Exercise2-MLP_BP

郭坤昌 2012522 计算机科学与技术专业

要求

以三层感知机为例，使用反向传播算法更新 MLP 的权重和偏置项。

MPL及其权重、偏置项定义如下：

Define S_w and S_b as:

$$\begin{aligned} S_w &= \sum_{c=1}^{C} \sum_{\boldsymbol{y}_i^M \in c} (\boldsymbol{y}_i^M - \boldsymbol{m}_c^M)(\boldsymbol{y}_i^M - \boldsymbol{m}_c^M)^T \\ S_b &= \sum_{c=1}^{C} n_c (\boldsymbol{m}_c^M - \boldsymbol{m}^M)(\boldsymbol{m}_c^M - \boldsymbol{m}^M)^T \end{aligned} \tag{1}$$

where $\boldsymbol{m}_c^M$ is the mean vector of $\boldsymbol{y}_i^M$ (the output of the ith sample from the cth class), $\boldsymbol{m}^M$ is the mean vector of the output $\boldsymbol{y}_i^M$ from all classes, n_c is the number of samples from the cth class. Define the discriminative regularization term $tr(S_w) - tr(S_b)$ and incorporate it into the objective function of the MLP:

$$E = \sum_i \sum_j \frac{1}{2}(\boldsymbol{y}_{i,j}^M - \boldsymbol{d}_{i,j})^2 + \frac{1}{2}\gamma(tr(S_w) - tr(S_b)). \tag{2}$$

where $\boldsymbol{y}_{i,j}^M$ is the jth element in the vector $\boldsymbol{y}_i^M$, $\boldsymbol{d}_{i,j}$ is the jth element in the label vector $\boldsymbol{d}_i$, tr denotes the trace of the matrix. Use the BP algorithm to update parameters $\boldsymbol{W}$ and $\boldsymbol{b}$ of the MLP.

模型定义及约定

1. M: 层数，约定输入层为第0层
2. N_i: 第i层神经元数量，N_0为输入特征个数
3. f: 激活函数，在$\mathbb{R}$上连续可导
4. w_{ij}^l: 连接第l层第i个神经元与第$l-1$层第j个神经元的权值，其中$l = 1, 2, \dots, M, i = 0, 1, \dots, N_i - 1, j = 0, 1, \dots, N_{i-1}$
5. $w_{i,N_{l-1}}^l$: 第l层第i个神经元的阈值
6. b_i^l: 第l层第i个神经元的偏置项，其中$l = 1, 2, \dots, M, i = 0, 1, \dots N_l - 1$
7. net_i^l: 第l层第i个神经元输入值，有

$$net_i^l = \sum_{k=0}^{N_{l-1}} w_{ik}^l y_k^{l-1} + b_i^{l-1}, l = 1, 2, \dots, M, i = 0, 1, \dots N_l - 1$$

8. y_i^l: 第l层第i个神经元输出值，有

$$y_i^l = \begin{cases} f(net_i^l), & l=1,2,\ldots,M, i=0,1,\ldots N_l-1 \\ x_i, & l=0, i=0,1,\ldots,N_0-1 \\ 1, & l=0,1,\ldots,M-1, i=N_l \end{cases}$$

9. E: 目标函数（详细的分析在后续推导部分给出）

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

其中

$$\begin{cases} S_w = \sum_{c=1}^{C} \sum_{y_i^M \in c} (y_i^M - m_c^M)(y_i^M - m_c^M)^T \\ S_b = \sum_{c=1}^{C} (m_c^M - m^M)(m_c^M - m^M)^T \end{cases}$$

10. η: 学习率
11. Δw_{ij}^l: 权值变化量

$$\Delta w_{ij}^l = -\eta \frac{\partial E}{\partial w_{ij}^l}, l=1,2,\ldots,M, i=0,1,\ldots,N_i-1, j=0,1,\ldots,N_{i-1}$$

12. Δb_i^l: 偏置项变化量

$$\Delta b_i^l = -\eta \frac{\partial E}{\partial b_i^l}, l=1,2,\ldots,M, i=0,1,\ldots,N_i-1$$

