交叉熵损失函数与 Adam 优化器的原理

作者姓名

2025 年 3 月 10 日

摘要

本文介绍了机器学习中常用的两种概念：交叉熵损失函数与 Adam 优化器。我们将详细讨论它们的数学原理，并通过图表来辅助理解，例如通过绘制 $-\log(x)$ 函数的曲线来直观展示交叉熵中 $-\log$ 部分的特性。此外，我们还会可视化 Adam 优化器的梯度更新路径，以帮助理解其自适应性特点。

1 引言

在机器学习和深度学习中，损失函数与优化器是模型训练的重要组成部分。交叉熵损失函数常用于分类任务，其目标是衡量预测概率分布与真实分布之间的差异；而 Adam 优化器则是一种基于梯度的一阶和二阶矩估计自适应调整学习率的方法。本文旨在系统地阐述这两个概念的原理，并借助图表进行说明。

2 交叉熵损失函数

交叉熵损失函数用于衡量模型输出的概率分布与真实标签分布之间的差异。在二分类问题中，其定义为：

$$L = -\left[y \log \hat{y} + (1-y) \log(1-\hat{y})\right], \tag{1}$$

其中 $y \in \{0, 1\}$ 表示真实标签，$\hat{y}$ 为模型预测的概率。

对于多分类问题，假设真实标签经过独热编码，交叉熵损失函数为：

$$L = -\sum_{i=1}^{C} y_i \log \hat{y}_i, \tag{2}$$

其中 C 为类别数，只有真实类别对应的 y_i 为 1，其余均为 0。

2.1 图示 1：$-\log(x)$-log(x)函数

下面的图表展示了函数 $f(x) = -\log(x)$ 在 $x \in (0, 1]$ 范围内的曲线：

2.2 图示 2：交叉熵随预测概率变化

为了更直观地理解交叉熵的行为，下面的图示展示了在二分类任务中，交叉熵损失随预测概率 $\hat{y}$ 的变化：

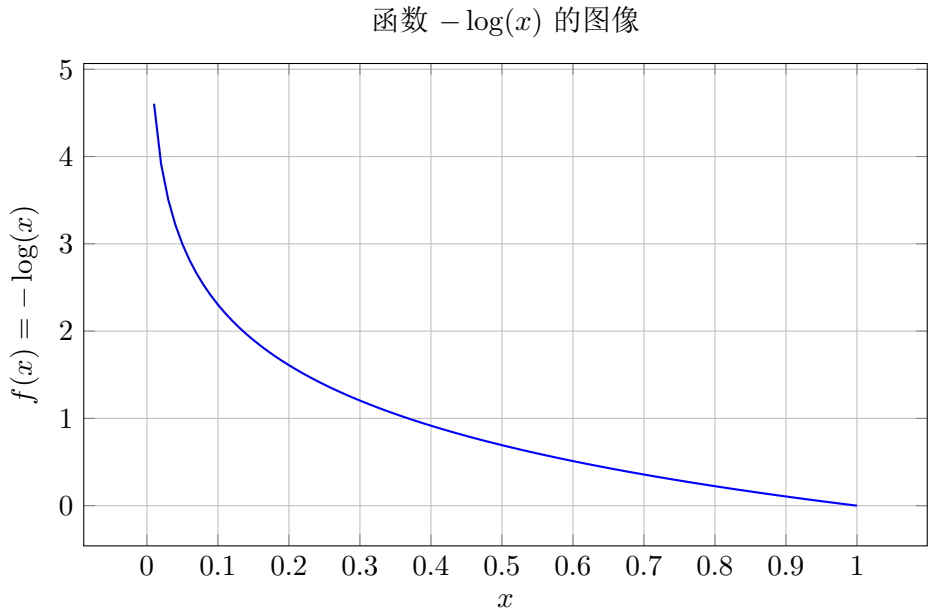

图 1: 函数 $-\log(x)$ 在区间 $(0.01, 1]$ 内的变化曲线

3 Adam 优化器

Adam 优化器结合了动量（Momentum）与 RMSProp 的思想，是一种自适应学习率的优化算法。其核心更新公式如下：

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \tag{3}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2, \tag{4}$$

