模式识别与机器学习 **30250293-1 (2023 春)**

第 13 章 (深度学习 1) 作业 - 参考答案

1 模型参数计算和反向传播

1.1 多层感知机做回归时的反向传播

我们用一个两层的多层感知机做 n 维特征到一维输出的回归，记 $\hat{y} = f(\boldsymbol{x}; \boldsymbol{W}, \boldsymbol{b}, \boldsymbol{U}, d)$，具体的前向计算过程为

$$\begin{aligned}
\boldsymbol{x} &\in \mathbb{R}^n \\
\boldsymbol{h} &= \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b} \in \mathbb{R}^m \\
\boldsymbol{a} &= \mathrm{ReLU}(\boldsymbol{h}) \in \mathbb{R}^m \\
\hat{y} &= \boldsymbol{U}\boldsymbol{a} + d \in \mathbb{R}
\end{aligned}$$

使用的损失函数为 MSE，即 $\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$。

对于给定的一个训练样本组 $(\boldsymbol{x}, y)$ 计算损失函数后，求

$$\frac{\partial \mathcal{L}}{\partial d}, \frac{\partial \mathcal{L}}{\partial \boldsymbol{U}}, \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}}, \frac{\partial \mathcal{L}}{\partial \boldsymbol{b}}$$

解：

以下使用将矩阵按行主序向量化矩阵的方式定义矩阵到向量空间的映射及其微分

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial d} &= \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial d} = -2(y - \hat{y})1 = 2(\hat{y} - y) \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{U}} &= \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial U} = 2(\hat{y} - y)(I_1 \otimes \boldsymbol{a}^\top) = 2(\hat{y} - y)\boldsymbol{a}^\top \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}} &= \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \boldsymbol{a}} \frac{\partial \boldsymbol{a}}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{W}} = 2(\hat{y} - y)\boldsymbol{U}\mathbb{I}(\boldsymbol{h} \geq 0)\boldsymbol{I}_m \otimes \boldsymbol{x}^\top \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}} &= \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \boldsymbol{a}} \frac{\partial \boldsymbol{a}}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{b}} = 2(\hat{y} - y)\boldsymbol{U}\mathbb{I}(\boldsymbol{h} \geq 0)I = 2(\hat{y} - y)\boldsymbol{U}\mathbb{I}(\boldsymbol{h} \geq 0)
\end{aligned}$$

1.2 卷积层的输出和参数的雅可比矩阵

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{w}} = \begin{bmatrix}
x_{1,1} & x_{1,2} & x_{2,1} & x_{2,2} \\
x_{1,2} & x_{1,3} & x_{2,2} & x_{2,3} \\
x_{2,1} & x_{2,2} & x_{3,1} & x_{3,2} \\
x_{2,2} & x_{2,3} & x_{3,2} & x_{3,3}
\end{bmatrix}$$

1.3 卷积神经网络参数计算

小明尝试自己搭建神经网络对高宽为 224×224 的三通道彩色图像进行分类，所需区分的类别是“猫、狗、鸡、鸭”四类。下表是网络结构。请填写空白部分:(卷积层的 stride 默认为 1，padding 默认为 0, Flatten 前的输出维度按照 [通道数, 高, 宽] 格式填写)

网络模块	输出维度	可训练参数个数
Identity	[3,224,224]	0
Conv2d(in_channels=3, out_channels=32, kernel_size=(3,3))	[32,222,222]	32*(3*3*3+1)=896
BatchNorm2d(num_features=32)	[32,222,222]	32*2=64
ReLU()	[32,222,222]	0
MaxPool2d(kernel_size=(2,2), stride=2)	[32,111,111]	0
Conv2d(in_channels=32, out_channels=1, kernel_size=(1,1))	[1,111,111]	1*(1*1*32+1)=33
AdaptiveAvgPool2d(output_size=(3,3))	[1,3,3]	0
Conv2d(in_channels=1, out_channels=1, kernel_size=(2,2))	[1,2,2]	1*(2*2+1)=5
Flatten	[4,]	0
Softmax	[4,]	0

2 Cifar-10 Mini 分类

Cifar-10 是一个有 10 类标签的迷你图像数据集 (https://www.cs.toronto.edu/~kriz/cifar.html)。本题中使用助教预先划分的此数据集的 3 类别子集进行作业训练，以保证本题只使用常规笔记本的算力就能较好地完成，完整地进行读取数据流程、训练网络、评价网络的流程训练。你需要做三个网络的训练，所以尽可能保证代码复用性以减少自己工作量

2.1 损失函数估计

本题为分类任务，均使用交叉熵作损失 $\mathcal{L}(\boldsymbol{y}_{\text{gt}}, \boldsymbol{y}_{\text{pred}}) = -\boldsymbol{y}_{\text{gt}}^{\top} \log(\boldsymbol{y}_{\text{pred}})$。不妨设网络在没有训练的时，对每个类别的预测概率都是相等的，且数据集内各类别数据是均衡的。在这样的假设下，你认为本题的 5 分类任务中，第一个 batch 的 loss 均值应该是多少？

$$\text{loss} = \sum_{i=1}^{5} -1/5 * \boldsymbol{e}_i^{\top} \log \frac{1}{5}\mathbf{1} = \log 5$$

作业批改时任意底数都算对了，但一般而言，我们与编程实践中保持一致即 $\log(\cdot) = \ln(\cdot)$ 以 e 为底，以上结果为 1.609，可以在后续编程中的第一次训练迭代观察到相应现象。