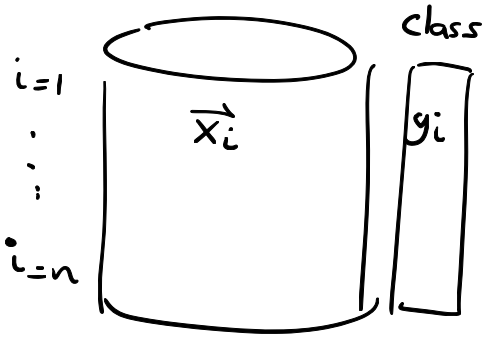
Logistic Regression

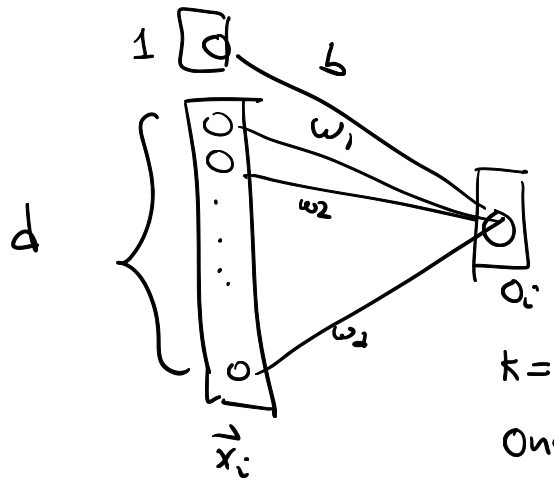


$\vec{x_i} \equiv$ d-dimensional input vector

$y_i \equiv$ categorical/symbolic target value

0 or 1

↑ -ve class ↖ +ve class



k = 1

One neuron for output

$$net\ o_i = \sum_{j=1}^{d} x_{ij}\, \omega_j' + b$$

parameters = $\vec{w} = (\omega_1, \omega_2 \ldots \omega_d) \equiv$ weight vector

$b \equiv$ bias

$f^{\circ} \simeq$ Sigmoid (σ)

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

$$o_i = \sigma\left(net_{o_i}\right) = \sigma\left(\vec{w}^T \vec{x}_i + b\right)$$

Unknowns: $\vec{w}, b$ $\qquad$ $f^{\circ} \equiv P(+ve)$

Likelihood $(\vec{w}, b) \equiv P(D \mid \vec{w}, b)$

Probability of observing the data given the parameters

$$= P(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \mid \vec{w}, b)$$

Joint probability

Assume that points are independent

$$= \prod_{i=1}^{n} P(\vec{x}_i \mid \vec{w}, b)$$

$$L(\vec{\omega}, b) = \prod_{i=1}^{n} \sigma\left(\vec{\omega}^T \vec{x}_i + b\right)^{y_i} \cdot \left(1 - \sigma\left(\vec{\omega}^T \vec{x}_i + b\right)\right)^{1-y_i}$$

$$P(+ve)^{y_i} \cdot P(-ve)^{1-y_i}$$

$$\max_{\vec{\omega}, b} \; L(\vec{\omega}, b)$$

sol: $\frac{\partial L(\vec{\omega}, b)}{\partial \vec{\omega}} = 0 \qquad \frac{\partial L(\vec{\omega}, b)}{\partial b} = 0$

$$\log L(\vec{\omega}, b) = \sum_{i=1}^{n} \log\left(\quad\right)$$

log likelihood

$$= \sum_{i=1}^{n} \Big[y_i \log\left(\sigma(\omega^T x_i + b)\right) + (1-y_i) \log\left(1 - \sigma(\omega^T x_i + b)\right) \Big]$$

$$1 - \sigma(\omega^T x_i + b) = \sigma\left(-(\omega^T x_i + b)\right)$$

$$\log L = \sum_{i=1}^{n} y_i \log\left(P(+ve)\right) + (1-y_i) \log\left(P(-ve)\right)$$

