Alexis Tabin
16-821-803

Computational Semantics for Natural Language Processing

Assignment 1

1. Skip-Gram

$w \in V_W$, a word, V_W word vocabulary, $w \to \vec{w} \in \mathbb{R}^d$

$c \in V_C$, a context, V_C context vocabulary, $c \to \vec{c} \in \mathbb{R}^d$

D, collection of observed word and context pairs

$\#(w,c)$, numbers of time (w,c) appears in D

$\#(w) = \sum_{c' \in V_C} \#(w,c')$ $\quad \#c = \sum_{w' \in V_W} \#(w',c)$

1. $P(D=1 \mid w,c) = \sigma(\vec{w} \cdot \vec{c}) = \frac{1}{1+\exp(-\vec{w}\cdot\vec{c})}$

$$\ell = \sum_{w \in V_W} \sum_{c \in V_C} \#(w,c) \log(\sigma(\vec{w}\cdot\vec{c}))$$

$$= \sum_{w \in V_W} \sum_{c \in V_C} \frac{\#(w,c)}{1+\exp(-\vec{w}\cdot\vec{c})}$$

Answers 1.1

An objective functions tries to optimize something. Here, we want to maximize $P(D=1 \mid w,c)$ for observed (w,c) pairs, ~~while maximizing~~ ~~$P(D=0)$~~. We can see that we could use this log likelihood function as the objective function, it would work, but it is really heavy on computation. It is heavy because for each time we need to compute ℓ, we need to go through every word, and for each word, we go though every context.

2. c_N, randomly drawn negative samples

for (w,c) we want to

1. maximize $P(D=1|w,c)$
2. while maximizing $P(D=0|w,c)$ for negative samples | i

k = # of samples

$$c_N \sim \frac{\#(c)}{|D|} = P_D(c)$$

Answers 1.2

Recall that $P(D=1|\omega,c) = \sigma(\vec{\omega}\cdot\vec{c}) = \dfrac{1}{1+\exp(-\vec{\omega}\cdot\vec{c})}$

$$\Rightarrow \quad P(D=0|w,c) = 1-P(D=1|w,c) = 1-\sigma(\vec{\omega}\cdot\vec{c})$$

$$\mathcal{L}(w,c) = P(D=1|w,c)\prod_{i=1}^{k \leftarrow \text{\# of samples}} P(D=0|w,c) \quad \text{by } |\, i$$

We calculate the log

$$\log \mathcal{L}(\omega,c) = \log\Big(P(D=1|\omega,c)\prod_{i=1}^{k} P(D=0|\omega,c)\Big) = \log(\sigma(\vec{\omega}\cdot\vec{c})) + \log\Big(\prod_{i=1}^{k} (1-\sigma(\vec{\omega}\cdot\vec{c})\Big)$$

$$= \log(\sigma(\vec{\omega}\cdot\vec{c})) + \sum_{i=1}^{k} \log(1-\sigma(\vec{\omega}\cdot\vec{c}))$$

multiply 2nd term and divise by k.

$$= \log(\sigma(\vec{\omega}\cdot\vec{c})) + \cancel{\tfrac{k}{k}}\ \underbrace{\sum_{i=1}^{k} \log(1-\sigma(\vec{\omega}\cdot\vec{c}))}_{k\cdot\frac{1}{k}} = \mathbb{E}_{c_N\sim P_D}\big(\log(1-\sigma(\vec{\omega}\cdot\vec{c}))\big)$$

ii

$$1-\sigma(\vec{\omega}\cdot\vec{c}) = 1-\frac{1}{1+e^{-\vec{\omega}\cdot\vec{c}}} = \frac{1+e^{-\vec{\omega}\cdot\vec{c}}-1}{1+e^{\vec{\omega}\cdot\vec{c}}} =$$

$= \cancel{\log}$

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

~~Start~~ Following of 1.2

with $\underline{ii}$, we have

$$\log h(w,c) = \log(\sigma(\vec{w}\cdot\vec{c})) + k \cdot \mathbb{E}_{c_N \sim P_D}\left(\log(\sigma(-\vec{w}\cdot\vec{c}))\right)$$

So we have for each pair the objective function we want to minimize :

$$h(\vec{w},\vec{c}) = \log(\sigma(\vec{w}\cdot\vec{c})) + k\, \mathbb{E}_{c_N \sim P_D}\left(\log(\sigma(-\vec{w}\cdot\vec{c}))\right)$$

For all pairs, it gives

$$l = \sum_{w \in V_W} \sum_{c \in V_c} \#(w,c)\left(\log \sigma(\vec{w}\cdot\vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}\left[\log(-\vec{w}\cdot\vec{c})\right]\right)$$

