Matrix factorization

distributional hypothesis?

w : which words does w appear together with within some distance?

- $\phi(w)$: a feature vector of w

  $\phi(w)_i$: the count of the word i appearing in the context of w.

  the co-occurrence vector of w.

- A co-occurrence matrix A

$$A = \begin{bmatrix} \phi^T(w_0) \\ \phi^T(w_1) \\ \vdots \\ \phi^T(w_V) \end{bmatrix} \in \mathbb{R}^{|V| \times |V|}$$

- compute/subtract the mean vector.

$$\mu = \frac{1}{|V|} \sum_{i=1}^{|V|} \phi(w_i)$$

$$\tilde{A} = \begin{bmatrix} \phi^T(w_0) - \mu^T \\ \phi^T(w_2) - \mu^T \\ \vdots \\ \phi^T(w_{|V|}) - \mu^T \end{bmatrix} \text{ : centered co-occurrence matrix}$$

- $\underline{U} \in \mathbb{R}^{|V| \times d}$, $\underline{d} \ll |V|$

  $\underline{V} \in \mathbb{R}^{|V| \times d}$

- PCA: $\min_{U,V} \underbrace{\|\tilde{A} - UV^T\|_F^2}$

$$= \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} \left(\tilde{\phi}(w_i)_j - U_i^T V_j\right)^2$$

word embedding

context embedding

extension to n-grams

$$\cancel{A = |V| \left\{ \overbrace{\begin{bmatrix} \quad & \quad & \quad \\ & & \\ & & \end{bmatrix}}^{|V|} \right.}$$

$$A = |V| \left\{ \overbrace{\begin{bmatrix} \phi & \checkmark & \checkmark & \checkmark & \checkmark \\ & & & & \\ & & & & \end{bmatrix}}^{|V|^n} \right.$$

$$\mu \; \begin{bmatrix} \quad & \quad & \quad & \quad & \quad \end{bmatrix}$$

$$\tilde{A} = \tilde{A} - \mu \mathbb{1}$$

$$\|\tilde{A} - UV^T\|_F^2 \qquad U \in \mathbb{R}^{|V| \times d}, \quad V \in \mathbb{R}^{|V|^n \times d} \Leftarrow \text{impractical to estimate!}$$

CBoW

$$V_{i\bar{j}} = V'_{ij_1} + V'_{ij_2} + \dots + V'_{ij_n}, \quad \text{where } \bar{j} = (j_1, j_2, \dots, j_n)$$

$$V' \in \mathbb{R}^{|V| \times d}$$

$$\min_{U, V'} \underbrace{\|\tilde{A} - UV^T\|_F^2}$$

$$\underbrace{\sum_{i=1}^{|V|} \sum_{j=1}^{|V|^n} \left(\tilde{\phi}(w_i)_{\bar{j}} - u_i^T \left(\sum_{n'=1}^{n} v'_{j_{n'}} \right) \right)^2}$$

beyond BoW

$$\sum_{i=1}^{|V|} \sum_{j=1}^{|V|^n} \left(\tilde{\phi}(w_i)_{\bar{j}} - u_i^T \overset{\text{n-gram}}{F_\theta}(\bar{j}) \right)^2 \qquad \tilde{\phi}(w_i)_{\bar{j}} = \phi(w_i)_{\bar{j}} - \mu_{\bar{j}}$$

$$= \sum^{|V|} \sum^{|V|^n} \left(\phi(w_i)_{\bar{j}} - \left(u_i^T F_\theta(\bar{j}) + \mu_{\bar{j}} \right) \right)^2$$

$$\sum_{i=1} \sum_{j=1} \Big(\;\; \Big)$$

A generative story

(A)

$v_D \in \mathbb{R}^d$: document embedding

$u_w \in \mathbb{R}^d$: word embedding of w

for the ith location in the document, which word should i write?

$$p(w^i = w \mid D) = \frac{\exp\{v_D^T u_w\}}{\sum_{w' \in V} \exp\{v_D^T u_{w'}\}}$$

vocabulary

$$\boxed{\frac{1}{N_D} \sum_{n=1}^{N_D} \sum_{t=1}^{T_n} \log p(w_t \mid D)}$$

(B)

$v_D \in \Delta^{d-1} \sim \text{Dirichlet}(\beta)$: mixture coefficients.

$u_w \in \mathbb{R}^k$: word embedding $|V|$

$u_d \in \mathbb{R}^k$: topic embedding d topics.

$$p(w_1 \cdots w_T, v_D, u_{d'}) = \prod_{t=1}^{T} \sum_{d'=1}^{d} v_D^{d'} p(w_t \mid u_{d'})$$

$$p(w_t \mid u_{d'}) = \frac{\exp\{u_{w_t}^T u_{d'}\}}{\sum_{w' \in V} \exp\{u_{w'}^T u_{d'}\}}$$

(C) autoregressive model

- Probabilistic modeling

- observations X
- hidden Z

· parameters Θ

$$\left\{ \begin{array}{l} \cdot\ p(X \mid z, \theta) \\ \cdot\ p(z) \\ \cdot\ p(\theta) \\ \cdots \end{array} \right\} \Longrightarrow$$

learning

$$\overset{\theta}{\cancel{p}}(\theta \mid D_{train} = \{x_1, \cdots, x_N\})$$

$$\overset{\theta}{\cancel{p}}(\{z_1 \cdots z_N\} \mid D_{train})$$

$$\int p(X_{test} \mid \theta, z)\, \overset{\theta}{\cancel{p}}(\theta, z \mid D_{train})\, d\theta, z = \underset{\Uparrow}{p(X_{test} \mid D_{train})}$$

predictive distribution