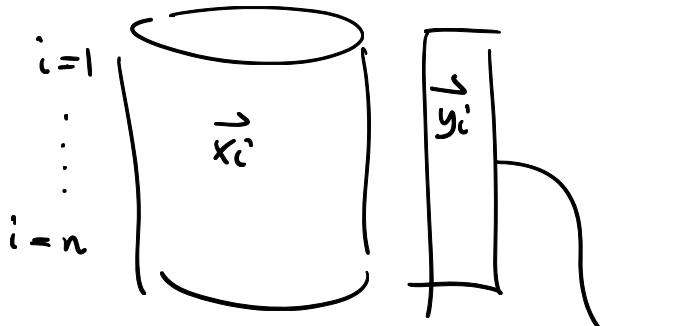
$= \quad -BCE$

binary cross entropy

$\max_{\vec{\omega}, b} \log L$ replace $\min_{\vec{\omega}, b} -\log L \equiv \underline{\underline{BCE}}$

↑ loss function

Multinomial Logistic Regression



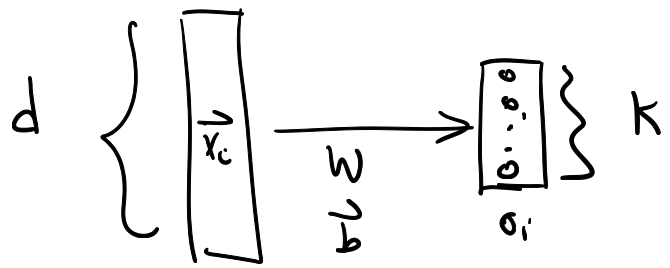
$y_i = \{ c_1 c_2 \ldots c_K \}$

K-classes

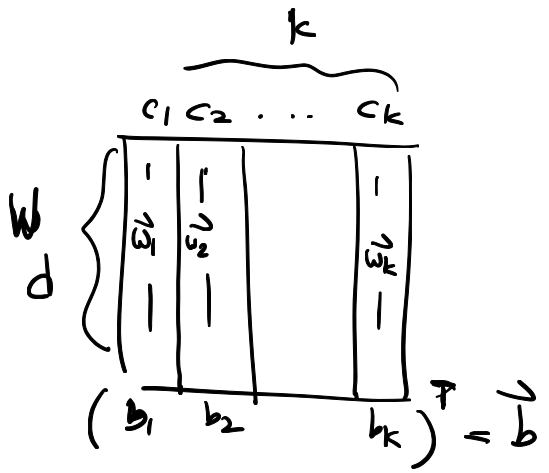
⇓

One-hot vectors

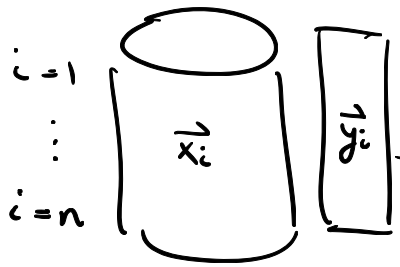
$c_1 = 1000$
$c_2 = 0100$
$c_3 = 0010$ e.g $k=4$
$c_4 = 0001$



W : dxk matrix



Multinomial Logistic Regression



$y_i \in \{ c_1, c_2, \ldots, c_K \}$

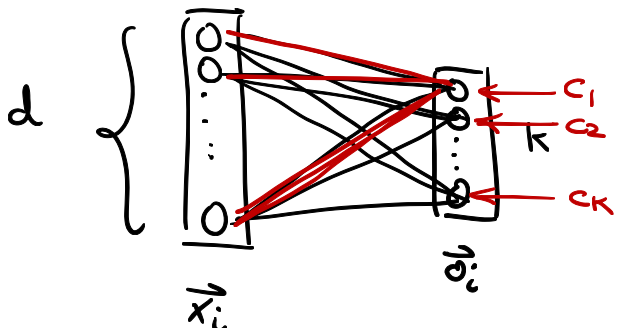
K categorical classes

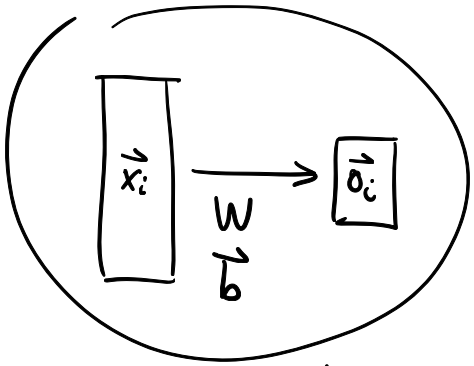
⇓

one-hot encoding

$K = 4$

$$\begin{cases} c_1 = 1000 \\ c_2 = 0100 \\ c_3 = 0010 \\ c_4 = 0001 \end{cases}$$

