ॐ MachineLearning

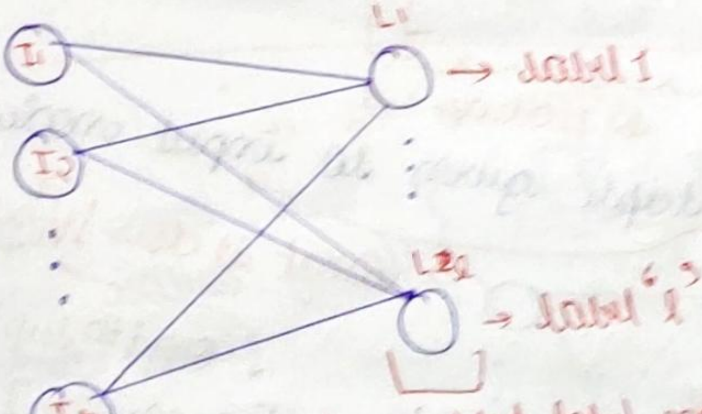
A	B	C	D
a_1	b_1	c_1	d_1
a_2	b_2	c_2	d_2
a_3	b_3	c_3	d_3
a_4	b_4	c_4	d_4
a_5	b_5	c_5	d_5
a_6	b_6	c_6	d_6

Lables
l_1
l_2
l_3
l_1
l_2
l_2

n tuples

← d features →

let us say there are 'l' lables
'd' features
'n' tuples



weighted label nodes each of these has weights for each feature

for each tuple we will have 'd' features so input node have it

ॐ Machine Learning

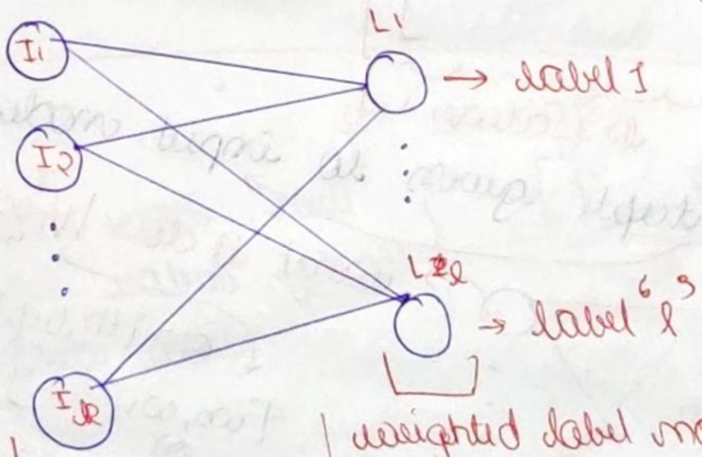
A	B	C	D
a_1	b_1	c_1	d_1
a_2	b_2	c_2	d_2
a_3	b_3	c_3	d_3
a_4	b_4	c_4	d_4
a_5	b_5	c_5	d_5
a_6	b_6	c_6	d_6

labels
l_1
l_2
l_3
l_1
l_2
l_2

n tuples

k features

let us say there are 'l' labels
'k' features
'n' tuples



weighted label nodes each of these has weights for each feature

for each tuple we will have 'k' features so input nodes have it

→ each L_1 label node has weights for each feature which are randomly initialized so

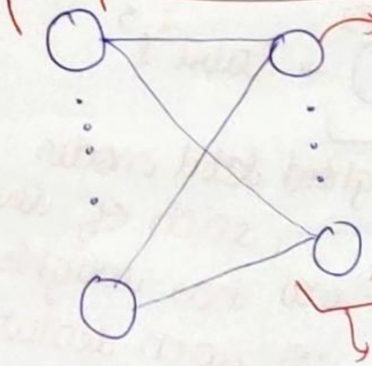
L_1 → outputs $W^T = [\omega_1, \omega_2 \dots \omega_k]$

