

1. Softmax - NLP

$$\text{Softmax}(x+c)_i = \frac{e^{x_i+c}}{\sum_j e^{x_j+c}} = \frac{e^{x_i} e^c}{e^c \sum_j e^{x_j}} = \text{softmax}(x)_i \quad (1) \quad (2)$$

$$\begin{aligned} \sum_{\omega \in W} y_{\omega} \log \hat{y}_{\omega} &= 0 + 0 + \dots + 0 + 1 \cdot \log \hat{y}_0 + \dots + 0 = \\ &= \log \hat{y}_0 = P(o=o | c=c) \end{aligned} \quad (3)$$

$$\frac{\partial \text{J}_{\text{naive-sm}}(V_c, o, U)}{\partial V_c} = -\frac{\partial}{\partial V_c} \left(-\log \frac{e^{u_o^T V_c}}{\sum_{\omega \in W} e^{u_{\omega}^T V_c}} \right) = \quad (4)$$

$$= \frac{\partial}{\partial V_c} \left(-u_o^T V_c + \log \sum_{\omega \in W} e^{u_{\omega}^T V_c} \right) = -u_o + \frac{1}{\sum_{\omega \in W} e^{u_{\omega}^T V_c}} \cdot \sum_{\omega \in W} \frac{\partial}{\partial V_c} (e^{u_{\omega}^T V_c}) =$$

$$= -u_o + \frac{1}{\sum_{\omega \in W} e^{u_{\omega}^T V_c}} \sum_{k \in W} e^{u_k^T V_c} \cdot u_k = \frac{\sum_{k \in W} e^{u_k^T V_c} \cdot u_k}{\sum_{\omega \in W} e^{u_{\omega}^T V_c}} - u_o =$$

$$= \sum_k P(k|c) u_k - u_o = \sum_k \hat{y}_k u_k - u_o$$

$$\frac{\partial}{\partial u_{\omega}} \text{J}_{\text{naive-sm}}(V_c, o, U) = \frac{\partial}{\partial u_{\omega}} \left(-u_o^T V_c + \log \sum_{k \in W} e^{u_k^T V_c} \right) = \quad (5)$$

$$= -V_c + \frac{1}{\sum_{k \in W} e^{u_k^T V_c}} \cdot e^{u_{\omega}^T V_c} \cdot V_c =$$

$$= V_c (P(\omega|c) - 1) = V_c (\hat{y}_{\omega} - y_{\omega})$$

$$\sigma'(x) = \frac{\partial}{\partial x} \left(\frac{1}{1+e^{-x}} \right) = -\frac{(1+e^{-x})^{-2} \cdot (-e^{-x})}{1} = \quad (6)$$

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

$$\frac{\partial \text{J}_{\text{naive-sm}}(V_c, o, U)}{\partial V_c} = -\frac{1}{\sigma(u_o^T V_c)} \sigma'(u_o^T V_c) (1-\sigma(u_o^T V_c)) u_o = \quad (7)$$

$$= -\sum_k \frac{1}{\sigma(u_k^T V_c)} \cdot \sigma'(-u_k^T V_c) (1-\sigma(-u_k^T V_c)) \cdot (-u_k) =$$

$$= u_o (\sigma(u_o^T V_c) - 1) + \sum_k u_k (1 - \sigma(-u_k^T V_c))$$

$$\frac{\partial \text{neg-sample}(V_c, 0, U)}{\partial U_0} = \frac{-1}{\sigma(u_0^T V_c)} \sigma(u_0^T V_c) (1 - \sigma(u_0^T V_c)) \cdot V_c$$

$$= -V_c (1 - \sigma(u_0^T V_c))$$

$$\frac{\partial \text{neg-sample}(V_c, 0, U)}{\partial U_K} = \frac{1}{\sigma(-u_K^T V_c)} \sigma(-u_K^T V_c) (1 - \sigma(-u_K^T V_c)) \cdot (-V_c)$$

$$= V_c (1 - \sigma(-u_K^T V_c)) = V_c \sigma(u_K^T V_c)$$

המשפט — כיצד נבנה את המודל — כיצד נבנה את המודל — כיצד נבנה את המודל

$$\frac{\partial J_{\text{skip-gram}}(V_c, w_{t-m}, \dots, w_{t+m}, U)}{\partial U} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial}{\partial U} J(V_c, w_{t+j}, U) \quad (1)$$

