机器学习作业二：MLP_BP推导

专业：信息安全 姓名：魏伯繁 学号：2011395

题目要求：根据所给资料推导反向传播算法

根据题目要求，给出Sw的表达式，并对其中的参数做一定解释

$$y_i^M表示第M层的输出$$

$$m_c^M表示第M层第c类样本的平均输出$$

$$S_w = \sum_{c=1}^{C} \sum_{y_i^M \in c} (y_i^M - m_c^M)(y_i^M - m_c^M)^T \tag{1}$$

根据题目要求，给出Sb表达式，并对其中的参数做一定解释

$$n_c表示第c类的样本数$$

$$m^M表示所有样本的第M层的平均输出$$

$$S_b = \sum_{c=1}^{C} n_c(m_c^M - m^M)(m_c^M - m^M)^T \tag{2}$$

根据题目要求，给出损失函数的表达式，并对参数做一定解释

$$y_{i,j}^M是向量y_i^M的第j个元素$$

$$d_{i,j}^M是标签d_i^M的第j个元素$$

$$tr表示矩阵的迹，\frac{1}{2}是为了简化结果而设计的$$

$$E = \sum_i \sum_j \frac{1}{2}(y_{i,j}^M - d_{i,j}^M)^2 + \frac{1}{2}\gamma(tr(S_w) - tr(S_b)) \tag{2}$$

根据上面所给出的分析，我们不难得知，本次推导所给出的条件并非像通常机器学习教程中反向推导的示例中一次只有一个样本的输出，而是一次输入多个样本，根据多个样本的反馈统一更新一次参数。

有了上面的分析，我们不妨假设一次输入k个样本，k个样本的值为：

$$I_1 = [x^{(1)}, x^{(2)}, x^{(3)}, \cdots, x^{(k)}] \tag{3}$$

假设每一次输入层有t1个结点，我们可以将上面的列向量扩展成矩阵形式：我们可以将其理解为输入k次，每次输入的是一个列向量，并且每个列向量要填满t1个输入神经元

$$I_1 = \begin{bmatrix} x_{1,1}^{(1)} & x_{1,2}^{(1)} & x_{1,3}^{(1)} \cdots & x_{1,k}^{(1)} \\ x_{2,1}^{(1)} & x_{2,2}^{(1)} & x_{2,3}^{(1)} \cdots & x_{2,k}^{(1)} \\ \cdots & & & \\ x_{t_1,1}^{(1)} & x_{t_1,2}^{(1)} & x_{t_1,3}^{(1)} \cdots & x_{t_1,k}^{(1)} \end{bmatrix}_{t_1 \times k} \tag{4}$$

接下来，我们不妨假设隐藏层有t2个神经元，则输入层与隐藏层之间的权重矩阵则可以表示为,其中下标的含义是，第一个下标为隐藏层的序号，第二个下表为输入层的序号。在这里说明一下上标的问题，由于参数是连接层1和层2的，所以上标选择1或者2都可以，在这里选择向后靠拢，所以权重的上标选择2

$$W_{12}=\begin{bmatrix} w_{1,1}^{(2)} & w_{1,2}^{(2)} & w_{1,3}^{(2)}\cdots & w_{1,t}^{(2)} \\ w_{2,1}^{(2)} & w_{2,2}^{(2)} & w_{2,3}^{(2)}\cdots & w_{2,t}^{(2)} \\ \cdots & & & \\ w_{t_2,1}^{(2)} & x_{t_2,2}^{(2)} & x_{t_2,3}^{(2)}\cdots & x_{t_2,t}^{(2)} \end{bmatrix}_{t_2\times t_1} \tag{5}$$

同样，隐藏层中还需要偏置项，其实偏置项只需要一个列向量就可以表示，但为了后面执行加法时方便，将偏执向量扩展为偏执矩阵

$$B_{12}=\begin{bmatrix} b_1^{(2)} & b_1^{(2)} & b_1^{(2)}\cdots & b_1^{(2)} \\ b_2^{(2)} & b_2^{(2)} & b_2^{(2)}\cdots & b_2^{(2)} \\ \cdots & & & \\ b_{t_2}^{(2)} & b_{t_2}^{(2)} & b_{t_2}^{(2)}\cdots & b_{t_2}^{(2)} \end{bmatrix}_{t_2\times k} \tag{5}$$