* we have constant or '0' weight

$$W^T = [\underbrace{\omega_0, \omega_1, \dots \omega_k}_{k \text{ features} + 1 \text{ constant}}]$$

* $$X = \underbrace{\begin{bmatrix} 1 & a_1 & b_1 & c_1 \\ 1 & a_2 & b_2 & c_2 \\ 1 & \vdots & \vdots & \vdots \\ \vdots & & & \\ \vdots & & & \end{bmatrix}}_{k \text{ feature} + 1} \Bigg\} \text{ n tuples (leave out label)}$$

* each tuple given to input node



what it does $W_1^T x_i$

on ith tuple

$[\omega_0, \omega_1, \dots]$

or

$[\omega_{1,0}, \omega_{1,1}, \omega_{1,2} \dots]$

it outputs value for each tuple for every so each row will have value for each label

$$\begin{bmatrix} \omega_{l,0} \\ \omega_{l,1} \\ \vdots \\ \omega_{l,k} \end{bmatrix}_{(k+1)\times 1} \begin{bmatrix} x_0 & x_1 & \dots & x_k \end{bmatrix}_{1\times(k+1)}$$

$$\boxed{\begin{bmatrix} \omega_{l,0} & \omega_{l,1}, & \dots \omega_{l,k} \end{bmatrix}_{(1*(k+1))} * \begin{bmatrix} x_0 \\ \vdots \\ x_k \end{bmatrix}_{(k+1)*1}}$$

this is for l, label

* each node will have values for all ~~xxxx~~ labels

$$\underbrace{[P_{i1}, P_{i2}, \dots P_{il}]}$$

values that label nodes return for ith row of the dataset

* Now what is the use of
$[\hat{p}_{i1}, \hat{p}_{i2}, \ldots \hat{p}_{il}]$?

which ever label has the higher probability will have more chance to belong to it

→ here comes the role of softmax

why softmax?

→ softmax function is a function that turns a vector of k real values into a k real vector which sums to 1

→ Input value might be positive, negative or zero. But transforms b/w 0 and 1

→ so it can be treated as probabilities

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}$$

1 example worth than 100 words

$\begin{bmatrix} 8 \\ 5 \\ 0 \end{bmatrix}$ be z

$e^8 = e^{z_0} = 2981.0$

$e^5 = e^{z_1} = 148.4$

$e^0 = e^{z_2} = 1$

softmax

↓

$$\sigma(\vec{z})_0 = \frac{2981.0}{2981.0 + 148.4 + 1} = \frac{2981.0}{3130.4} = 0.95$$

$\sigma(\vec{z})_1 = \frac{148.4}{3130.4} = 0.047$

$\sigma(\vec{z})_2 = \frac{1}{3130.4} = 0.0003$

so label 0 has probability of 95%.

→ so now $[s_{i1}, s_{i2}, \ldots s_{il}]$ are values we got for ith row for each label

give this to softmax.

$$P_{i1} = \frac{e^{s_{i1}}}{Z} \qquad Z = \sum_{j=1}^{l} e^{s_{ij}}$$

returns probability that ith row belongs to label 1

we have 'l' labels so their summation

$P_i = [P_{i1} \ P_{i2} \ldots P_{il}]$

probabilities of ith row belonging to each label.

$Y_i = [y_{i1} \ y_{i2} \ldots y_{il}]$

this is what actually the label

if ith tuple has l_1 lables

$Y_i = [1 \ 0 \ 0 \ldots 0]$

Now comes KL divergence

it quantifies or shows how much are the probabilities differ in both the vectors

Kind of loss function which we need to reduce.

$$KL(P\|Q) = -\sum_{j=0}^{l} \cancel{P} P_{ij} \log\left(\frac{Q_{ij}}{P_{ij}}\right)$$

ith row
jth label
probability

now we have ground truth as
$\vec{Y_i}$ for i th row
$\vec{P_i}$ for ith row as probs generated.

$$KL(\vec{Y_i}\|\vec{P_i}) = -\sum_{j=0}^{l} P_{ij} \log\left(\frac{y_{ij}}{P_{ij}}\right)$$

