模式识别与机器学习 30250293-1 (2023 春)

# 第 14 章 (深度学习 2) 作业

## 1 越时反向传播 (BPTT, Back Propagation Through Time)

考虑一个用最后时间步的输出做分类 logit 的文本二分类问题, 网络前向的过程为

$$\begin{aligned}
\boldsymbol{h}_0 &= \boldsymbol{0} \in \mathbb{R}^n \\
\boldsymbol{x}_t &\in \mathbb{R}^n,\ t = 1, 2, \ldots, l \\
\boldsymbol{h}_t &= \mathrm{ReLU}(\boldsymbol{b} + \boldsymbol{W}\boldsymbol{h}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t) \in \mathbb{R}^n,\ t = 1, 2, \ldots, l \\
y_{\text{pred}} &= \mathrm{Sigmoid}(\boldsymbol{V}\boldsymbol{h}_l + d) \in \mathbb{R},
\end{aligned}$$

使用交叉熵作为分类损失。考虑 batchsize=1 的 SGD 进行网络训练，且样本的标签为 $y_{\text{gt}}$ 时，损失函数具体为

$$\mathcal{L}(y_{\text{pred}}, y_{\text{gt}}) = -(y_{\text{gt}} \log(y_{\text{pred}}) + (1 - y_{\text{gt}}) \log(1 - y_{\text{pred}}))$$

求

$$\frac{\partial \mathcal{L}}{\partial y_{\text{pred}}}, \frac{\partial \mathcal{L}}{\partial d}, \frac{\partial \mathcal{L}}{\partial \boldsymbol{V}}, \frac{\partial \mathcal{L}}{\partial \boldsymbol{h}_t}, \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}}$$

解：

$$\frac{\partial \mathcal{L}}{\partial y_{\text{pred}}} = -\left(\frac{y_{\text{gt}}}{y_{\text{pred}}} - \frac{1 - y_{\text{gt}}}{1 - y_{\text{pred}}}\right) = \frac{y_{\text{pred}} - y_{\text{gt}}}{y_{\text{pred}}(1 - y_{\text{pred}})}$$

$$\frac{\partial \mathcal{L}}{\partial d} = \frac{\partial \mathcal{L}}{\partial y_{\text{pred}}} \frac{\partial y_{\text{pred}}}{\partial d} = \frac{\partial \mathcal{L}}{\partial y_{\text{pred}}} y_{\text{pred}}(1 - y_{\text{pred}})1 = y_{\text{pred}} - y_{\text{gt}}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{V}} = \frac{\partial \mathcal{L}}{\partial y_{\text{pred}}} \frac{\partial y_{\text{pred}}}{\partial \boldsymbol{V}} = \frac{\partial \mathcal{L}}{\partial y_{\text{pred}}} y_{\text{pred}}(1 - y_{\text{pred}})\boldsymbol{h}_l^\top = (y_{\text{pred}} - y_{\text{gt}})\boldsymbol{h}_l^\top$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{h}_t} &= \frac{\partial \mathcal{L}}{\partial y_{\text{pred}}} \frac{y_{\text{pred}}}{\partial \boldsymbol{h}_l} \frac{\partial \boldsymbol{h}_l}{\partial \boldsymbol{h}_{l-1}} \cdots \frac{\partial \boldsymbol{h}_{t+1}}{\partial \boldsymbol{h}_t} \\
&= \left(\frac{y_{\text{pred}} - y_{\text{gt}}}{y_{\text{pred}}(1 - y_{\text{pred}})}\right) (y_{\text{pred}}(1 - y_{\text{pred}})\boldsymbol{V}) (\mathrm{diag}(\mathbb{I}(\boldsymbol{h}_l > 0))\boldsymbol{W}) \ldots (\mathrm{diag}(\mathbb{I}(\boldsymbol{h}_{t+1} > 0))\boldsymbol{W}) \\
&= (y_{\text{pred}} - y_{\text{gt}})\boldsymbol{V} \prod_{i=t+1}^{l} (\mathrm{diag}(\mathbb{I}(\boldsymbol{h}_i > 0))\boldsymbol{W})
\end{aligned}$$

当 $t = l$ 时退化为 $\frac{\partial \mathcal{L}}{\partial \boldsymbol{h}_l} = (y_{\text{pred}} - y_{\text{gt}})\boldsymbol{V}$

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}} &= \sum_{t=1}^{l} \frac{\partial \mathcal{L}}{\partial \boldsymbol{h}_t} \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{W}} \\
&= \sum_{t=1}^{l} \frac{\partial \mathcal{L}}{\partial \boldsymbol{h}_t} \operatorname{diag}(\mathbb{I}(\boldsymbol{h}_t > 0)) \left(\boldsymbol{I}_n \otimes \boldsymbol{h}_{t-1}^\top\right) \\
&= \sum_{t=1}^{l-1} \left[ (y_{\text{pred}} - y_{\text{gt}}) \boldsymbol{V} \left( \prod_{i=t+1}^{l} \operatorname{diag}(\mathbb{I}(\boldsymbol{h}_i > 0)) \boldsymbol{W} \right) \operatorname{diag}(\mathbb{I}(\boldsymbol{h}_t > 0)) \left(\boldsymbol{I}_n \otimes \boldsymbol{h}_{t-1}^\top\right) \right] \\
&+ (y_{\text{pred}} - y_{\text{gt}}) \boldsymbol{V} \operatorname{diag}(\mathbb{I}(\boldsymbol{h}_l > 0)) \left(\boldsymbol{I}_n \otimes \boldsymbol{h}_{l-1}^\top\right)
\end{aligned}
$$