

## 3.6 softmax回归的从零开始重点摘录与练习解答

### 问题解答

1、本节直接实现了基于数学定义 softmax 运算的 softmax 函数。这可能会导致什么问题？提示：

尝试计算  $e^{50}$  的大小

解：当输入向量中的元素较大时，exp 计算可能会出现数值溢出 (无穷大，NaN) 的问题；

而当输入向量中的元素较小时，exp 计算可能会出现数值下溢 (分母为0) 的问题。

2、本节中的函数 cross\_entropy 是根据交叉熵损失函数的定义实现的。它可能有什么问题？提示：

考虑对数的定义域。

解：当预测标签向量中的元素接近0时，log 计算可能会出现数值溢出 (无穷大，NaN) 的问题。

3、请想一个解决方案来解决上述两个问题。

解：对于第一问题，处理方法如下：

采用减去最大值操作使得 exp 计算的最大输入为 0，排除了数值溢出的可能性，即令

$$\mathbf{Z} = \mathbf{O} - \max(o_i)$$

则

$$\text{softmax}(\mathbf{Z}) = \frac{\exp(z_j)}{\sum_{k=1}^q \exp(z_k)}$$

同时使得分母  $\geq 1$  ( $\exp(0) = 1$ )，排除了分母数值下溢导致被零除的可能。

对于第二个问题的处理方法如下：

采用减去最大值操作使得交叉熵损失计算的最大输入为  $(\min(o_j) - \max(o_i))$ ，不会超出数据类型容许的最大数字，排除了数值溢出的可能性。

$$l(y^{(i)}, \hat{y}^{(i)}) = -\log \hat{y}_j = -\log \frac{\exp(z_j)}{\sum_{k=1}^q \exp(z_k)} = \log \sum_{k=1}^q \exp(z_k) - z_j$$

4、返回概率最大的分类标签总是最优解吗？例如，医疗诊断场景下可以这样做吗？

解：返回概率最大的分类标签并不总是最优解。

概率最大的分类标签只是基于模型的预测结果，可能存在误差。模型的预测结果可能受到多种因素的影响，如数据质量、特征选择、模型选择等。因此，仅仅依靠概率最大的分类标签可能会导致错误的判断。

在医疗诊断等领域，决策可能需要综合考虑多个因素，如患者的病史、症状、实验室检查结果等。仅仅依靠概率最大的分类标签可能无法充分考虑这些因素，导致不准确的诊断结果。概率最大的分类标签可以作为参考，但不能作为唯一的依据。

5、假设我们使用 **softmax** 回归来预测下一个单词，可选取的单词数目过多可能会带来哪些问题？

解：1) 计算复杂度增加：**softmax** 回归的计算复杂度与可选取的单词数目成正比。当可选取的单词数目过多时，计算 **softmax** 函数的指数项和可能会变得非常耗时，导致模型训练和推理的效率下降。

2) 参数空间增大：**softmax** 回归的参数矩阵的大小与可选取的单词数目成正比。当可选取的单词数目过多时，模型需要学习更多的参数，导致模型的参数空间变得非常庞大，增加了模型训练的难度和模型的存储需求。

3) 数据稀疏性增加：当可选取的单词数目过多时，训练数据中每个单词的出现频率可能会变得非常低，导致数据的稀疏性增加。这会使得模型难以准确地估计每个单词的概率分布，可能导致模型的预测性能下降。

4) 模型泛化能力下降：当可选取的单词数目过多时，模型可能会过于依赖训练数据中出现频率较高的单词，而忽略了其他单词的特征。这会导致模型的泛化能力下降，对于训练数据中未出现或出现频率较低的单词的预测效果较差。

因此，在使用 **softmax** 回归进行下一个单词的预测时，需要权衡可选取的单词数目，选择适当的单词数量以平衡计算复杂度、参数空间、数据稀疏性和模型泛化能力。