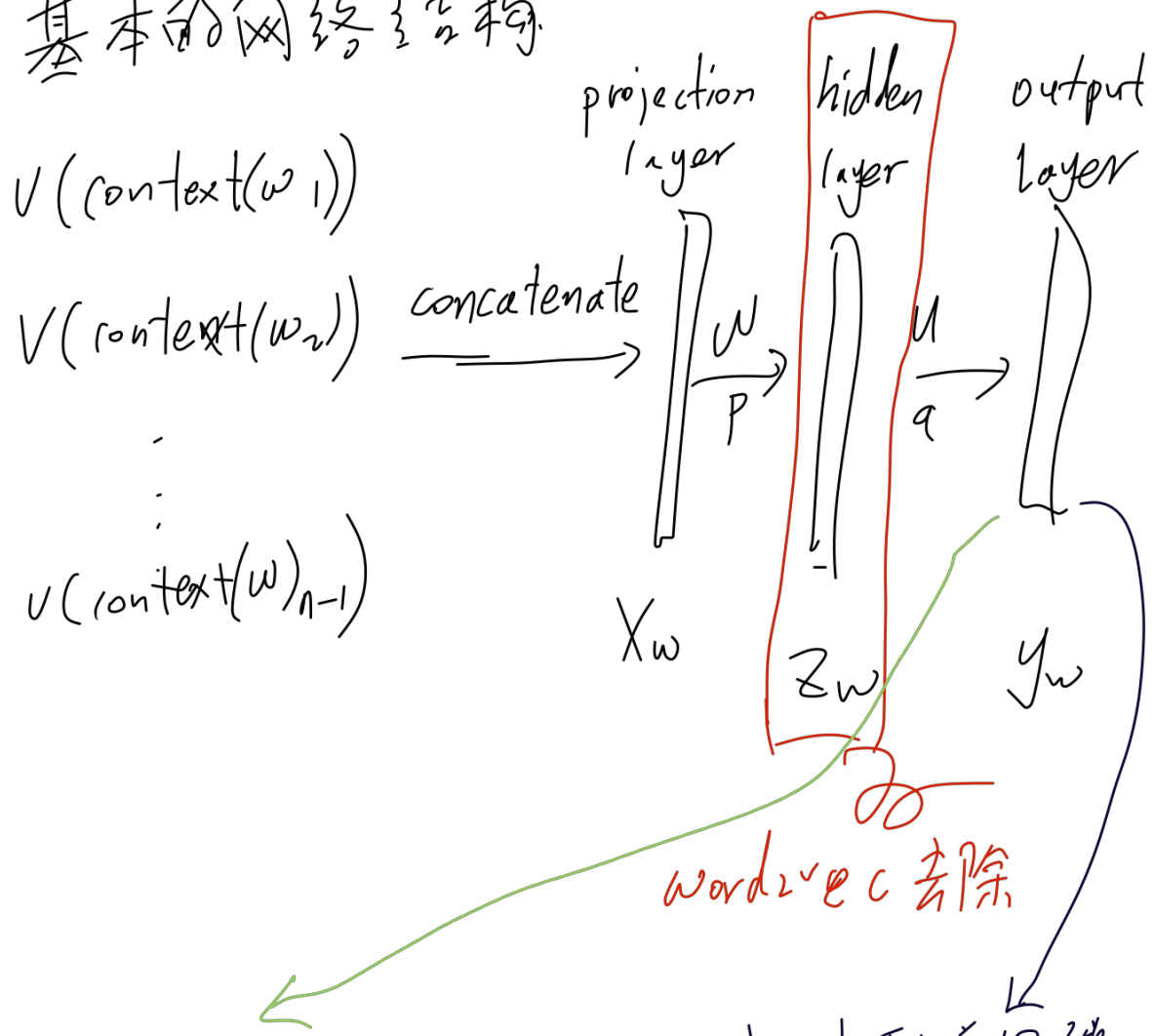
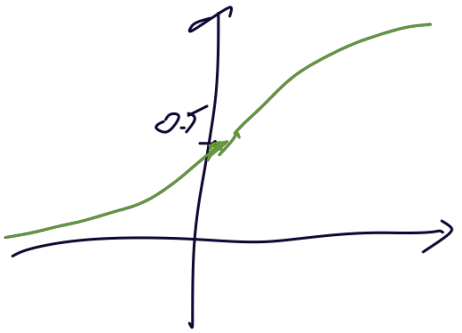
基本的网络结构


