Computational Semantics for Natural Language Processing

Alexis Tabin
16-821-803

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

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

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

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

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

Answers 1.1

An objective functions tries to optimize something. Here, we want to maximize $P(D=1 \mid w,c)$ for ~~on~~ 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 $P(D=1 \mid w,c)$
2. while maximizing $P(D=0 \mid w,c)$ for negative samples $\quad \mid i$

k = # of samples

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

Answers 1.2

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

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

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

We calculate the log

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

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

multiply 2nd term and divide by k

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

ii

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

$= \cancel{\log}$

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

~~Sta~~ Following of 1.2

with ii, we have

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

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

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

For all pairs, it gives

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

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

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

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

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

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

→ we then plug that back in the loss function :

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

answer →

3

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

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

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

$\rightsquigarrow$ ~~$\ell(x) = \#(w,c)\cdot\sigma(-x)$~~

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

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

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

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

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

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

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

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

as we saw in 1.1

$$\rightsquigarrow \#(w,c)\,\cancel{\sigma(x)}\,\underbrace{(1-\sigma(x))}_{\sigma(-x)}\frac{1}{\cancel{\sigma(x)}} + \frac{h\#(w)\#(c)}{|D|}\sigma(x) = 0$$

⋮ → with trivial adjustments

result →

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

rewritten

following of 1.4

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

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

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

↳ this equation has 2 solutions:

$y_1 = -1 \Rightarrow e^x = -1 \Rightarrow$ 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. $M = W^* \cdot C^{*T}$?

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

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

Computational Semantics for Natural Language Processing

Alexis Tabin
16 - 821 - 803

Assignment 1

2. Multi-prototype Word Embeddings

- (w, c), pair of a word and a context
- N_w, word senses (prototypes
- $W \in \mathbb{R}^{|V_w| \times N_w \times d}$
- $h_w \in \{1, \dots, N_w\}$, $h_w = h$ ~~means~~ if w means of h^{th} prototype
- $\pi_{wh} = P(h_w = h | w)$

We model
$$P(D=1 | w, c) = \sum_{h=1}^{N_w} P(h_w = h | w) P(D=1 | w, h_w = h, c)$$
$$= \sum_{h=1}^{N_w} \pi_{wh} P(D=1 | w, h_w = h, c)$$

likelihood log-likelihood

2.1. $P(D)$, $L(D) = ?$

$$P(D) = \prod_{(w,c) \in D} P_r[D=1 | w, c]$$
$$= \prod_{(w,c) \in D} \sum_{h=1}^{N_w} P(h_w = h | w) P(D=1 | w, h_w = h, c)$$
$$L(D) = \log(P(D)) = \log\left(\prod_{(w,c)} \sum_{h=1}^{N_w} \pi_{wh} P(D=1 | w, h_w = h, c)\right)$$
$$= \sum_{(w,c) \in D} \log\left(\sum_{h=1}^{N_w} \pi_{wh} P(D=1 | w, h_w = h, c)\right)$$

2.2 $\quad M_{whc} \in \{0,1\} := \mathbb{1}\{h_w = h\}$

$P(D, M)$, $\log(P(D,M)) = L(D,M)$

$\triangleright\ P(D,M) = P(M|D)\,P(D)$

$P(M|D) = P(M=0|D)\,P(M=0) + P(M=1|D)\,P(M=1)$

$$\prod_{(w,c)\in D}^{D} \sum_{h=1}^{N_w} \pi_{wh}\, P(D=1|w, h_w = h, c) \cdot P(M|D)$$

$$= \prod_{(w,c)\in D} \prod_{h=1}^{N_w} \pi_{wh}\, P(D=1|w, h_w=h, c)^{M_{whc}}$$

$L(D,M) =$

$$\triangleright\ \log(P(D,M)) = \log \prod_{(w,c)} \prod_{h=1}^{N_w} \pi_{wh}\, P(D=1|w, h_w=h, c)^{M_{whc}}$$

$$= \sum_{(w,c)} \sum_{h=1}^{N_w} M_{whc} \log(P(D=1|w, h_w=h, c)) + \log(\pi_{wh})$$