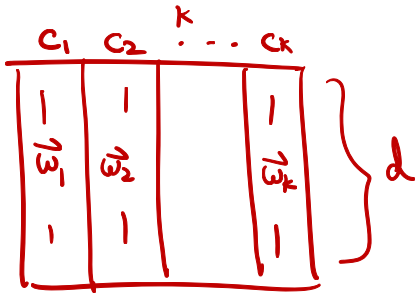




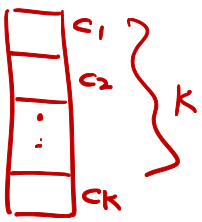
$W = d \times K$ matrix

$\vec{b} = K$-dim vector

$W =$



$\vec{b}$



$P(c_1 | \vec{x}_i)$? $P(c_j | \vec{x}_i)$?

prob of class c_j given $\vec{x}_i$

$$\overrightarrow{net_o} = W^T \vec{x}_i + \vec{b}$$

$(k \times d)(d \times 1)$ $k \times 1$

$k \times 1$

$$\begin{bmatrix} net_{o_1} \\ net_{o_2} \\ \vdots \\ net_{o_k} \end{bmatrix} \quad k\text{-dim}$$

$$\vec{o} = \overset{o}{f}(\overrightarrow{net_o}) = \text{softmax}(net_{o_i} \mid net_{o_1}\ net_{o_2} \ldots net_{o_k})$$

output vector k-dim

activation function

$\forall\ i = 1 \ldots n$

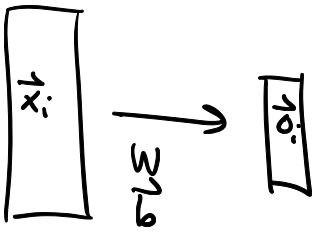
net_o

$$= \begin{bmatrix} P(c_1 | \overrightarrow{net_o}) \\ P(c_2 | \overrightarrow{net_o}) \\ \vdots \\ P(c_k | \overrightarrow{net_o}) \end{bmatrix}$$

$$P(c_j | \vec{x}_i) = P(c_j | \overrightarrow{net_o}) = \text{softmax}(net_{o_j} \mid \overrightarrow{net_o})$$

$$= \frac{e^{net_{o_j}}}{\sum_{a=1}^{k} e^{net_{o_a}}}$$

$$\sum_{j=1}^{k} P(c_j | x_i) = 1$$



Unknown parameters are $W, \vec{b}$

MLE : Maximum likelihood Estimation

$$\max_{\substack{W \\ \vec{b}}} L(W, \vec{b}) = \prod_{i=1}^{n} P(\vec{x}_i \mid \vec{w}, b)$$

assuming all points are independent

$$= \prod_{i=1}^{n} \left[\prod_{j=1}^{k} P(\vec{x}_i \mid \vec{w}, b)^{y_{ij}} \right]$$

e.g. $\vec{y}_i = (\underbrace{1}_{j}\ 0\ 0\ 0)$ only one entry is 1, rest are 0

(1 2 3 4)

true target

$\vec{y}_2 = (0\ 0\ \underbrace{1}_{j}\ 0)$: $y_{23} = 1$

(1 2 3 4)

$$\max_{W, \vec{b}} \log L(W, \vec{b}) = \sum_{i=1}^{n} \sum_{j=1}^{k} y_{ij} \log \left(P(x_i \mid W, \vec{b}) \right)$$

0	0	1	0

$\vec{y}_i$

vs.