于是，根据上面的定义，我们可以给出隐藏层的输入Z2

$$Z_2=[z^{(1)},z^{(2)},z^{(3)},\cdots,z^{(k)}] \tag{6}$$

Z2的计算公式为,其中*表示矩阵乘法

$$Z_2=W_{12}*I_1+B_{12} \tag{7}$$

同理可以将Z2进行扩写为矩阵

$$Z_2=\begin{bmatrix} z_{1,1}^{(2)} & z_{1,2}^{(2)} & z_{1,3}^{(2)}\cdots & z_{1,k}^{(2)} \\ z_{2,1}^{(2)} & z_{2,2}^{(2)} & z_{2,3}^{(2)}\cdots & z_{2,k}^{(2)} \\ \cdots & & & \\ z_{t_2,1}^{(2)} & z_{t_2,2}^{(2)} & z_{t_2,3}^{(2)}\cdots & z_{t_2,k}^{(2)} \end{bmatrix}_{t_1\times k} \tag{8}$$

由于题目中未给出激活函数，于是我们选取最常用的激活函数sigmoid

$$f(x)=\frac{1}{1+e^{-x}} \tag{9}$$

由于sigmoid函数的良好性质，我们可以给出其导数的表达式

$$f'(x)=y(1-y) \tag{10}$$

接着，我们对矩阵Z2的每一个元素都运用sigmoid激活函数，得到隐藏层向输出层的输入

$$f(Z_2)=O_2=\begin{bmatrix} o_{1,1}^{(2)} & o_{1,2}^{(2)} & o_{1,3}^{(2)}\cdots & o_{1,k}^{(2)} \\ o_{2,1}^{(2)} & o_{2,2}^{(2)} & o_{2,3}^{(2)}\cdots & o_{2,k}^{(2)} \\ \cdots & & & \\ o_{t_2,1}^{(2)} & o_{t_2,2}^{(2)} & o_{t_2,3}^{(2)}\cdots & o_{t_2,k}^{(2)} \end{bmatrix}_{t_2\times k} \tag{11}$$

假设输出层的神经元个数为t3，仿照W12的定义可以给出W23的定义

$$W_{23}=\begin{bmatrix} w_{1,1}^{(3)} & w_{1,2}^{(3)} & w_{1,3}^{(3)}\cdots & w_{1,t_2}^{(3)} \\ w_{2,1}^{(3)} & w_{2,2}^{(3)} & w_{2,3}^{(3)}\cdots & w_{2,t_2}^{(3)} \\ \cdots & & & \\ w_{t_3,1}^{(3)} & x_{t_3,2}^{(3)} & x_{t_3,3}^{(3)}\cdots & x_{t_3,t_2}^{(3)} \end{bmatrix}_{t_3\times t_2} \tag{12}$$

同理定义隐藏层向输出层的偏置项：

$$B_{23}=\begin{bmatrix} b_1^{(3)} & b_1^{(3)} & b_1^{(3)}\cdots & b_1^{(3)} \\ b_2^{(3)} & b_2^{(3)} & b_2^{(3)}\cdots & b_2^{(3)} \\ \cdots & & & \\ b_{t_3}^{(3)} & b_{t_3}^{(3)} & b_{t_3}^{(3)}\cdots & b_{t_3}^{(3)} \end{bmatrix}_{t_3\times k} \tag{13}$$

同理可以给出输出层的输入矩阵Z3的计算公式

$$Z_3=W_{23}*O_2+B_{23} \tag{14}$$

同理给出Z3的矩阵表示

$$Z_3=\begin{bmatrix} z_{1,1}^{(3)} & z_{1,2}^{(3)} & z_{1,3}^{(3)}\cdots & z_{1,k}^{(3)} \\ z_{2,1}^{(3)} & z_{2,2}^{(3)} & z_{2,3}^{(3)}\cdots & z_{2,k}^{(3)} \\ \cdots & & & \\ z_{t_3,1}^{(3)} & z_{t_3,2}^{(3)} & z_{t_3,3}^{(3)}\cdots & z_{t_3,k}^{(3)} \end{bmatrix}_{t_3\times k} \tag{15}$$

同样对Z3矩阵应用sigmoid函数可以得到矩阵A3，也就是最终的输出矩阵

$$A_3=f(Z_3)\begin{bmatrix} a_{1,1}^{(3)} & a_{1,2}^{(3)} & a_{1,3}^{(3)}\cdots & a_{1,k}^{(3)} \\ a_{2,1}^{(3)} & a_{2,2}^{(3)} & a_{2,3}^{(3)}\cdots & a_{2,k}^{(3)} \\ \cdots & & & \\ a_{t_3,1}^{(3)} & a_{t_3,2}^{(3)} & a_{t_3,3}^{(3)}\cdots & a_{t_3,k}^{(3)} \end{bmatrix}_{t_3\times k} \tag{16}$$