As $\vec{Y_i} = \underbrace{[1, 0, 0 \ldots]}$

something just like this so we can have that class label prob where it is not '0'

suppose say label is 1 P_{i1} is all

we want

$KL(\vec{y_i} \,||\, \vec{P_i})$ $\underbrace{- \ln P_{ij}}$ $+ \lambda \overbrace{\sum_{t=1}^{L} \|W_t\|^2}^{\text{regularization}}$

(LOSS FUNCTION)

where j is actual label of the tuple

Now minimize the loss and adjust weights $\underbrace{W_1^T \ldots W_l^T}$ $[w_{11}, w_{21} \ldots w_{1k}]$

Story Boils DOWN

$$\underbrace{\sum_{i=1}^{N} KL(\vec{y}_i \| \vec{P}_i)}_{1} + \alpha \sum_{t=1}^{L} \|W_t\|^2$$

for all tuples

$$= \sum_{i=1}^{N} \sum_{j=1}^{L} \underbrace{I(y_{ij} == 1)}(-\ln P_{ij}) + (\quad)$$

don't worry it only means Identification function to recognize only the labels where y_{ij} is 1

$$-\ln P_{ij} = \frac{e^{W_j^T x_i}}{Z}$$

x_i th row of the set.

j th label weight

$$Z = \sum_{k=1}^{L} e^{W_k^T x_i}$$

Now reduce loss $W_1^T, W_2^T \ldots W_L^T$

let us first write derivation for

$W_j^T = [\omega_{j1} \ \omega_{j2} \ldots \ \omega_{jk}]$

for som ω_{jk} as all are variables.

$$\frac{\partial L}{\partial W_{jk}} = -\sum_{i=1}^{n} \sum_{j'=1}^{L} I(y_{ij}=1) \frac{1}{P_{ij'}} \frac{\partial (P_{ij'})}{\partial W_{jk}}$$

$$+ \lambda \frac{\partial}{\partial \omega_{jk}} \left[\sum_{m=1}^{K} \omega_{jm}^2 \right] = 2\omega_{jk}$$

$$\frac{\partial (P_{ij'})}{\partial \omega_{jk}} = \underbrace{I(j=j')}_{} \left[\frac{e^{\omega_j^T x_i} \cdot x_{ik}}{z} - \ldots \right]$$

this only appears when this

$$\frac{e^{\omega_j^T x_i} \cdot e^{\omega_j^T x_i} \cdot x_{ik}}{z^2}$$

$$= I(j=j') \cdot \frac{e^{\omega_j^T x_i} \cdot x_{ik}}{z} - \frac{\left(e^{W_j^T x_i}\right)^2 \cdot x_{ik}}{z^2}$$

$$= \frac{P_{ij} * x_{ik} * I(j=j')}{z} - \frac{(P_{ij})^2 \cdot x_{ik}}{z^2}$$

$$= \frac{P_{ij} \cdot x_{ik}}{z} \left(I(j=j') - \frac{P_{ij}}{z} \right)$$

$$P_{ij} = \frac{e^{\overbrace{W_j^T x_i}}}{\sum_{k=1}^{L} e^{W_k^T x_i}} \quad [W_{j1}\ W_{j2} \dots W_{jk}]$$

features total are 'k'

$i \to$ row

$j \to$ label

$$\frac{1}{P_{ij}} \cdot \frac{\partial}{\partial W_{jk}} \left(\frac{e^{W_j^T x_i}}{Z} \right)$$

some $[\ \bigcirc\ ] \to W_j$

k-th feature weight.