$$\frac{\partial J_{\text{se}}(V_c, w_{t-m}, \dots, w_{t+m}, U)}{\partial V_c} = - \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial}{\partial V_c} J(V_c, w_{t+j}, U) \quad (2)$$

$$\left. \frac{\partial J_{\text{se}}(V_c, w_{t-m}, \dots, w_{t+m}, U)}{\partial V_c} \right|_{w \neq c} = 0 \quad (3)$$

$$L(\theta) = \prod_{t=1}^T \prod_{j=-m}^m p_{\theta}(w_{t+j} | w_t)$$

ל 4

אבל גם פה (כפי שראינו) לא ידוע לנו מהי הפונקציה p_{θ}

$$\arg \max_{\theta} L(\theta) = \arg \max_{\theta} \prod_{t=1}^T \prod_{j=-m}^m p_{\theta}(w_{t+j} | w_t)^{\#(w_{t+j}, w_t)}$$

אם ניקח את הלוג.

$$= \arg \max_{\theta} \sum_{t=1}^T \sum_{j=-m}^m \#(w_{t+j}, w_t) \cdot \log p_{\theta}(w_{t+j} | w_t)$$

אם ניקח את הלוג.

$$= \arg \max_{\theta} \sum_{j=-m}^m \#(w_{t+j}, w_t) \log p_{\theta}(w_{t+j} | w_t)$$

$$\leq p_{\theta}(w_{t+j} | w_t) \quad \sum_{j=-m}^m p_{\theta}(w_{t+j} | w_t) = 1$$

$$= \arg \max_{\theta, \lambda} \left[\sum_{j=-m}^m \#(w_{t+j}, w_t) \cdot \log p_{\theta}(w_{t+j} | w_t) \right] - \lambda \left(\sum_{j=-m}^m p_{\theta}(w_{t+j} | w_t) - 1 \right)$$

$$\frac{\partial}{\partial p_{\theta}(0|c)} \Rightarrow \frac{\#(0, c)}{p_{\theta}(0|c)} - \lambda = 0$$

$$\Rightarrow p_{\theta}(0|c) = \frac{\#(0, c)}{\lambda}$$

$$\sum_{i=-m}^m p_{\theta}(0_i | c) = 1 = \sum_{i=-m}^m \frac{\#(0_i, c)}{\lambda} \Rightarrow \lambda = \sum_{i=-m}^m \#(0_i, c)$$

אם כן

$$p_{\theta}(0|c) = \frac{\#(0, c)}{\sum_{i=-m}^m \#(0_i, c)}$$

{ house, bank }

"house house house"

"ban/c ban/c ban/c"

Person's name: _____

$$P(\text{house} | \text{house}) = 1$$

$$p(\text{house} | \text{bank}) = 0$$

$$p(\text{bunk} | \text{bank}) = 1$$

$$p(\text{bank} | \text{house}) = 0$$

$$p(o|c) = \frac{\exp(u_o^T \cdot v_c)}{\sum_w \exp(u_w^T \cdot v_c)}$$

c = bank o = house

$$p(\mathbf{b}|\mathbf{c}) = \frac{\exp(\mathbf{u}_0^T \mathbf{v}_c)}{\sum \exp(\mathbf{u}_w^T \mathbf{v}_c)} \geq 0$$

∴ $p(\text{house} | \text{bank}) \neq 0$

$$\text{relu}(\vec{x}) = \max(0, \vec{x}) \Rightarrow \geq 0$$

5

$$\text{relu}(\vec{x}_1)^T \cdot \text{relu}(\vec{x}_2) \geq 0$$

threshold = 0.5

אם נניח $\text{sigmoid}(0) = 0.5$ ונניח $\text{relu}(x) = \max(0, x)$

$$\forall \vec{x}_1, \vec{x}_2 \quad \text{sigmoid}(\text{relu}(\vec{x}_1) \cdot \text{relu}(\vec{x}_2)) \geq 0.5$$

