



深度学习

理论与实践

左元



目录

I

深度学习

1	引言	15
1.1	监督学习	16
1.1.1	回归和分类问题	17
1.1.2	输入	18
1.1.3	机器学习模型	18
1.1.4	深度神经网络	19
1.1.5	结构化输出	19
1.2	无监督学习	21
1.2.1	生成模型	21
1.2.2	隐变量 (latent variables)	23
1.2.3	联系监督学习和无监督学习	24
1.3	强化学习	24
1.4	一个简单的例子	25
1.4.1	合成数据	25
1.4.2	线性模型	26
1.4.3	误差函数	26
1.4.4	模型复杂度	27
1.4.5	正则化	29
1.4.6	模型选择	30
2	数学基础	33
2.1	线性代数	33
2.1.1	标量和向量	33
2.1.2	向量运算	33
2.1.3	向量范数	34
2.1.4	矩阵和张量	35
2.2	微积分	37
2.2.1	导数	37

2.2.2	偏导数和梯度	39
2.2.3	机器学习中常见函数求导	41
2.3	矩阵微积分	43
2.3.1	线性变换的求导	43
2.3.2	矩阵的逐点运算以及导数	47
2.3.3	矩阵转置的求导	47
2.3.4	矩阵 Reshape 的求导	48
2.3.5	黑塞矩阵 (Hessian Matrix)	49
2.4	数值优化	49
2.4.1	数值优化要解决的问题	49
2.4.2	凸函数	49
2.4.3	梯度下降法	51
2.5	自动微分	59
2.5.1	数值微分和符号微分的缺点	59
2.5.2	反向传播算法	60
2.6	概率论	62
2.6.1	概率	62
2.6.2	概率分布	63
2.6.3	均匀分布	64
2.6.4	二项分布 (伯努利分布)	65
2.6.5	多项分布	65
2.6.6	正态分布 (高斯分布)	65
2.7	信息论	65
2.7.1	熵	65
3	一元线性回归	67
3.1	玩具数据集	67
3.2	线性回归的理论知识	68
3.3	线性回归的实现	69
3.4	梯度下降法为什么可以找到损失函数的最小值呢?	72
3.5	如何求导? (反向传播算法)	73
3.5.1	数值微分的实现	74
3.5.2	反向传播算法	76
4	分类问题：以手写数字识别为例	77
4.1	先用玩具数据集研究一下分类问题	78
4.1.1	准备训练数据集	78
4.1.2	神经网络模型结构的设计	78
4.1.3	如何设计损失函数?	83
4.1.4	如何让损失函数最小?	85
4.2	准备训练数据集	88
4.3	神经网络模型结构的设计	90
4.4	如何设计损失函数?	94
4.5	如何让损失函数最小?	96
5	卷积神经网络：将手写数字识别准确率拉满！	99

6	损失函数：均方误差损失和交叉熵损失的由来	101
7	神经网络的学习：使用梯度下降法使损失函数最小化	103
8	PyTorch 简介	105
8.1	什么是 PyTorch	105
8.1.1	PyTorch 的三大核心组件	105
8.1.2	定义深度学习	106
8.1.3	安装 PyTorch	108
8.2	理解张量	109
8.2.1	标量、向量、矩阵和张量	109
8.2.2	张量数据类型	110
8.2.3	常见的 PyTorch 张量操作	110
8.3	将模型视为计算图	112
8.4	轻松实现自动微分	113
8.5	实现多层神经网络	117
8.6	设置高效的数据加载器	120
8.7	典型的训练循环	124
8.8	保存和加载模型	127
8.9	使用 GPU 优化训练性能	127
8.9.1	在 GPU 设备上运行 PyTorch	128
8.9.2	单个 GPU 训练	129
8.9.3	使用多个 GPU 训练	130
8.10	小结	131
9	Transformer：从零实现大语言模型	133
10	监督学习	135
10.1	监督学习概述	135
10.2	线性回归示例	136
10.2.1	一维（1D）线性回归模型	136
10.2.2	损失	136
10.2.3	训练	138
10.2.4	测试	139
10.2.5	最小二乘法	139

II

大语言模型

11	理解大语言模型	147
11.1	什么是大语言模型	147
11.2	大语言模型的应用	148
11.3	构建和使用大语言模型的各个阶段	149
11.4	Transformer 架构介绍	149
11.5	利用大型数据集	150

11.6	深入剖析 GPT 架构	151
11.7	构建大语言模型	152
11.8	小结	152
12	大语言模型的架构	153
13	处理文本数据	155
13.1	对文本分词	155
13.1.1	BPE 背后的核心理念	155
13.1.2	一个简单的 BPE 实现	159
13.1.3	BPE 实现逐步讲解	168
13.2	创建词嵌入查找表	171
13.2.1	词嵌入深入讨论	171
13.2.2	词嵌入的反向传播	175



List of Figures

1.1 机器学习是人工智能中的一个子领域，将数学模型拟合到观测数据上。机器学习大致可分为监督学习、无监督学习和强化学习。深度神经网络对这些领域的每一个都有贡献	16
1.2 回归和分类问题	17
1.3 机器学习模型。该模型代表了一系列关系，将输入（儿童的年龄）与输出（儿童的身高）关联在一起。具体关系是通过训练数据集来选择的，这些训练数据由输入/输出对（橙色点）组成。当我们训练模型时，会在可能的关系中搜索一个能很好地描述数据的关系。这里，训练后的模型是青色曲线，可用来计算任何年龄的儿童的身高	19
1.4 具有结构化输出的监督学习任务。在每个案例中，输出都有复杂的内部结构或语法。某些情况下，许多输出和输入兼容	20
1.5 针对图像的生成模型。左边两张图像是由训练有猫图片的模型生成的。这些不是真实的猫，而是概率模型的样本。右边两张图像是由训练有建筑物图像的模型生成的。	21
1.6 文本数据生成模型合成的短篇故事。该模型描述了一个概率分布，为每个输出字符串分配一个概率。从模型中抽样可以创建遵循训练数据（这里是短篇故事）统计特征但之前从未见过的字符串	22
1.7 图像修复。在原始图像（左）中，男孩被金属电缆遮挡。这些不希望出现的区域（中）被移除，生成模型合成了一个新图像（右），在这个过程中保持剩余像素不变	22
1.8 条件文本合成。给定一段初始文本（黑色部分），文本的生成模型可以通过合成“缺失”的剩余部分来可信地延续字符串。由 GPT3 生成	22
1.9 人脸的变化。人脸大约有 42 块肌肉，因此可用大约 42 个数字来描述同一个人在相同光照下的图像中的大部分变化。一般来说，图像，音乐和文本的数据集可以用较少的隐变量来描述，尽管将这些变量与特定的物理机制联系起来通常更困难。	23
1.10 隐变量。许多生成模型使用深度学习模型来描述低维“潜”变量与观察到的高维数据之间的关系。隐变量具有简单的概率分布。因此，可通过从隐变量的简单分布中采样，然后使用深度学习模型样本映射到观察数据空间来生成新的样本	23
1.11 图像插值。在每一行中，左右的图像是真实的，中间的三个图像代表由生成模型创建的一系列插值。这些插值的基础生成模型已经学会所有图像都可以由一组隐变量创建。通过找到这两个真实图像的隐变量，在它们之间插值，然后使用这些中间变量创建新图像，从而生成视觉上可信又混合了两个原始图像特征的中间结果	24
1.12 由标题“时代广场上玩滑板的泰迪熊”生成的多幅图像。由 DALL-E 2 生成	24
1.13 强化学习的策略网络。将深度神经网络融入强化学习的一种方法是使用它们定义从状态（棋盘上的位置）到动作（可能的移动）。这种映射被称为策略	25

1.14 一个由 $N = 10$ 个数据点组成的训练集，以蓝色圆点显示，其中每个数据点包含了输入变量 x 及其对应的目标变量 t 的观测值。绿色曲线显示了用来生成数据的函数 $\sin(2\pi x)$ 。我们的目标是在不知道绿色曲线的情况下，预测新的输入变量 x 所对应的目标变量 t 的值。	26
1.15 平方和误差函数的几何解释[该误差函数对应来自函数 $y(x, \mathbf{w})$ 的每个数据点的位移（如垂直的绿色箭头所示）平方和的一半]	27
1.16 具有不同阶数的多项式图示。多项式如红色曲线所示。这里通过最小化平方和误差函数来拟合训练数据集	27
1.17 由式 1.3 定义的均方根误差图（在训练集和独立的测试集上对 M 的各个值进行评估）	28
2.1 导数是切线	37
2.2 求导数的原理	38
2.3 左图为曲面图，右图为等高线图	39
2.4 方向导数	40
2.5 梯度 (gradient)	41
2.6 sigmoid, relu, tanh	43
2.7 凸函数, 凹函数等	50
2.8 可微函数示例	51
2.9 不可微函数示例	52
2.10 凸函数与非凸函数示例图	52
2.11 具有鞍点的半凸 (semi-convex) 函数	53
2.12 $z = x^2 - y^2$ 的鞍点示意图	54
2.13 $f(x) = 0.5x^2 + y^2$ 示意图	54
2.14 梯度下降法步骤	55
2.15 不同学习率的对比	57
2.16 梯度下降法尝试逃离鞍点示意图	58
2.17 没有逃离鞍点	58
2.18 计算图	61
2.19 计算图	61
3.1 使用的玩具数据集	68
3.2 线性回归的示例	69
3.3 梯度下降法	70
3.4 训练后的模型	72
3.5 梯度下降法示意图	73
3.6 曲线 $y = f(x)$ 和通过其两点的直线	73
3.7 比较真的导数、前向差分近似和中心差分近似	74
4.1 ReLU 神经元	79
4.2 ReLU 激活函数	79
4.3	80
4.4 softmax 函数	82
4.5 mnist 数据集的前 5 行，形状为 5x785	89
4.6 第一行的图片	89
4.7 ReLU 神经元	91
4.8 ReLU 激活函数	91
4.9 用于解决手写数字分类的神经网络结构，其中蓝色部分为需要学习的参数	92
4.10 softmax 函数	94
8.1 PyTorch 的三大核心组件包括作为计算基础构建块的张量库、用于模型优化的自动微分引擎以及深度学习工具函数，这使得实现和训练深度神经网络模型更加容易	106
8.2 深度学习是机器学习的一个子类别，专注于实现深度神经网络。机器学习是人工智能的一个子类别，涉及从数据中学习的算法。人工智能是一个更广泛的概念，指的是机器人能够执行通常需要人类智能水平的任务	107