推导

改写目标函数

1. 改写S_w和S_b

目标函数中增加了一个正则化项，它是由最终分类器输出结果的散度矩阵（类内散度矩阵S_w和类间散度矩阵S_b）求迹组成的。猜测这里的神经网络运用了Fisher线性判别的思想，达到在分类问题上提取特征的目的。

> *该部分使用另一种形式的散度矩阵，详细证明过程参照[Fisher线性判别散度矩阵Sb,Sw 另一种表达形式的证明]*

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

其中m_c^M是所有判别为c类的y_i^M的均值向量，以n_c表示判别为c类的y_i^M的均值向量个数，则

$$m_c^M = \frac{1}{n_c} \sum_{y_i^M \in c} y_i^M$$

S_w可以改写为

$$S_w = \frac{1}{2} \sum_{i,j} A_{i,j}^w (y_i^M - y_j^M)(y_i^M - y_j^M)^T$$

其中

$$A_{i,j}^w = \begin{cases} \frac{1}{n_c}, & y_i^M \in c, y_j^M \in c \\ 0, & y_i^M \in c_1, y_j^M \in c_2, c_1 \neq c_2 \end{cases}$$

对于类间散度

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

其中m^M是y_i^M在所有类上判别的均值向量，以这里所有最终分类器数目等于N_M，则

$$m^M = \frac{1}{N_M} \sum_{i=0}^{N_M - 1} y_i^M$$

S_b可以改写为

$$S_w = \frac{1}{2} \sum_{i,j} A_{i,j}^b (y_i^M - y_j^M)(y_i^M - y_j^M)^T$$

其中

$$A_{i,j}^b = \begin{cases} \frac{1}{N_M} - \frac{1}{n_c}, & y_i^M \in c, y_j^M \in c \\ \frac{1}{N_M}, & y_i^M \in c_1, y_j^M \in c_2, c_1 \neq c_2 \end{cases}$$

这样

$$\begin{aligned} tr(S_w) - tr(S_b) &= tr(S_w - S_b) \\ &= tr(\frac{1}{2} \sum_{i,j} A_{i,j}^w (y_i^M - y_j^M)(y_i^M - y_j^M)^T - \frac{1}{2} \sum_{i,j} A_{i,j}^b (y_i^M - y_j^M)(y_i^M - y_j^M)^T) \\ &= \frac{1}{2} \sum_{i,j} (A_{i,j}^w - A_{i,j}^b) tr((y_i^M - y_j^M)(y_i^M - y_j^M)^T)) \end{aligned}$$

2. 改写目标函数E

在目标函数中$\sum_i \sum_j \frac{1}{2}(y_{ij}^M - d_{ij})^2$可改写为$\sum_{k=0}^{N_M - 1} \frac{1}{2}(y_k^M - d_k)^2$，则最终目标函数改写为如下公式

$$E = \sum_{k=0}^{N_M - 1} \frac{1}{2}(y_k^M - d_k)^2 + \frac{1}{2}\gamma(tr(S_w) - tr(S_b))$$

3. 根据目标函数计算$\frac{\partial E}{\partial y_k^M}$

$$\begin{aligned} \frac{\partial E}{\partial y_k^M} &= y_k^M - d_k + \frac{1}{2}\gamma(\frac{1}{2} \sum_{j=0}^{N_M - 1} 4(A_{k,j}^w - A_{k,j}^b)(y_k^M - y_j^M)) \\ &= y_k^M - d_k + \gamma(\sum_{j=0}^{N_M - 1} (A_{k,j}^w - A_{k,j}^b)(y_k^M - y_j^M)) \end{aligned}$$

显然，散度矩阵求迹再求导的结果，与最终计算结果的分类有关。特别地，对y_k^M求偏导时，该项计算结果与其他结果的分类有关。

计算权值和偏置项变化量

1. 链式法则变换

$$
\begin{aligned}
\Delta w_{ij}^l &= -\eta \frac{\partial E}{\partial w_{ij}^l} \\
&= -\eta \frac{\partial E}{\partial net_i^l} \frac{\partial net_i^l}{\partial w_{ij}^l} \\
\Delta b_i^l &= -\eta \frac{\partial E}{\partial b_i^l} \\
&= -\eta \frac{\partial E}{\partial net_i^l} \frac{\partial net_i^l}{\partial b_i^l}
\end{aligned}
$$

