עיבוד שפה טבעית – תרגיל בית 1 – אייל באן ונדב גת

<u>שאלה 2</u>

.1

$$softmax(x+c)_i = \frac{\exp(x_i + c)}{\sum_j \exp(x_j + c)} = \frac{e^{x_i} \cdot e^c}{\sum_j (e^{x_j} \cdot e^c)} = \frac{e^{x_i} \cdot e^c}{e^c \cdot \sum_j e^{x_j}} = \frac{e^{x_i}}{\sum_j e^{x_j}} = softmax(x)_i$$

.2

$$-\sum_{w \in W} y_w \log(\widehat{y_w}) = -\sum_{w \neq o} 0 \cdot \log(\widehat{y_w}) - 1 \cdot \log(\widehat{y_o}) = -\log(\widehat{y_o})$$

.3

$$\begin{aligned}
\nabla_{v_c} - \log P(O = o|C = c) &= \nabla_{v_c} - \log\left(\frac{e^{u_o^T v_c}}{\sum_{w \in W} e^{u_w^T v_c}}\right) \\
&= \nabla_{v_c} - \log\left(e^{u_o^T v_c}\right) + \log\left(\sum_{w \in W} e^{u_w^T v_c}\right) \\
&= \nabla_{v_c} - u_o^T v_c + \log\left(\sum_{w \in W} e^{u_w^T v_c}\right) = -u_o^T + \frac{\nabla_{v_C}\left(\sum_{w \in W} e^{u_w^T v_c}\right)}{\sum_{w' \in W} e^{u_{w'}^T v_c}} \\
&= -u_o^T + \frac{\sum_{w \in W} u_w^T \cdot e^{u_w^T v_c}}{\sum_{w' \in W} e^{u_{w'}^T v_c}} = -u_o^T + \sum_{w \in W} \frac{u_w^T \cdot e^{u_w^T v_c}}{\sum_{w' \in W} e^{u_{w'}^T v_c}} \\
&= -u_o^T + \sum_{w \in W} u_w^T \cdot P(O = w'|C = c) = -u_o^T + \sum_{w \in W} \hat{y}_w u_w^T \\
&= -Uy + U\hat{y}
\end{aligned}$$

.4

במקרה בו $w' \neq o$:

$$\begin{aligned}
\nabla_{u_{w'}} - \log P(O = o \mid C = c) &= \nabla_{u_{w'}} - \log\left(\frac{\exp(u_o^T \cdot v_c)}{\sum_{w \in W} \exp(u_w^T \cdot v_c)}\right) \\
&= \nabla_{u_{w'}}\left(-u_o^T \cdot v_c + \log\left(\sum_{w \in W} \exp(u_w^T \cdot v_c)\right)\right) \\
&= \frac{1}{\sum_{w \in W} \exp u_w^T \cdot v_c} \cdot \nabla_{u_{w'}}\left(\sum_{w \in W} \exp(u_w^T \cdot v_c)\right) = \frac{v_c \cdot \exp(u_{w'}^T \cdot v_c)}{\sum_{w \in W} \exp(u_w^T \cdot v_c)} \\
&= \widehat{y_{w'}} v_c
\end{aligned}$$

במקרה בו $w' = o$:

$$\nabla_{u_o} -\log P(O=o \mid C=c) = \nabla_{u_o} -\log\left(\frac{\exp(u_o^T v_c)}{\sum_{w\in W}\exp(u_w^T v_c)}\right)$$
$$= \nabla_{u_o}\left(-u_o^T v_c + \log\left(\sum_{w\in W}\exp(u_w^T v_c)\right)\right)$$
$$= -v_c + \frac{1}{\sum_{w\in W}\exp(u_w^T v_c)}\cdot v_c \cdot \exp(u_o^T v_c) = (\widehat{y_o}-1)v_c$$

.5

$$\nabla_x \sigma(x) = \nabla_x \frac{1}{1+e^{-x}} = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}}\cdot\frac{e^{-x}}{1+e^{-x}} = \sigma(x)\cdot\frac{e^{-x}}{1+e^{-x}}$$
$$= \sigma(x)\cdot\frac{1+e^{-x}-1}{1+e^{-x}} = \sigma(x)\cdot\left(1-\frac{1}{1+e^{-x}}\right) = \sigma(x)-\sigma^2(x)$$