0.1	0.2	0.5	0.2

$\vec{o}_i$

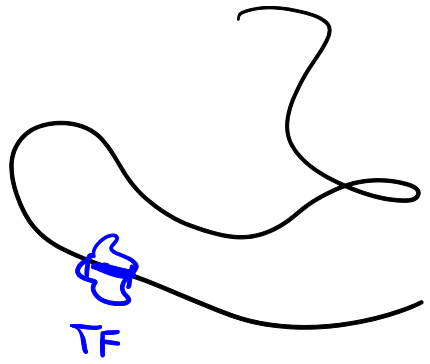
$\equiv -CE$

- Cross Entropy

$$\min_{w,\vec{b}} \; -\log L(w,\vec{b}) \equiv CE \text{ loss}$$

$$\nabla_w = \frac{\partial CE}{\partial w} \qquad \nabla_{\vec{b}} = \frac{\partial CE}{\partial \vec{b}}$$

TF proteins → pattern → biological word



DNA / Protein sequence

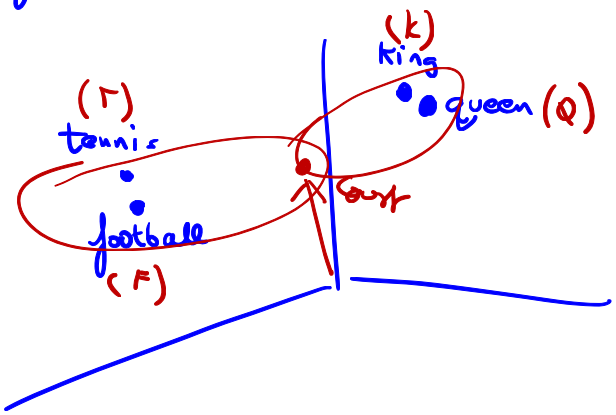
1) what are the words?

2) how to represent them numerically?

Word embedding

Context independent

One word One vector

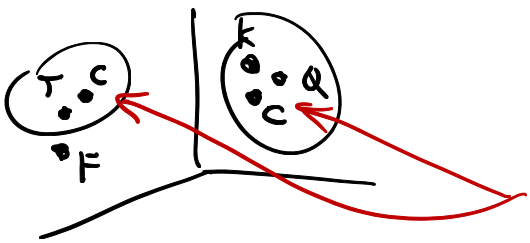


court?
↓
multiple sense/meanings

d-dim space

$$\vec{w}_{court} = \begin{bmatrix} K/Q \\ \hline T/F \end{bmatrix} \} \text{ d-dim}$$

Context dependent



$\vec{w}^1_{court}$: vector close to K/Q

$\vec{w}^2_{court}$: vector close to tennis

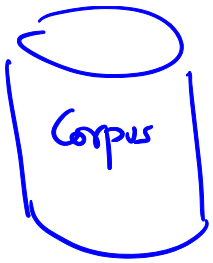
Task: Word → Vector representation

d-dim

d = 300 ←
d = 768 ←
⋮

Word2Vec : word to vector

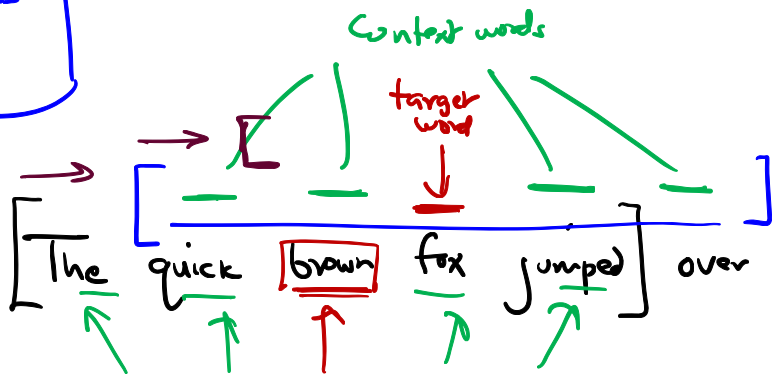
(skip gram)



A very large collection of text : Entire Wikipedia

(web crawls)

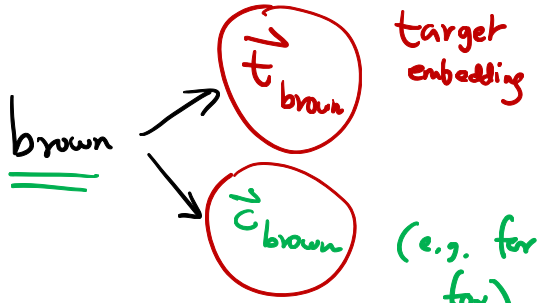
(all digitised books, etc.)



