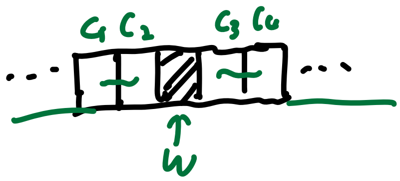
Lecture 05 & 06 Word Embeddings

✯ Skip-Gram model : →



- $C[w] = \{c_1, c_2, c_3, c_4\}$

Given a target word w, predict context c.

Text: "The quick brown fox jumps over the lazy dog."

→ $\text{Text} = [w_1, w_2, \dots, w_T]$ T tokens.

Goal: for each w, maximize the corpus probability:

$$\arg\max_{\theta} \prod_{w \in \text{Text}} \left[\prod_{c \in C[w]} p(c \mid w; \theta) \right] \quad (1)$$

Define $D := \{ (w, c) : w \in \text{Text}, c \in C[w] \}$.

$$\Rightarrow \quad \arg\max_{\theta} \prod_{(w,c) \in D} p(c \mid w; \theta). \quad (2)$$

How can we properly define $p(c \mid w; \theta)$?

Using softmax !

$$p(c|w;\theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}},$$

- v_c: vector representation of word c.
- v_w: vector representation of word w.
- C: set of all context words.
- V: set of all target words.

To maximize (2), take log first:

$$\arg\max_{\theta} \sum_{(w,c) \in D} \log p(c|w;\theta)$$

$$:= \sum_{(w,c) \in D} \left(\log e^{v_c \cdot v_w} - \log \sum_{c' \in C} e^{v_{c'} \cdot v_w} \right) \quad (3)$$

We hope:

By maximizing (3), we can get $v_1, v_2, \cdots v_{|V|}$, which are good embeddings in the sense that similar words will have similar vetors!

But, what is time complexity of maximizing (3)

$|D| = |V| \cdot \text{window} \cdot |C| \approx \text{window} \cdot |V|^2$: this only one epoch of using SGD.

Reducing time complexity:

1. hierarchical softmax
2. negative sampling ✓

* negative sampling:

Idea: find parameters θ such that the probabilities that all of observations indeed came from the data:

Given pair (w, c) from $V \times C$, what is the probability that (w, c) is indeed from $V \times C$?

First try:

If I can find θ such that $p(w, c; \theta)$ is really large given $(w, c) \in D$, then this θ maybe good ?

To make this concrete, let

$p(D=1|w,c;\theta)$: prob. that (w,c) is from D,

then we should optimize the following:

$$\arg\max_{\theta} \prod_{(w,c)\in D} p(D=1|w,c;\theta) \quad (4)$$

take log and using $p(D=1|w,c;\theta) = \frac{1}{1+e^{-v_c \cdot w_w}}$

$$\Rightarrow \arg\max_{\theta} \sum_{(w,c)\in D} \log \frac{1}{1+e^{-v_c \cdot v_w}}$$

By setting, $v_c = v_w$ and $v_c \cdot v_w = K$, $K \geq 40$.

then one can optimize (4)!

To prevent all vectors from having same value,

by disallowing some (w,c), to let the model

know :

$$\begin{cases} (w,c)\in D \rightarrow p(D=1|w,c;\theta) \text{ should be high!} \\ (w,c)\notin D \rightarrow p(D=1|w,c;\theta) \text{ should be low!} \end{cases}$$

We call $(w, c) \notin D$ a negative pair.

How to generate a negative pair?

We have at most window $\cdot |V|$ "positive" pairs.

What is the probability that randomly generate $(w, c) \in V \times C$ is a "positive" pair?

- $\dfrac{\text{window} \cdot |V|}{|V| \cdot |C|} \approx \dfrac{\text{window}}{|C|}$ quite small!

Say, D' = set of negative samples! then

the goal is

$$\arg\max_{\theta} \prod_{(w,c) \in D} p(D=1 \mid w, c; \theta) \cdot \underbrace{\prod_{(w,c) \in D'} p(D=0 \mid c, w; \theta)}_{1 - p(D=1 \cdots)}.$$

Take log and we have:

$$= \arg\max_{\theta} \sum_{(w,c) \in D} \log \frac{1}{1+e^{-v_c \cdot v_w}} + \sum_{(w,c) \in D'} \log \frac{1}{1+e^{v_c \cdot v_w}}$$

$$= \arg\max_{\theta} \sum_{(w,c) \in D} \log \sigma(v_c \cdot v_w) + \sum_{(w,c) \in D'} \log \sigma(-v_c \cdot v_w).$$

Remark:

① For each $(w,c) \in D$, we generate k negatives.

$$(w,c) \sim P_{words}(w) \cdot \frac{P_{contexts}(c)^{3/4}}{Z}$$

Z: constant.