.6

$$\nabla_{v_c} J_{neg-sample}(v_c,o,U) = \nabla_{v_c} -\log\big(\sigma(u_o^T v_c)\big) - \sum_{k=1}^{K}\log\big(\sigma(-u_k^T v_c)\big)$$
$$= \nabla_{v_c}\big(-\log\big(\sigma(u_o^T v_c)\big)\big) + \nabla_{v_c}\left(\sum_{k=1}^{K} -\log\big(\sigma(-u_k^T v_c)\big)\right)$$
$$= -\frac{\nabla_{v_c}\sigma(u_o^T v_c)}{\sigma(u_o^T v_c)} + \sum_{k=1}^{K} -\frac{\nabla_{v_c}\sigma(-u_k^T v_c)}{\sigma(-u_k^T v_c)}$$
$$= -\frac{u_o^t\sigma(u_o^T v_c) - u_o^t\sigma^2(u_o^T v_c)}{\sigma(u_o^T v_c)} + \sum_{k=1}^{K} -\frac{-u_k^T\sigma(-u_k^T v_c) + u_k^T\sigma^2(-u_k^T v_c)}{\sigma(-u_k^T v_c)}$$
$$= -u_o^T + u_o^T\sigma(u_o^T v_c) + \sum_{k=1}^{K} u_k^T - u_k^T\sigma(-u_k^T v_c) =$$
$$u_o^T(\sigma(u_o^T v_c)-1) + \sum_{k=1}^{K} u_k^T\big(1-\sigma(-u_k^T v_c)\big)$$

$$\nabla_{u_o} J_{neg-sample}(v_c,o,U) = \nabla_{u_o} -\log\big(\sigma(u_o^T v_c)\big) - \sum_{k=1}^{K}\log\big(\sigma(-u_k^T v_c)\big)$$
$$= -\frac{1}{\sigma(u_o^T v_c)}\big(v_c\cdot(\sigma(u_o^T v_c)-\sigma^2(u_o^T v_c))\big) = -\frac{v_c\cdot\sigma(u_o^T v_c) - v_c\cdot\sigma^2(u_o^T v_c)}{\sigma(u_o^T v_c)}$$
$$= v_c\cdot\sigma(u_o^T v_c) - v_c = v_c(\sigma(u_o^T v_c)-1)$$

$$\nabla_{u_k} J_{neg-sample}(v_c,o,U) = \nabla_{u_k} -\log\big(\sigma(u_o^T v_c)\big) - \sum_{k=1}^{K}\log\big(\sigma(-u_k^T v_c)\big)$$
$$= 0 - \frac{\nabla_{u_k}\sigma(-u_k^T v_c)}{\sigma(-u_k^T v_c)} = \frac{v_c\sigma(-u_k^T v_c) - v_c\sigma^2(-u_k^T v_c)}{\sigma(-u_k^T v_c)} = v_c - v_c\sigma(-u_k^T v_c)$$
$$= v_c\big(1-\sigma(-u_k^T v_c)\big)$$

החישוב עבור *loss* זה הוא יותר יעיל כיוון שדורש חישוב עם פחות פרמטרים – *k+1* במקום $2|W|$ וקטורים.

.7

$$\partial J_{skip-gram}(v_c, w_{t-m}, \dots, w_{t+m}, U)/\partial U = \partial \sum_{\substack{-m \le j \le m \\ j \ne 0}} J\left(v_c, w_{t+j}, U\right)/\partial U$$

