HW5

PB20111689 蓝俊玮

5.1

如果采用的是线性激活函数 $f(\boldsymbol{x})=\boldsymbol{w}^T\boldsymbol{x}$ 作为神经元的激活函数，那么我们可以知道，对于任意一个隐藏层单元，都有输入为 $\sum_i w_ix_i$，那么经过激活函数之后，得到的输出仍然为 $\sum_i w_i'x_i$，即经过激活函数之后得到的结果仍然是线性拟合。则这样的激活函数无法拟合出更加复杂的模型，只能做线性模型的训练学习。因此线性激活函数的缺陷就是会使神经网络的训练结果变差，无法拟合除线性模型以外的更复杂的模型。

T2

对于函数 $\frac{\exp(x_i)}{\sum_{j=1}^C\exp(x_j)}$ 和函数 $\log\sum_{j=1}^C\exp{(x_j)}$ 来说，当 $x_i\to\infty\mid x_j\to\infty$ 时，这时候对于 softmax 函数来说，分子会趋向于正无穷，导致 $\exp$ 函数的计算值过大，从而发生溢出；而对于第二个函数而言，也会导致 $\exp$ 函数的计算值过大，从而出现溢出现象。当 $x_j\to-\infty$ 时，会导致 softmax 函数的分母趋向于 0，从而导致整个 softmax 发生除以 0 的溢出现象；而对第二个函数来说，会导致 $\log$ 函数趋向于负无穷，同样会发生数值溢出现象。

要解决数值溢出问题，可以取 $M=\max(x_i),\ \forall i\in[1,C]$，然后在计算时采用 $\frac{\exp(x_i-M)}{\sum_{j=1}^C\exp(x_j-M)}$ 来表示即可，同理，解决下溢出时可以取 $M=\min(x_i),\ \forall i\in[1,C]$，然后在计算时采用 $\frac{\exp(x_i-M)}{\sum_{j=1}^C\exp(x_j-M)}$ 来表示即可。

T3

令：

$$f(x_i)=\frac{\exp{(x_i)}}{\sum_{j=1}^C\exp{(x_j)}}$$
$$g(x_i)=\log{(f(x_i))}$$

则：

$$\frac{\partial f(x_i)}{\partial x_k}=\frac{-\exp{(x_i)}\exp{(x_k)}}{(\sum_{j=1}^C\exp{(x_j)})^2}=-f(x_i)f(x_k)\quad if\ k\neq i$$

$$\frac{\partial f(x_i)}{\partial x_k}=\frac{\exp{(x_i)}(\sum_{j=1}^C\exp{(x_j)})-\exp{(x_i)}\exp{(x_k)}}{(\sum_{j=1}^C\exp{(x_j)})^2}=f(x_i)-f(x_i)f(x_k)\quad if\ k=i$$

$$\frac{\partial g(x_i)}{\partial x_k}=\frac{\partial g(x_i)}{\partial f(x_i)}\frac{\partial f(x_i)}{\partial x_k}=\frac{1}{f(x_i)}\frac{\partial f(x_i)}{\partial x_k}$$
$$\Rightarrow\frac{\partial g(x_i)}{\partial x_k}=-f(x_k)\quad if\ k\neq i$$
$$\Rightarrow\frac{\partial g(x_i)}{\partial x_k}=1-f(x_k)\quad if\ k=i$$

所以可以得到

$$\frac{\partial f(x_i)}{\partial x_k} = f(x_i)(\delta_{ik} - f(x_k))$$
$$\frac{\partial g(x_i)}{\partial x_k} = \delta_{ik} - f(x_k)$$
$$\frac{\partial f(x_i)}{\partial \boldsymbol{x}} = [f(x_i)(\delta_{i1} - f(x_k)),\ f(x_i)(\delta_{i2} - f(x_k)),\ \ldots,\ f(x_i)(\delta_{iC} - f(x_k))]$$
$$\frac{\partial g(x_i)}{\partial \boldsymbol{x}} = [\delta_{i1} - f(x_k),\ \delta_{i2} - f(x_k),\ \ldots,\ \delta_{iC} - f(x_k)]$$

T4

隐藏层单元有

$$x_1.in = 0.6 \times 0.2 + 0.2 \times 0.3 = 0.18$$
$$x_1.out = ReLU(0.18) = 0.18$$
$$x_2.in = 0.1 \times 0.2 + 0.7 \times 0.3 = 0.23$$
$$x_2.out = ReLU(0.23) = 0.23$$
$$y.in = 0.5 \times 0.18 + 0.8 \times 0.23 = 0.274$$
$$\hat{y} = y.out = ReLU(0.274) = 0.274$$

误差传播为

$$E(W) = \frac{1}{2}(y-\hat{y})^2 = 0.5 \times (0.5-0.274)^2 = 0.025538$$
$$g_y = (y.out - y) \times ReLU'(y.in) = -0.226$$
$$\frac{\partial E(W)}{\partial w_1} = g_y \times x_1.out = -0.04068$$
$$w_1 = w_1 - \alpha \times \frac{\partial E(W)}{\partial w_1} = 0.54068$$
$$\frac{\partial E(W)}{\partial w_2} = g_y \times x_2.out = -0.05198$$
$$w_2 = w_2 - \alpha \times \frac{\partial E(W)}{\partial w_2} = 0.85198$$

$$g_{x_1} = g_y \times w_1 \times ReLU'(x_1.in) = -0.113$$
$$g_{x_2} = g_y \times w_2 \times ReLU'(x_2.in) = -0.1808$$
$$\frac{\partial E(W)}{\partial v_{A1}} = g_{x_1} \times A = -0.0226$$
$$v_{A1} = v_{A1} - \alpha \times \frac{\partial E(W)}{\partial v_{A1}} = 0.6226$$
$$\frac{\partial E(W)}{\partial v_{A2}} = g_{x_2} \times A = -0.03616$$
$$v_{A2} = v_{A2} - \alpha \times \frac{\partial E(W)}{\partial v_{A2}} = 0.13616$$
$$\frac{\partial E(W)}{\partial v_{B1}} = g_{x_1} \times B = -0.0339$$
$$v_{B1} = v_{B1} - \alpha \times \frac{\partial E(W)}{\partial v_{B1}} = 0.2339$$
$$\frac{\partial E(W)}{\partial v_{B2}} = g_{x_2} \times B = -0.05424$$
$$v_{B1} = v_{B1} - \alpha \times \frac{\partial E(W)}{\partial v_{B1}} = 0.75424$$

则更新后的输出值为

$$x_1'.in = 0.6226 \times 0.2 + 0.2339 \times 0.3 = 0.19469$$
$$x_1'.out = ReLU(0.19469) = 0.19469$$
$$x_2'.in = 0.13616 \times 0.2 + 0.75424 \times 0.3 = 0.253504$$
$$x_2'.out = ReLU(0.253504) = 0.253504$$
$$y'.in = 0.54068 \times 0.19469 + 0.85198 \times 0.253504 = 0.32$$
$$y'.out = ReLU(0.32) = 0.32$$

则参数更新后的平方损失为

$$E(W) = \frac{1}{2}(y-\hat{y})^2 = 0.0162$$

则经过一次误差反向传播后，平方损失值下降了。