HW7

Problem 1. Bayes

Posterior $p|Y \sim \text{Dirichlet}(Y+\alpha)$

$$f(p_1 = v_1, \cdots, p_n = v_n \mid Y = y) = \frac{1}{B(y+\alpha)} \prod_{i=1}^{n} v_i^{y_{(i)}+\alpha_i - 1}$$

Proof:

In this model, there are 3 distributions:

$$\begin{cases} \text{Prior } p \sim \text{Dirichlet}(\alpha) & \text{where } p \in R^n, \ \alpha \in R^n \\ \text{Observed } Y|p,M \sim \text{Multi}(p, M) & \text{where } Y \in N^n, \ M \in R \quad Y = \sum_{i=1}^{M} w_i \\ \text{Observed } w_i | p \sim \text{Discrete}(p), \ i \in [M] & \text{where } w_i \in N^n \text{ is 1-sparse} \end{cases}$$

n is the number of words in the vocabulary (a bag of words)

M is the number of words in a document

$p_j \in (0,1)$ is the probability that word i occurs in a document $i \in [M], j \in [n]$

$$P((w_i)_{(j)} = 1) = p_j$$

α is the parameter of Prior Dirchlet Distribution

Y is a random vector with multinomial distribution:

$$\mathbb{P}(Y = y \mid p_1 = v_1, \cdots, p_n = v_n) = M! \prod_{i=1}^{n} \frac{p_i^{y_{(i)}}}{y_{(i)}!}$$

where $y_{(j)}$ is the number of times word j occurs in a document $j \in [n]$

w_i is a random vector generated i.i.d from Discrete(p)

$$\mathbb{P}(w_1, w_2, \cdots, w_M \mid p) = \prod_{i=1}^{n} p_i^{y_{(i)}}$$

By Bayes Rule, posterior is

$$\mathbb{P}(p|Y) = \frac{\mathbb{P}(Y|p)\,\mathbb{P}(p)}{\mathbb{P}(Y)}$$

$$\propto \mathbb{P}(Y|p)\,\mathbb{P}(p)$$

$$\propto \mathbb{P}(w_1, w_2, \cdots, w_m|p)\ \mathbb{P}(p)$$

$$= \left(\prod_{i=1}^{n} p_i^{\,y(i)}\right)\left(\frac{1}{B(\alpha)}\prod_{i=1}^{n} p_i^{\,\alpha_i - 1}\right)$$

$$\propto \prod_{i=1}^{n} p_i^{\,y(i)+\alpha_i-1}$$

Thus, $\mathbb{P}(p|Y) \sim \text{Dirichlet}\ (Y+\alpha)$

$$f(p_1 = v_1, \cdots, p_n = v_n | Y = y) = \frac{1}{B(y+\alpha)}\prod_{i=1}^{n} v_i^{\,y(i)+\alpha_i-1}$$

Problem 2 : word2vec as PCA

1) plug in $x = v_w^T v_c$ to $l(w,c)$

$$l(w,c) = \#(w,c) \log(\sigma(v_w^T v_c)) + k\,\#(w)\frac{\#(c)}{|D|} \log(\sigma(-v_w^T v_c))$$

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

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

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

$$\frac{\partial l}{\partial x} = \#(w,c) \frac{\sigma'(x)}{\sigma(x)} + k\,\#(w)\frac{\#(c)}{|D|} \frac{\sigma'(-x)}{\sigma(-x)}$$

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

$$= \#(w,c) \frac{e^{-x}}{1+e^{-x}} - k\,\#(w)\frac{\#(c)}{|D|} \frac{e^{x}}{1+e^{x}}$$

$$= \#(w,c) \frac{1}{e^{x}+1} - k\,\#(w)\frac{\#(c)}{|D|} \frac{e^{x}}{e^{x}+1}$$

set $\frac{\partial l}{\partial x} = 0$

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

let $y = e^{x}$

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

$$\begin{cases} y_1 = -1 & \text{(invalid)} \\ y_2 = \dfrac{\#(w,c)}{k\cdot\#(w)\cdot\frac{\#(c)}{|D|}} = \dfrac{\#(w,c)\cdot|D|}{\#(w)\,\#(c)}\cdot\dfrac{1}{k} & \text{(valid)} \end{cases}$$

$$x = \log y_2 = \log\left(\frac{\#(w,c)\cdot|D|}{\#(w)\,\#(c)}\cdot\frac{1}{k}\right) = \log\frac{\#(w,c)\,|D|}{\#w\cdot\#(c)} - \log k$$

2) Implicit matrix factorization of skip-gram

Since the association metric is defined as $x = PMI(w,c) - \log k$

Where PMI is the well-known pointwise mutual information matrix

$$PMI(w,c) = \log\left(\frac{\#(w,c)\,|D|}{\#(w)\,\#(c)}\right) \in R^{|V|\times|V|}$$

The skip-gram embeddings obtained by optimizing the local objective are equivalent to factorizing matrix $M \in R^{|V|\times|V|}$

$$M = W \cdot C^T$$

where $W \in R^{|V|\times d}$ is the word embedding matrix
$C \in R^{|V|\times d}$ is the context embedding matrix

M is shifted positive PMI matrix $\boxed{M = SPPMI_k(w,c) = \max(PMI(w,c) - \log k, 0)}$

Motivation for rank-d SVD of $M \in R^{|V|\times|V|}$

Working directly with matrix PMI has 2 computational challenges

① The matrix is ill-defined

Because rows of matrix PMI contain many entries of word-context pairs (w,c) that were never observed in the corpus

$$PMI(w,c) = \log 0 \rightarrow -\infty$$

② The matrix is dense

Because the high dimensions of the matrix $(|V| \times |V|)$

it's a major practical issue.

But there are still advantages to working with dense low-dimensional vectors, such as improved computational efficiency and better generalization.

Thus we use truncated SVD to achieve the optimal rank d factorization with respect to L_2 loss

$$M_d = \underset{M' \mid rank(M') = d}{\arg\min} \| M' - M \|_F^2 \Rightarrow \boxed{M_d = U_d \Sigma_d V_d^T}$$

SVD factorizes matrix $M \in R^{|V|\times|V|}$ into : $M = U \Sigma V^T$

Where $U \in R^{|V|\times|V|}$ is an orthonormal matrix, with columns of left singular vectors

$\Sigma \in R^{|V|\times|V|}$ is a diagonal matrix with diagonal entries of singular values

$V \in R^{|V|\times|V|}$ is an orthonormal matrix, with columns of right singular vectors

The matrix $M_d = U_d \Sigma_d V_d^T$ is the rank d matrix that best approximate the original matrix M by minimizing the reconstruction error

where $U_d \in R^{|M| \times d}$ the columns of U_d are the top d left singular vectors of matrix M.

$\Sigma_d \in R^{d \times d}$ is the diagonal matrix formed from the top d singular values.

$V_d \in R^{|V| \times d}$ the columns of V_d are the top d right singular vectors of matrix M.