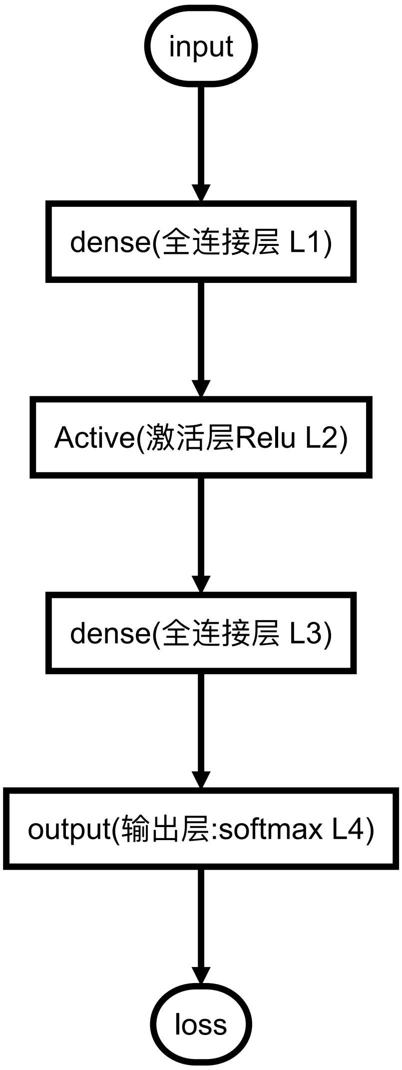
反向传播推倒

北京大学
地球与空间科学学院 陈宁 1601210216

本推倒考虑单个全连接隐藏层且激活函数为Relu，输出层为softmax，损失选择cross-entropy，为了灵活性，将全连接隐藏层拆分为两个层次——全连接dense层和激活层active层,即全连接层是一个线性变换不包含非线性因素。网络形式如下：



其中input为$X^{(0)}$,经过L_1数据变为$X^{(1)}$，经过L_2数据变为$X^{(2)}$经过L_3数据变为$X^{(3)}$经过L_4数据变为$X^{(4)}$，Forward过程为：

$$score = X^{(3)} = X^{(2)} * W_2 = Relu(X^{(1)}) * W_2 = Relu(X^{(0)} * W_1) * W_2$$

$$output = X^{(4)} = softmax(score)$$

$$loss = cross\_entropy(output, Y)$$

反向传播过程：

（1） L_4层 Loss的反向传播

$$L = -\frac{1}{N}\sum_{i=1}^{N} log(\frac{e^{x_{ik}}}{\sum_j e^{x_{ij}}})$$

$$= -\frac{1}{N}\sum_{i=1}^{N}[x_{ik} - log\sum_j e^{x_{ij}}]$$

其中N为batchsize,k为对应的true label，因此梯度为

$$\frac{\partial L}{\partial X_{im}^{(3)}} = \frac{1}{N}\sum_{i=1}^{N}\frac{e^{x_{im}}}{\sum_j e^{x_{ij}}} \qquad (m \neq k)$$

$$\frac{\partial L}{\partial X_{ik}^{(3)}} = \frac{1}{N}\sum_{i=1}^{N}[\frac{e^{x_{ik}}}{\sum_j e^{x_{ij}}} - 1] \qquad m = k$$

(2)L_3层反向传播过程:分别对W和X进行反向传播

$$dL = tr[(\frac{\partial L}{\partial X^{(3)}})^T \bullet d(X^{(2)}W_2)]$$

$$= tr[(\frac{\partial L}{\partial X^{(3)}})^T \bullet X^{(2)}dW_2] + tr[W_2(\frac{\partial L}{\partial X^{(3)}})^T dX^{(2)}]$$

根据微分定义、标量对矩阵求导的定义对比可以得到：

$$dW_2 = \frac{\partial L}{\partial W_2} = X^{(2)T} \bullet \frac{\partial L}{\partial X^{(3)}}$$

$$dX^{(2)} = \frac{\partial L}{\partial X^{(2)}} = \frac{\partial L}{\partial X^{(3)}} \bullet W_2^T$$

(3)对激活层反向传播过程 $Relu = max(0, x)$

$$\frac{\partial L}{\partial X_{ij}^{(1)}} = \begin{cases} dX_{ij}^{(2)} \bullet 1 & \text{对于}X_{ij}^{(1)} > 0\text{的部分特征} \\ 0 \quad otherwise \end{cases}$$

(4)对L_1全连接层的反向传播只需要对W_1梯度

$$\frac{\partial L}{\partial W_1} = X^{(0)T} \bullet dX^{(1)}$$