$P_{words}(w)$: unigram distribution of words

$P_{contexts}(c)$: unigram distribution of contexts.

② $k \cdot |D| \approx |D'|$.

then, we can get the loss

$$-\sum_{t=1}^{T} \underbrace{\left[\log \sigma(c_{pos} \cdot w_t) + \sum_{j=1}^{k} \log \sigma(-c_{neg_j} \cdot w_t)\right]}$$

$J(c_{pos}, w_t, c_{neg_{1:k}}; \theta)$: $k+2$ vectors

for each.

$$SGD: \theta^{t+1} = \theta^{t} - \eta_t \cdot \nabla J(c_{pos}, w_t, c_{neg_{1:k}}; \theta^t).$$

☆ Training skig-gram of negatiie sampling

Setting :

- V : set of target words
- C : set of context words (Usually, $V = C$).
- $\theta = [W_{in}, W_{out}]$, where
- $W_{in} = W_{input}$: hidden layer. projection layer
- $W_{out} = W_{output}$: output layer.

(Recall our goal is given W_{in}, we predict context).

Objective :

$$\min_{\theta=[W_{in}, W_{out}]} -\sum_{t=1}^{T}\left[\log\sigma(W_t^T \cdot C_{pos}) + \sum_{i=1}^{k}\log\sigma(-W_t^T \cdot C_{neg_i})\right],$$

W_t : current target word (t).

C_{neg_i} : i-th negative word. $i = 1, 2, \cdots k$.

C_{pos} : the context word.

$$J(C_{pos}, C_{neg_{1:k}}, w) =$$

$$-\left[\log\sigma(W^T \cdot C_{pos}) + \sum_{i=1}^{k}\log\sigma(-W^T \cdot C_{neg_i})\right]$$

SGD:

$$\theta^{t+1} = \theta^t - \eta_t \cdot \nabla J(\theta^t).$$

given each pair $(w, c_{pos}, c_{neg_{1:k}})$, the gradient:

① $$\frac{\partial J}{\partial c_{pos}} = -\frac{1}{\sigma(w^T \cdot c_{pos})} \cdot \sigma'(w^T \cdot c_{pos}) \cdot w$$

Recall $\sigma'(x) = \sigma(x)(1-\sigma(x))$

$$= [\sigma(w^T \cdot c_{pos}) - 1] \cdot w.$$

② $\frac{\partial J}{\partial c_{neg_i}} = ?$ How.

Imaging $w^T \cdot c_{pos} \Leftrightarrow (-w^T) \cdot c_{neg_i}$,

so, replacing w by $-w$ on $\frac{\partial J}{\partial c_{pos}}$, you will get: $\frac{\partial J}{\partial c_{neg_i}}$, we have:

$$\frac{\partial J}{\partial c_{neg_i}} = [\sigma(-w^T c_{neg_i}) - 1](-w)$$

$$= [1 - \sigma(-w^T c_{neg_i})] \cdot w$$

$= \sigma(W^T \cdot C_{neg_i}) \cdot W, \quad i = 1, 2, \cdots k.$

Recall that: $\sigma(-x) = 1 - \sigma(x)$.

③ $\frac{\partial J}{\partial W}$, pretty easy,

$W^T \cdot C_{pos} = C_{pos}^T \cdot W$: two varibles have no difference, the same:

$W^T \cdot C_{neg_i} = C_{neg_i}^T \cdot W$. We have:

$$\frac{\partial J}{\partial W} = [\sigma(W^T \cdot C_{pos}) - 1] \cdot C_{pos} + \sum_{i=1}^{k} [\sigma(C_{neg_i} \cdot W)] \cdot C_{neg_i}.$$

Then, we have the updates of W, C_{pos}, $C_{neg_{1:k}}$ at time t:

SGD:

$$c_{pos}^{t+1} = c_{pos}^{t} - \eta_t \cdot [\sigma(c_{pos}^{t} \cdot w^t) - 1] \cdot w^t$$

$$c_{neg_i}^{t+1} = c_{neg}^{t} - \eta_t \cdot [\sigma(c_{neg_i}^{t} \cdot w^t)] \cdot w^t$$

$$w^{t+1} = w^t - \eta_t \cdot \Big[[\sigma(c_{pos}^{t} \cdot w^t) - 1] \cdot c_{pos} + \sum_{i=1}^{k} [\sigma(c_{neg_i} \cdot w^t)] \cdot c_{neg_i}$$