202-2-5211 עיבוד שפה טבעית — תרגיל 2

מועד הגשה: 6 בפברואר 2024, 13:59

משקל התרגיל: 8 נקודות מהציון הסופי. מספר הנקודות של כל שאלה מופיע בסוגריים.

ניתן להגיש את התרגיל בצמדים. העבודה צריכה להתבצע בצמד באופן רוחבי, ולא ע"י חלוקת השאלות בין חבר.ות.י הצמד. ניתן בנוסף להתייעץ עם סטודנטים.ות אחרים.ות אך בסופו של דבר על הפתרון להיות של המגיש.ה.ים.ות בלבד. מומלץ להתחיל לעבוד על התרגיל **מוקדם**.

1 רגולריזציה (1)

[חומרי עזר ממוקדים: פרק 5.7 בספר של סטנפורד, פרק 2.5 בספר של אייזנסטיין.]

פיתוח משוואת המטרה עבור פרמטרים של רגרסיה לוגיסטית עם רגולריזציית L_2 היא כדלקמן:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left[\left[\sum_{i=1}^{m} \log p[y^{(i)}|x^{(i)}] \right] - \gamma \sum_{j=1}^{n} \theta_j^2 \right],$$

כאשר m הוא מספר הדוגמאות בדאטאסט, n הוא מספר הפרמטרים במודל הרגרסיה, $p[y^{(i)}|x^{(i)}]$ מחושב על־ידי פונקציית הסיגמויד, ו־ γ הוא קבוע חיובי.

הוכיחו **אלגברית** כי עבור דאטאסט נתון, אם $\hat{\theta}_L$ הינו הפתרון האופטימלי ללא אלמנט הרגולריזציה, ו־ $\hat{\theta}_R$ הינו הפתרון האופטימלי עם אלמנט הרגולריזציה, אזי מתקיים תמיד

$$\|\hat{\boldsymbol{\theta}}_R\|_2^2 \leq \|\hat{\boldsymbol{\theta}}_L\|_2^2.$$

2 סכימות תיוג רצפים (3)

2.1 חלקי דיבר

בסכימת התיוג לחלקי דיבר של מאגר־העצים־של־פן (PTB) קיים תג בשם TO, שמטרתו לשמש עבור כל המופעים של המילה to. יש שיטענו שלמילה הזו יש למעשה שני תפקידים נבדלים מאוד ועליה לקבל תג שונה בהתאם לכל אחד מהם. לראייה יביאו את השיפוטים הבאים של דוברי אנגלית ילידים, כאשר הסימון כוכבית (*) משמעו משפט שאינו תקין לדובר השפה.

- She's gonna talk about him.
- (*) She's gonna the beach.

הסבירו כיצד המידע הזה תומך בטיעון, והציעו חלק דיבר חלופי (מתוך הסט המקובל של PTB או מתוך הסט האוניברסלי שלמדנו בכיתה) עבור התפקיד של המילה to שמשלב אותה בקטגוריה מוכרת.

2.2 ישויות מקוננות

הזכרנו בשיעור על זיהוי ישויות (NER) את קיומן של ישויות מקוננות בטקסטים, לדוגמה "הועד הפועל של רשות השידור", שבה גם הביטוי כולו וגם "רשות השידור" הינן ישויות העומדות בפני עצמן, והראינו כי סכימת BIO סטנדרטית אינה מספיקה לטפל במקרים כאלה. הציעו סכימה **במסגרת תיוג רצפים** שתאפשר טיפול בישויות מקוננות, או לחלופין הוכיחו באופן קונסטרוקטיבי שלא ניתן לבצע זאת (כלומר, שלכל סכימה מוצעת נוכל ליצור דוגמה נגדית שאינה בת־תיוג במסגרתה).

3 מימוש מתייג רצפים (4)

הורידו את מחברת הקוד ממודל ועיקבו אחר הוראות הביצוע וההגשה על־גביה.

4 העשרה / הכנה: הגדרת בעיות כתיוג רצפים (0)

ללא ציון, ללא בדיקה. נדון בזה בשיעורי החזרה.

תופעת **חילוף הקוד** (code switching), או לעתים **עירוב הקוד** (code mixing), מתייחסת למצב בו דובר.ת רב־לשוני.ת משתמש.ת במילים משתי שפות שונות בתוך אותו מבע (משפט). להלן כמה דוגמאות מעברית־אנגלית (העברית מתועתקת לטובת נוחות הכתב. האות x משמשת לציון ההגה ח, או כ לא־דגושה, ו־ š היא ההגה ש):

alon lakax et ha shoe off.

we limroxed the chocolate on the cake.

ma hi nag'a be?

ha witch kova šeli muxan.

1. הצביעו על קושי אפשרי בניתוח תחבירי של משפט מעורב־קוד, בליווי דוגמא שתמחיש חוסר יכולת להוציא עץ תלויות תקין. ניתן להשתמש בכל זוג שפות שהוא, לרבות בדוגמאות לעיל, ובלבד שהדוגמא תהיה מובנת ודקדוקית במידה סבירה עבור דובריהן. **שימו לב:** הכוונה בשאלה זו אינה ליכולת לזהות את המילים עצמן. הניחו מנתח תחבירי שיודע מה שפת המקור של כל מילה ומה חלק הדיבר שלה.

גישה יותר צנועה להתמודדות עם עירוב קוד היא להתייחס למבע מעורב כאל בעיית תיוג רצפים. בהינתן מבע מעורב, נתייג כל מילה לשפת המקור שלה.

2. האם יש צורך בסכימת תיוג מסוג BIO עבור הבעיה כפי שהוגדרה? נמקו.

3. הציעו סכימה בעלת **ארבעה** תגים לפחות להתמודדות עם תופעות שונות בדוגמאות לעיל, והדגימו את הצורך בכולם.

4. הניחו שלרשותנו דאטא עירוב קוד עברי־אנגלי הנתון בתעתיק כמו בדוגמאות לעיל ומתויג לפי סכימת התיוג שהגדרתסן בסעיף הקודם. אנו מפתחים מערכת תיוג רצפים מבוססת־פיצ'רים עבור הבעיה (למשל, CRF). הציעו ארבע משפחות פיצ'רים (feature templates) למערכת התיוג, והסבירו בקצרה כיצד כל משפחה יכולה לעזור לפעולת החיזוי.