$\boxed{3.}$ Assuming $\vec{w}\cdot\vec{c}$ independent, $l(w,c) = ?$, let's rewrite l

$$l = \sum_w \sum_c \#(w,c)\left(\log \sigma(\vec{w}\cdot\vec{c})\right) + \sum_w \sum_c \#(w,c)\left(k \cdot \mathbb{E}_{c_N \sim P_D}\left[\log \sigma(-\vec{w}\cdot\vec{c}_N)\right]\right)$$

$$= \sum_w \sum_c \#(w,c)\left(\log \sigma(\vec{w}\cdot\vec{c})\right) + \sum_{w \in V_W} \#(w)\left(k \cdot \mathbb{E}_{c_N \sim P_D}\left[\log \sigma(-\vec{w}\cdot\vec{c}_N)\right]\right)$$

recall the expectation term : $\mathbb{E}_{c_N \sim P_D}\left[\log(\sigma(-\vec{w}\cdot c_N))\right] = \sum_{c_N \in C} \frac{\#(c_N)}{|D|} \log \sigma(-\vec{w}\cdot c_N)$

$$= \frac{\#(c)}{|D|} \log \sigma(-\vec{w}\cdot\vec{c}) + \sum_{c_N \in V_c \setminus \{c\}} \frac{\#(c_N)}{|D|} \log \sigma(-\vec{w}\cdot c_N)$$

→ we then plug that back in the loss function :

$$l = \sum_w \sum_c \#(w,c) \log(\sigma(\vec{w}\cdot\vec{c})) + \sum_w \#w \cdot k \cdot (\ldots)$$ answer →

We can see how to express $\ell(w,c)$:

$$\ell(w,c) = \#(w,c)\log(\vec{w}\cdot\vec{c}) + k\cdot\#w\cdot\frac{\#c}{|D|}\log(-\vec{w}\cdot\vec{c})$$

$\boxed{1.4}$ $x = \vec{w}\cdot\vec{c}$, we want $x^* = \max\limits_{x} \ell$

$$\rightsquigarrow \cancel{\ell(x) = \#(w,c)\cdot\sigma(-x)}$$

$$\rightsquigarrow \ell(x) = \#(w,c)\cdot\log(\sigma(x)) + \#(w)\cdot k\cdot\frac{\#c}{|D|}\log(\sigma(-x))$$

$$x^* = \max_{x}(\ell) \rightsquigarrow \frac{\partial \ell(x)}{\partial x} = 0$$

$$\rightsquigarrow \#(w,c)\,\sigma'(x)\cdot\frac{1}{\sigma(x)} + \frac{\#(w)\cdot k\cdot\#(c)}{|D|}\cdot\sigma'(-x)\frac{1}{\sigma(-x)} = 0$$

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

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

$$= (1+e^{-x})^{-2}e^{-x}$$

$$= \frac{1}{e^{x}+e^{-2x}}$$

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

as we saw in 1.1

$$\rightsquigarrow \#(w,c)\,\cancel{\sigma(x)}\underbrace{(1-\sigma(x))}_{\sigma(-x)}\frac{1}{\cancel{\sigma(x)}}$$

$$+ \frac{k\#(w)\#(c)}{|D|}\sigma(x) = 0$$

$\vdots$ → with trivial adjustments

result →

$$\Rightarrow \boxed{e^{2x} - \left(\frac{\#(w,c)}{k\,\#(w)\,\frac{\#(c)}{|D|}} - 1\right)e^{x} - \frac{\#(w,c)}{k\,\#(w)\,\frac{\#c}{|D|}} = 0}$$

rewritten

following of 1.4

$$\rightarrow e^{2x} - \left(\frac{\#(w,c)}{k\,\#(w)\,\frac{\#(c)}{|D|}} - 1 \right) e^{x} - \frac{\#(w,c)}{k\,\#(w)\,\frac{\#c}{|D|}} = 0$$

let's define $y = e^{x}$

$$\Rightarrow y^2 - by - c = 0$$

↳ this equation has 2 solutions:

$$y_1 = -1 \Rightarrow e^{x} = -1 \Rightarrow \text{not possible}$$

$$y_2 = \frac{\#(w,c)}{k \cdot \#(w) \cdot \frac{\#c}{|D|}}$$

Remember, we have $y = e^{x} = e^{\vec{w}\cdot\vec{c}}$

$$\Rightarrow \vec{w}\cdot\vec{c} = \log(y)$$

$$x^* =$$

$$\Rightarrow \vec{w}\cdot\vec{c} = \log\left(\frac{\#(w,c)\cdot|D|}{\#(w)\cdot\#(c)} \frac{1}{k} \right) = \text{argmax}\ \ell$$

$\boxtimes$

5. $k=1$

$$x^* = \log\left(\frac{\#(w,c)\cdot|D|}{\#(w)\cdot\#(c)} \cdot \cancel{\frac{1}{k}}\right)$$

$$= \log\left(\frac{\#(w,c)}{|D|}\cdot\frac{|D|}{\#(w)}\cdot\frac{|D|}{\#(c)}\right)$$

$$= \log\left(\frac{P_D(w,c)}{P_D(c)\,P(w)}\right) \underset{\text{by definition}}{=} PMI(w,c)$$

6. $M = W^* \cdot C^{*T}$?

$$M^{SGNS}_{ij} = W_i \cdot C_j = \vec{w}_i \cdot \vec{c}_j = PMI(w_i, c_j) - \log k$$

For negative-sampling values $k > 1$, SGNS is factorizing a shifted PMI matrix $M^{PMI_k} = M^{PMI} - \log k$