2.3 E-step: $Q(\Theta) := \mathbb{E}_{M|D}\, l(D,M)$

$$\mathbb{E}_{M|D} M_{whc} = \mu_{whc}$$

$$Q(\Theta) = \mathbb{E}_{M|D}\, l(D,M) = \mathbb{E}\left[\sum_{(w,c)\in D} \sum_{h=1}^{N_w} M_{whc} \log(P(D=1|\ldots)) + \log(\pi_{w\ldots}\right.$$

$$= \sum_{(w,c)\in D} \sum_{h=1}^{N_w} \mathbb{E}_{M|D} M_{whc} \log\left(P(D=1|w, h_w=h, c) + \log(\pi_{wh})\right)$$

$$= \sum_{(w,c)\in D} \sum_{h=1}^{N_w} \mu_{whc} \log\left(P(D=1|w, h_w=h, c) + \log(\pi_{wh})\right)$$

parameters Θ: π_{wh} for $h = 1$ to N_w

~~and μ_{whc}~~

~~…~~ 2.4.

parameters Θ are $\{\pi_{wh'}\}_{h'=1}^{N_w}$

2.4 write μ_{whc} as a $\{\pi_{wc'}\}_{k'=1}^{N_w}$

and $P(\underset{=1}{D} \mid w, h_w = k', c)$

D $\mu_{whc} = \mathbb{E}_{M|D} M_{whc} = 1 \cdot P(M = 1 \mid D)$ ~~$+ 0 \cdot P(M = 0 \mid D)$~~

$M = 1$ if $h_w = h$ (2.2)

$$= P(h_w = h \mid D, w) = \frac{P(D \mid w, h_w = h) \underbrace{P(h_w = h \mid w)}_{\pi_{wh}}}{P(D \mid w)}$$

$$= \frac{\pi_{wh} P(D \mid w, h_w = h)}{P(D \mid w)}$$

for $D = 1$, we have : $\mu_{whc} = \dfrac{\pi_{wh} P(D \mid w, h_w = h)}{\sum_{h=1}^{N_w} P(D = 1 \mid w, h_w = h, c)}$

2.5 | M-step

$\triangleright\ \pi^*_{whc} := \arg\max_{\pi_{whc}} Q(\Theta)$ parameter

$$Q(\Theta) \overset{!}{=}_{2.3} \sum_{(w,c)\in D} \sum_{h=1}^{N_w} \Big(\underbrace{\mu_{whc}\Big(\log\big(P(D=1 \mid w, h_w = h, c)\big)\Big)}_{\text{no } \pi_{whc}} + \log(\pi_{whc}) \cdot$$

$$\Rightarrow \arg\max_{\pi_{whc}} Q(\Theta) = \arg\max_{\pi_{whc}} \left(\sum_{(w,c)} \sum \mu_{whc} \log(\pi_{whc}) \right)$$

Lagrangian constraint: $\sum^{N_w} \pi_{whc} = 1$

$$\Rightarrow \mathcal{L}(\pi_{whc}; \lambda) = \sum_{(w,c)} \sum^{N_w} \mu_{whc} \log(\pi_{whc}) + \lambda\left(1 - \sum_{h=1}^{N_w} \pi_{whc}\right)$$

now we derive the lagrangian

$\frac{\partial L}{\partial \pi_{whc}} = 0 \iff$ and with a few calculations, we have

$\pi^*_{whc} = C \cdot \sum_{(w,c)} \mu_{whc}$ where c is just a scaling cst

Computational Semantics for Natural Language Processing

Alexis Tabin
16-821-803

Assignment 1

Hierarchical Softmax & Huffmann Coding

D, document $\quad$ V, vocabulary $\quad$ $p(w) = \frac{\# w}{|D|}$, word frequency

\# of times w appears in D

$|D|$, # of words $\quad$ $|V|$, voc size

$$P_w(c) = \frac{\exp\{(s(w,c))\}}{\sum_{c \in V} \exp\{(s(w,c))\}}, \quad O(|V|) \text{ for each pair } (w,c)$$

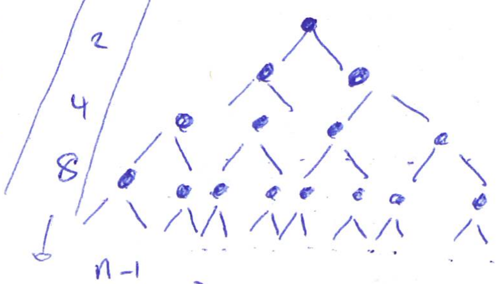
approximated by

$$P_w(c) = \prod_{j=1}^{L(c)-1} P_w((c,j) \to (c,j+1))$$

depth of node c in tree

node of depth j on road to leaf c

prob of transition



3.1 binary tree, $|V| = 2^{10} = 1024$

To compute $P_w(c)$, we need to go all the way down to the leaf, the way is composed of 9 edges here.

