עיבוד שפה טבעית - תרגיל 1

מגיש: עידן גרינשטיין , ת.ז: 307960682

2 Understanding word2vec

(a)

$$\text{softmax}(x+c)_i = \frac{e^{x_i+c}}{\sum_{j=1}^{n} e^{x_j+c}} = \frac{e^{x_i}\cdot e^{c}}{e^{c}\sum_{j=1}^{n} e^{x_j}} =$$

$$= \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} = \text{softmax}(x)_i$$

זה נכון לכל $n \geq i \geq 1$ ולכן מתקיים:

$$\text{softmax}(x) = \text{softmax}(x+c)$$

(b)

$$-\sum_{w\in W} y_w \log(\hat{y}_w) = \Big(-\sum_{\substack{w\in W\\ w\neq o}} y_w \log \hat{y}_w\Big) - y_o \log(\hat{y}_o) =$$

$$= \Big(-\sum_{\substack{w\in W\\ w\neq o}} 0\cdot \log \hat{y}_w\Big) - 1\cdot \log(\hat{y}_o) = -\log \hat{y}_o$$

(c)

$$J(v_c, o, U) = -\log(\hat{y}_o) = -\log\left(\frac{e^{u_o^T v_c}}{\sum_{w\in W} e^{u_w^T v_c}}\right) =$$

$$= -u_o^T v_c + \log\Big(\sum_{w\in W} e^{u_w^T v_c}\Big)$$

נפעיל גרדיאנט לפי v_c:

$$\nabla J = -u_o + \frac{1}{\sum_{w\in W} e^{u_w^T v_c}} \cdot \left(\sum_{w\in W} e^{u_w^T v_c} \cdot u_w \right) =$$

$$= -u_o + \sum_{w\in W} \left(\frac{e^{u_w^T v_c}}{\sum_{w\in W} e^{u_w^T v_c}} \cdot u_w \right) =$$

$$= -u_o + \hat{y}_o u_o + \sum_{w\in W\setminus\{o\}} \hat{y}_w u_w = U(\hat{y} - y)$$

(d) עבור $w=o$:

$$\nabla J = -v_c + \frac{1}{\sum_{w\in W} e^{u_w^T v_c}} \left(e^{u_o^T v_c} \cdot v_c \right) =$$

$$= -v_c + \hat{y} v_c = (\hat{y}_w - 1) v_c$$

עבור $w \neq o$:

$$\nabla J = \frac{1}{\sum_{w\in W} e^{u_w^T v_c}} \left(e^{u_w^T v_c} \cdot v_c \right) = \hat{y}_w v_c$$

(e) עבור $x \in \mathbb{R}$:

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

כמוכן מתקיים:

$$1 - \sigma(x) = \frac{1+e^{-x}-1}{1+e^{-x}} = \frac{e^{-x}}{1+e^{-x}} = \sigma(-x)$$

$$\nabla \sigma(x) = \sigma(x)(1-\sigma(x))$$

כעת, עבור $x \in \mathbb{R}^d$, וקטור במימד d, נקבל:

$$D_\sigma = \begin{pmatrix} \sigma(x_1)(1-\sigma(x_1)) & & \\ & \ddots & \\ & & \sigma(x_d)(1-\sigma(x_d)) \end{pmatrix}$$

כהנחה שפונקציית הסיגמואיד על וקטור פועלת איבר-איבר.

(f)

$$\frac{\partial J}{\partial V_c} = -\frac{1}{\sigma(U_o^T V_c)} \cdot \sigma(U_o^T V_c)(1-\sigma(U_o^T V_c))U_o -$$

$$-\sum_{k=1}^{K} -\frac{1}{\sigma(-U_k^T V_c)} \cdot \sigma(-U_k^T V_c)(1-\sigma(-U_k^T V_c))U_k =$$

$$= \boxed{-(1-\sigma(U_o^T V_c))U_o + \sum_{k=1}^{K} \sigma(U_k^T V_c)U_k}$$

$$\frac{\partial J}{\partial U_o} = -\frac{1}{\sigma(U_o^T V_c)} \cdot \sigma(U_o^T V_c)(1-\sigma(U_o^T V_c))V_c =$$

$$= \boxed{-(1-\sigma(U_o^T V_c))V_c}$$

$$\frac{\partial J}{\partial U_k} = \frac{1}{\sigma(-U_k^T V_c)} \cdot \sigma(-U_k^T V_c)(1-\sigma(-U_k^T V_c))V_c =$$

$$= \boxed{\sigma(U_k^T V_c) \cdot V_c}$$

פונקציית המחיר של ה-naive softmax דורשת מעבר על כל המילון (ביצוע מכפלות פנימיות וסכימה) לעומת פונקציית המחיר הנוכחית דורשת מעבר על K+1 מילים לכל זוג (V_c, U_o).