Window size
w = 5

target / center word

4 context words



Vocabulary V

$$T = \begin{bmatrix} \vec{t}_1 \\ \vec{t}_2 \\ \vdots \\ \vec{t}_{|V|} \end{bmatrix} \qquad C = \begin{bmatrix} \vec{c}_1 \\ \vec{c}_2 \\ \vdots \\ \vec{c}_{|V|} \end{bmatrix}$$

$|V| \times d$ $\qquad$ $|V| \times d$

d-dim target and context embedding $\forall\ i = 1 \ldots |V|$ for all words in vocab

Unknown parameters

Use logistic regression to learn T and C matrices

[The quick brown fox jumped] over the fence

target

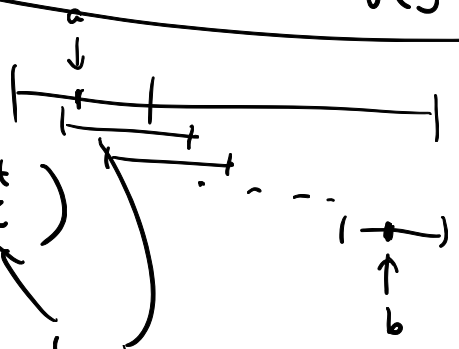
$$P(\text{context} \mid \text{target}) = P(\text{the}, \text{quick}, \text{fox}, \text{jumped} \mid \text{brown})$$

joint prob

assume the words are independent

$$= P(\text{the} \mid \text{brown}) \times P(\text{quick} \mid \text{brown}) \times P(\text{fox} \mid \text{brown}) \times P(\text{jumped} \mid \text{brown})$$

$\forall$ all windows of size w



$$P = \prod_{i=\lfloor w/2 \rfloor}^{L-\lfloor w/2 \rfloor+1} \left(\prod_{i-\lfloor w/2 \rfloor \le j \le i+\lfloor w/2 \rfloor} P(w_j^c \mid w_i^t) \right)$$