General case:

$$O(\log|V| - 1)$$
$$= O(\log|V|)$$

nbr nodes: $\sum_{i=0}^{n-1} 2^i$, here $n = 10$

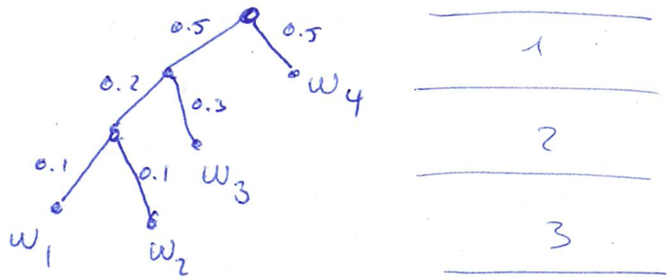
$\Rightarrow$ $\boxed{\# \text{ inner nodes} = 1023}$

Total # nodes = inner nodes + leaf nodes
= 1023 + 1024 = 2047

\# of edge = 2046

3.2 $|V| = 4$

Intuitively, we have :



$$\mathbb{E}[L(c)] = 3 \cdot 0.1 + 3 \cdot 0.1 + 2 \cdot 0.3 + 1 \cdot 0.5 = 1,7$$

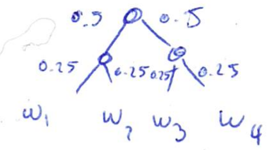
- ~~$\mathbb{E}[L(c)] \geq \mathbb{E}[-\log(p(w_i))]$~~

$$L(w_i) \geq -\log(p(w_i))$$

no clue

- worst case : let take the last example.

the worst case is this one :



: $\mathbb{E}[L(c)] = 4 \cdot (0.25 \cdot 2)$
$= 2 = \log_2 4$

general case : $\mathbb{E}[L(c)] = O(\log|V|)$

Computational Semantics for Natural Language Processing

Alexis Tabin
16 - 821 - 803

Assignment 1

4. FastText Embeddings

4.1 a) 40'000

b) 2e6

4.2 a) i)

doesn't depend on u_c

$$\frac{\partial \mathcal{L}}{\partial u_c} = \frac{\partial}{\partial u_c}\left(-\log(\sigma(s(w_t, w_c))) - \cdots\right)$$

$$= \frac{\partial}{\partial u_c} - \log\left(\sigma\left(\sum_g z_g^T u_c\right)\right)$$

$$= -\frac{1}{\sigma\left(\sum_g z_g^T u_c\right)} \frac{\partial}{\partial u_c} \sigma\left(\sum_g z_g^T u_c\right)$$

$$= -\frac{1}{\cancel{\sigma\left(\sum_g z_g^T u_c\right)}} \left(1 - \sigma\left(\sum_g z_g^T u_c\right)\right) \cancel{\sigma\left(\sum_g z_g^T u_c\right)} \sum_g z_g$$

$$= \left(\sigma\left(\sum_g z_g^T u_c\right) - 1\right) \sum_g z_g$$

a) ii) $$\frac{\partial L}{\partial u_n} = \frac{\partial}{\partial u_n} \underbrace{-\log(\dots)}_{=0} - \sum_{u \in N_a} \log(\sigma(-s(w_+, w_n)))$$

$$= \frac{\partial L}{\partial u_n} - \sum_{n \in N_a} \log(\sigma(-s(w_+, w_n)))$$

Same as i) but with a negative sign

$$\Rightarrow \frac{\partial L}{\partial u_n} = (1 - \sigma(-s(w_+, w_n)) \sum_g z_g$$

b) $$\frac{\partial L}{\partial z_g} = \cancel{\frac{\partial L}{\partial u_c} \frac{\partial L}{\partial u_n} = (1 - \sigma(-s(w_+, w_n)) \sum_g z_g + \sum_{u \in N_a} (1 - \sigma(}$$

u_c

sorry for the mess

$$= \sum_{u \in N_a} (1 - \sigma(-s(w_+, w_u)) u_n$$

$$+ (\sigma(s(w_+, w_c) - 1) u_c$$