8.3 监督学习的预测建模工作流程包括一个训练阶段，在该阶段中，模型在训练数据集中带标签的示例上进行训练。训练好的模型随后可用于预测新观测数据的标签	108
8.4 不同秩的张量。这里零维对应于秩 0，一维对应于秩 1，二维对应于秩 2。一个由 3 个元素组成的三维向量仍然是秩为 1 的张量	109
8.5 逻辑回归的前向传播作为一个计算图。输入特征 x_1 与模型权重 w_1 相乘，并在加上偏置后通过激活函数 σ 传递。损失是通过比较模型输出 a 与给定标签 y 来计算的	113
8.6 在计算图中计算损失梯度的最常见方法是从右向左应用链式法则，这也称为“反向模型自动求导”或“反向传播”。我们从输出层（或损失本身）开始，向后通过网络一直到输入层。这么做是为了计算损失相对于网络中每个参数（权重和偏置）的梯度，从而为训练过程中如何更新这些参数提供信息	114
8.7 一个具有两个隐藏层的多层感知机。每个节点表示各自层中的一个单元。为了方便展示，这里每层都只有几个节点	117
8.8 PyTorch 实现了 Dataset 类和 DataLoader 类。Dataset 类用于实例化定义如何加载每条数据记录的对象。DataLoader 类负责处理数据的打乱和组装成批次	121
10.1 线性回归模型。对于给定的参数 $\Phi = [\phi_0, \phi_1]^T$ ，模型根据输入（ x 轴）对输出（ y 轴）进行预测。不同的截距 ϕ_0 和斜率 ϕ_1 的选择会改变这些预测结果（青色、橙色和灰色直线）。线性回归模型（式 (10.4)）定义了一组输入/输出关系（直线），而参数决定了该组中的具体成员（特定的直线）	136
10.2 线性回归的训练数据、模型和损失。图(b) ~ (d) 分别展示了具有不同参数的线性回归模型。根据截距和斜率参数 $\Phi = [\phi_0, \phi_1]^T$ 的选择，模型误差（橙色虚线）可能更大或者更小。损失 \mathcal{L} 是这些误差平方的总和	137
10.3 针对图 10.2(a) 的数据集的线性回归模型的损失函数	138
10.4 线性回归训练。训练目标是找到对应于最小损失的截距和斜率参数	139
12.1 GPT-2 和 Qwen3 的架构图	154
13.1 分词器	156
13.2 tiktoken 分词器	157
13.3 训练示例 ID 为 1 的向量表示	172
13.4 训练示例 ID 为 2 的向量表示	172
13.5 查找一批 ID 的向量表示	173
13.6 对第一个训练示例的底层计算过程	174
13.7 对第二个训练示例的底层计算过程	174



List of Tables

1.1 不同阶数多项式的系数 w^* 。注意观察随着多项式阶数的增加，系数的变化幅度是如何急剧	29
1.2 正则化参数 λ 取不同值时， $M = 9$ 的多项式模型的系数 w^* 。注章， $\ln \lambda = -\infty$ 对应未采用正则化的模型，也就是图 1.7 右下图中的拟合结果。可见，随着 λ 值的增大，系数的量级会减小	30
2.1 常见函数的导数	38
2.2 导数的求导法则	39
2.3 概率的计算	62
12.1 训练数据：输入-预测目标对	153

深度学习

I

1.3	强化学习	24
1.4	一个简单的例子	25
2	数学基础	33
2.1	线性代数	33
2.2	微积分	37
2.3	矩阵微积分	43
2.4	数值优化	49
2.5	自动微分	59
2.6	概率论	62
2.7	信息论	65
3	一元线性回归	67
3.1	玩具数据集	67
3.2	线性回归的理论知识	68
3.3	线性回归的实现	69
	梯度下降法为什么可以找到损失函数的最小值	
3.4	呢?	72
3.5	如何求导? (反向传播算法)	73
4	分类问题: 以手写数字识别为例	77
4.1	先用玩具数据集研究一下分类问题	78
4.2	准备训练数据集	88
4.3	神经网络模型结构的设计	90
4.4	如何设计损失函数?	94
4.5	如何让损失函数最小?	96
	卷积神经网络: 将手写数字识别准确率拉满!	
5	损失函数: 均方误差损失和交叉熵损失的由来	99
	神经网络的学习: 使用梯度下降法使损失函数最小化	
6	由来	101
7	神经网络的学习: 使用梯度下降法使损失函数最小化	103
8	PyTorch 简介	105
8.1	什么是 PyTorch	105
8.2	理解张量	109
8.3	将模型视为计算图	112
8.4	轻松实现自动微分	113
8.5	实现多层神经网络	117
8.6	设置高效的数据加载器	120
8.7	典型的训练循环	124
8.8	保存和加载模型	127
8.9	使用 GPU 优化训练性能	127
8.10	小结	131
	Transformer: 从零实现大语言模型	
9	133	
10	监督学习	135
10.1	监督学习概述	135
10.2	线性回归示例	136



1. 引言

在学术界，深度学习的发展历史极不寻常。一小群科学家坚持不懈地在一个看似没有前途的领域工作了 25 年，最终使得一个领域发生了技术革命并极大地影响了人类社会。研究者持续地探究学术界或者工程界中深奥且难以解决的问题，通常情况下这些问题无法得到根本性的解决。但深度学习领域是个例外，尽管广泛的怀疑仍然存在，但 **Yoshua Bengio**、**Geoff Hinton** 和 **Yann LeCun** 等人的系统性努力最终取得了成效！

🔥 深度学习历史

以 **ChatGPT** 为代表的人工神经网络的逆袭之旅，在整个科技史上也算得上跌宕起伏。它曾经在流派众多的人工智能界内部屡受歧视和打击。不止一位天才先驱以悲剧结束一生：1943 年，沃尔特·皮茨（Walter Pitts）在与沃伦·麦卡洛克（Warren McCulloch）共同提出神经网络的数学表示时才 20 岁，后来因为与导师维纳失和而脱离学术界，最终因饮酒过度于 46 岁辞世；1958 年，30 岁的弗兰克·罗森布拉特（Frank Rosenblatt）通过感知机实际实现了神经网络，而 1971 年，他在 43 岁生日那天溺水身亡；反向传播的主要提出者大卫·鲁梅尔哈特（David Rumelhart）则正值盛年（50 多岁）就罹患了罕见的不治之症，1998 年开始逐渐失智，最终在与病魔斗争十多年来离世……

一些顶级会议以及明斯基这样的学术巨人都曾毫不客气地反对甚至排斥神经网络，逼得辛顿等人不得不先后采用“关联记忆”“并行分布式处理”“卷积网络”“深度学习”等中性或者晦涩的术语为自己赢得一隅生存空间。

辛顿自己从 20 世纪 70 年代开始，坚守冷门方向几十年。从英国到美国，最后立足曾经的学术边陲加拿大，他在资金支持匮乏的情况下努力建立起一个人数不多但精英辈出的学派。

直到 2012 年，他的博士生伊尔亚·苏茨克维等在 **ImageNet** 比赛中用新方法一飞冲天，深度学习才开始成为 AI 的显学，并广泛应用于各个产业。2020 年，他又在 OpenAI 带队，通过千亿参数的 **GPT-3** 开启了大模型时代。

人工智能（**artificial intelligence**, AI）是一种旨在构建模拟人类智能行为的系统。它涵盖了多种方法，包括基于逻辑、搜索和概率推理的方法。机器学习是 AI 的一个子集，通过将数学模型拟合到观察到的数据上来学习做出决策。这一领域已经经历了爆炸性增长，导致现在机器学习几乎被误认为 AI 的同义词了。

深度神经网络是一种机器学习模型，当它拟合到数据时，被称为深度学习。目前，深度学习是最强大、最实用的机器学习模型，经常出现在日常生活中。例如，使用自然语言处理算法翻译文本、使用计算机视觉系统在互联网上搜索特定物体的图像，或者通过语音识别界面与数字助手对话，都是司空见惯的事情。所有这些应用都是由深度学习驱动的。

机器学习方法大致可以分为三个领域：监督学习、无监督学习和强化学习。目前，这三个领域的前沿方法都依赖于深度学习（图 1.1）。

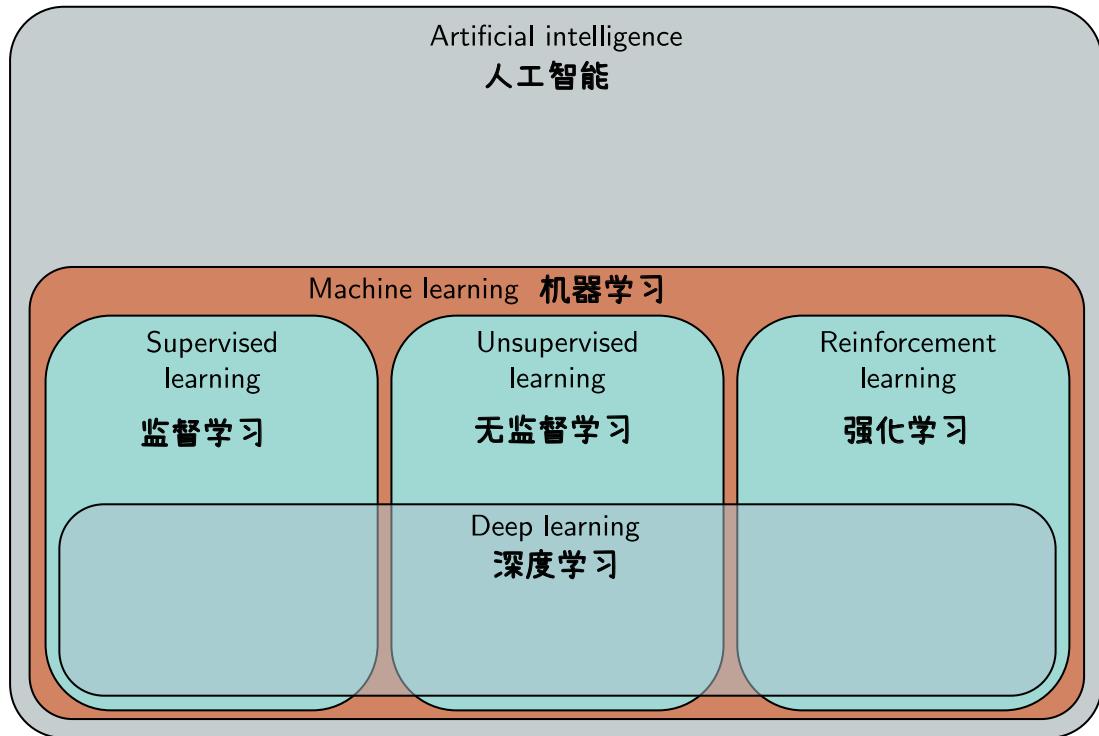


图 1.1 机器学习是人工智能中的一个子领域，将数学模型拟合到观测数据上。机器学习大致可分为监督学习、无监督学习和强化学习。深度神经网络对这些领域的每一个都有贡献

1.1 监督学习

监督学习模型定义了从输入数据到输出预测的映射。接下来将讨论输入、输出、模型本身，以及“学习”这个动作对一个模型究竟意味着什么。

1.1.1 回归和分类问题



图 1.2 回归和分类问题

图 1.2 展示了几个回归和分类问题。在每个案例中，都有一个有意义的实际输入（句子、声音文件、图片等），并将其编码为一个数值向量。这个向量构成了模型的输入。模型将输入映射到一个输出向量，然后将其“转换”并返回一个有意义的实际预测。目前，我们专注于输入和输出，并将模型视为黑盒，它接收一个数值向量并返回另一个数值向量。

图 1.2(a)中的模型基于输入特征(如房屋的平方英尺数和卧室数量)预测房价。因为模型返回一个连续数值(而不是一个类别分配),所以这是回归问题。相比之下,图 1.2(b)中的模型以一个分子的化学结构为输入,并预测它的熔点和沸点。因为它预测不止一个数值,所以是二元回归问题。

图 1.2(c)中的模型接收一个包含餐厅评论的文本字符串作为输入,并预测评论是正面的还是负面的。因为模型试图将输入分配到两个类别中的一个,所以这是二元分类问题。输出向量包含输入属于每个类别的概率。图 1.2(d)和图 1.2(e)展示了多分类问题。这里,模型将输入分配到 $N > 2$ 个类别中的一个。在图 1.2(d)的案例中,输入是一个音频文件,模型预测它包含的音乐类型。在图 1.2(e)的案例中,输入是一张图片,模型预测它包含的物体。每种情况下,模型都返回一个大小为 N 的向量,其中包含 N 个类别的概率。

1.1.2 输入

图 1.2 中的输入数据的类型差异很大。在房价预测例子中,输入的是包含了表征房屋特征值的固定长度的向量。这是一组表格数据,因为这组数据没有内部结构;如果我们改变输入的顺序并构建一个新模型,我们希望模型的预测结果保持不变。

