Ex 7-1

(a) "I like cats, cats like to catch mice".

symbols $D = \{$ cats, dogs, mice, like, hate, to, catch, . $\}$.

local representation:

$$\text{cats} = e_1 = (1\,0\,0\,0\,0\,0\,0\,0)^T$$
$$\text{dogs} = e_2 = (0\,1\,0\,0\,0\,0\,0\,0)^T$$
$$\text{mice} = e_3 = (0\,0\,1\,0\,0\,0\,0\,0)^T$$
$$\text{like} = e_4 = (0\,0\,0\,1\,0\,0\,0\,0)^T$$
$$\text{I} = e_5 = (0\,0\,0\,0\,1\,0\,0\,0)^T$$
$$\text{to} = e_6 = (0\,0\,0\,0\,0\,1\,0\,0)^T$$
$$\text{catch} = e_7 = (0\,0\,0\,0\,0\,0\,1\,0)^T$$
$$. = e_8 = (0\,0\,0\,0\,0\,0\,0\,1)^T.$$

⇒ disadvatages:
- inefficient for large symbol sets (one hot, sparse)

bag of symbols:

$$S = \begin{pmatrix} 0&0&1&0&1&0&0&0 \\ 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&1 \\ 0&1&0&0&0&1&0&0&0 \\ 1&0&0&0&0&0&0&0 \\ 0&0&0&0&0&1&0&0 \\ 0&0&0&0&0&0&1&0 \\ 0&0&0&1&0&0&0&0 \end{pmatrix} \Rightarrow S = \begin{pmatrix} 2 \\ 0 \\ 1 \\ 2 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \Rightarrow$$

disadvatages:
- does not preserve sequence orders.

(b) one hot encoded vectors are orthogonal, witch makes cosine similarity equals to zero: $\cos(x,y) = \frac{x^T y}{\|x\|\,\|y\|}$.

distributed representations encode several features as dense real-valued vectors, each of these features captures some of the syntactic and semantic information in each word (= meaningful similarities

Ex 7-2

loss: mean squared error

(a) encoder: sigmoid

decoder: linear

$$L(x, \hat{x}) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - (U\sigma(Wx_i + b_e) + b_d))^2$$

where $\hat{x} = D(E(x))$ $\quad x_i, \hat{x}_i \in \mathbb{R}^{d\times 1}$ $\quad W \in \mathbb{R}^{m\times d}$

$b_e \in \mathbb{R}^{m\times 1}$ $\quad U \in \mathbb{R}^{d\times m}$ $\quad b_d \in \mathbb{R}^{d\times 1}$.

(b) bias = 0 $\Rightarrow$ $L(x, \hat{x}) = \frac{1}{N}\sum_{i=1}^{N}(x_i - U\sigma(Wx_i))^2$

$$= \frac{1}{N}\sum_{i=1}^{N}(x_i - UWx_i)^2$$

$$= \frac{1}{N}\|X - UWX\|^2$$

(c)

(d) non linearity, convlayer, mini batches.

Ex 7-3.

(a) energy function probability dist: $P(v,h) = \frac{1}{Z}\exp\{-E(v,h)\}$

minimize of $E(v,h)$ equivalent to maximize: $-E(v,h)$.

$-E(v,h) = +b^T v + c^T h + v^T W h$.

preference of v; preference of h; preference of connecton of v and h.

preference

b, c positive $\Rightarrow$ v, h = 1

negative $\Rightarrow$ v, h = 0

W positive $\Rightarrow$ v, h = 1

negative $\Rightarrow$ v or h = 0

Ex 7-2

(b)

$$p(h|v) = \frac{p(h,v)}{p(v)} = \frac{\frac{1}{Z}\exp\{-E(v,h)\}}{\sum_h \left(\frac{1}{Z}\exp\{-E(v,h)\}\right)}$$

$$= \frac{\exp\{b^T v + c^T h + v^T W h\}}{\sum_h \exp\{b^T v + c^T h + v^T W h\}}$$

$$= \frac{\exp\{b^T v\}\cdot\exp\{c^T h + v^T W h\}}{\exp\{b^T v\}\cdot\sum_h \exp\{c^T h + v^T W h\}} \quad (= Z') = p(v)\cdot\exp\{$$

$$= \frac{\exp\{c^T h + v^T W h\}}{Z'} = \frac{1}{Z'}\exp\left\{\sum_j (c_j h_j + v^T W_j h_j)\right\}$$

$$= \frac{1}{Z'}\prod_j \exp\{c_j h_j + v^T W_j h_j\}$$

$$= \prod_j \sqrt[n]{\frac{1}{Z'}}\cdot\exp\{c_j h_j + v^T W_j h_j\}$$

Therefore $p(h_j|v)$ are independent.

$$p(h_j=0|v) = \frac{p(h_j=0,v)}{p(h_j=0,v)+p(h_j=1,v)} = \frac{\sqrt[n]{\frac{1}{Z'}}\cdot\exp\{0\}}{\sqrt[n]{\frac{1}{Z'}}\left(\exp\{0\}+\exp\{c_j + v^T W_j\}\right.}$$

$$= \frac{1}{1+\exp\{c_j + v^T W_j\}} = \sigma(c_j + v^T W_j)$$

$$p(h_j=1|v) = \frac{\exp\{c_j + v^T W_j\}}{1+\exp\{c_j + v^T W_j\}} = \sigma(c_j + v^T W_j)$$

$$p(h|v) = \prod p(h_j=1|v) = \prod_j \sigma(c_j + v^T W_j)$$

(c) $W = \begin{pmatrix} -0.384 & -2.295 & -0.872 & -1.219 \\ -1.258 & -2.219 & -0.154 & -1.581 \end{pmatrix}$

$b = \begin{pmatrix} 1.350 \\ 2.266 \\ 0.299 \\ 1.545 \end{pmatrix}$ $\quad c = \begin{pmatrix} 3.077 \\ 3.205 \end{pmatrix}$

$v = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix}$

$p(h_1|v) = \text{sigmoid}(c_1 + v^T W_{:1}) = 0.306.$

$p(h_2|v) = 0.136.$

(d) $p(v_i|h)$ $\quad p(v_1|h) = 0.794$

$p(v_2|h) = 0.906$

$p(v_3|h) = 0.574$

$p(v_4|h) = 0.824.$