至此为止，我们将所需要的参数、矩阵、输入全部梳理完成，并且完成了从输入层到输出层的前向传播的过程，接下来我们将对前向传播的逆过程--反向传播参数更新进行推导，需要更新的参数其实包括四个矩阵W12、W23、B12、B23

推导更新W23以及B23，由链式求导法则，我们可以得出：根据前面的定义，我们可以得知，第二项为对sigmoid函数求导，第三项为矩阵对矩阵求导

$$\frac{\partial E}{\partial W_{23}}=\frac{\partial E}{\partial A_3}\cdot\frac{\partial A_3}{\partial Z_3}\cdot\frac{\partial Z_3}{\partial W_{23}} \tag{17}$$

在第一项中需要完成损失函数对A3矩阵的求导:

当不考虑损失项时，我们可以得到:

$$E_1=\sum_i\sum_j\frac{1}{2}(y_{i,j}^M-d_{i,j}^M)^2 \tag{18}$$

将其带换成我们使用的符号推导即为：因为最后的结果d只在输出层有效，所以可以去掉dij上面代表层数的上标

$$E_1=\sum_i\sum_j\frac{1}{2}(a_{i,j}^3-d_{i,j})^2 \tag{19}$$

由于其实际的含义为在每一个Z的项中对A中的对应位置求导，所以我们可以得到

$$\frac{\partial E_1}{\partial A_3}=\begin{bmatrix} a_{1,1}^{(3)}-d_{1,1} & a_{1,2}^{(3)}-d_{1,2} & a_{1,3}^{(3)}d_{1,3}\cdots & a_{1,k}^{(3)}d_{1,k} \\ a_{2,1}^{(3)}-d_{2,1} & a_{2,2}^{(3)}-d_{2,2} & a_{2,3}^{(3)}-d_{2,3}\cdots & a_{2,k}^{(3)}-d_{2,k} \\ \cdots & & & \\ a_{t_3,1}^{(3)}-d_{t_3,1} & a_{t_3,2}^{(3)}-d_{t_3,2} & a_{t_3,3}^{(3)}-d_{t_3,3}\cdots & a_{t_3,k}^{(3)}-d_{t_3,k} \end{bmatrix}_{t_3\times k} \tag{20}$$

接下来我们考虑损失项Sw，由公式1可知，对单个输入(以第一个输入为例)的处理时我们可以将该矩阵转换为如下形式，并且假设该向量属于第c类且是在针对隐藏层-输出层进行更新

$$S_w = \begin{bmatrix} a_{11}^{(3)} - m_{c1}^{(3)} \\ a_{21}^{(3)} - m_{c2}^{(3)} \\ a_{31}^{(3)} - m_{c3}^{(3)} \\ \cdots \\ a_{t_3 1}^{(3)} - m_{c_{t_3}}^{(3)} \end{bmatrix}_{t_3 \times 1} \quad \begin{bmatrix} a_{11}^{(3)} - m_{c1}^{(3)} & a_{21}^{(3)} - m_{c2}^{(3)} & a_{31}^{(3)} - m_{c3}^{(3)} & \cdots & a_{t_3 1}^{(3)} - m_{c_{t_3}}^{(3)} \end{bmatrix}_{1 \times t_3} \tag{21}$$

在计算式，不必计算其完整的计算结果，因为我们只关心矩阵的迹的值，所以只需获取对角线上元素的和

$$\sum_{i=1}^{t_3} (a_{i1}^{(3)} - m_{ci}^{(3)})^2 \tag{22}$$

于是，我们完整考虑每一个输入的可能可以写出tr（Sw）的表达式为：

$$tr(S_w) = \sum_{j=1}^{k} \sum_{i=1}^{t_3} (a_{ij}^{(3)} - m_{c_j j}^{(3)})^2 \tag{23}$$

如果要对Sw的迹求偏导，还需要获得其平均值的表达形式，根据定义可以给出

$$m_{c_j j}^{(3)} = \frac{1}{n_c} \cdot \sum_{a_{ij} \in c} a_{i,j}^{(3)} \tag{24}$$

于是可以根据23以及24的表达式完成对E2的求导,由于这里涉及的表达式过多，不便于直接书写在矩阵中，我们先求E2对一个矩阵因子Yij的求导结果

$$\frac{\partial E_2}{\partial a_{ij}^{(3)}} = 2(a_{ij}^{(3)} - m_{c_j j}^{(3)}) - (1 - \frac{2}{n_c}) \tag{25}$$