相反,在餐厅评论的例子中,输入是一段文字。输入的长度可能根据评论中的单词数量而变化,在该例中输入顺序就很重要:“我的妻子吃了鸡肉”与“鸡肉吃了我的妻子”的含义完全不同。文本必须在传递给模型之前编码成数值形式。在该例中,使用大小为 10000 的固定词汇表,并简单地将单词的索引串联起来。

对于音乐分类的例子,输入向量可能是固定大小的(如 10 秒的片段),但维度非常高(包含许多条目)。数字音频通常以 44.1kHz 的频率采样并用 16 位整数表示,因此 10 秒的片段将由 441000 个整数组成。显然,监督学习模型必须能够处理相当庞大的输入数据。在图像分类的例子中的输入(由每个像素处的 RGB 值组成)也很庞大。此外,它的结构自然是二维的;即使在输入向量中不相邻,上下相邻的两个像素也是密切相关的。

最后,考虑预测分子的熔点和沸点的模型输入。一个分子可能包含不同数量的原子。另外,这些原子还可通过不同方式进行连接。所以这种情况下,模型必须同时将分子的几何结构和组成分子的原子作为输入。

1.1.3 机器学习模型

到目前为止,我们可将机器学习模型视作黑盒,它接收输入向量并返回输出向量。但这个黑盒里面究竟是什么呢?联想一下根据年龄预测孩子身高的模型(图 1.3)。机器学习模型是一个数学函数,它描述了平均身高如何随年龄变化(图 1.3 中的青色曲线)。将年龄代入这个函数时,它就会返回身高。例如,如果年龄是 10 岁,预测身高将是 139 厘米。



图 1.3 机器学习模型。该模型代表了一系列关系，将输入（儿童的年龄）与输出（儿童的身高）关联在一起。具体关系是通过训练数据集来选择的，这些训练数据由输入/输出对（橙色点）组成。当我们训练模型时，会在可能的关系中搜索一个能很好地描述数据的关系。这里，训练后的模型是青色曲线，可用来计算任何年龄的儿童的身高

更准确地说，该模型表示一系列将输入映射到输出的函数（即，不同的青色曲线族）。特定的函数（曲线）是使用训练数据（输入/输出对）选择的。在图 1.3 中，这些对用橙色点表示，可以看到用该模型（即青线）来描述这些数据就非常合理。当谈论训练或拟合一个模型时，我们的意思是在可能的函数（可能的青色曲线）之间搜索，来找到那个可以最准确地描述训练数据的函数。

因此，图 1.2 中的模型需要用标记的“输入/输出对”进行训练。例如，音乐分类模型需要大量音频片段，其中人类专家已经标记了每个片段的流派。这些输入/输出对在训练过程中扮演教师或监督者的角色，由此产生了“监督学习”这个术语。

1.1.4 深度神经网络

本教程将重点介绍深度神经网络，这是一种特别实用的机器学习模型，也是函数，可以表示输入和输出之间极其广泛的关系族，并可通过遍历这个关系族找到训练数据之间的关系。

深度神经网络可以处理非常大，长度可变且包含各种内部结构的输入。它们可以输出单个实数值（回归），多个数值（多元回归）或两个及更多类别的概率（分别为二元分类和多元分类）。正如我们将在下一节看到的，它们的输出也可能非常大，长度可变且包含内部结构。想象具有这些性质的函数可能十分困难，但你现在应该尝试暂时放下怀疑。

1.1.5 结构化输出

图 1.4(a)描绘了一个用于语义分割的多元分类模型。这里，输入图像的每个像素都被分配一个二元标签，指示它属于牛本身还是属于背景。图 1.4(b)展示了一个单目深度估计模型，该模型的输入是一幅街景图像，输出是每个像素的深度。这两种情况下，输出都是高维且结构化的。然而，这种结构与输入密切相关，并且可供利用；如果一个像素被标记为“牛”，那么与其具有相似 RGB 值的邻居可能也有相同的标签。

图 1.4(c) ~ (e)描绘了三组具有复杂结构的输出但与输入联系不那么密切的模型。图 1.4(c)展示的模型的输入是音频文件，输出是该文件中语音的文字转录。图 1.4(d)是翻译模型，输入是中文文本，而输出是法文文本。图 1.4(e)描述了一种极具挑战性的任务，其输入是描述性文本，而模型的输出是必须与此描述相符的图像。



图 1.4 具有结构化输出的监督学习任务。在每个案例中，输出都有复杂的内部结构或语法。某些情况下，许多输出和输入兼容

原则上，后三个任务都可在标准的监督学习框架下实现，但由于下面两个原因使它们变得更加困难。首先，输出可能确实是模糊的；从一个中文句子到法文句子有多种有效的翻译方式，任何标题都可能与多种图像相符。其次，输出包含相当多的结构；并不是所有单词或者字符都能构成有效的中文

和法文句子，也不是所有的 RGB 的组合都能构成合理的图像。除了学习映射，我们还必须遵循输出的“语法”。

幸运的是，这种“语法”可在不需要输出标签的情况下进行学习。例如，可通过学习大量文本数据的统计信息来学习如何形成有效的中文句子。这为后文中的无监督学习模型打下了基础。

1.2 无监督学习

不使用输入数据对应的输出数据标签来构建模型的过程称为无监督学习；没有输出标签意味着没有“监督”。无监督学习的目标并非是学习从输入到输出的映射，而是描述或理解输入数据的结构。就像监督学习一样，数据可能具有非常不同的特征：可能是离散的或连续的，低维的或高维的，长度固定的或可变的。

1.2.1 生成模型

本教程关注的是生成无监督模型，这类模型能合成新的数据样本，这些样本在统计学上与训练数据无法区分。一些生成模型明确描述了输入数据的概率分布，并从这个分布中采样生成新的样本。另一些模型只是学习生成新样本的机制，而不是描述它们的分布。

最先进的生成模型可以合成极其逼真但与训练样本不同的样本。它们在生成图像（图 1.5）和文本（图 1.6）方面特别成功。它们还可以在一定约束条件下合成数据，即预先设定某些输出（称为条件生成），例如图像修复（图 1.7）和文本补全（图 1.8）。事实上，现代文本生成模型如此强大，以至于它们看起来是智能的。给定一段文本然后提出一个问题，模型通常可以通过生成文档最可能的结尾部分来“填补”缺失的答案。然而，实际上，模型只是了解语言的统计信息，并不理解答案的意义。



图 1.5 针对图像的生成模型。左边两张图像是由训练有猫图片的模型生成的。这些不是真实的猫，而是概率模型的样本。右边两张图像是由训练有建筑物图像的模型生成的。

The moon had risen by the time I reached the edge of the forest, and the light that filtered through the trees was silver and cold. I shivered, though I was not cold, and quickened my pace. I had never been so far from the village before, and I was not sure what to expect. I had been walking for hours, and I was tired and hungry. I had left in such a hurry that I had not thought to pack any food, and I had not thought to bring a weapon. I was unarmed and alone in a strange place, and I did not know what I was doing.

I had been walking for so long that I had lost all sense of time, and I had no idea how far I had come. I only knew that I had to keep going. I had to find her. I was getting close. I could feel it. She was nearby, and she was in trouble. I had to find her and help her, before it was too late.

图 1.6 文本数据生成模型合成的短篇故事。该模型描述了一个概率分布，为每个输出字符串分配一个概率。从模型中抽样可以创建遵循训练数据（这里是短篇故事）统计特征但之前从未见过的字符串



图 1.7 图像修复。在原始图像（左）中，男孩被金属电缆遮挡。这些不希望出现的区域（中）被移除，生成模型合成了一个新图像（右），在这个过程中保持剩余像素不变

I was a little nervous before my first lecture at the University of Bath. It seemed like there were hundreds of students and they looked intimidating. I stepped up to the lectern and was about to speak when something bizarre happened.

Suddenly, the room was filled with a deafening noise, like a giant roar. It was so loud that I couldn't hear anything else and I had to cover my ears. I could see the students looking around, confused and frightened. Then, as quickly as it had started, the noise stopped and the room was silent again.

I stood there for a few moments, trying to make sense of what had just happened. Then I realized that the students were all staring at me, waiting for me to say something. I tried to think of something witty or clever to say, but my mind was blank. So I just said, "Well, that was strange," and then I started my lecture.

图 1.8 条件文本合成。给定一段初始文本（黑色部分），文本的生成模型可以通过合成“缺失”的剩余部分来可信地延续字符串。由 GPT3 生成

1.2.2 隐变量 (latent variables)

一些 (但并非全部) 生成模型利用了这样一个观察结果：数据的内在维度可能比观察到的变量维度总数所暗示的要少。例如，有效且有意义的英语句子数量远少于通过随机抽取单词并排列组合生成的字符串数量。同样，真实世界的图像只是通过对每个像素随机抽取 RGB 值而创建的图像中极少的一部分，这是因为图像是由物理过程产生的 (见图 1.9)。



图 1.9 人脸的变化。人脸大约有 42 块肌肉，因此可用大约 42 个数字来描述同一个人在相同光线下图像中的大部分变化。一般来说，图像、音乐和文本的数据集可以用较少的隐变量来描述，尽管将这些变量与特定的物理机制联系起来通常更困难。

这引出了一个观点，即可用更少数量的隐变量来描述每个数据样本。这里，深度学习的作用是描述这些隐变量与数据之间的映射。获得的隐变量可以有一个简单的概率分布。通过从这个分布中抽样并传递给深度学习模型，可以创建新的样本 (图 1.10)。

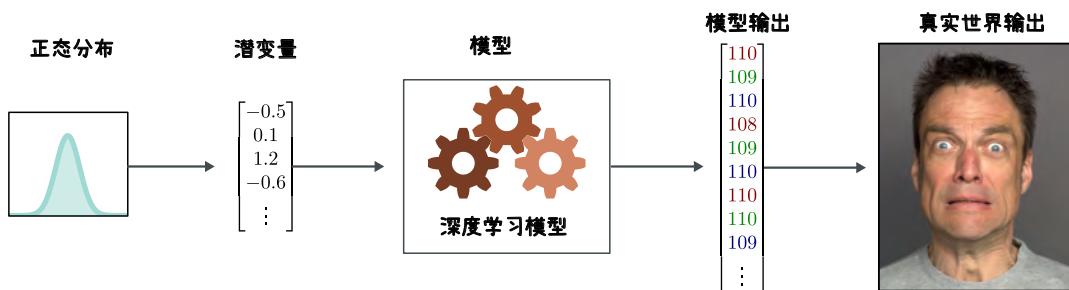


图 1.10 隐变量。许多生成模型使用深度学习模型来描述低维“潜”变量与观察到的高维数据之间的关系。隐变量具有简单的概率分布。因此，可通过从隐变量的简单分布中采样，然后使用深度学习模型样本映射到观察数据空间来生成新的样本。

这些模型开启了操纵真实数据的全新方法。例如，考虑找出支撑两个真实样本的隐变量。可通过在它们的潜在表示之间插值，并将中间位置映射回数据空间来实现这些样本之间的插值 (图 1.11)。



图 1.11 图像插值。在每一行中，左右的图像是真实的，中间的三个图像代表由生成模型创建的一系列插值。这些插值的基础生成模型已经学会所有图像都可以由一组隐变量创建。通过找到这两个真实图像的隐变量，在它们之间插值，然后使用这些中间变量创建新图像，从而生成视觉上可信又混合了两个原始图像特征的中间结果

1.2.3 联系监督学习和无监督学习

具有隐变量的生成模型也可促进监督学习模型的发展，特别是在输出具有结构时（见图 1.4）。例如，考虑学习如何预测与标题对应的图像。可以不直接将文本输入映射到图像上，而是学习解释文本的隐变量与解释图像的隐变量之间的关系。