a = all possible windows

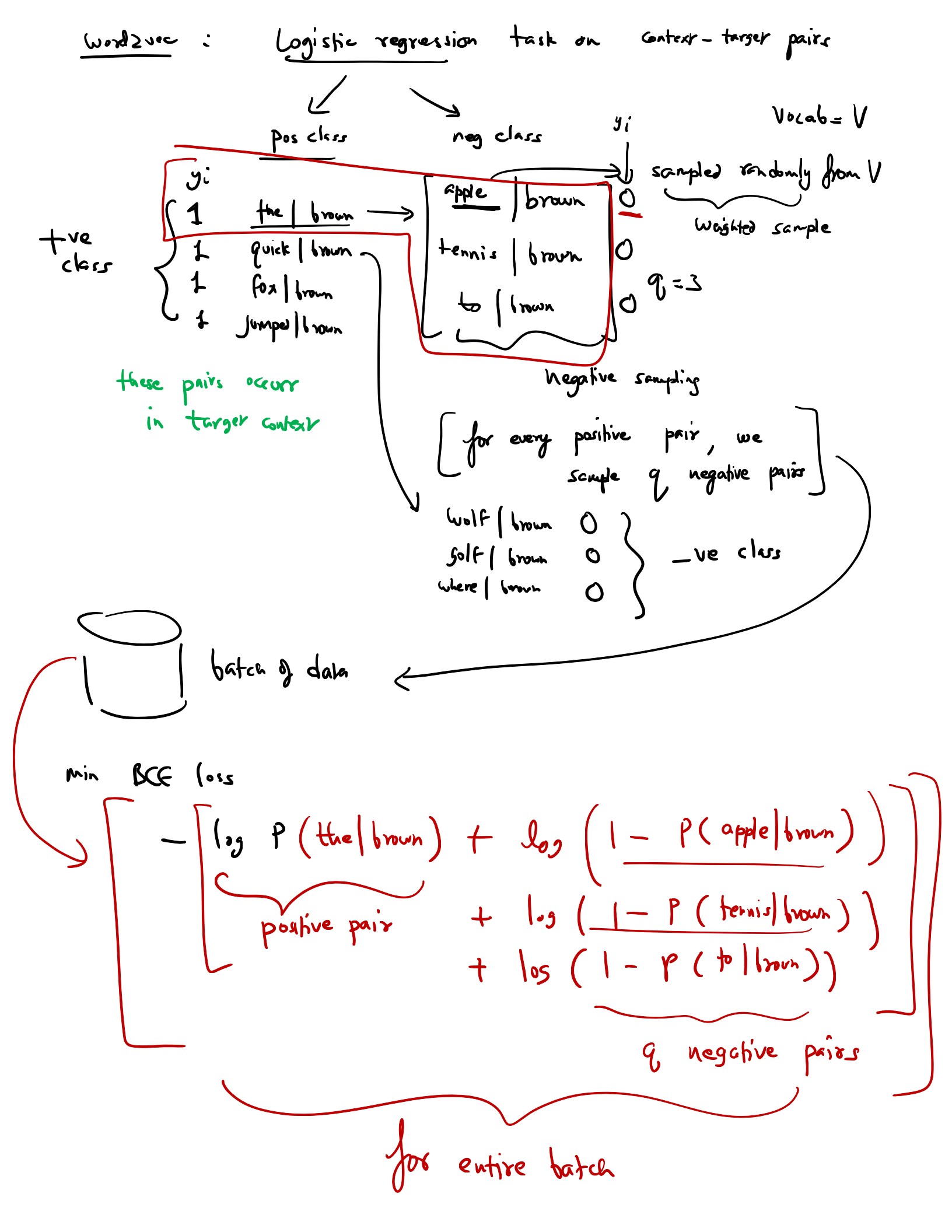
context

target

all context–target pairs in that window

$$-\log L = -\sum_{i=\text{window}} \sum_{j=\text{context words}} \log\left(P(w_j^c \mid w_i^t)\right)$$

BCE loss


word2vec : Logistic regression task on context - target pairs
pos class
neg class
yi
Vocab = V
sampled randomly from V
Weighted sample
yi
1 the | brown
1 quick | brown
1 fox | brown
1 jumped | brown
+ve class
apple | brown 0
tennis | brown 0
to | brown 0
q = 3
these pairs occurr in target context
Negative sampling
[for every positive pair, we sample q negative pairs]
wolf | brown 0
golf | brown 0
where | brown 0
-ve class
batch of data
min BCE loss
- [log P(the|brown) + log (1 - P(apple|brown))
positive pair
+ log (1 - P(tennis|brown))
+ log (1 - P(to|brown))]
q negative pairs
for entire batch

$$P(\text{the} \mid \text{brown}) = \sigma\left(\vec{c}_{\text{the}}^{\,T} \vec{t}_{\text{brown}}\right)$$

$\vec{c}_{\text{the}}$: d

$\vec{t}_{\text{brown}}$: d

Unknown parameters

dot product

sigmoid

$$\nabla_T = \frac{\partial BCE}{\partial T} \qquad \nabla_C = \frac{\partial BCE}{\partial C}$$

$$T = \begin{bmatrix} \vec{t}_1 \\ \vdots \\ \vec{t}_{|V|} \end{bmatrix} \quad |V| \times d$$

$$C = \begin{bmatrix} \vec{c}_1 \\ \vec{c}_2 \\ \vdots \\ \vec{c}_{|V|} \end{bmatrix} \quad |V| \times d$$

word → $\vec{t}_{\text{word}}$

word → $\vec{c}_{\text{word}}$

$$\vec{w} = \frac{\vec{t}_w + \vec{c}_w}{2}$$

average of target & context vectors

embedding for the word $\vec{w}$