令

$$
\delta_i^l = -\frac{\partial E}{\partial net_i^l}
$$

根据

$$
net_i^l = \sum_{k=0}^{N_{l-1}} w_{ik}^l y_k^{l-1} + b_i
$$

计算得

$$
\begin{aligned}
\frac{\partial net_i^l}{\partial w_{ij}^l} &= y_j^{l-1} \\
\frac{\partial net_i^l}{\partial b_i} &= 1
\end{aligned}
$$

则

$$
\begin{aligned}
\Delta w_{ij}^l &= \eta \delta_i^l y_j^{l-1} \\
\Delta b_i^l &= \eta \delta_i^l
\end{aligned}
$$

2. 求取δ_i^l

当$l = M$时，此时为输出层

$$
\begin{aligned}
\delta_i^l &= \delta_i^M \\
&= -\frac{\partial E}{\partial net_i^M} \\
&= -\frac{\partial E}{\partial y_i^M} \frac{\partial y_i^M}{\partial net_i^M}
\end{aligned}
$$

由

$$
y_i^l = f(net_i^l)
$$

得

$$\frac{\partial y_i^l}{\partial net_i^l} = f^{'}$$

又根据之前计算得到$\frac{\partial E}{\partial y_k^M}$，

$$\frac{\partial E}{\partial y_k^M} = y_k^M - d_k + \gamma(\sum_{j=0}^{N_M-1}(A_{k,j}^w - A_{k,j}^b)(y_k^M - y_j^M))$$

因此可以计算出δ_i^M（为了简洁直到最终结果，均不将$\frac{\partial E}{\partial y_k^M}$展开）

$$\begin{aligned}\delta_i^M &= -\frac{\partial E}{\partial y_i^M}\frac{\partial y_i^M}{\partial net_i^M} \\ &= -\frac{\partial E}{\partial y_i^M} f^{'}\end{aligned}$$

当$l < M$时，此时为隐含层

$$\begin{aligned}\delta_i^l &= -\frac{\partial E}{\partial net_i^l} \\ &= -\sum_{k=0}^{N_{l+1}-1}\frac{\partial E}{\partial net_k^{l+1}}\frac{\partial net_k^{l+1}}{\partial net_i^l} \\ &= -\sum_{k=0}^{N_{l+1}-1}\delta_k^{l+1}\frac{\partial net_k^{l+1}}{\partial y_i^l}\frac{\partial y_i^l}{\partial net_i^l}\end{aligned}$$

由

$$net_k^{l+1} = \sum_{i=0}^{N_l} w_{kj}^{l+1} y_i^l + b_k^{l+1}$$

得

$$\frac{\partial net_k^{l+1}}{\partial y_i^l} = w_{ki}^{l+1}$$

又

$$\frac{\partial y_i^l}{\partial net_i^l} = f^{'}$$

因此

$$\begin{aligned}\delta_i^l &= -\sum_{k=0}^{N_{l+1}-1}\delta_k^{l+1}\frac{\partial net_k^{l+1}}{\partial y_i^l}\frac{\partial y_i^l}{\partial net_i^l} \\ &= -f^{'}\sum_{k=0}^{N_{l+1}-1}\delta_k^{l+1} w_{ki}^{l+1}\end{aligned}$$

综上

$$\delta_i^l = \begin{cases} -\frac{\partial E}{\partial y_i^M} f^{'}, & l = M \\ -f^{'} \sum_{k=0}^{N_{l+1}-1} \delta_k^{l+1} w_{ki}^{l+1}, & l < M \end{cases}$$

3. 权值和偏置量更新

本题为3层感知机，即$M = 3$则

(这里为了与幂区分，将表示层数的上标括起)