这样做有三个优点。首先，由于输入和输出的维度较低，我们可能需要更少的文本/图像对来学习这种映射。其次，我们更可能生成看起来合理的图像；任何合理的隐变量值都应该产生像合理样本的东西。最后，如果在两组隐变量之间的映射或从隐变量到图像的映射中引入随机性，那么可生成多个都能被标题很好地描述的图像（见图 1.12）。



图 1.12 由标题“时代广场上玩滑板的泰迪熊”生成的多幅图像。由 DALL-E 2 生成

1.3 强化学习

机器学习的最后一个领域是强化学习。这一范式引入了“智能体”概念，智能体存在于一个世界中，可在每个时间步执行特定的动作。这些动作会改变系统的状态，但这种改变并不总是确定的。执行动作也会产生奖励，强化学习的目标是让“智能体”学会选择能带来高奖励的动作。

一个难点是奖励可能在动作执行后一段时间才出现，因此奖励与动作的关联并不直接。这称为时间信用分配问题。随着智能体的学习，它必须在探索和利用已有知识之间进行权衡；也许智能体已经学会了如何获得适度的奖励；它应该遵循这种策略（利用已知知识），还是应该尝试用不同的动作来检验是否有进一步提升的可能性（探索其他机会）？

两个例子

第一个例子考虑训练一个人形机器人的移动。机器人在任何时候都可执行有限数量的动作（移动各个关节），这些动作会改变机器人的世界状态（它的姿势）。我们可能会因为机器人到达障碍赛中

的检查点奖励它。要到达每个检查点，它必须执行许多动作，且当它收到奖励时，不清楚哪些动作与奖励有关，哪些是无关的。这就是时间信用分配问题的一个例子。

第二个例子是学习下棋。同样，智能体在任何时候都有一组有效的动作（移动棋子）。然而，这些动作以非确定方式改变系统状态；对于选择的任何动作，对手都可能做出许多不同的反馈。这里，可能会在吃掉棋子时设置奖励，或仅在游戏结束时获得单一奖励来设置奖励结构。在后一种情况下，时间信用分配问题是极端的；系统必须学习它所执行的大量动作，并了解哪些对成功或失败起到了关键作用。

探索-利用权衡在这两个例子中也很明显。机器人可能已经发现，它可以通过侧卧并用一条腿推的方式向前移动。这种策略会推动机器人，并带来奖励，但比最佳解决方案（双腿平衡行走）要慢得多。因此，它面临着两种选择：利用它已经知道的东西（如何笨拙地滑过地板）或探索更多的动作空间（可能加快移动速度）。同样，在国际象棋的例子中，智能体可能学会了一系列合理的开局。它应该利用这些知识还是探索不同的开局顺序？

或许，深度学习如何融入强化学习框架并不明显。有几种可能的方法，但一种技术是使用深度网络构建从观察到的世界状态到动作的映射。这称为策略网络。在机器人例子中，策略网络将学习从其传感器测量数据到关节运动的映射。在国际象棋例子中，策略网络将学习从棋盘的当前状态到移动选择的映射（图 1.13）。

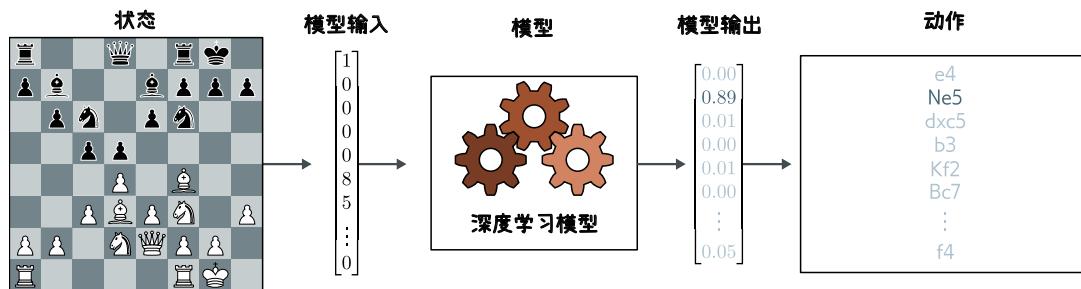


图 1.13 强化学习的策略网络。将深度神经网络融入强化学习的一种方法是使用它们定义从状态（棋盘上的位置）到动作（可能的移动）。这种映射被称为策略

1.4 一个简单的例子

我们举一个简单的例子——用多项式拟合一个小型合成数据集。这也是一个监督学习问题，在这个问题中，我们希望根据输入的变量值，对目标变量进行预测。

1.4.1 合成数据

我们用 x 表示输入变量，用 t 表示目标变量，并假设这两个变量在实数轴取值连续。给定一个训练集，其中包含 N 个 x 的观测值，记作 x_1, \dots, x_N ，还包含相应 t 的观测值，记作 t_1, \dots, t_N 。我们的目标是根据 x 的某个新值来预测相应的 t 的值。机器学习的一个关键目标是对以前没有见过的输入进行准确预测，这种能力成为泛化能力（generalization）。

我们可以通过从正弦函数采样生成的合成数据集来说明这一点。下图展示了由 $N = 10$ 个数据点组成的训练数据集，其中输入值 x_n ($n = 1, \dots, N$) 是通过在区间 $[0, 1]$ 上均匀采样生成的。对应的目标值 t_n 则是先计算每个 x 所对应的函数 $\sin(2\pi x)$ 的值，然后向每个数据点添加少量随机噪声（由高斯分布控制）得到的。通过这种方式生成数据，我们可以捕获许多现实世界数据集的一个重要特性——它们具有我们希望了解的潜在规律，但个别观测值会被随机噪声干扰。这种噪声可能源于它们固有的随机过程（例如放射性衰变），但更常见的原因是存在未被观测到的变异源。

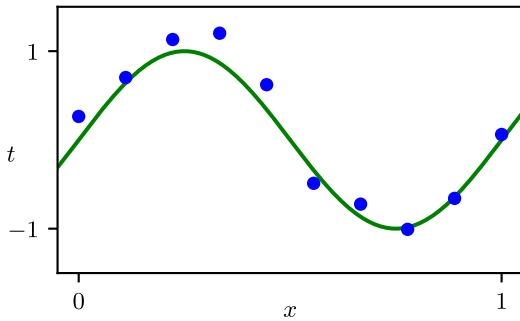


图 1.14 一个由 $N = 10$ 个数据点组成的训练集, 以蓝色圆点显示, 其中每个数据点包含了输入变量 x 及其对应的目标变量 t 的观测值。绿色曲线显示了用来生成数据的函数 $\sin(2\pi x)$ 。我们的目标是在不知道绿色曲线的情况下, 预测新的输入变量 x 所对应的目标变量 t 的值

在这个示例中, 我们事先知道真正生成数据是通过一个正弦函数。在机器学习的实际应用中, 我们的目标是在有限的训练数据集中发现隐藏的规律。不过, 了解数据的生成过程有助于我们阐明机器学习中的一些重要概念。

1.4.2 线性模型

我们的目标是利用这个训练数据集来预测输入变量的新值 \hat{x} 所对应的目标变量的值 \hat{t} , 这涉及到隐式地尝试发现潜在的函数 $\sin(2\pi x)$ 。这本质上是一个十分困难的问题, 因为我们必须从有限的数据集推广到整个函数。此外, 观测数据受到噪声干扰, 因此对于给定的 \hat{x} , \hat{t} 的适当取值存在不确定性。概率论提供了一个以精确和定量的方式来表达这种不确定性的框架。

从数据中学习概率是机器学习的核心!

现在让我们先从一种相对非正式的方式出发, 考虑一种基于曲线拟合的简单方法。我们将使用多项式函数来拟合数据, 其形式如下:

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M = \sum_{j=0}^M w_j x^j \quad (1.1)$$

其中 M 是多项式的阶数 (`order`), x^j 表示 x 的 j 次幂。多项式系数 w_0, \dots, w_M 统称为向量 \mathbf{w} 。注意, 尽管多项式函数 $y(x, \mathbf{w})$ 是关于 x 的非线性函数, 但它也是系数 \mathbf{w} 的线性函数。在上面的式子中, 像这个多项式一样, 关于未知参数呈线性的函数具有重要的特性, 同时也存在明显的局限性, 它们被称为线性模型 (`linear model`)。

1.4.3 误差函数

多项式系数的值将通过拟合训练数据来确定, 这可以通过最小化误差函数 (`error function`) 来实现, 该误差函数度量了对于任意给定的 \mathbf{w} , 函数 $y(x, \mathbf{w})$ 与训练数据集中数据点之间的拟合误差。有一个使用广泛的简单误差函数, 即每个数据点 x_n 的预测值 $y(x_n, \mathbf{w})$ 与相应目标值 t_n 之间的差的平方和的二分之一:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 \quad (1.2)$$

其中引入了系数 $1/2$ 是为了后续计算方便。后面我们同样会推导这个误差函数。注意, 这个误差函数是非负的, 当且仅当函数 $y(x, \mathbf{w})$ 正好通过每个训练数据点时, 其值等于 0。平方和误差函数的几何解释如下图所示。

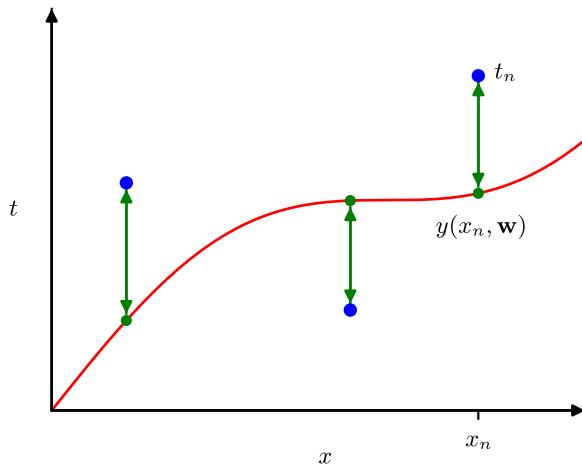


图 1.15 平方和误差函数的几何解释 [该误差函数对应来自函数 $y(x, \mathbf{w})$ 的每个数据点的位移 (如垂直的绿色箭头所示) 平方和的一半]

我们可以通过选择能够使 $E(\mathbf{w})$ 尽可能小的 \mathbf{w} 值来解决曲线拟合问题。因为平方和误差函数是系数 \mathbf{w} 的二次函数，其对系数的导数是系数 \mathbf{w} 的线性函数，所以该误差函数的最小化有一个唯一解，记作 \mathbf{w}^* ，可以通过解析形式求得解析解。最终的多项式由函数 $y(x, \mathbf{w}^*)$ 给出。

1.4.4 模型复杂度

我们还面临选择多项式的阶数 M 的问题，这将引出模型比较或者模型选择这一重要概念。在下图中，我们展示了 4 个拟合实例，它们分别使用阶数 $M = 0, 1, 3, 9$ 的多项式来拟合训练数据集的数据。