(g)

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

$$= \sum_{\substack{-m \le j \le m \\ j \neq 0}} \frac{\partial J}{\partial U}(v_c, w_{t+j}, U)$$

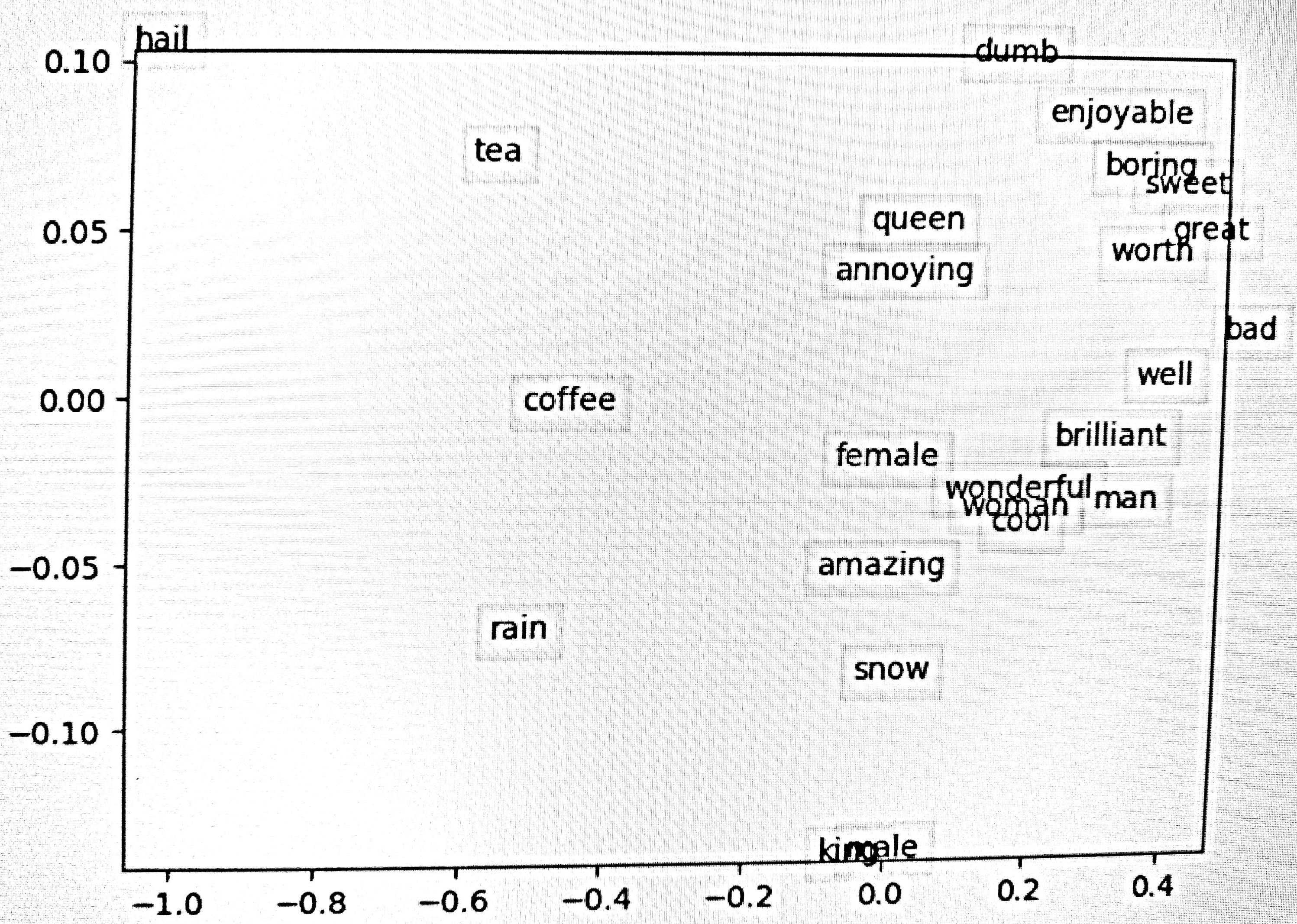
$$\partial J_{skip\text{-}gram}(v_c, w_{t-m}, \dots, w_{t+m}, U) / \partial v_c =$$

$$= \frac{\partial J}{\partial v_c}(v_c, w_{t+j}, U)$$

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

מכיוון שהביטוי אינו תלוי ב-V_w כאשר $w \neq c$.