$j = j'$

$$\left(\frac{x_{ik} e^{W_j^T x_i} \cdot Z - e^{W_j^T x_i} \cdot}{Z^2} \right.$$

$$\sum_{t=1}^{L} \frac{\partial\, e^{W_t^T x_i}}{\partial W_{jk}} \qquad \frac{- e^{W_j^T x_i} \cdot x_{jk} e^{\infty}}{Z^2}$$

$$\frac{\partial L}{\partial W_{j'k}} = -\sum_{i=1}^{N} \sum_{j=1}^{L} \frac{I(j=j')}{P_{ij}} \left(\frac{x_{ik} e^{W_j^T x_i} \cdot Z - e^{W_j^T x_i}}{} \right)$$

$$\left(\frac{x_{ik} e^{W_j^T x_i} Z - e^{W_j^T x_i} \cdot e^{W_j^T x_i} x_{jk}}{Z^2} \right)$$

$$+ I(j \neq j') \left(0 - \frac{e^{W_j^T x_i} \cdot x_{j'k} \cdot e^{W_j^T x_i}}{Z^2} \right)$$

$$\left(x_{ik} \cdot p_{ij} - (p_{ij})^2 \cdot x_{ik}\right) I(j=j')$$

$$- I(j \neq j') \left(p_{ij'} \cdot p_{ij}\right) x_{ik}$$

$$= p_{ij} x_{ik} \left(I(j=j')(1-p_{ij}) - I(j \neq j')(p_{ij'}) \right)$$

$$\boxed{\frac{\partial p_{ij'}}{\partial w_{jk}}} = \begin{cases} p_{ij} x_{ik} (1-p_{ij}) & j = j' \\ p_{ij} x_{ik} p_{ij'} & j \neq j' \end{cases}$$

$$w_{jk} - \alpha \frac{\partial \, loss}{\partial w_{jk}}$$

A	B	C

P	N

Now Version 2

Features K →

Labels L →

↓ data described by N

DATASET

Features F ↓

PRIOR WORDS

$(N \times K)$ — there are no labels (here also means the same)

$(F \times L)$ — what are features? features are nothing but words

$I(g_f \in x_i)$ → identity function

g_f — feature f; x_i — tuple i

0 → if x_i tuple has no g_f feature

1 → if it has

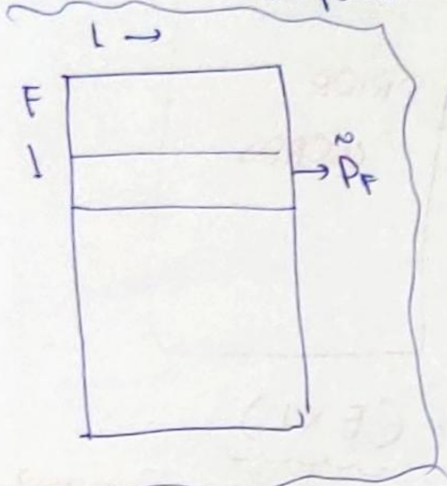
Suppose

	Pos	Neutral	Neg	$\tilde{P}_f$
Excellent	0.9	0.05	0.05	→ all sum up to one

Excellent → feature

this what $F \times L$ has.

$\tilde{P}_f \rightarrow$ denotes probability vector for feature 'f' of the F×L Matrix or prior knowledge



$$C_f = \underbrace{\sum_{i=1}^{N} I(g_f \in x_i)}_{\text{count of all the files having 'f' feature}}$$

We have got P_i [ ] having probabilities of each tuple for each label

$$\tilde{P}_f = \underbrace{\frac{1}{C_f} \sum_{i=1}^{N} P_i \, I(g_f \in x_i)}$$

Sum of all probabilities vectors which has feature 'f'

---

total files which have feature 'f'

$\hat{p}$

pos	neg	neu	
			ith

(N×L)