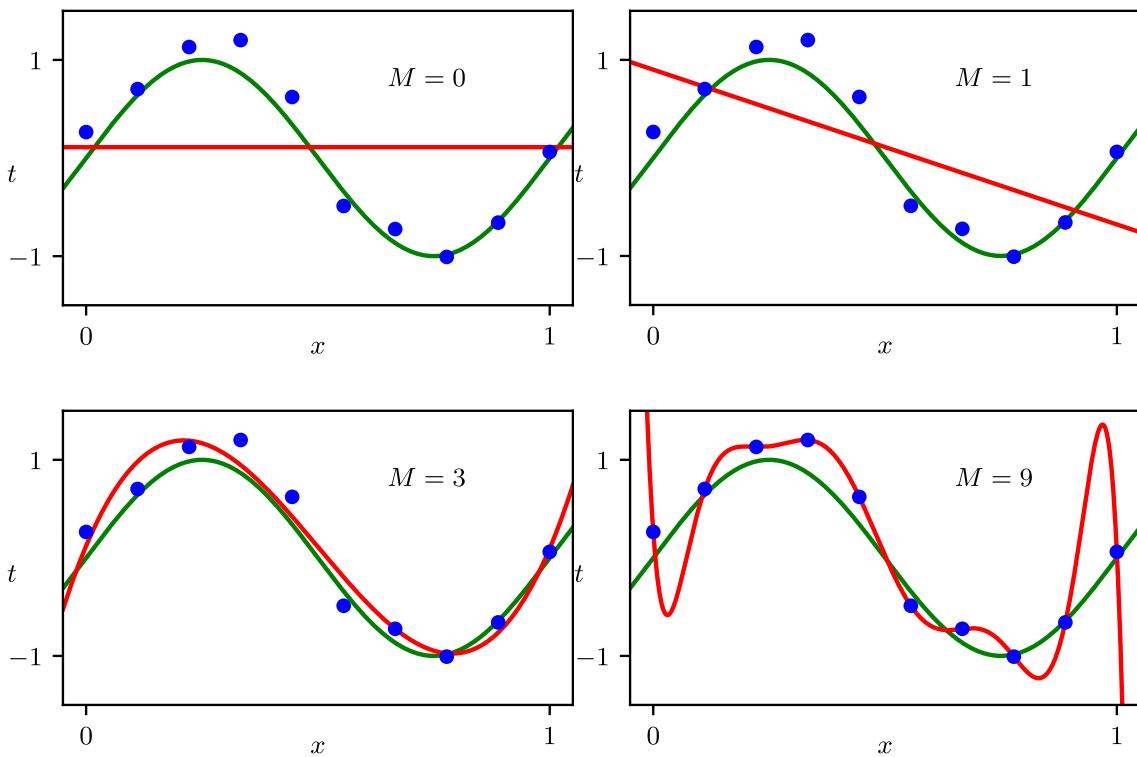


图 1.16 具有不同阶数的多项式图示。多项式如红色曲线所示。这里通过最小化平方和误差函数来拟合训练数据集

可以发现，常数($M = 0$)和一阶($M = 1$)多项式对数据的拟合较差，因此对函数 $\sin(2\pi x)$ 的表示较差。三阶($M = 3$)多项式似乎对函数 $\sin(2\pi x)$ 给出了最佳拟合。高阶($M = 9$)多项式得到了一个对训练数据完美的拟合。事实上，这个多项式曲线精确地穿过了每一个数据点，使得误差 $E(\mathbf{w}^*) = 0$ 。然而，拟合出的曲线却出现了剧烈的波动，完全不能反映出函数 $\sin(2\pi x)$ 的真实形态。这种现象称为过拟合(**over-fitting**)。

我们的目标是让模型获得良好的泛化能力，使其能够对新的数据做出准确的预测。为了定量地探究泛化性能与模型复杂度 M 之间的依赖关系，我们可以引入一个独立的测试集。该测试集包含100个数据点，其生成方式与训练集相同。针对每一个 M 值，我们不仅可以计算出模型在训练集上的残差 $E(\mathbf{w}^*)$ [式(1.2)]，还可以计算出其在测试集上的残差 $E(\mathbf{w}^*)$ 。与评估误差函数 $E(\mathbf{w})$ 相比，有时使用均方根(Root Mean Square, RMS)误差更为方便，均方根误差定义如下：

$$E_{\text{RMS}} = \sqrt{\frac{1}{N} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2} \quad (1.3)$$

公式中的 $1/N$ 是为了让不同大小的数据集能够在相同的基准下进行比较，而求平方根则是为了确保 E_{RMS} 是在与目标变量相同的尺度上(以相同的单位)进行测量的。下图展示了不同 M 值的训练集和测试集的RMS误差图。测试集上的RMS误差反映了我们根据新观测数据 x 预测其对应 t 值的能力。从下图中可以看出，当 M 值较小时测试集误差较大，这是因为此时的多项式模型灵活性不足，无法捕捉函数 $\sin(2\pi x)$ 中的振荡。当 M 取值在[3, 8]时，测试集误差较小，同时这些模型也能合理地表示出数据的生成函数 $\sin(2\pi x)$ ，如上图中 $M = 3$ 时所示。

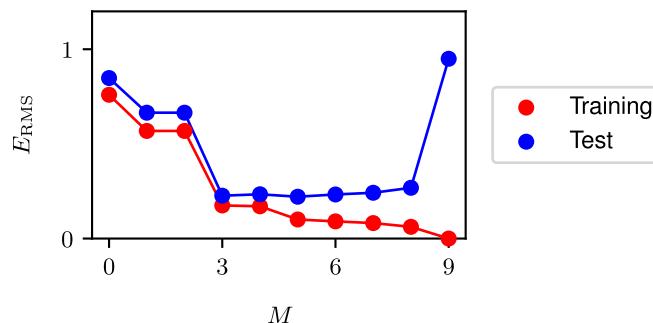


图 1.17 由式 1.3 定义的均方根误差图(在训练集和独立的测试集上对 M 的各个值进行评估)

当 $M = 9$ 时，训练集误差变为0。这是符合预期的，因为该多项式包含10个自由度(对应10个系数 w_0, \dots, w_9)，所以可以精确地调整到训练集中的10个数据点。然而，如图1.7和图1.8所示，测试集误差变得极大，函数 $y(x, \mathbf{w}^*)$ 表现出剧烈振荡。

这可能看起来很矛盾，因为一个给定阶数的多项式包含了所有更低阶的多项式作为特例。因此， $M = 9$ 的多项式理应能够产生至少与 $M = 3$ 的多项式一样好的结果。此外，我们或许会认为，预测新数据的最佳模型就应该是生成这些数据的真实函数 $\sin(2\pi x)$ 本身(我们后续将验证这一点)。同时，我们知道 $\sin(2\pi x)$ 的幂级数展开式中包含了所有阶数的项，所以我们会很自然地推断，随着模型复杂度 M 的增加，预测效果应该会持续提升。

通过观察表1.1中不同阶数多项式拟合得到的系数 \mathbf{w}^* ，我们可以更深入地了解这个问题。我们注意到，随着 M 的增加，系数的幅度越来越大。特别是当 $M = 9$ 时，为了让对应的多项式曲线能精准地穿过每一个数据点，这些系数被精细地调整到了很大的正值或负值。但在数据点之间，尤其是在数据范围的两端附近，曲线却出现了大幅度的摆动，正如我们在图1.7中看到的那样。直观地看，当多项式模型具有较大的 M 值时，它变得更加灵活，从而更容易受到目标值上随机噪声的影响，并过度拟合了这些噪声。

	$M = 0$	$M = 1$	$M = 3$	$M = 9$
w_0^*	0.11	0.90	0.12	0.26
w_1^*		-1.58	11.20	-66.13
w_2^*			-33.67	1,665.69
w_3^*			22.43	-15,566.69
w_4^*				76,321.23
w_5^*				-217,389.15
w_6^*				370,628.48
w_7^*				-372,051.47
w_8^*				202,540.70
w_9^*				-46,080.94

表 1.1 不同阶数多项式的系数 w^* 。注意观察随着多项式阶数的增加，系数的变化幅度是如何急剧

为进一步探究该现象，我们观察随着数据集大小的变化，模型学习效果的相应变化，如图 1.9 所示。可以看出，当模型复杂度固定时，数据集越大，过拟合现象就越不明显。换言之，数据量越大，我们就能用越复杂（即更灵活）的模型去拟合数据。

经典统计学中有一条常用的启发式经验：训练数据点的数量应至少是模型中可学习参数数量的若干倍（比如 5 倍或 10 倍）。然而，我们在继续探讨深度学习后会发现，即使模型参数的数量远远超过训练数据点的数量，也一样可以获得非常出色的结果。

1.4.5 正则化

根据可用训练集的大小来限制模型中参数的数量，其结果有些不尽如人意。而根据待解决问题的复杂性来选择模型的复杂性似乎更合理。作为限制参数数量的替代方案，正则化（regularization）技术经常被用于控制过拟合现象，它通过向误差函数添加一个惩罚项来抑制系数取值过大。最简单的惩罚项采用所有系数的平方和的形式，误差函数变为

$$\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2 \quad (1.4)$$

其中 $\|\mathbf{w}\|^2 = \mathbf{w}^T \mathbf{w} = w_0^2 + w_1^2 + \dots + w_M^2$ ，并且系数 λ 控制着正则化项与平方和误差项之间的相对重要性。注意，正则化项中通常不包含系数 w_0 ，因为如果包含 w_0 ，就会导致最终结果收到目标变量所选原点的影响。当然也可以包含 w_0 ，但需要单独为其配置一个正则化系数。同样，我们可以求得式 (1.4) 的精确解析解。在神经网络领域，这种方法称为权重衰减（weight decay），因为神经网络中的参数通常称为权重，而这种正则化手段会促使这些权重向零衰减。

图 1.10 展示了采用正则化误差函数 [式(1.4)] 对 9 阶 ($M=9$) 多项式进行拟合的结果，所用数据集与之前相同。观察发现，当取 $\ln \lambda = -18$ 时，过拟合现象被有效抑制，此时的拟合曲线与目标函数 $\sin(2\pi x)$ 相当接近。反之，若 λ 取值过大，则会导致欠拟合，如图 1.10 中 $\ln \lambda = 0$ 的情形。表 1.2 给出了不同 λ 值下拟合得到的多项式系数。这些数值表明，正则化确实发挥了预期作用，有效减小了系数的幅度。

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
w_0^*	0.26	0.26	0.11
w_1^*	-66.13	0.64	-0.07
w_2^*	1,665.69	43.68	-0.09
w_3^*	-15,566.61	-144.00	-0.07
w_4^*	76,321.23	57.90	-0.05
w_5^*	-217,389.15	117.36	-0.04
w_6^*	370,626.48	9.87	-0.02
w_7^*	-372,051.47	-90.02	-0.01
w_8^*	202,540.70	-70.90	-0.01
w_9^*	-46,080.94	75.26	0.00

表 1.2 正则化参数 λ 取不同值时, $M = 9$ 的多项式模型的系数 w^* 。注章, $\ln \lambda = -\infty$ 对应未采用正则化的模型, 也就是图 1.7 右下图中的拟合结果。可见, 随着 λ 值的增大, 系数的量级会减小

通过绘制训练集和测试集的 RMS 误差 [式(1.3)] 与 $\ln \lambda$ 的关系, 可以看出正则化项对泛化误差的影响, 如图 1.11 所示。可以看到, λ 现在控制了模型的有效复杂性, 从而决定了过拟合的程度。

1.4.6 模型选择

在本例中, λ 作为一个超参数 (**hyper parameter**), 它的值在基于误差函数最小化来确定模型参数 w 的过程中始终保持不变。需要注意的是, 我们不能通过同时对 w 和 λ 最小化误差函数的方式来简单地确定 λ 的取值, 因为这样会导致 λ 趋近于 0, 从而产生一个在训练集上误差极小甚至为零的过拟合模型。类似地, 多项式的阶数 M 也是模型的一个超参数, 单纯地优化训练集误差关于 M 的取值会导致 M 过大, 同样会引发过拟合问题。因此, 我们需要找到一种有效的方法来确定这些超参数的合理取值。有一种简单的思路, 即将已有的数据集划分为训练集和验证集 (**validation set**) [也称为保留集 (**hold-out set**) 或开发集 (**development set**)] , 其中训练集用于确定模型系数 w , 而我们最终选择在验证集上误差最小的模型。如果使用有限规模的数据集多次迭代模型设计, 也可能会出现对验证集的过拟合现象。为此, 通常需要预留出测试集, 用于对最终选定模型的性能进行评估。