③ (ש) מהגרף שהתקבל ניתן לראות שעיקר הקיבוצים
הם של מילים בעלות משמעות סינטקטית שונה
ולא בעלי משמעות סמנטית שונה. לדוגמא,
נראה שיש קיבוץ של שמות תואר כגון
brilliant, wonderful אשר הינן מילים בעלי משמעות
סינטקטית שונה. לעומת זאת, מילים בעלי משמעות
סמנטית שונה כגון queen, king רחוקים זה
מזה. אולי הגדלת החלון, או הגדרת חלון
בצורה שתתפוס משמעות של מילים בצורה טובה יותר
ויגרום לצירוף סמנטי יותר טוב


hail
dumb
enjoyable
tea
boring
sweet
queen
great
worth
annoying
bad
well
coffee
brilliant
female
wonderful
woman
man
cool
amazing
rain
snow
king
male
0.10
0.05
0.00
−0.05
−0.10
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4

(4) (א) נסתכל על הפונקציה $J(\theta)$:

$$J(\theta) = \log L(\theta) = \sum_{t=1}^{T} \sum_{-m \le j \le m,\ j \ne 0} \log P_\theta(w_{t+j} \mid w_t)$$

נמצא מקסימום של J והוא יהיה גם המקסימום של L, מכיוון ש-$\log$ הינה פונקציה מונוטונית.

מתקיים:

$$J = \sum_{c \in V} \sum_{o \in V} \#(c,o) \log P_\theta(o \mid c)$$

לכל מילת מרכז c, נמצא את ההסתברויות שממקסמות את:

$$\sum_{o \in V} \#(c,o) \log P_\theta(o \mid c)$$

נגדיר:

$$f = \sum_{i=1}^{n} \#(c,o_i) \log x_i$$

כאשר n הוא מספר המילים אשר מופיעות בסמוך ל-c ולכל מילה כזאת ניתן אינדקס בין 1 ל-n, $x_i = P_\theta(o_i \mid c)$.

בנוסף נגדיר את האילוץ:

$$g = \sum_{i=1}^{n} x_i - 1$$

מכיוון ש- $\{x_i\}_{i=1}^{n}$ הסתברויות.

כעת נשתמש בכופלי לגרנז' כדי למצוא מקסימום:

$$\nabla f = \left(\frac{\#(c,o_1)}{x_1}, \dots, \frac{\#(c,o_n)}{x_n} \right)$$

$$\nabla g = (1, \dots, 1)$$

נמצא פתרון ל:

$$\begin{cases} \nabla f = \lambda \nabla g \\ g = 0 \end{cases}$$

$$\begin{cases} \frac{\#(c,o_1)}{x_1} = \lambda \\ \vdots \\ \frac{\#(c,o_n)}{x_n} = \lambda \\ x_1 + \dots + x_n = 1 \end{cases} \Rightarrow \quad \lambda = \sum_{i=1}^{n} \#(c,o_i) \qquad x_i = \frac{\#(c,o_i)}{\sum_{i=1}^{n} \#(c,o_i)}$$

זהו נקודת מקסימום מכיוון ש-f היא חסומה מלמעלה ולכן קיבלנו את הדרוש, כלומר הפתרון האופטימלי מתקבל כאשר:

$$P_\theta(o|c) = \frac{\#(o,c)}{\sum_{o'} \#(o',c)}$$

(b) נניח המילון שלנו הינו $V=\{A,B\}$ והקורפוס שלנו הינו "AB".
אזי ההסתברויות האופטימליות הינן:

$$P(B|A)=1 \quad , \quad P(A|B)=1$$

אבל:

$$P_\theta(A|B) = \frac{e^{O_a \cdot C_b}}{e^{O_a \cdot C_b} + e^{O_b C_b}} < 1$$

מכיוון ש- $e^x \neq 0$ לכל $x \in \mathbb{R}$ ולכן לא ניתן להגיע להסתברויות האופטימליות במודל זה.

(5) (a) נשים לב שמתקיים $\sigma(x)\geq 0.5$ לכל
$x\geq 0$. בנוסף, $relu(x_1)^T\cdot relu(x_2)$ הינו אי
שלילי ולכן המודל שלנו תמיד ינבא תוצאה
חיובית.

(ii) הסיכוי המקסימלי של המודל על מדגם
האימון הוא $0.25=\frac{1}{3+1}$.

(b) תיקון פשוט לבעיה זו יכול להיות
הרכבת פונקציית ה-relu, אך השארת ה-relu
והעברה של אותם הערכים מ-$[0.5,1]$ ל-$[0,1]$
על ידי הפונקציה $2x-1$ הפועלת
על $relu(x_1)^T\cdot relu(x_2)$.