其中 β_1 和 β_2 分别为动量和均方根的衰减因子。

为了修正偏差：

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1-\beta_2^t}. \tag{5}$$

最终的参数更新：

$$\theta_t = \theta_{t-1} - \frac{\alpha}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t. \tag{6}$$

3.1 图示：Adam 优化器梯度更新轨迹

下面的图示展示了 Adam 在参数空间中的自适应更新路径：

4 结论

本文介绍了交叉熵损失函数和 Adam 优化器的基本原理，并通过数学公式和图表加深理解。希望这些可视化示例能帮助更直观地理解其数学本质。

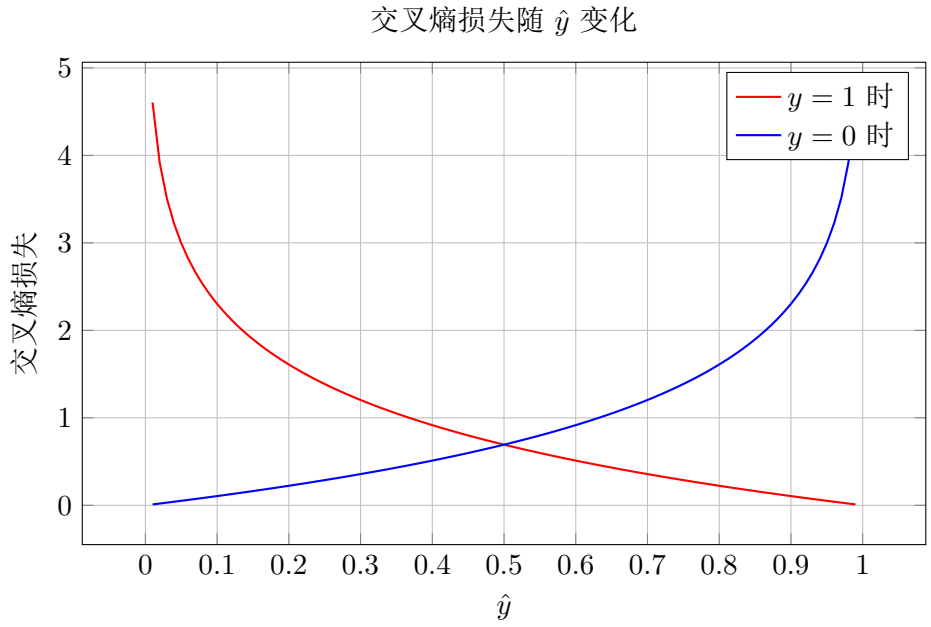

图 2: 交叉熵损失随预测概率 $\hat{y}$ 变化的曲线

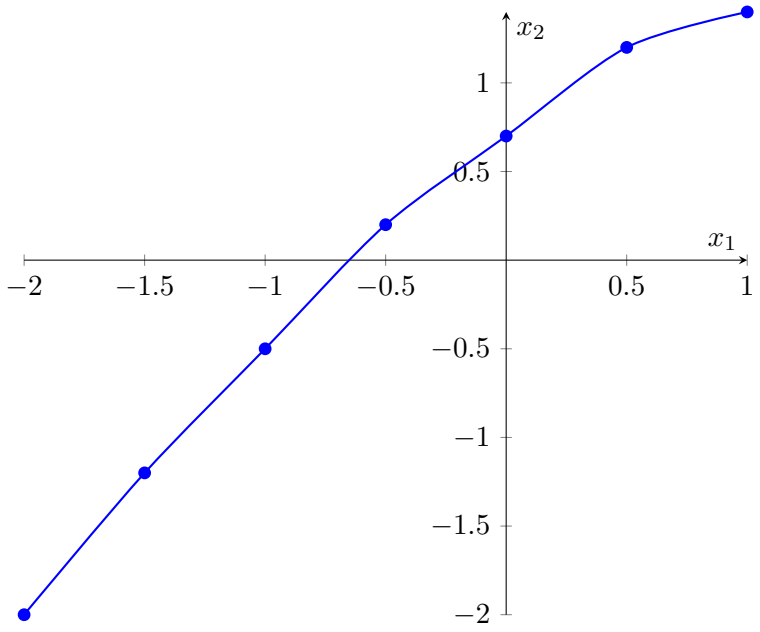

图 3: Adam 在参数空间中的更新路径示意图