שאלות 4-5

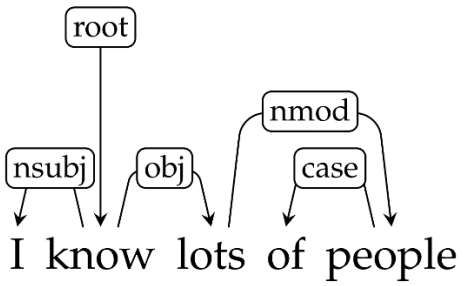
<u>שאלה 4 (20 נקודות):</u>

א. (10 נק') הגדירו את פונקציית הניקוד (score) שבעזרתה ה-MST Parser נותן ניקוד לעצי תלויות אפשריים עבור משפט נתון. הגדירו במדויק את כל המושגים המשמשים אתכם בתשובתכם.

ב. (10 נק') תנו דוגמא לתכונית (feature) אותה לא ניתן להגדיר בפונקציית הניקוד שהגדרתם בסעיף א'. הסבירו מדוע לא ניתן להגדיר אותה. (אין צורך להסביר מדוע קידוד התכונית שהצעתם עשוי להיות מועיל)

שאלה 5 (20 נקודות):

עבור המשפט הבא ועץ התלויות הנתון, רשמו סדרת מעברים (sequence of transitions) אשר מניבה את עץ התלויות הזה, ע״פ מערכת המעברים arc standard וע״פ מערכת המעברים arc eager (ראו נספח).



בהצלחה!