אם x_1, x_2 הם מספרים, אז $\text{relu}(x_1) \cdot \text{relu}(x_2) \geq 0$

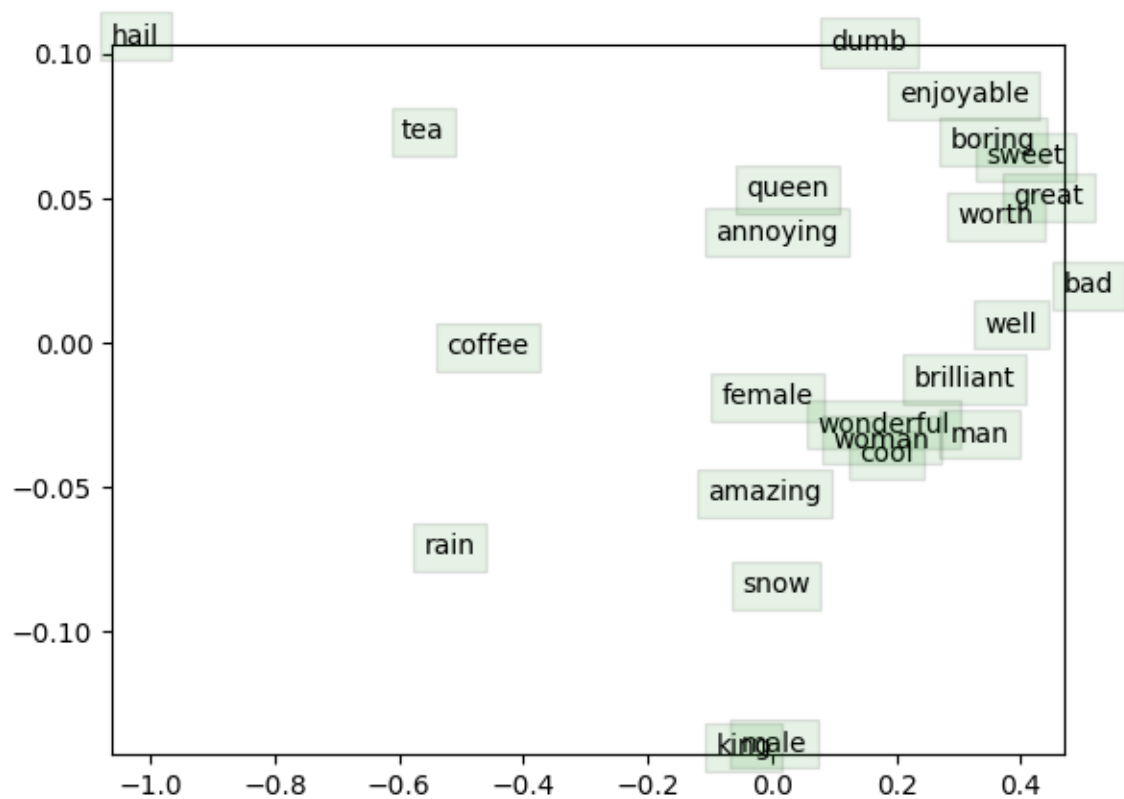
אם x_1, x_2 הם וקטורים, אז $\text{relu}(x_1) \cdot \text{relu}(x_2) \geq 0$ כי כל ערך ב- relu הוא לא שלילי.

$$\text{Acc}(M) \geq 0.25$$

$$p(\text{paraphrase} | x_1, x_2) = \text{sigmoid}(\vec{x}_1^T \cdot \vec{x}_2)$$

$$0 < \vec{x}_1^T \cdot \vec{x}_2 < 1$$

אם \vec{x}_1 ו- \vec{x}_2 הם וקטורים, אז $\vec{x}_1^T \cdot \vec{x}_2$ הוא סקלר. אם \vec{x}_1 ו- \vec{x}_2 הם וקטורים, אז $\vec{x}_1^T \cdot \vec{x}_2$ הוא סקלר.



There is somewhat of a cluster for all the adjectives. Also, 'king' and 'male' are clustered in a meaningful manner (though 'queen' and 'female' are not), etc. A good way to improve the results would be to increase the data so that more occurrences would appear.