$\tilde{p}$

pos	neg	neu

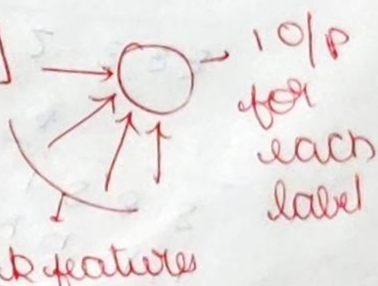
(F×L)

	f	
	1	ith row

(N×K)

$$\underbrace{\sum_{f=1}^{F} KL(\hat{P}_f \,||\, \tilde{P}_f)}_{\text{loss function}} + \underbrace{\lambda \sum_{l=1}^{L} ||W_l||^2}_{\text{weights init randomly}}$$

$W_l = [w_{l1} \; w_{l2} \cdots w_{lk}]$



We do differentiation for one label and 1 feature weight keep in mind repeat for L×K times

$$\frac{\partial O}{\partial \omega_{lk}} = \frac{\partial}{\partial \omega_{lk}}\left(\sum_{f=1}^{F}\left(\underbrace{\sum_{l'=1}^{L} \hat{P}_{fl'} \ln \frac{\hat{P}_{fl'}}{\tilde{P}_{fl'}}}_{\text{for each label for feature } f}\right)\right)$$

subscript

$$+ \frac{\partial}{\partial \omega_{lk}} \lambda \sum_{l'=1}^{L} \sum_{k'=1}^{K} \omega_{l'k'}^{2}$$

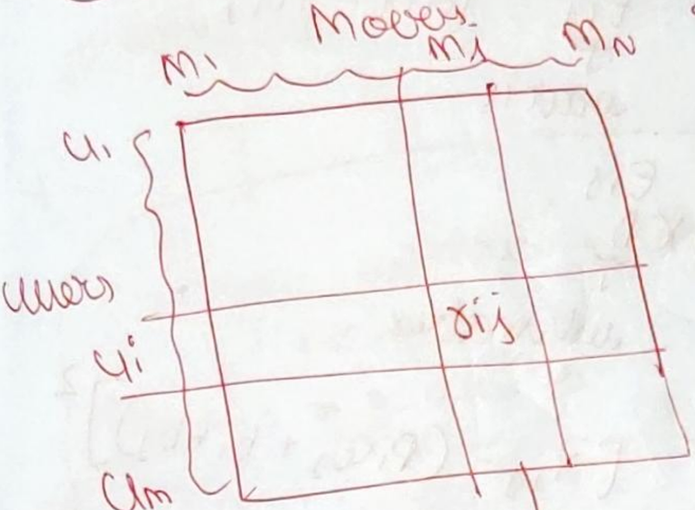
$$= \sum_{f=1}^{F} \sum_{l'=1}^{L} \left(\ln \frac{\hat{P}_{fl'}}{\tilde{P}_{fl'}} + 1\right) \frac{\partial}{\partial \omega_{lk}} \hat{P}_{fl'}$$

$$\lambda \cancel{\sum_{l'=1}^{L}} \cancel{\sum_{k'=1}^{K}}$$

$$\boxed{\lambda \, 2 \omega_{lk}}$$

$$\hat{P}_{tl'} = \frac{e^{W^T x_j}}{Z}$$

WORD2VEC



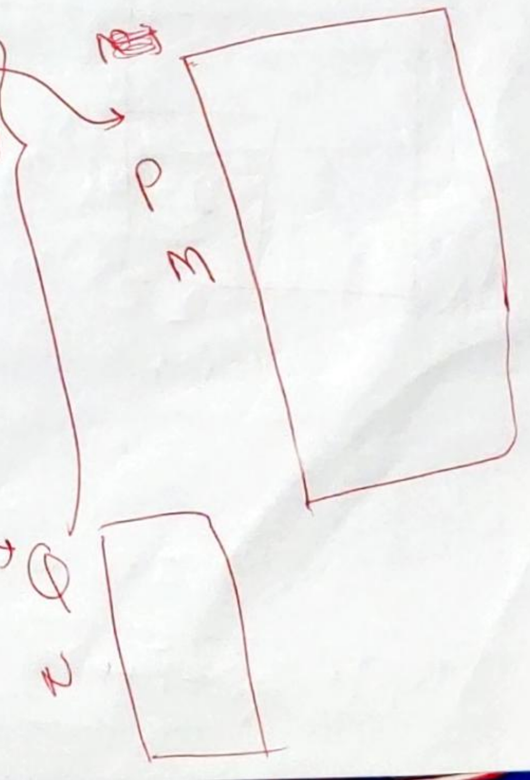
may user may not see each movie so some values missing

