机器学习经典算法推导

黎雷蕾

2018 年 8 月 1 日

目录

1 BP 算法 **2**

1.1 Δw 更新公式推导 3

1.2 $\Delta\theta$ 更新公式推导 5

Chapter 1

BP 算法

假设样本输入为 x_i，对应的标签为 y_i。

假设第 h 个隐藏层输入为：$\alpha_h = \sum_i v_{ih} x_i$，输出为：$b_h$。

假设第 j 个输出层输入为：$\beta_j = \sum_j w_{hj} b_h$，输出为：$\hat{y}_j = f(\beta_j - \theta_j)$，其中 $f(x)$ 为激活函数，这里取 sigmoid，θ 为阈值。

综上，我们可以建立从输入到输出的的联系:

$$\begin{cases} \alpha_h = \sum_i v_{ih} x_i \\ b_h = f_1(\alpha_h - \gamma_h) \\ \beta_j = \sum_j w_{hj} b_h \\ \hat{y}_j = f(\beta_j - \theta_j) \end{cases} \tag{1.1}$$

对于某个样本 k，计算均方误差，为了方便，增加一个系数 $\frac{1}{2}$:

$$E_k = \frac{1}{2} \sum_j (\hat{y}_j^k - y_j^k)^2 \tag{1.2}$$

我们进行随机梯度下降，满足:

$$g(x + \Delta x) < g(x) \tag{1.3}$$

根据泰勒展开式展开到一阶:

$$\begin{aligned} f(x + \Delta x) &\approx f(x) + \Delta x f'(x) + \frac{1}{2!} \Delta x f^{(2)}(x) + \cdots + \frac{1}{n!} \Delta x f^{(n)}(x) \\ &\approx f(x) + \Delta x f'(x) \end{aligned} \tag{1.4}$$

那么公式 1.3可以转化为:

$$\begin{aligned} g(x+\Delta x) < g(x) \Rightarrow g(x) + \Delta x g'(x) < g(x) \\ \Rightarrow \Delta x g'(x) < 0 \end{aligned} \tag{1.5}$$

我们只需让 $\Delta x g'(x)$ 趋近于一个接近零的极小负数即可，引入学习速率 η，由梯度下降的公式及偏导数的定义:

$$v \leftarrow v + \Delta v \tag{1.6}$$

$$\Delta x = -\eta \frac{\partial L}{\partial x} \tag{1.7}$$

其中，L 就是我们定义的损失函数，x 就是需要优化的参数了。

1.1 Δw 更新公式推导

对于某个样本 k，由公式 1.7我们可以推出 Δw 的优化公式:

$$\Delta w_{hj} = -\eta \frac{\partial E_k}{\partial w_{hj}} \tag{1.8}$$

由公式 1.1我们可以知道:w 先影响 β 再影响 $\hat{y}$ 最后影响 E，这就构成了一个误差逆向传播链，那么由链式法则可以知道:

$$\frac{\partial E_k}{\partial w_{hj}} = \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} \cdot \frac{\partial \beta_j}{\partial w_{hj}} \tag{1.9}$$

因为 $\beta_j = \sum_j w_{hj} b_h$，那么:

$$\begin{aligned} \frac{\partial \beta_j}{\partial w_{hj}} &= \frac{\partial \left(\sum_j w_{hj} b_h \right)}{\partial w_{hj}} \\ &= \frac{\partial \left(w_{h1} b_h + w_{h2} b_h + \cdots + w_{hj} b_h \right)}{\partial w_{hj}} \\ &= b_h \end{aligned} \tag{1.10}$$

接下来推导 $\frac{\partial E_k}{\partial \hat{y}_j^k}$:

$$\begin{aligned}\frac{\partial E_k}{\partial \hat{y}_j^k} &= \frac{\left(\frac{1}{2}\sum_j(\hat{y}_j^k - y_j^k)^2\right)}{\partial \hat{y}_j^k} \\ &= \frac{\partial\left(\frac{1}{2}\left((\hat{y}_1^k - y_1^k)^2 + (\hat{y}_2^k - y_2^k)^2 + \cdots + (\hat{y}_j^k - y_j^k)^2\right)\right)}{\partial \hat{y}_j^k} \\ &= \frac{\partial\left(\frac{1}{2}(\hat{y}_j^k - y_j^k)^2\right)}{\partial \hat{y}_j^k} \\ &= \hat{y}_j^k - y_j^k\end{aligned} \tag{1.11}$$

往下继续推导 $\frac{\partial \hat{y}_j^k}{\partial \beta_j}$，这里需要借用一个 sigmoid 函数的特殊性质:

$$f'(x) = f(x)(1 - f(x)) \tag{1.12}$$

借用 sigmoid 函数的性质，我们进行如下推导:

$$\begin{aligned}\frac{\partial \hat{y}_j^k}{\partial \beta_j} &= \frac{\partial f(\beta_j^k - \theta_j)}{\partial \beta_j} \\ &= f'(\beta_j^k - \theta_j) \\ &= f(\beta_j^k - \theta_j)(1 - f(\beta_j^k - \theta_j)) \\ &= \hat{y}_j^k(1 - \hat{y}_j^k)\end{aligned} \tag{1.13}$$

我们简化偏导数的表达公式:

$$\begin{aligned}g_j &= \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} \\ &= (\hat{y}_j^k - y_j^k)\hat{y}_j^k(1 - \hat{y}_j^k)\end{aligned} \tag{1.14}$$

至此，综合公式 1.9、1.11、1.13、1.10、1.14我们可以求出 Δw_{hj} 的更新公式了。

$$\begin{aligned}\Delta w_{hj} &= -\eta g_j b_h \\ &= -\eta(\hat{y}_j^k - y_j^k)\hat{y}_j^k(1 - \hat{y}_j^k)b_h\end{aligned} \tag{1.15}$$

这说明 w 的更新完全可以由输出 $\hat{y}$，label y 和上一层的输出 b_h 进行更新了。

1.2 $\Delta\theta$ 更新公式推导

同理，我们可以由公式 1.7来确定 θ_j 的更新公式:

$$
\begin{aligned}
\Delta\theta_j &= -\eta\frac{\partial E_k}{\partial \theta_j} \\
&= -\eta\frac{\partial\left(\frac{1}{2}\sum_j(\hat{y}_j^k - y_j^k)^2\right)}{\partial\theta_j} \\
&= -\eta\frac{\partial\left(\frac{1}{2}\sum_j(f(\beta_j^k - \theta_j) - y_j^k)^2\right)}{\partial\theta_j} \\
&= -\eta\left((f(\beta_j^k - \theta_j) - y_j^k)\cdot f'(\beta_j^k - \theta_j)\right) \\
&= -\eta\left((\hat{y}_j^k - y_j^k)\cdot f(\beta_j^k - \theta_j)\cdot(1 - f(\beta_j^k - \theta_j))\right) \\
&= -\eta\left((\hat{y}_j^k - y_j^k)\cdot \hat{y}_j^k\cdot(1 - \hat{y}_j^k)\right) \\
&= -\eta g_j
\end{aligned} \tag{1.16}
$$