在某些实际应用场景中, 可用于模型训练和测试的数据量往往较为有限。为了构建性能良好的模型, 我们希望尽量充分地利用一切可获取的数据进行训练。然而, 如果验证集的规模过小, 则会导致对模型预测性能的评估存在较大偏差。交叉验证 (**cross-validation**) 技术为解决这一困境提供了一种有效途径, 如图 1.12 所示。该方法允许将 $(S - 1)/S$ 的数据用于模型训练, 与此同时, 利用全部数据来评估模型的性能。当数据资源极度匮乏时, 还可以考虑 $S = N$ 的极端情况, 其中 N 表示数据点的总数, 这时的交叉验证就演变成了留一法 (**leave-one-out**)。

交叉验证的主要缺点在于所需的训练次数增加了 S 倍, 这对于训练过程本身计算成本较高的模型来说是一个大问题。交叉验证等使用独立数据评估性能的技术还存在另一个问题, 即对于单个模型可能存在多个复杂度超参数 (例如, 可能有多个正则化超参数)。在最坏的情况下, 探索这些超参数设置的最佳组合可能需要指数级数量的训练次数。现代机器学习的前沿领域需要非常大的模型和大规模训练数据集。因此, 超参数设置的探索空间有限, 很大程度上依赖于从小模型获得的经验和启发式方法。

这个将多项式模型拟合到基于正弦函数生成的合成数据集的简单示例, 阐释了机器学习领域的一些核心概念, 在后续章节我们将进一步用这个示例进行讨论。然而, 现实中的机器学习应用往往要复杂得多: 首先, 用于模型训练的数据集规模可能会非常庞大, 数据量通常会高出几个数量级; 其次, 模型的输入变量的数量通常也会显著增加, 例如, 在图像分析领域, 输入变量的规模可达百万量级, 并且还可能伴随着多个输出变量。在这些应用中, 将输出映射到输入的可学习函数通常由

一类被称为神经网络的模型来表征，这些模型往往具有海量的参数，其数量甚至可以达到千亿的规模。此时，误差函数将表现为这些模型参数的复杂非线性函数，无法再通过解析解的方式进行最小化，而必须借助迭代优化的方法，利用误差函数关于模型参数的导数信息来逐步逼近最优解。这对计算机硬件提出了更高的要求，并且不可避免地会引入相当高的计算成本。



2. 数学基础

2.1 线性代数

2.1.1 标量和向量

1. 标量 (scalar)

标量是一个单独的数，只有大小。

2. 向量 (vector)

向量由标量组成，有大小有方向。

- 行向量：

$$[2 \ 5 \ 8] \quad (2.1)$$

- 列向量：

$$\begin{bmatrix} 2 \\ 5 \\ 8 \end{bmatrix} \quad (2.2)$$

2.1.2 向量运算

1. 向量转置：列向量转置结果为行向量

$$\begin{aligned} \mathbf{x} &= \begin{bmatrix} 2 \\ 5 \\ 8 \end{bmatrix} \\ \mathbf{x}^T &= [2 \ 5 \ 8] \end{aligned} \quad (2.3)$$

2. 向量相加：对应元素相加

$$\begin{bmatrix} 2 \\ 5 \\ 8 \end{bmatrix} + \begin{bmatrix} 1 \\ 3 \\ 7 \end{bmatrix} = \begin{bmatrix} 3 \\ 8 \\ 15 \end{bmatrix} \quad (2.4)$$

3. 向量与标量相乘：标量与向量每个元素相乘

$$3 \times \begin{bmatrix} 2 \\ 5 \\ 8 \end{bmatrix} = \begin{bmatrix} 6 \\ 15 \\ 24 \end{bmatrix} \quad (2.5)$$

4. 向量内积：又称向量点乘，两向量对应元素乘积之和，结果为标量

$$\langle \mathbf{x}, \mathbf{y} \rangle = \left\langle \begin{bmatrix} 2 \\ 5 \\ 8 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \\ 7 \end{bmatrix} \right\rangle = 2 + 15 + 56 = 73 \quad (2.6)$$

两向量之间夹角表示为

$$\cos \theta = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \sqrt{\langle \mathbf{y}, \mathbf{y} \rangle}} \quad (2.7)$$

2.1.3 向量范数

范数（norm）是具有“长度”概念的函数。

1. L_0 范数

$$\|\mathbf{x}\|_0 = \text{非零元素的个数} \quad (2.8)$$

例如：

$$\mathbf{x} = \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix} \quad (2.9)$$

$$\|\mathbf{x}\|_0 = 2$$

2. L_1 范数

$$\|\mathbf{x}\|_1 = \sum_{i=1}^m |x_i| = |x_1| + \cdots + |x_m| \quad (2.10)$$

例如：

$$\mathbf{x} = \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix} \quad (2.11)$$

$$\|\mathbf{x}\|_1 = 0 + 2 + 1 = 3$$

3. L_2 范数

$$\|\mathbf{x}\|_2 = \left(\sum_{i=1}^m |x_i|^2 \right)^{\frac{1}{2}} = \sqrt{|x_1|^2 + \cdots + |x_m|^2} \quad (2.12)$$

例如：

$$\mathbf{x} = \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix} \quad (2.13)$$

$$\|\mathbf{x}\|_2 = \sqrt{0 + 4 + 1} = \sqrt{5}$$

4. L_p 范数

$$\|x\|_p = \left(\sum_{i=1}^m |x_i|^p \right)^{\frac{1}{p}} = (|x_1|^p + \dots + |x_m|^p)^{\frac{1}{p}} \quad (2.14)$$

2.1.4 矩阵和张量

2.1.4.1 矩阵的概念

一个 $m \times n$ 的矩阵 (`matrix`) 是一个有 m 行 n 列元素的矩形阵列，用 $\mathbb{R}^{m \times n}$ 表示所有 $m \times n$ 实数矩阵的向量空间。

$$\begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 4 & 8 \end{bmatrix} \in \mathbb{R}^{3 \times 2} \quad (2.15)$$

1. 方阵：行数等于列数的矩阵

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \in \mathbb{R}^{2 \times 2} \quad (2.16)$$

2. 对角矩阵：主对角线以外元素全为0的方阵：

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 9 \end{bmatrix} \quad (2.17)$$

3. 单位矩阵：主对角线元素全为1的对角矩阵

$$\mathbf{I}_{3 \times 3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (2.18)$$

2.1.4.2 矩阵转置

1. 矩阵转置运算

矩阵 $\mathbf{A} \in \mathbb{R}^{m \times n}$ 的转置是一个 $n \times m$ 的矩阵，记为 \mathbf{A}^T 。其中的第 i 个行向量是原矩阵的第 i 个列向量。或者说，转置矩阵 \mathbf{A}^T 第 i 行第 j 列的元素是原矩阵 \mathbf{A} 第 j 行第 i 列的元素。

$$\begin{aligned} [\mathbf{A}^T]_{ij} &= a_{ij} \\ \mathbf{A} &= \begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 4 & 8 \end{bmatrix} \in \mathbb{R}^{3 \times 2} \\ \mathbf{A}^T &= \begin{bmatrix} 1 & 3 & 4 \\ 2 & 5 & 8 \end{bmatrix} \in \mathbb{R}^{2 \times 3} \end{aligned} \quad (2.19)$$

2. 矩阵转置的性质

$$\begin{aligned} (\mathbf{A}^T)^T &= \mathbf{A} \\ (\mathbf{A} + \mathbf{B})^T &= \mathbf{A}^T + \mathbf{B}^T \\ (k\mathbf{A})^T &= k\mathbf{A}^T \\ (\mathbf{AB})^T &= \mathbf{B}^T \mathbf{A}^T \end{aligned} \quad (2.20)$$

2.1.4.3 矩阵乘法

1. 矩阵乘法运算

两个矩阵的乘法仅当矩阵 A 的列数和矩阵 B 的行数相等时才能定义, 如 $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, 它们的乘积 $AB \in \mathbb{R}^{m \times p}$:

$$[AB]_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} = \sum_{k=1}^n a_{ik}b_{kj} \quad (2.21)$$

例如:

$$\begin{bmatrix} 1 & 0 & 2 \\ -1 & 3 & 1 \end{bmatrix} \cdot \begin{bmatrix} 3 & 1 \\ 2 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 \times 3 + 0 \times 2 + 2 \times 1 & 1 \times 1 + 0 \times 1 + 2 \times 0 \\ (-1) \times 3 + 3 \times 2 + 1 \times 1 & (-1) \times 1 + 3 \times 1 + 1 \times 0 \end{bmatrix} = \begin{bmatrix} 5 & 1 \\ 4 & 2 \end{bmatrix} \quad (2.22)$$

特别的, 矩阵和单位矩阵相乘等于矩阵本身。

$$\begin{aligned} AI &= A \quad (A \in \mathbb{R}^{m \times n}, I \in \mathbb{R}^{n \times n}) \\ IA &= A \quad (I \in \mathbb{R}^{n \times n}, A \in \mathbb{R}^{n \times m}) \end{aligned} \quad (2.23)$$

例如

$$AI = \begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 4 & 8 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 \times 1 + 2 \times 0 & 1 \times 0 + 2 \times 1 \\ 3 \times 1 + 5 \times 0 & 3 \times 0 + 5 \times 1 \\ 4 \times 1 + 8 \times 0 & 4 \times 0 + 8 \times 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 4 & 8 \end{bmatrix} \quad (2.24)$$

2. 矩阵乘法的性质

矩阵乘法满足结合律、左分配律和右分配律。不满足交换律即 $AB \neq BA$ 。

- 结合律: 若 $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, $C \in \mathbb{R}^{p \times q}$, 则 $(AB)C = A(BC)$ 。
- 左分配律: 若 $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{m \times n}$, $C \in \mathbb{R}^{p \times q}$, 则 $(A+B)C = AC + BC$ 。
- 右分配律: 若 $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, $C \in \mathbb{R}^{n \times p}$, 则 $A(B+C) = AB + AC$ 。

2.1.4.4 矩阵的逆

对于方阵 A , 如果存在另一个方阵 A^{-1} , 使得 $AA^{-1} = I$ 成立, 此时 $A^{-1}A = I$ 也同样成立。称 A^{-1} 为 A 的逆矩阵。例如:

$$AA^{-1} = \begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix} \cdot \begin{bmatrix} -5 & 2 \\ 3 & -1 \end{bmatrix} = \begin{bmatrix} 1 \times (-5) + 2 \times 3 & 1 \times 2 + 2 \times (-1) \\ 3 \times (-5) + 5 \times 3 & 3 \times 2 + 5 \times (-1) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I \quad (2.25)$$

2.1.4.5 矩阵的 Hadamard 积

矩阵 $A \in \mathbb{R}^{m \times n}$ 和矩阵 $B \in \mathbb{R}^{m \times n}$ 的 Hadamard 积记作 $A \odot B$, 它是两个矩阵对应元素的乘积, 是一个 $m \times n$ 的矩阵。

$$(A \odot B)_{ij} = a_{ij}b_{ij} \quad (2.26)$$

2.1.4.6 张量 (tensor)

张量 (tensor) 可视为多维数组, 是标量, 1维向量和2维矩阵的 n 维推广。

例如: 3维张量, 形状是 $3 \times 3 \times 2$

$$\begin{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 4 & 8 \end{bmatrix} & \begin{bmatrix} 3 & 2 \\ 1 & 6 \\ 7 & 3 \end{bmatrix} & \begin{bmatrix} 5 & 6 \\ 9 & 1 \\ 2 & 4 \end{bmatrix} \end{bmatrix} \quad (2.27)$$