$$\begin{cases} \delta_i^{(3)} = -f^{'} \dfrac{\partial E}{\partial y_i^M} \\ \\ \delta_i^{(2)} = -f^{'} \displaystyle\sum_{k=0}^{N_3-1} \delta_k^{(3)} w_{ki}^{(3)} \\ \qquad = -f^{'} \displaystyle\sum_{k=0}^{N_3-1} (-\dfrac{\partial E}{\partial y_k^M} f^{'}) w_{ki}^{(3)} \\ \qquad = (-f^{'})^2 \displaystyle\sum_{k=0}^{N_3-1} \dfrac{\partial E}{\partial y_k^M} w_{ki}^{(3)} \\ \\ \delta_i^{(1)} = -f^{'} \displaystyle\sum_{k=0}^{N_2-1} \delta_k^{(2)} w_{ki}^{(2)} \\ \qquad = -f^{'} \displaystyle\sum_{k=0}^{N_2-1} ((f^{'})^2 \sum_{t=0}^{N_3-1} \dfrac{\partial E}{\partial y_t^M} w_{ti}^{(3)}) w_{ki}^{(2)} \\ \qquad = (-f^{'})^3 \displaystyle\sum_{k=0}^{N_2-1} (\sum_{t=0}^{N_3-1} \dfrac{\partial E}{\partial y_t^M} w_{ti}^{(3)}) w_{ki}^{(2)} \end{cases}$$

由之前计算得

$$\Delta w_{ij}^l = \eta \delta_i^l y_j^{l-1}$$
$$\Delta b_i^l = \eta \delta_i^l$$

更新权值

$$
\begin{cases}
\begin{aligned}
\Delta w_{ij}^{(3)} &= \eta\delta_i^{(3)} y_j^{(2)} \\
&= \eta(-f^{'})y_j^{(2)} \frac{\partial E}{\partial y_i^M}, i=0,1,\ldots,N_3-1, j=0,1,\ldots,N_2 \\
\\
\Delta w_{ij}^{(2)} &= \eta\delta_i^{(2)} y_j^{(1)} \\
&= \eta y_j^{(1)}(-f^{'})^2 \sum_{k=0}^{N_3-1} \frac{\partial E}{\partial y_k^M} w_{ki}^{(3)}, i=0,1,\ldots,N_2-1, j=0,1,\ldots,N_1 \\
\\
\Delta w_{ij}^{(1)} &= \eta\delta_i^{(1)} y_j^{(0)} \\
&= \eta x_j(-f^{'})^3 \sum_{k=0}^{N_2-1}(\sum_{t=0}^{N_3-1} \frac{\partial E}{\partial y_t^M} w_{ti}^{(3)}) w_{ki}^{(2)}, i=0,1,\ldots,N_1-1, j=0,1,\ldots,N_0 \\
\\
\Delta b_i^{(3)} &= \eta\delta_i^{(3)} \\
&= \eta(-f^{'})\frac{\partial E}{\partial y_i^M}, i=0,1,\ldots,N_3-1 \\
\\
\Delta b_i^{(2)} &= \eta\delta_i^{(2)} \\
&= \eta(-f^{'})^2 \sum_{k=0}^{N_3-1} \frac{\partial E}{\partial y_k^M} w_{ki}^{(3)}, i=0,1,\ldots,N_2-1 \\
\\
\Delta b_i^{(1)} &= \eta\delta_i^{(1)} \\
&= \eta(-f^{'})^3 \sum_{k=0}^{N_2-1}(\sum_{t=0}^{N_3-1} \frac{\partial E}{\partial y_t^M} w_{ti}^{(3)}) w_{ki}^{(2)}, i=0,1,\ldots,N_1-1
\end{aligned}
\end{cases}
$$

其中

$$
\frac{\partial E}{\partial y_k^M} = y_k^M - d_k + \gamma(\sum_{j=0}^{N_M-1}(A_{k,j}^w - A_{k,j}^b)(y_k^M - y_j^M))
$$

且有

$$
A_{i,j}^w = \begin{cases} \frac{1}{n_c}, & y_i^M \in c, y_j^M \in c \\ 0, & y_i^M \in c_1, y_j^M \in c_2, c_1 \neq c_2 \end{cases}
$$
$$
A_{i,j}^b = \begin{cases} \frac{1}{N_M} - \frac{1}{n_c}, & y_i^M \in c, y_j^M \in c \\ \frac{1}{N_M}, & y_i^M \in c_1, y_j^M \in c_2, c_1 \neq c_2 \end{cases}
$$

其中n_c表示判别为c类的y_i^M的均值向量个数

参考文献

[Fisher线性判别散度矩阵Sb,Sw 另一种表达形式的证明]