V(context(w1))
V(context(w2))
V(context(w)n-1)
concatenate
projection layer
hidden layer
output layer
W
P
U
q
Xw
Zw
yw
word2vec去除
hierarchical softmax
对几十万个单词做.
softmax
$S_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$
得到每个单词的概率

2.1 sigmoid

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



$$\sigma'(x) = \sigma(x)(1-\sigma(x))$$

$$\log'(\sigma(x)) = 1-\sigma(x)$$

$$\log'(1-\sigma(x)) = -\sigma(x)$$

逻辑回归做二分类，用 sigmoid.

$$x = (x_1\ x_2 \cdots\ x_n)^T$$

$$h_\theta(x) = (\theta_0 + \theta_1 x_1 + \theta_2 x_2 \cdots\ \theta_n x_n)$$

$$= (\theta_0 x_0 + \theta_1 x_1 \cdots + \theta_n x_n) \quad (x_0 = 1)$$

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

样本 x 的分值 $h_\theta(x)$

$$y=\begin{cases}0, & h_\theta(x)<0.5\\ 1, & h_\theta(x)\geq 0.5\end{cases}$$

θ是参数，需更新. 损失函数:

$$J(\theta)=\frac{1}{m}\sum_{i=1}^{m}cost(x_i,y_i)$$

$$cost(x_i,y_i)=\begin{cases}-\log h_\theta(x_i) & y_i=1\\ -\log(1-h_\theta(x_i)) & y_i=0\end{cases}$$

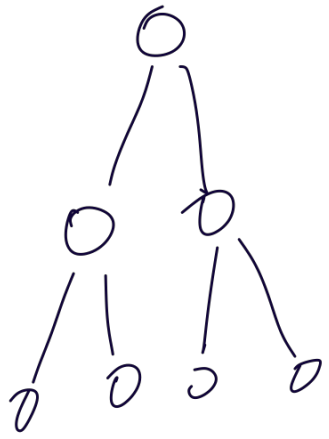
$$P(A|B)=\frac{P(AB)}{P(B)}\qquad P(B|A)=\frac{P(AB)}{P(A)}$$

Bayes

$$P(A|B) = \frac{P(A)\cdot P(B|A)}{P(B)}$$

两个条件概率

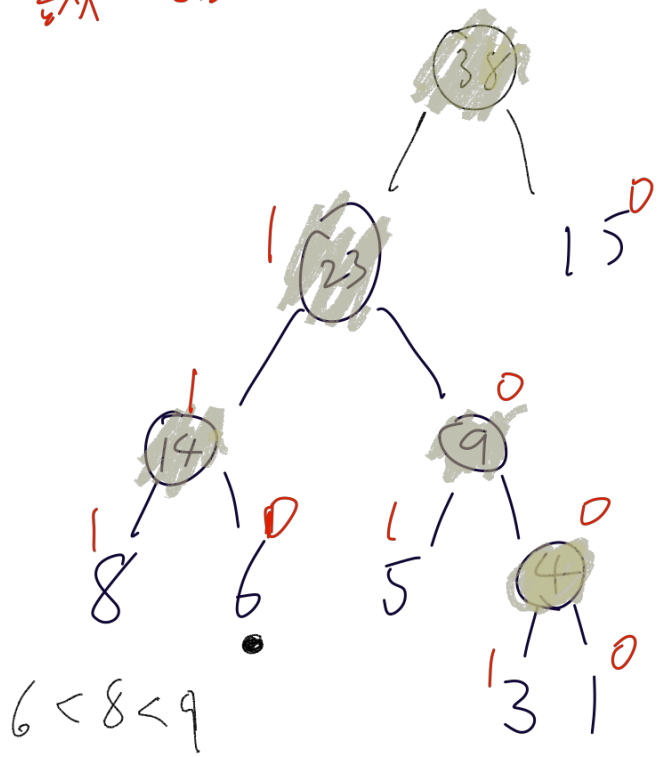
Huffman 树是二叉树.



