Homework 3

软 31 万璐瑶 2023012127

2025 年 5 月 15 日

1 简答题

1. 什么是交叉熵（Cross Entropy）？在学习一个类别分布（Categorical Distribution）时，使用交叉熵作为损失函数比绝对值损失函数（Absolute Error，$L_{abs} = |y_i - \hat{y}_i|$）有什么好处？

   交叉熵是对两个概率分布之间差异的度量，特别是在分类问题中。使用交叉熵作为损失函数可以更好地捕捉模型输出的概率分布与真实分布之间的差异，从而提高模型的性能。

2. 多层感知机（Multilayer Perceptron）相比线性模型有哪些优势？相较于训练浅而宽的神经网络（“宽度学习”），训练相对窄而深的神经网络有什么好处？

   通过引入隐藏层和非线性激活函数（如 ReLU、Sigmoid），MLP 能够学习输入与输出之间的非线性映射。根据通用近似定理，单隐藏层的 MLP（宽度足够大）可以逼近任意连续函数，而线性模型仅能表达线性超平面。同时线性模型依赖人工特征工程，而 MLP 可以自动学习特征表示，减少了人工设计特征的需求。浅而宽的网络依靠单层或少数层中堆叠大量神经元，直接学习输入到输出的复杂映射，但缺乏逐级抽象能力。训练相对窄而深的神经网络有助于捕捉数据中的层次特征和抽象表示，深度网络可以通过多层非线性变换来学习复杂的特征表示，从而提高模型的表达能力和泛化性能。同时，浅层网络的梯度分散，宽网络中大量神经元共享梯度信号，可能导致优化方向模糊；深层网络的梯度聚焦，可以通过残差连接等技术缓解梯度消失问题，促进更深层网络的训练。

3. 卷积（Convolution）和互相关（Cross-correlation）分别是什么意思？在卷积神经网络中，卷积核通常进行的是卷积还是互相关操作？

   卷积对应的函数是：

   $$(f * g)(x) = \int_{-\infty}^{\infty} f(t)g(x-t)dt$$

   互相关对应的函数是：

   $$(f \star g)(x) = \int_{-\infty}^{\infty} f(t)g(x+t)dt$$

   卷积是对模板 g 先进行翻转后再滑动，而互相关直接是对模板 g 进行滑动，无需翻转。

   在 CNN 中，卷积核通常进行的是互相关操作。

4. 批量大小（Batch Size）对于优化器 (比如随机梯度下降) 影响巨大。为了减小现存占用，小宣提出将每次前向传播的批量大小减半，梯度累积两次再进行反向传播。请问这种方法能确保训练得到的模型效果参数一致吗（假设随机状态、batch 划分、dropout 的神经元相同）？若有影响，请指出原因（例如优化器、模型中的某些层）; 若无影响，请论证。

5. 为什么说残差连接（Residual Connection）有利于训练更深层的深度网络？残差链接能够缓解梯度消失（Gradient Vanishing）的问题吗

   因为残差连接通过令 $H(x) = F(x) + x$, 使得网络可以学习到 $F(x)$，而不是直接学习 $H(x)$，从而使得网络更容易优化。

   残差连接通过输出与中间层的参数的直连可以缓解梯度消失的问题，它允许梯度直接通过跳跃连接传播到更早的层，避免了梯度不稳定，从而减小了梯度消失的风险，使学习效果更稳定。