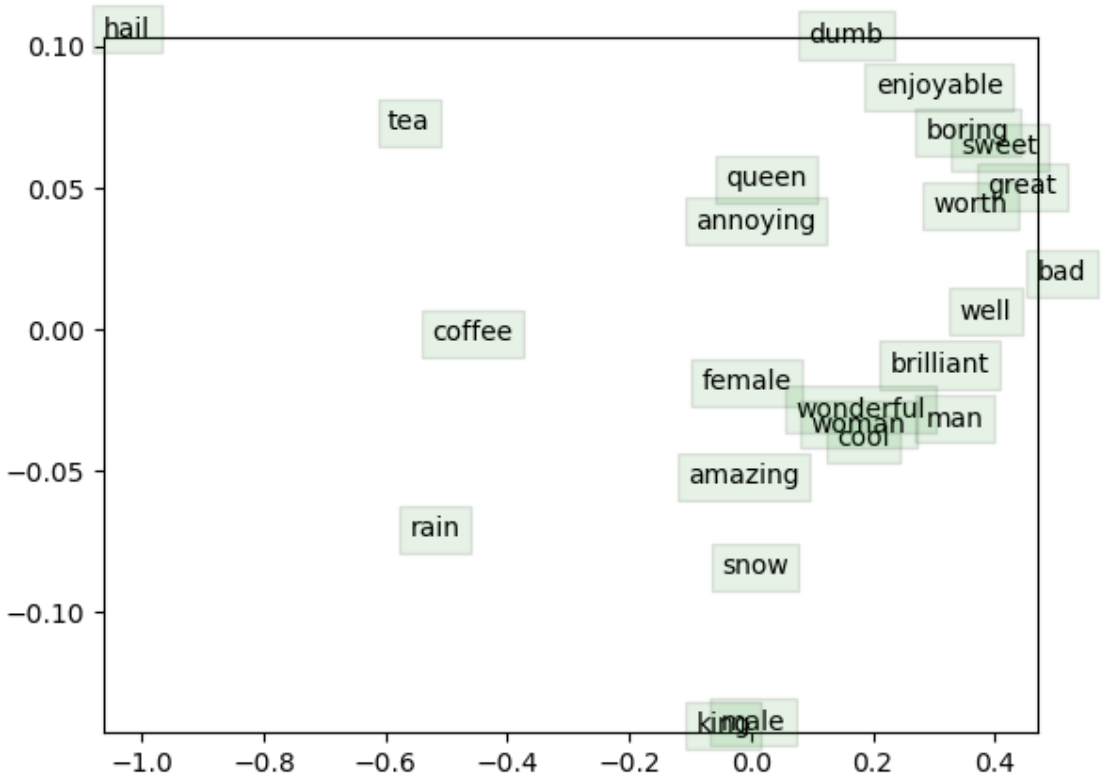
$$= \sum_{\substack{-m \le j \le m \\ j \ne 0}} \partial J\left(v_c, w_{t+j}, U\right)/\partial U$$

$$\partial J_{skip-gram}(v_c, w_{t-m}, \dots, w_{t+m}, U)/\partial v_c = \partial \sum_{\substack{-m \le j \le m \\ j \ne 0}} J\left(v_c, w_{t+j}, U\right)/\partial v_c$$

$$= \sum_{\substack{-m \le j \le m \\ j \ne 0}} \partial J\left(v_c, w_{t+j}, U\right)/\partial v_c$$

$$\partial J_{skip-gram}(v_c, w_{t-m}, \dots, w_{t+m}, U)/\partial v_w = 0$$

<u>שאלה 3</u>



In this plot, we can clearly see that all the adjectives formed a cluster in the top right of the plot. The fact that the adjectives cluster together can be explained by the fact that they are used

נשים לב כי בגרף זה, כל הפעלים התקבצו יחדיו בפינה הימנית העליונה. ניתן להסביר זאת משום שכל אלה יש תפקיד תחבירי זהה (פועל) ולכן יופיעו במקומות דומים במשפט. כמו כן, מרבית הפעלים המופיעים ברשימה הם פעלים "גנריים", כלומר יכולים לתאר מגוון רחב מאוד של שמות עצם, כך שאין להם השפעה סמנטית גדולה.

כמו כן, מרבית שמות העצם נמצאים מחוץ למקבץ הפעלים וכן במרחק יחסית גדול אחד מהשני. ניתן להסביר זאת בכך שלעומת הפעלים הגנריים, לשמות העצם ישנה נטייה סמנטית. כלומר שם עצם כמו "hail" סביר שיופיע יחד עם מילים אחרות הקשורות לחורף ולמשקעים, ולכן לא סביר שיהיה דומה למילה "*coffee*".