$M \times n$ matrix

$r_{ij} \rightarrow$ rating of m_j by u_i

$\underbrace{u_i}_{\text{user i}} \longrightarrow p_i$ transformed vector

$\underbrace{m_j}_{\text{movie j}} \rightarrow q_j$



$$\hat{r}_{ij} = p_i a_j^T + \underbrace{(b_i + b_j)}_{\text{biases}}$$

predicted

$\underset{\text{true}}{r_{ij}}$

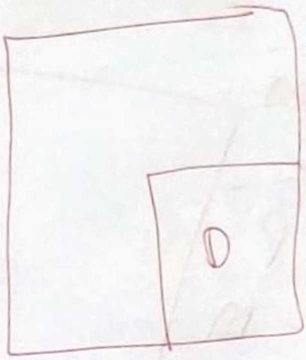
$$\underset{\text{error is}}{E_{ij}} = [r_{ij} - \hat{r}_{ij}]^2$$

$$\text{Loss} = \frac{1}{N_D} \sum_{(i,j) \in D} E_{ij}$$

N_D: no of entries in D

D: sub matrix

$$= \frac{1}{N_D} \sum_{(i,j) \in D} [r_{ij} - (p_i a_j^T + b_i + b_j)]^2$$



D

Machine Translation

source sentence → paired → target translation

learning Representations of words using unlabelled data

[w1, ... wn]

wT

↓

make embeddings of size 5

(example)

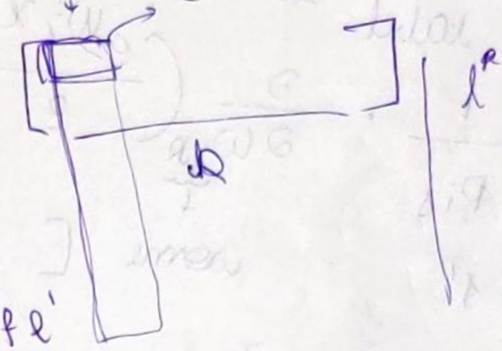
w1 w2 w3 w4 w5 → PIVOT

P(w3 | w1 w2 w4 w5)

given these words

appearing pivot

$$\frac{\partial}{\partial \omega_{jk}} \left(\sum_{f=1}^{F} \sum_{l'=1}^{L} \hat{P}_{fl'} \ln \frac{\hat{P}_{fl'}}{\tilde{P}_{fl'}} \right)$$



$$\frac{\partial \hat{P}_{fl'}}{\partial \omega_{jk}} \ln \frac{\hat{P}_{fl'}}{\tilde{P}_{fl'}}$$

$$+ \hat{P}_{fl'} \frac{\tilde{P}_{fl'}}{\hat{P}_{fl'}} \frac{1}{\tilde{P}_{fl'}} \frac{\partial}{\partial \omega_{jk}} \left(\hat{P}_{fl'} \right)$$

$$\left(\ln \frac{\hat{P}_{fl'}}{\tilde{P}_{fl'}} + 1 \right) \frac{\partial}{\partial \omega_{jk}} (\hat{P}_{fl'})$$

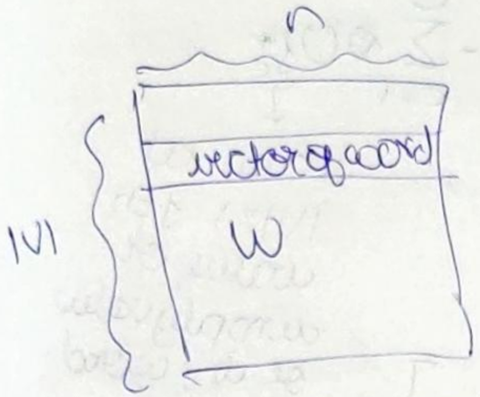
$$\frac{1}{C_f} \sum_{i=1}^{N} I_f (f \in x_i) P_{il'}$$