2.1.4.7 正定矩阵 (Positive Definite Matrix)

对于 $n \times n$ 实对称矩阵 A :

A 是正定矩阵, 如果对于所有非零向量 $\mathbf{x} \in \mathbb{R}^n$, 都有:

$$\mathbf{x}^T \mathbf{A} \mathbf{x} > 0 \quad (2.28)$$

\mathbf{A} 是半正定矩阵，如果对于所有非零向量 $\mathbf{x} \in \mathbb{R}^n$ ，都有：

$$\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0 \quad (2.29)$$

2.2 微积分

2.2.1 导数

2.2.1.1 导数的概念

导数(derivative)是微积分中的一个概念。函数在某一点的导数是指这个函数在这一点附近的变化率(即函数在这一点的切线斜率)。导数的本质是通过极限的概念对函数进行局部的线性逼近。

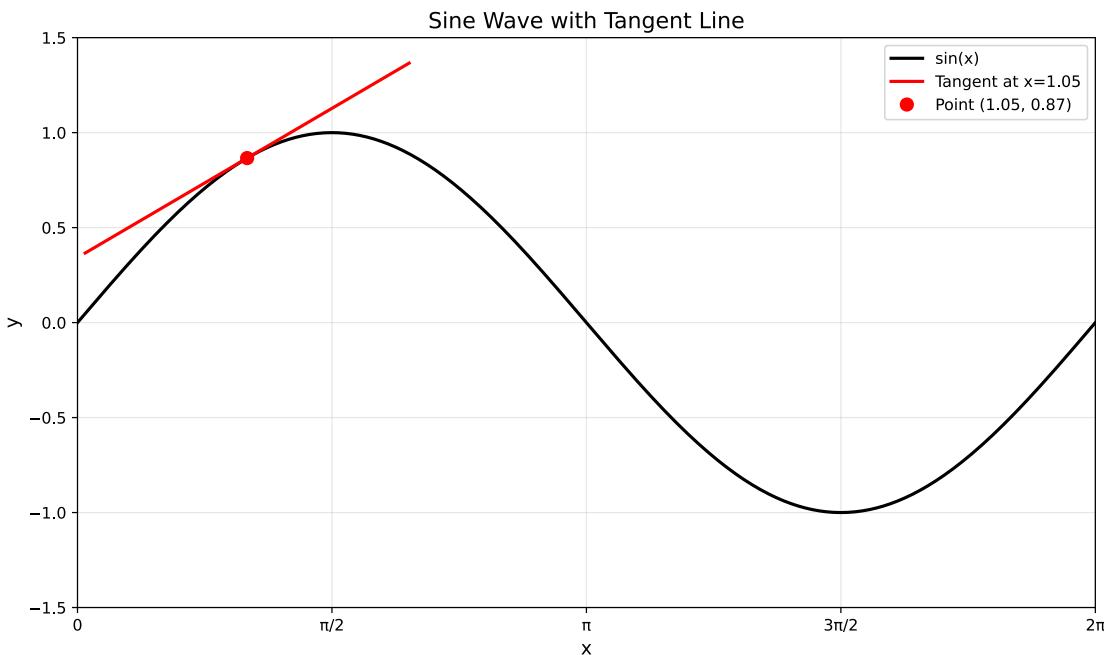


图 2.1 导数是切线

当函数 f 的自变量在一点 x_0 上产生一个增量 Δx 时，函数输出值的增量 Δy 与自变量增量 Δx 的比值在 Δx 趋于0时的极限如果存在，即为 f 在 x_0 处的导数，记作 $f'(x_0)$ 、 $\frac{df}{dx}(x_0)$ 或 $\frac{df}{dx}|_{x=x_0}$ 。

$$f'(x_0) = \lim_{\Delta x \rightarrow 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} \quad (2.30)$$

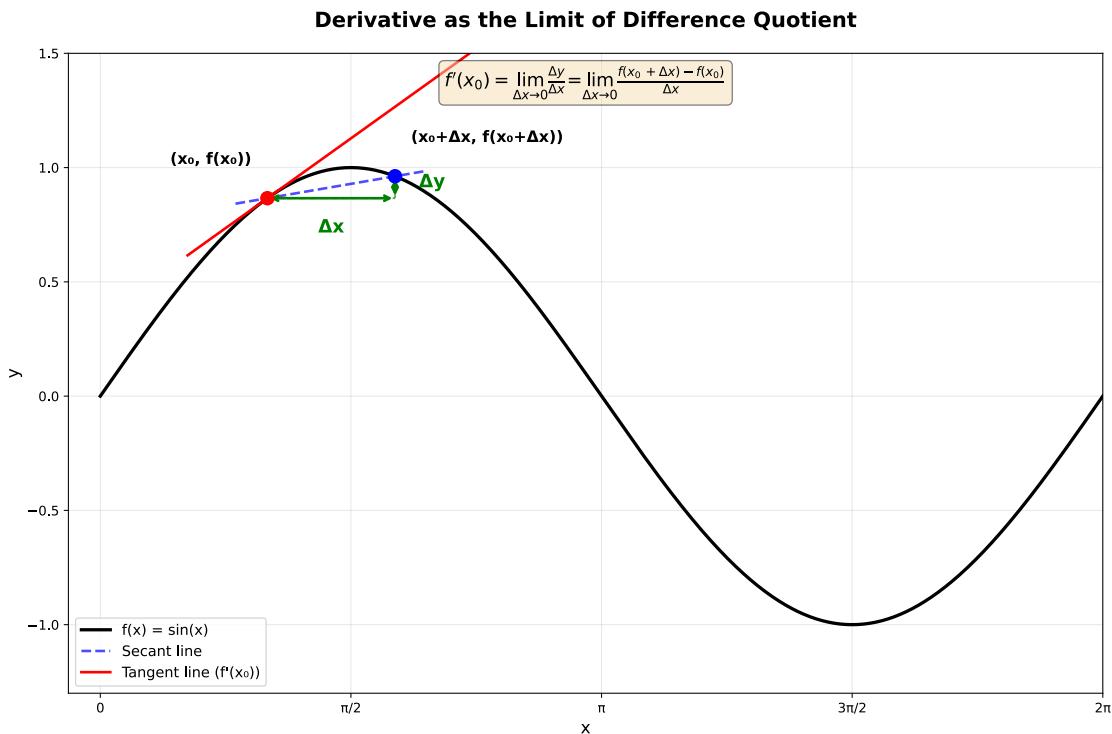


图 2.2 求导数的原理

可以将上面求导数的原理，直接转换成程序

数值微分

```
1 def derivative(f, x):
2     delta_x = 1e-5 # Δx
3     deriv = (f(x + delta_x) - f(x)) / delta_x #  $\frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$ 
4     return deriv
```

Python

2.2.1.2 基本函数的导数

在机器学习和深度学习中， $\log x = \log_e x$ ，我们后面也用这种表示法。

说明	公式	例子
常数的导数	$(C)' = 0$	$(3)' = 0$
幂函数的导数	$(x^\alpha)' = \alpha x^{\alpha-1}$	$(x^3)' = 3x^2$
指数函数的导数	$(a^x)' = a^x \log a$	$(3^x)' = 3^x \log 3$
	$(e^x)' = e^x$	——
三角函数的导数	$(\sin x)' = \cos x$	——
	$(\cos x)' = -\sin x$	——

表 2.1 常见函数的导数

2.2.1.3 导数的求导法则

说明	公式
两函数之和求导	$(f + g)' = f' + g'$
两函数之积求导	$(fg)' = f'g + fg'$
两函数之商求导	$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$
复合函数求导 (链式求导法则)	若 $f(x) = h(g(x))$, 则 $\frac{df}{dx}(x) = \frac{df}{dh} \cdot \frac{dh}{dg} \cdot \frac{dg}{dx}$

表 2.2 导数的求导法则

例如, 求函数 $f(x) = x^4 + \sin(x^2) - \log(x)e^x + 7$ 在 $x = 3$ 处的导数。

$$\begin{aligned} f'(x) &= (x^4 + \sin(x^2) - \log(x)e^x + 7)' \\ &= 4x^{4-1} + \cos(x^2) \cdot 2x - \left(\frac{e^x}{x} + \log(x)e^x \right) + 0 \\ &= 4x^3 + 2x \cos(x^2) - \frac{e^x}{x} - \log(x)e^x \end{aligned} \quad (2.31)$$

所以

$$f'(3) = 108 + 6 \cos(9) - \frac{e^3}{3} - \log(3)e^3 \quad (2.32)$$

2.2.2 偏导数和梯度

2.2.2.1 偏导数

如果函数 f 的自变量并非单个元素, 而是多个元素, 例如:

$$f(x, y) = x^2 + xy + y^2 \quad (2.33)$$

我们可以绘制出函数的曲面图和等高线图

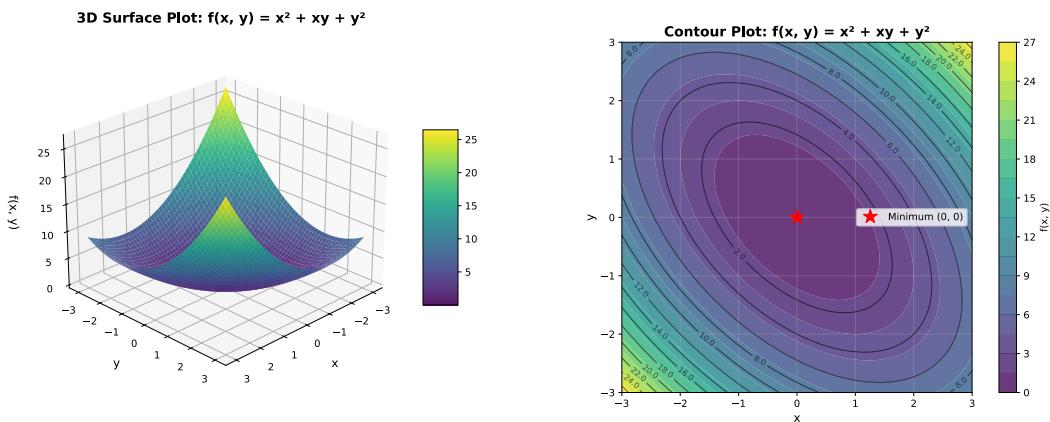


图 2.3 左图为曲面图, 右图为等高线图

可将其中一个元素 x 看作常数, 此时 f 可看作关于另一元素 y 的函数。

$$f_x(y) = x^2 + xy + y^2 \quad (2.34)$$

在 $x = a$ 固定的情况下，可计算 f_x 关于 y 的导数：

$$f_{x=a}'(y) = a + 2y \quad (2.35)$$

这种导数称为偏导数，一般记作：

$$\frac{\partial f}{\partial y}(x, y) = x + 2y \quad (2.36)$$

更一般地来说，一个多元函数 $f(x_1, x_2, \dots, x_n)$ 在点 (a_1, a_2, \dots, a_n) 处对 x_i 的偏导数定义为：

$$\frac{\partial f}{\partial x_i}(a_1, a_2, \dots, a_n) = \lim_{\Delta x_i \rightarrow 0} \frac{f(a_1, \dots, a_i + \Delta x_i, \dots, a_n) - f(a_1, \dots, a_i, \dots, a_n)}{\Delta x_i} \quad (2.37)$$

2.2.2.2 方向导数

偏导数可以看作是多元函数 f 沿某个自变量轴方向的变化率。

如果我们任意选取一个方向 l ，那么在某个点 (x_0, y_0) 处，二元函数 $f(x, y)$ 沿着这个方向的变化率可以用极限定义为：

$$\frac{\partial f}{\partial l}(x_0, y_0) = \lim_{\Delta l \rightarrow 0} \frac{f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0)}{\Delta l} \quad (2.38)$$