מהסיבות שציינו לעיל, ציפינו לקבל תוצאה דומה למה שקיבלנו, כלומר ש"דימיון" בין מילים יקבע לפי הקרבה הסינטקטית שלהם. אמנם לדעתנו, צריכה להיות השפעה גדולה יותר גם על הסמנטיקה של המילים. לדוגמה, *"enjoyable"* ו-*"boring"* אמורות להיות די רחוקות אחת מן השנייה כי המשמעות שלהן הפוכה, אמנם הן קרובות מאוד בייצוג הוקטורי שכן הן יופיעו במקומות מאוד דומים.

שאלה 4

א. נשים לב כי מספר הפעמים שהביטוי $p_\theta(o \mid c)$ מופיע ב-$\mathcal{L}(\theta)$ הוא בדיוק מספר הפעמים שהמילה o מופיעה בחלון של c (כלומר ב-$2m$ המילים הכי קרובות אליה) - $\#(c,o)$, כיוון שהוא מופיע פעם אחת בדיוק בכל פעם שזוג המילים הופיעו יחד. נגדיר בעיית אופטימיזציה, שבה α_o ייצג את $p_\theta(o \mid c)$:

$$\max \sum_o \#(c,o) \cdot \log \alpha_o$$
$$s.t. \left(\sum_o \alpha_o\right) - 1 = 0, \alpha_o \geq 0$$

הורדנו את הסכום במכנה בפונקציית המטרה כיוון שהוא קבוע ולכן לא ישפיע על הערך בו יתקבל המקסימום. נגדיר לגרנז'יאן:

$$f(\alpha,\lambda) = \sum_o \#(c,o) \cdot \log \alpha_o - \lambda + \lambda \sum_o \alpha_o$$

$$\nabla_{\alpha_o} f(\alpha,\lambda) = \lambda + \frac{\#(c,o)}{\alpha_o} = 0 \Rightarrow \alpha_o = -\frac{\#(c,o)}{\lambda}$$

וכעת כדי שהפרמטרים α אכן ייצגו הסתברות הם חייבים להיסכם לאחד, כלומר:

$$\lambda = -\sum_{o'} \#(c, o')$$

ואז מתקיים:

$$\sum_{o} \alpha_o = \sum_{o} -\frac{\#(c,o)}{-\sum_{o'} \#(c,o')} = \frac{\sum_{o} \#(c,o)}{\sum_{o'} \#(c,o')} = 1$$

כנדרש. קיבלנו:

$$\alpha_o = p(o \mid c) = \frac{\#(c,o)}{\sum_{o'} \#(c,o')}$$

כלומר גם מקיים את תנאי הא"ש הנדרש בבעיית האופטימיזציה, וגם מוכיח את הנדרש.

ב. נגדיר:

$$W = \{a, b, c, d\}, \qquad C = \{ab, bc, cd\}$$

ואז הפתרון האופטימלי יהיה (לפי סעיף א'):

$$p(a \mid b) = p(c \mid b) = \frac{1}{2}, \qquad p(c \mid d) = 1, \qquad p(a \mid d) = 0$$

נניח כי U, V וקטורים שמתקבלים כתוצאה מאלגוריתם *skipgram* שעבורם מתקיימים שני התנאים הראשונים. אז מתקיים:

$$p(a \mid b) = \frac{\exp(u_1 v_2)}{\sum_{i=1}^{4} \exp(u_i v_2)} = \frac{\exp(u_3 v_2)}{\sum_{i=1}^{4} \exp(u_i v_2)} = p(c \mid b) \Rightarrow u_1 = u_3$$

כאשר במעבר האחרון הנחנו ש-$v_2 \neq 0$ (u_i הוא הסקלר המייצג את המילה ה-i בוקטור U, v_i כנ"ל ב-V). אבל אז יתקיים גם:

$$p(a \mid d) = \frac{\exp(u_1 v_4)}{\sum_{i=1}^{4} \exp(u_i v_2)} = \frac{\exp(u_3 v_4)}{\sum_{i=1}^{4} \exp(u_i v_2)} = p(c \mid d)$$

ובפרט לא יתקיימו שני התנאים האחרונים של הפתרון האופטימלי.