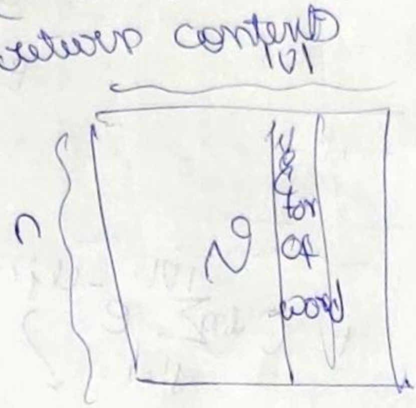
$$\cancel{\frac{e^{\omega_{l'}^{T} x_i}}{\cdot}}$$

$$x_{ik} P_{il'} \left([I(j = l') - P_{ij}] \right)$$

CBOW (continuous bag of words)

skipgram model (center word given return contexts)





$h = \cancel{x^T} W x^T$

$\underbrace{}$ k^{th} row of W
$n \times 1$
( ) matrix

input and some x_k is 1 remaining are 0

$u_c = v^T \cdot h$

$n \times |V|$ $\quad n \times 1$

$1 \times n$ $\quad 1 \times n$ $\quad |V| \times 1$

$1 \times |V|$

$|V| \times n$ $\quad |V| \times 1$

$n \times 1$

$u_c = v^T h$

$$\boxed{u_c = v^T W^T x}$$

$\cancel{n \times |V|}$ $\quad n \times |V|$ $\quad |V| \times 1$

$|V| \times n$ $\quad n \times 1$ $\quad$ loss =

$|V| \times 1$

$$\boxed{u_{cj} = v'^T_{wj} h}$$

jth word in vocabulary

one hot vectors of content words

$$E = -\sum_{c=1}^{C} u_{j_c^*}$$

value at $|V| \times 1$ jth value or simply value of the word.

$$+ C \log \sum_{j'=1}^{|V|} e^{-u_{j'}}$$

$$\left[\quad \right]_{|V| \times 1}$$

$$\frac{\partial E}{\partial u_{c,j}} = \underbrace{y_{c,j}}_{\text{output that we get}} - \underbrace{t_{c,j}}_{\text{ground truth}}$$

$E \cdot I_j$ column vector represents row vec sum of prediction errors across context word panel for current word.

Glove (Globalvector)

co occurance matrix is written

$\frac{P_{ik}}{P_{jk}}$	k = solid	k = gas	k = water	k = f
$P(k \mid ice)$	1.9×10^{-4}	6.6×10^{-5}	3.3×10^{-3}	1.7×10^{-5}
$P(k \mid steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$\frac{P(k \mid ice)}{P(k \mid steam)}$	8.9	8.5×10^{-2}	1.36	0.96

$P_{ik} \rightarrow$ probably that word k occuring given i

P_{jk} probability that word k occurring j

$P_{ik} = \frac{X\_ik}{X\_i}$ $\rightarrow$ seeing words i, k together $\rightarrow$ seeing i alone in corpus

i, j content words k is norm word

$F(w_i\, w_j\, \overline{w_k}) = \frac{P_{ik}}{P_{jk}}$

$\boxed{F(\underbrace{w_i - w_j}, \overline{w_k})} = \frac{P_{ik}}{P_{jk}} \rightarrow$ scalar

linear difference between the words.

vector value.