然后，我们考虑因子项Sb，可以按照花间Sw的思路对Sb进行展开，在这里忽略矩阵相乘的推导步骤，直接给出tr(Sb)的结果,同样以隐藏层和输出层为例

$$tr(S_b) = \sum_{c=1}^{C} \sum_{i=1}^{t_3} n_c (m_{ci}^{(3)} - m_c^{(3)}) \tag{26}$$

利用和处理Sw一样的思路处理Sb，其中mci的值已经由上式给出，下面给出mc的值

$$m_c^{(3)} = \sum_{c=1}^{C} \frac{1}{n_c} \cdot \sum_{a_{ij} \in c} a_{i,j}^{(3)} \tag{27}$$

于是，我们可以联立式(24)以及式(27)得到如下推导：

$$\frac{\partial E_3}{\partial a_{i,j}^{(3)}} = n_c \cdot ((1 - \frac{2}{n_c}) + (\frac{C}{n_c})) \tag{28}$$

又由于：

$$E = E_1 + E_2 + E_3 \tag{29}$$

我们可以得出：针对每一个矩阵元素的偏导结果

$$\frac{\partial E}{\partial a_{i,j}} = \frac{\partial E_1}{\partial a_{i,j}} + \frac{\partial E_2}{\partial a_{i,j}} + \frac{\partial E_3}{\partial a_{i,j}} \tag{30}$$

代入式20、25、28的结果我们可以得到，我们将这个值起一个新的名字叫做α

$$\alpha_{i,j} = \frac{\partial E}{\partial a_{i,j}} = a_{i,j}^{(3)} - d_{i,j} + 2(a_{ij}^{(3)} - m_{c_j j}^{(3)}) - (1 - \frac{2}{n_c}) + n_c \cdot ((1 - \frac{2}{n_c}) + (\frac{C}{n_c})) \tag{31}$$

于是，最困难的一部分，也就是也就是整个链式求导的第一项就已经完成了，最后的矩阵表示为(注意区分这里减号后的位alpha符号，由式31给出，并不是矩阵A中的元素a)

$$F_{23}=\frac{\partial E}{\partial A_3}=\begin{bmatrix} a_{1,1}^{(3)}-\alpha_{1,1} & a_{1,2}^{(3)}-\alpha_{1,2} & a_{1,3}^{(3)}-\alpha_{1,3}\cdots & a_{1,k}^{(3)}-\alpha_{1,k} \\ a_{2,1}^{(3)}-\alpha_{2,1} & a_{2,2}^{(3)}-\alpha_{2,2} & a_{2,3}^{(3)}-\alpha_{2,3}\cdots & a_{2,k}^{(3)}-\alpha_{2,k} \\ \cdots & & & \\ a_{t_3,1}^{(3)}-\alpha_{t_3,1} & a_{t_3,2}^{(3)}-\alpha_{t_3,2} & a_{t_3,3}^{(3)}-\alpha_{t_3,3}\cdots & a_{t_3,k}^{(3)}-\alpha_{t_3,k} \end{bmatrix}_{t_3\times k} \tag{32}$$

然后我们求取A3关于Z3的偏导，这一部分的偏导依赖于对sigmoid函数的求导，于是我们可以很轻易的给出矩阵：

$$S_{23}=\frac{\partial A_3}{\partial Z_3}=\begin{bmatrix} a_{1,1}^{(3)}\cdot(1-a_{1,1}^{(3)}) & a_{1,2}^{(3)}\cdot(1-a_{1,2}^{(3)}) & a_{1,3}^{(3)}\cdot(1-a_{1,3}^{(3)})\cdots & a_{1,k}^{(3)}\cdot(1-a_{1,k}^{(3)}) \\ a_{2,1}^{(3)}\cdot(1-a_{2,1}^{(3)}) & a_{2,2}^{(3)}\cdot(1-a_{2,2}^{(3)}) & a_{2,3}^{(3)}\cdot(1-a_{2,3}^{(3)})\cdots & a_{2,k}^{(3)}\cdot(1-a_{2,k}^{(3)}) \\ \cdots & & & \\ a_{t_3,1}^{(3)}\cdot(1-a_{t_3,1}^{(3)}) & a_{t_3,2}^{(3)}\cdot(1-a_{t_3,2}^{(3)}) & a_{t_3,3}^{(3)}\cdot(1-a_{t_3,3}^{(3)})\cdots & a_{t_3,k}^{(3)}\cdot(1-a_{t_3,k}^{(3)}) \end{bmatrix}_{t_3\times k} \tag{33}$$