例子:

15, 8, 6, 5, 3, 1.

我数数看已经证明过了



6 < 8 < 9

词频越大的词离根结点越近.

Huffman 编码

是不等长编码，利用霍夫曼树

设计的二进制前缀码

word2vec 也用到了霍夫曼编码.

$$P(text|voice) = \frac{P(voice|text)\,P(text)}{P(voice)}$$

$P(voice|text)$：声学模型　$P(text)$：语言模型

LM: $P(w_1 w_2 \cdots w_n)$

$= P(w_n) \cdot P(w_n | w_1 \cdots w_{n-1})$

$= P(w_n)$.

$$P(A|B) = \frac{P(AB)}{P(B)}$$

$P(w_1 w_2 w_3)$

$= P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_2)$.

$P(w_1 w_2 w_3 \cdots w_n)$

$= P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_2) \cdots P(w_n|w_{n-1})$

词典个数为 N. 对于长度为 T 的句子，

每个位置有 N 种可能，$\underbrace{N \cdot N \cdot N \cdots N}_{T} = N^T$

每种可能要计算 T 个参数，共需 $T \cdot N^T$

内存开销很大

n-gram，决策树，最大熵模型、最大熵马尔科夫、条件随机场，神经网络 构建语言模型。

n-gram 模型。

$O(N^n)$ 词典大小 $N=20000$

n 是 ngram. = 1, 2, 3, 4.

$p(w_1 w_2 w_3 w_4 w_5)$ 3-gram

$= p(w_1) \cdot p(w_2|w_1) \cdot p(w_3|w_1 w_2) \cdot p(w_4|w_2 w_3)$

$\cdot p(w_5|w_3 \cdot w_4)$ ？对不对。

对问题建模，构造目标函数，进行优化.

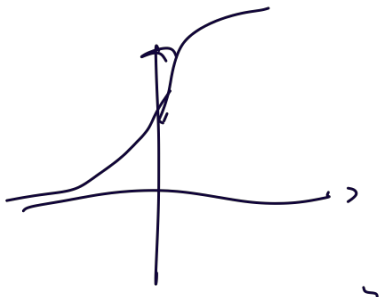
最大似然

$$\prod_{w \in C} P(w|context(w))$$

最大对数似然

$$L = \sum_{w \in C} \log P(w|Context(w))$$

tanh 双曲正切函数，作为隐藏层的激活函数



词向量
- one hot representation → 01
- distribution representation
  - 将概念分担给每个单词

Word2vec
- CBOW → hierarchical softmax
- skip-gram → negative sampling

CBOW

目标函数：

$$L=\sum_{w\in C}\log p(w|context(w))$$

相对于神经网络的语言模型，

1' 去掉了隐藏层。

2' 用 Huffman 树替代 softmax。

$$p(w|context(w))=\prod_{j=2}^{l^w} p(d_j^w|x_w,\theta_{j-1}^w)$$

故 L 是关于 θ 和 w 的函数。

求 L 关于 θ 和 w 的偏导并更新。

$$\theta_{j-1}^w := \theta_{j-1}^w+\eta\left[\frac{\partial L(w,j)}{\partial \theta_{j-1}^w}\right]$$

$$v(\tilde{w}) := v(\tilde{w})+\eta\sum_{j=2}^{l^w}\frac{\partial L(w,j)}{\partial w}$$

skip-gram 目标函数

$$L = \sum_{w \in C} \log p(context(w) | w)$$