$F((w_i - w_j)^T \cdot \overline{w_k}) = \frac{F(w_i^T \cdot \overline{w_k})}{F(w_j^T \cdot \overline{w_k})}$

$F(w_i^T \cdot \overline{w_k}) \rightarrow P_{ik} \rightarrow \frac{X\_ik}{X\_i}$

$F(w_j^T \cdot \overline{w_k}) \rightarrow P_{jk} \rightarrow \frac{X\_jk}{X\_j}$

replace F with exponential

$e^{w_i^T \cdot \overline{w_k}} = P\_ik$

$\boxed{\log(x_i) + w_i^T \overline{w_k} = \log(X\_ik)}$

bias term for w_i, w_k

$$w_i^T \cdot \bar{w}_k + b_i + \bar{b}_k = \log(X_{ik})$$

$$J = \sum_{i,j} \underbrace{f(X_{ij})}_{\text{weight function}} \left(w_i^T \cdot \bar{w}_k + b_i + \bar{b}_k - \log(X_{ik})\right)$$

<u>hierarchical softmax</u>

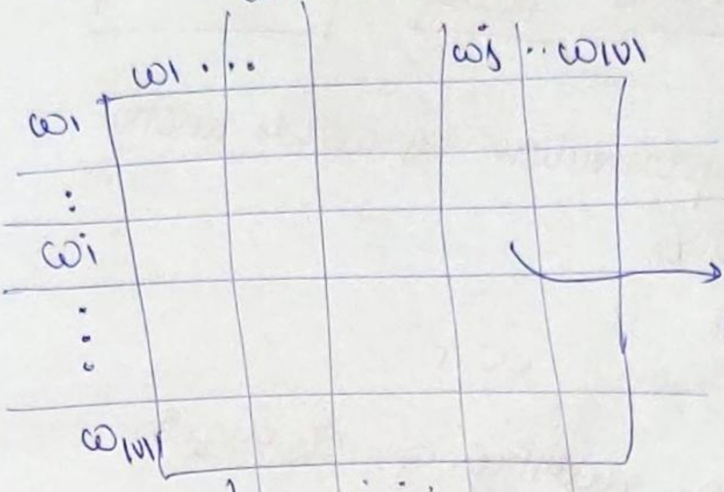
$w_1 \; w_2 \; w_3 \; w_4 \; w_5$

$$\sum\sum \ln P(w_1|w_3) + \ln P(w_2|w_3) + \ln P(w_4|w_3) + \ln P(w_5|w_3)$$

$$P\left(\frac{w_i}{w_j}\right) = \frac{e^{V_{w_i}^T V_{w_j}}}{\sum_{i=1}^{|V|} e^{V_{w_i}^T V_{w_j}}}$$

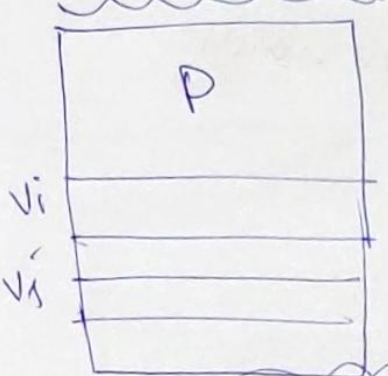
$$\ln P = V_{w_i}^T V_{w_j} - \ln\left(\sum_{i=1}^{|V|} e^{V_{w_i}^T V_{w_j}}\right)$$

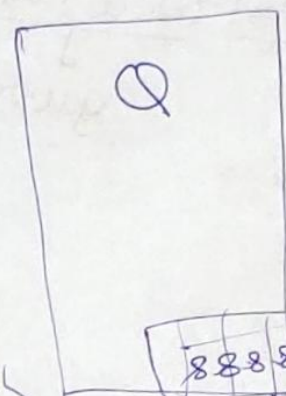
Objective: to learn representation for words using unlabelled data



$P\left(\frac{w_i}{w_j}\right)$ probability of w_i occurring given w_j occurred

~~100~~ K





$$P\left(\frac{w_i}{w_j}\right) = \frac{e^{v_i^T v_j}}{\sum_{i'=1}^{|V|} e^{v_{i'}^T v_j}}$$