## Section 2 (skip-gram with negative sampling (SGNS) setup)

skip-gram model : stream of words $w_i \in [N]$, contexts $c_i \in [N]^L := [M]$

observed in $n$ pairs $(w_i, c_i)$ (so $c_i$ is fixed window around $w_i$)

and looking for word representations $W_1, \dots, W_N$,

context representations $C_1, \dots, C_M$

to maximise

$$\sum_{i=1}^{n} f(w_i, c_i ; W_1, \dots, W_N, C_1, \dots, C_M)$$

where

$$f(k, l ; \text{—}) = \log\{\sigma(W_k^T C_l)\} + t \cdot E\{\log \sigma(-W_k^T C_L)\}$$

where $L \sim \text{categorical}\{\frac{n_1^{(c)}}{n}, \dots, \frac{n_M^{(c)}}{n}\}$

and $\sigma(x) = \frac{1}{1+e^{-x}}$ is the standard logistic function $\boxed{\text{logit}}$

which, by the way, satisfies $\sigma(-x) = 1 - \sigma(x)$, and $\sigma^{-1}(p) = \log(\frac{p}{1-p})$

Binary logistic regression

$$\mathbb{P}(Y=1 | X) = \sigma(\theta^T X) \; ; \; \mathbb{P}(Y=0 | X) = 1 - \sigma(\theta^T X) = \sigma(-\theta^T X)$$

Interpretation :

Fix $C_1, \dots, C_M$, and replace $t \cdot E\{\log \sigma(-W_k^T C_L)\}$

with $\log\{\sigma(-W_e^T C_{L_1})\} + \dots + \log(\sigma(-W_e^T C_{L_t}))$

$L_i$ i.i.d. replicates of $L$

Then optimising

$$\log\{\sigma(W_k^T C_l)\} + \log\{\sigma(-W_k^T C_{L_1})\} + \dots + \log(\sigma(-W_k^T C_{L_t}))$$

for $W_k$ is fitting a logistic classifier to recognise $C_l$ as valid (class 1)

and $C_{L_i}$ as invalid (class 0)

## Section 3 (Implicit matrix factorization)

Remember we are looking to maximize

$$\sum_{i=1}^{n} f(w_i, c_i ; W_1, \ldots, W_N, C_1, \ldots, C_M)$$

where

$$f(k, l ; \text{—}) = \log\{\sigma(W_k^T C_l)\} + t \cdot E\{\log \sigma(-W_k^T C_L)\}$$

where $L \sim \text{categorical}\left\{\frac{n_1^{(c)}}{n}, \ldots, \frac{n_M^{(c)}}{n}\right\}$

Let $M_{kl} = W_k^T \cdot C_l$. Then can write objective as optimisation

of $\sum_{kl}$ (terms involving $kl$)

where each term in the sum is

$$n_{kl}^{(wc)} \log\{\sigma(M_{kl})\} + t \cdot n_k^{(w)} \cdot \frac{n_l^{(c)}}{n} \cdot \log\{\sigma(-M_{kl})\}$$

Now look to maximise for each $p = \sigma(M_{kl})$:

$$n_{kl}^{(wc)} \log(p) + t \cdot n_k^{(w)} \cdot \frac{n_l^{(c)}}{n} \log(1-p)$$

$$\hat{p} = \frac{n_{kl}^{(wc)}}{n_{kl}^{(wc)} + \frac{t\, n_k^{(w)} n_l^{(c)}}{n}}$$

$$\hat{M}_{kl} = \text{logit}(\hat{p}) = \log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \log\left\{ n \frac{n_{kl}^{(wc)}}{t\, n_k^{(w)} n_l^{(c)}} \right\}$$

$$= \log\left\{\frac{n_{kl}^{(wc)}/n}{(n_k^{(w)}/n)\,(n_l^{(c)}/n)}\right\} - \log(t).$$

For discrete variables $X, Y$, the pointwise mutual information matrix is

$$\log \left\{ \frac{P(x,y)}{P(x)\,P(y)} \right\},$$ and so $\hat{M}$ can be viewed as a shifted empirical PMI.

Obviously can find decomposition $\hat{M}_{kl} = \hat{W}_k^T \cdot \hat{C}_l$

if $\hat{W}_k, \hat{C}_l$ are high-dimensional enough (e.g. via SVD)

So the claim is SGNS is factorizing $\hat{M}$, but I don't know how to make this statement formal if you admit $\hat{W}_k, \hat{C}_l$ aren't arbitrarily high-dimensional.

Just because:

$\hat{M} = \hat{W}\hat{C}^T$, where $\hat{W}, \hat{C} = \arg\max$(objective function) (unconstrained)

doesn't mean:

$\hat{M} \approx \hat{W}\hat{C}^T$, where $\hat{W}, \hat{C} = \arg\max$(obj) (constrained)

A weighted matrix factorization?

SVD of $\hat{M}$ gives $\bar{M} = \bar{W}\bar{C}^T$ to minimize $|\hat{M} - M|_F^2 = \sum |\hat{M}_{ij} - M_{ij}|^2$

However, we see that objective

$$\overbrace{n_{kl}^{(wc)}}^{a} \log\{\sigma(M_{kl})\} + \overbrace{t \cdot n_k^{(w)} \cdot \frac{n_l^{(c)}}{n}}^{b} \cdot \log\{\sigma(-M_{kl})\}$$

is not a fixed function of $\hat{M}_{kl} = \log\left(\frac{a}{b}\right)$

and in fact for $\log\left(\frac{a}{b}\right)$ fixed, i.e. $\frac{n_{kl}}{n_k\, n_l}$ fixed,

contribution is $x$ times higher if word $k$ or context $l$ appears $x$ more times.

Performance metrics

- rank correlation between word pairs (human ascribed similarity) (Spearman correlation)
  and corresponding cosine similarity $s(u,v) = \frac{u^T v}{|u|\,|v|}$
- analogy "$\overset{a}{\text{Paris}}$ is to $\overset{a^*}{\text{France}}$ as $\overset{b}{\text{Tokyo}}$ is to $\overset{b^*}{\underline{\text{Japan}}}$"

  hidden

  found using argmax $s(b^*, a^*)\, s(b^*, b) / \left[ s(b^*, a) + \varepsilon \right]$

  e.g. $s(\text{Japan}, \text{France})\; s(\text{Japan}, \text{Tokyo}) \,/\, s(\text{Japan}, \text{Paris})$