最后是Z3关于W3的偏导，我们可以根据式14得到：其中，E为单位阵，运算符代表克罗内克积

$$\frac{\partial Z_3}{\partial W_{23}}=E\bigotimes O_2^T \tag{34}$$

于是，我们联立式32、33以及34即可得到正确的链式求导值，即有：(注意，这里的乘法不是矩阵乘，而是矩阵的内积，即对应元素相乘)

$$\begin{gathered} W_{23}=W_{23}-\eta\cdot\frac{\partial E}{\partial W_{23}} \\ \text{其中，}\frac{\partial E}{\partial W_{23}}\text{的值为：}(F_{23}\cdot A_{23})(E\bigotimes O_2^T) \end{gathered} \tag{35}$$

对于W23的更新已经完成了，下面我们更新对B23的更新：仔细观察式14，其实该部分与W23的更新的区别仅仅在于式17的最后一项，于是我们可以利用之前的结果得到新的B23的值，为：

$$\begin{gathered} B_{23}=B_{23}-\eta\cdot\frac{\partial E}{\partial B_{23}} \\ \text{其中，}\frac{\partial E}{\partial B_{23}}\text{的值为：}(F_{23}\cdot A_{23})(E) \end{gathered} \tag{36}$$

同理，我们可以得到更新W12和B12的推导：

$$\frac{\partial E}{\partial W_{12}}=\frac{\partial E}{\partial A_3}\frac{\partial A_3}{\partial Z_3}\frac{\partial Z_3}{\partial O_2}\frac{\partial O_2}{\partial Z_2}\frac{\partial Z_2}{\partial W_{12}} \tag{37}$$

其中，前两项我们在上面的推导中已经涉及，由式32及33给出，新引入的偏导方程式中的第一个和第三个皆为矩阵对矩阵的求导，使用对矩阵求导的公式计算器克罗内克积可得到，新引入的第二项为对sigmoid函数求导

我们模仿从输出层到隐藏层的推导进行公式推导：

$$\begin{gathered} \text{由公式：}Z_3=W_{23}*O_2+B_{23}\text{可得} \\ \frac{\partial Z_3}{\partial O_2}=(E\bigotimes W_{23}^T) \end{gathered} \tag{38}$$

再推导sigmoid函数的求导：

$$S_{12}=\frac{\partial O_2}{\partial Z_2}=\begin{bmatrix} o_{1,1}^{(2)}\cdot(1-o_{1,1}^{(2)}) & o_{1,2}^{(2)}\cdot(1-o_{1,2}^{(2)}) & o_{1,3}^{(2)}\cdot(1-o_{1,3}^{(2)})\cdots & o_{1,k}^{(2)}\cdot(1-o_{1,k}^{(2)}) \\ o_{2,1}^{(2)}\cdot(1-o_{2,1}^{(2)}) & o_{2,2}^{(2)}\cdot(1-o_{2,2}^{(2)}) & o_{2,3}^{(2)}\cdot(1-o_{2,3}^{(2)})\cdots & o_{2,k}^{(2)}\cdot(1-o_{2,k}^{(2)}) \\ \cdots & & & \\ o_{t_2,1}^{(2)}\cdot(1-o_{t_2,1}^{(2)}) & o_{t_2,2}^{(2)}(1-o_{t_2,2}^{(2)}) & o_{t_2,3}^{(2)}(1-o_{t_3,1}^{(2)})\cdots & o_{t_2,k}^{(2)}(1-o_{t_2,k}^{(2)}) \end{bmatrix}_{t_2\times k} \tag{39}$$

最后一个元素我们也予以推导：

由条件我们可得：$Z_2 = W_{12} * I_1 + B_{12}$

$$\frac{\partial Z_2}{\partial W_{12}} = (E \bigotimes I_1^T) \tag{40}$$

最终完成联立，得到最终的结果：

$$\frac{\partial E}{\partial W_{12}} = F_{23} \cdot S_{23} \cdot (E \bigotimes W_{23}^T) \cdot S_{12} \cdot (E \bigotimes I_1^T) \tag{41}$$

同理可以推导对于B12的偏导

$$\frac{\partial E}{\partial B_{12}} = F_{23} \cdot S_{23} \cdot (E \bigotimes W_{23}^T) \cdot S_{12} \cdot (E) \tag{42}$$

至此，三层感知机的反向传播算法就已经推到完成