כעת, אם מתקיים $v_2 = 0$ אז:

$$\forall 1 \leq j \leq 4. \frac{\exp(u_j v_2)}{\sum_{i=1}^{4} \exp(u_i v_2)} = \frac{1}{4}$$

בסתירה לקיום שני התנאים הראשונים. לכן בכל מקרה לא יתכן שאלגוריתם *skipgram* יחזיר את הפתרון האופטימלי במקרה המתואר.

שאלה 5

א. ראשית נבחין כי לכל $x \in \mathbb{R}^d$, מתקיים כי $relu(x) \in \mathbb{R}_+^d$ כאשר $\mathbb{R}_+$ הוא הממשיים האי-שליליים. כלומר מתקבל וקטור שבכל קורדינטה הוא אי-שלילי. אם כך, לכל $x_1, x_2 \in \mathbb{R}^d$ מתקיים ש- $relu(x_1)^T relu(x_2) \geq 0$ שכן מכפלה של וקטורים אי-שליליים היא אי-שלילית.

משום שהמודל הוא בינארי, עבור $p(the\ pair\ is\ a\ paraphrase|x_1, x_2) \geq 0.5$ המודל מתייג 1, כלומר קובע שהזוג הוא פרפרזה. נבחין כי לכל $a \geq 0$ מתקיים $\sigma(a) \geq 0.5$ ולכן לכל $x_1, x_2 \in \mathbb{R}^d$ מתקיים $p(the\ pair\ is\ a\ paraphrase|x_1, x_2) = \sigma\big(rely(x_1)^T relu(x_2)\big) \geq 0.5$ ולכן לכל זוג x_1, x_2, המודל יקבע שהם פרפרזה.

קיבלנו שעל *dataset* המכיל 25% דוגמאות חיוביות וכל היתר שליליות, המודל יקבע כי כל דוגמה היא פרפרזה, קרי יצדק על 25% ועל 75% הנוספות ישגה. לכן הדיוק המירבי שניתן להשיג הוא 25%.

ב. ניתן להשתמש במודל הבא:

$$p(the\ pair\ is\ a\ paraphrase|x_1, x_2) = 2(\sigma(relu(x_1)^T relu(x_2) - 0.5)$$

ראשית נבחין כי במקרה הזה המסווג לא תמיד יחזיר 1 כי ישנם קלטים עבורם $p < 0.5$.

עבור שני וקטורים x, y דומים, ניתן להניח כי עבור קורדינטה i, אם $x_i > 0$ אזי גם $y_i > 0$. לעומת זאת עבור וקטורים x', y' שונים סביר שעבור הקורדינטה i, אין תלות ולכן הסיכוי שגם $x_i, y_i > 0$ הוא $\frac{1}{4}$. אם כך נצפה כי עבור וקטורים דומים נקבל תוצאה גבוהה. בעוד שעבור וקטורים שונים נצפה שעבור *75%* מהקורדינטות ב-x', y' לפחות אחד מן הוקטורים שלילי ולכן יתאפסו תחת ה-*relu*, ובכך "תורמות" 0 למכפלה.

מודל אלטרנטיבי הוא:

$$p(the\ pair\ is\ a\ paraphrase|x_1, x_2) = \sigma(x_1^T x_2)$$

גם במקרה זה נבחין כי קיימים קלטים שעליהם המודל לא מחזיר 1.

עבור וקטורים x, y דומים נצפה כי ישנה התאמה בסימן השליליות של כל קורדינטה i ולכן נקבל מכפלה $x_1^T x_2 > 0$ ואף גבוהה. לעומת זאת עבור שני וקטורים x', y' שונים, נצפה כי אין התאמה בין סימן השליליות של כל קורדינטה ולכן רק ב50% מהקורדינטות נקבל מכפלה חיובית ואילו בשאר מכפלה שלילית. במקרה זה סביר כי התוצאה לא תהיה חיובית גבוהה.