עיבוד שפה טבעית - תרגיל בית 2

מגישים: דביר בן זיקרי 315409508

עמית זולן 207299033

1. על מנת להוכיח $\|\widehat{\theta_R}\|_2^2 \leq \|\widehat{\theta_L}\|_2^2$ נסתכל על ההגדרה הכללית של פונקציית המטרה של $\widehat{\theta_R}$ עם אלמנט רגולרזיציה L_2:

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

כאשר עבור $\widehat{\theta_L}$ –

$$\widehat{\theta_L} = argmax_\theta \left[\sum_{i=1}^{m} \log p\left[y^{(i)} | \, x^{(i)}\right] \right]$$

מאחר וזהו הפתרון ללא אלמנט הרגולריזציה.

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

$$\widehat{\theta_R} = argmax_\theta [p(\theta) - q(\theta)], \qquad \widehat{\theta_L} = argmax_\theta [p(\theta)]$$

אזי מאחר ו$\widehat{\theta_L}$ ו$\widehat{\theta_R}$ הינם פתרונות אופטימליים בלי ועם אלמנט הרגולריזציה בהתאמה, לכן

$$(1)\ p\left(\widehat{\theta_L}\right) \geq p\left(\widehat{\theta_R}\right)$$
$$(2)\ p\left(\widehat{\theta_R}\right) - q\left(\widehat{\theta_R}\right) \geq p\left(\widehat{\theta_L}\right) - q\left(\widehat{\theta_L}\right)$$

לפי הגדרת בעיות האופטימזציה הנ"ל.

נוכיח את הטענה בשלילה:

נניח כי $\|\widehat{\theta_R}\|_2^2 > \|\widehat{\theta_L}\|_2^2$ כלומר,

$$(3)\ \ q\left(\widehat{\theta_R}\right) > q\left(\widehat{\theta_L}\right)$$

זאת מאחר ש $\gamma > 0$ בבעיית האופטימזציה עם אלמנט הרגולריזציה L_2.

אזי:

$$p(\widehat{\theta_R}) - q(\widehat{\theta_R}) \overset{(3)}{\lesssim} p(\widehat{\theta_R}) - q(\widehat{\theta_L}) \overset{(1)}{\lesssim} p(\widehat{\theta_L}) - q(\widehat{\theta_L})$$

כלומר קיבלנו כי

$$p(\widehat{\theta_R}) - q(\widehat{\theta_R}) \leq p(\widehat{\theta_L}) - q(\widehat{\theta_L})$$

בסתירה לכך ש $\widehat{\theta_R}$, הינו הפתרון האופטימלי לבעיית האופטימזציה עם אלמנט הרגולריזציה L_2.

.2

.2.1 במשפטים הנ"ל קיימת המילה באנגלית – *gonna* שהינה עגה לצמד המילים *going to*, וכן למילה *to* יש שני חלקי דיבר שונים:

- *She's gonna talk about him.* – המילה *to* הינה *Part* מצמד המילים *to talk*.
- *She's gonna the beach* – המילה *to* הינה *ADP* (מילת יחס) המכוונת לאן היא הולכת.

מאחר והמילה *to* מסתתרת בתוך המילה *gonna* נוצר משפט לא תקין כאשר *to* הינה מילת יחס ואיננה *part*.

ולכן שני המשפטים הנ"ל תומכים בטיעון שהתיוג *TO* למילה *to* איננו מספיק ולכן אנו מציעים את התיוג *ADP* כאשר *to* הינה מילת יחס.

ניתן להשאיר את התיוג *TO* כאשר הינה *PART* אך אפשר גם להחליפה ב*PART*.

2.2. בבעית זיהוי ישויות, הקלט שלנו הינו מסמך בעל *n* מילים, והפלט הינו *n* תגים, כלומר, תג לכל מילה. בעיית ישויות מקוננות הינה תופעה בה מופיעה ישות בתוך ישות ובכך אין אנו מזהים את הישות המקוננת.
על מנת לפתור את הבעיה הזו יש צורך בדרך לתייג מילה כך שתהייה משוייכת ליותר מישות אחת.
ניתן להוכיח שאין דרך לבצע זו על ידי הסתכלות על הבעיה באופן אינדוקטיבי:
נניח כי קיים משפט באורך *n* שמייצג ישות ומילה בתוך המשפט שמשוייכת ל$k > 0$ ישויות, אם ניקח את הישות הגדולה (כל המשפט), ונקונן אותה כחלק מישות חדשה, כך שנקבל כי המילה כעת משוייכת ל$k + 1$ ישויות –נקבל כי המילה צריכה תג שמשייך אותה ל $k + 1$ ישויות, אם נמשיך כך באופן אינדוקטיבי נקבל כי נצטרך תג לכל $k + i$ הישויות שהמילה משוייכות אליהן בכל שלב i, ולכן אם נסתכל על שלב $i + 1$ נצטרך שוב תג נוסף. כלומר אם נשאיף את *i* לאינסוף נצטרך אינסוף תגים.

3. מצ"ב כמחברת.
4.

- אחד מאיתנו הינה קצין בסדיר ולכן רואה את רוב ההרצאות באופן מוקלט, והשני רואה הרצאות פעם נוספת כדי לחדד את הידע, ולכן נשמח אם תוכל לשים את המצלמה (מהזום) בצד השני של המצגת מאחר והיא מסתירה לפעמים את הכותרת ואף חלק מהטקסט במצגות.
- נשמח אם תוכל לציין יותר דוגמאות בחלקי הבלשנות של ההרצאות בין אם תוך כדי ההרצאה או לאחר מכן במודל.