这里， Δl 就是沿方向 l 的微小改变量， Δx 和 Δy 与 Δl 的关系为：

$$\begin{aligned} \Delta x &= \Delta l \cdot \cos \alpha \\ \Delta y &= \Delta l \cdot \cos \beta \end{aligned} \quad (2.39)$$

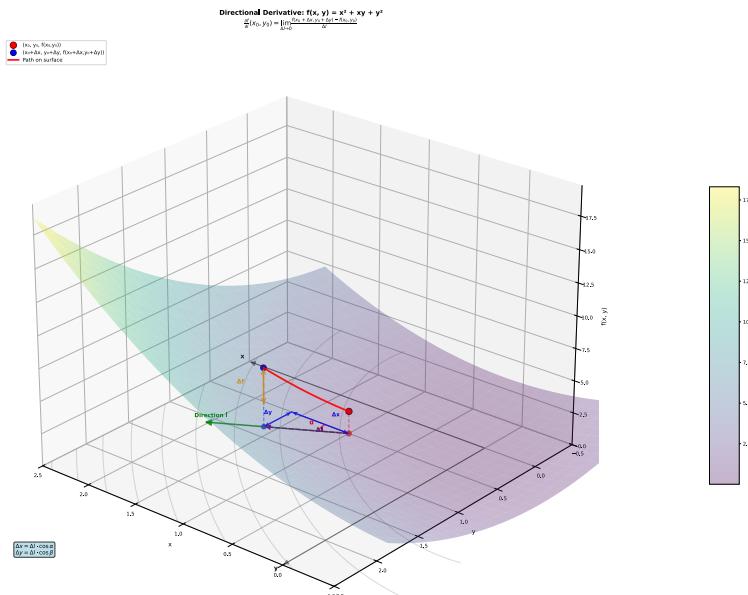


图 2.4 方向导数

根据全微分公式，上式可以表示为：

$$\frac{\partial f}{\partial l}(x_0, y_0) = f_x(x_0, y_0) \cos \alpha + f_y(x_0, y_0) \cos \beta \quad (2.40)$$

其中 $f_x(x_0, y_0), f_y(x_0, y_0)$ 表示点 (x_0, y_0) 处 f 对 x, y 的偏导数; $\cos \alpha, \cos \beta$ 是方向 l 的方向余弦, 即 l 方向的单位方向向量可以表示为 $\mathbf{I}_0 = (\cos \alpha, \cos \beta)$ 。

这个“沿某个方向的变化率”, 就被称为 $f(x, y)$ 沿方向 l 的方向导数。

2.2.2.3 梯度 (gradient)

多元函数 $f(x_1, \dots, x_n)$ 关于每个变量 x_i 都有偏导数 $\frac{\partial f}{\partial x_i}$, 在点 $a = (a_1, a_2, \dots, a_n)$ 处, 这些偏导数定义出一个向量:

$$\nabla f(a) = \left[\frac{\partial f}{\partial x_1}(a) \quad \frac{\partial f}{\partial x_2}(a) \quad \cdots \quad \frac{\partial f}{\partial x_n}(a) \right] \quad (2.41)$$

这个向量称为 f 在点 a 的梯度, 记作 $\nabla f(a)$ 或者 $\text{grad}\{f(a)\}$ 。

例如: $f(x, y) = x^2 + xy + y^2$ 在 $(1, 1)$ 处的梯度为 $[3 \ 3]$ 。

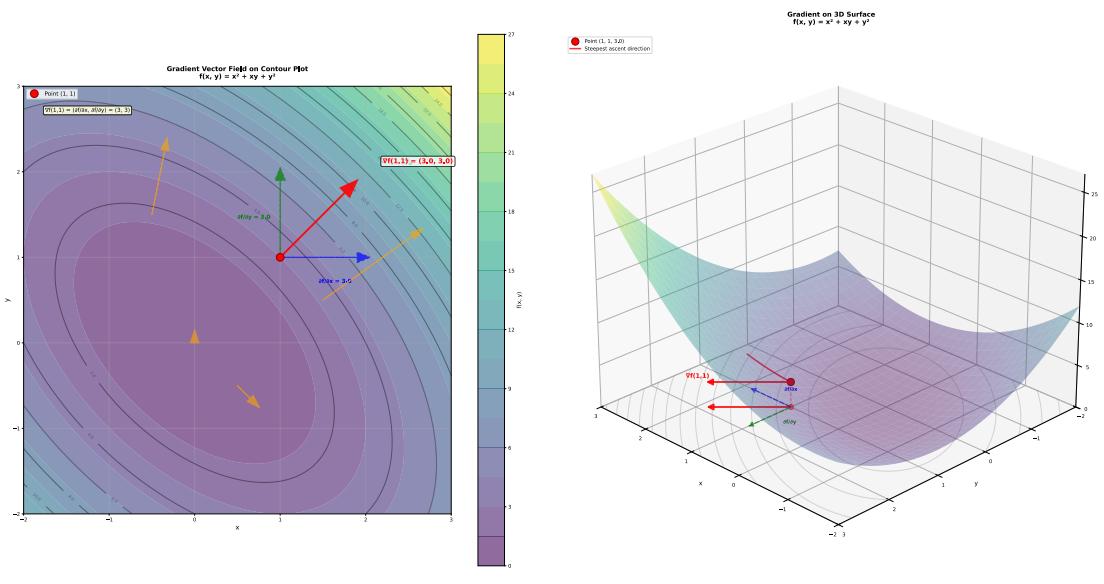


图 2.5 梯度 (gradient)

梯度向量表示的方向, 就是函数在这一点处, 方向导数取最大值的方向。换句话说, 梯度的方向, 就是函数值变化最快的方向。

2.2.3 机器学习中常见函数求导

1. Sigmoid 函数求导

Sigmoid 函数定义如下:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2.42)$$

求导过程如下

$$\begin{aligned}
\sigma'(x) &= \frac{0 \cdot (1 + e^{-x}) - 1 \cdot e^{-x} \cdot (-1)}{(1 + e^{-x})^2} \\
&= \frac{1}{1 + e^{-x}} \cdot \frac{e^{-x}}{1 + e^{-x}} \\
&= \frac{1}{1 + e^{-x}} \cdot \left(1 - \frac{1}{1 + e^{-x}}\right) \\
&= \sigma(x)(1 - \sigma(x))
\end{aligned} \tag{2.43}$$

2. ReLU 函数求导

$$\text{ReLU}(x) = \max(x, 0) \tag{2.44}$$

求导结果如下

$$\text{ReLU}'(x) = \begin{cases} 1, & x > 0 \\ 0, & x < 0 \end{cases} \tag{2.45}$$

注意！0点不可导。

3. Tanh 函数求导

\tanh 函数定义如下

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{2.46}$$

求导过程如下

$$\begin{aligned}
\tanh'(x) &= \frac{\frac{d(e^x - e^{-x})}{dx} \cdot (e^x + e^{-x}) - \frac{d(e^x + e^{-x})}{x} \cdot (e^x - e^{-x})}{(e^x + e^{-x})^2} \\
&= \frac{(e^x + e^{-x})^2 - (e^x - e^{-x})^2}{(e^x + e^{-x})^2} \\
&= 1 - \frac{(e^x - e^{-x})^2}{(e^x + e^{-x})^2} \\
&= 1 - (\tanh(x))^2
\end{aligned} \tag{2.47}$$

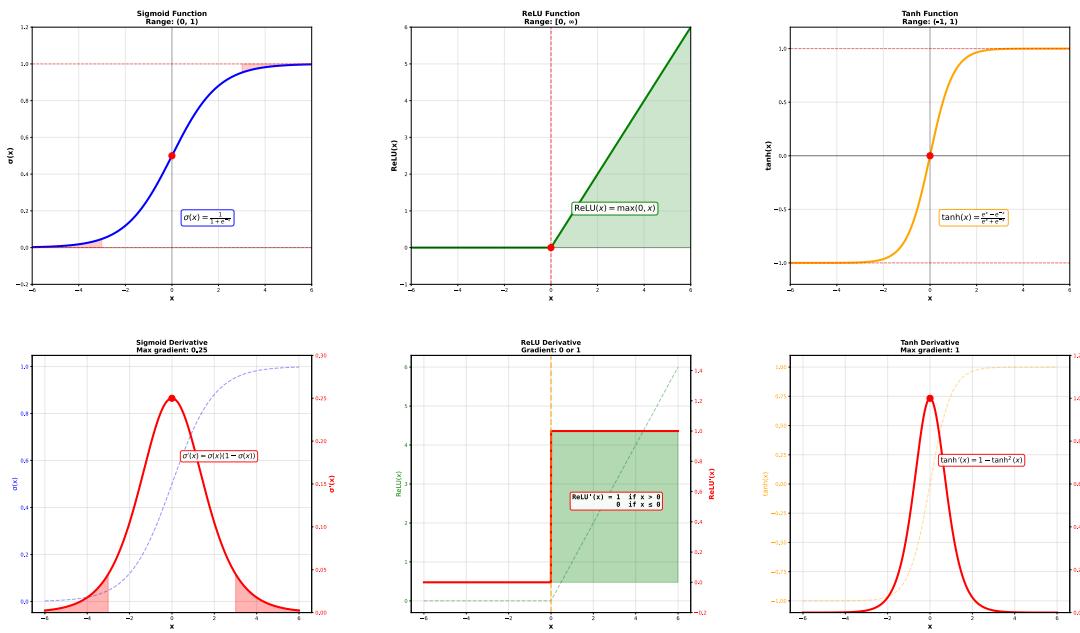


图 2.6 sigmoid, relu, tanh

2.3 矩阵微积分

矩阵求导的本质就是标量函数对变量的每个元素逐个求导，只是写成了向量、矩阵的形式。

⚡ 注意！

只有标量函数对标量变量的导数具有现实意义，矩阵只是一种表示方法。所以向量对向量的求导，向量对矩阵的求导，矩阵对矩阵的求导等等，都必须从标量函数 \mathcal{L} 开始进行计算，并针对每个标量变量进行求导，最后可以整理成矩阵形式！

2.3.1 线性变换的求导

线性变换的定义为

$$\mathbf{Y}_{m \times p} = \mathbf{W}_{m \times n} \mathbf{X}_{n \times p} + \mathbf{b}_{m \times 1} \quad (2.48)$$

其中 $\mathbf{b}_{m \times 1}$ 会被广播为形状 $m \times p$ 。

例如

$$\begin{aligned}
 Y &= \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \end{bmatrix} \\
 &= \begin{bmatrix} 7 & 10 \\ 15 & 22 \end{bmatrix} + \underbrace{\begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix}}_{\text{广播以后}} \\
 &= \begin{bmatrix} 9 & 12 \\ 16 & 23 \end{bmatrix}
 \end{aligned} \quad (2.49)$$

现在我们思考一个问题

$$\begin{aligned}\frac{\partial \mathbf{Y}}{\partial \mathbf{W}} &=? \\ \frac{\partial \mathbf{Y}}{\partial \mathbf{X}} &=? \\ \frac{\partial \mathbf{Y}}{\partial \mathbf{b}} &=?\end{aligned}\tag{2.50}$$

我们知道矩阵对矩阵求导没有实际意义, 必须从标量函数开始求导。所以我们需要一个标量函数, 也就是

$$\mathcal{L}(\mathbf{Y}) = \mathcal{L}(\mathbf{W}_{m \times n} \mathbf{X}_{n \times p} + \mathbf{b}_{m \times 1})\tag{2.51}$$

标量对矩阵求导, 就是对矩阵的每个元素求偏导数后组成的矩阵。

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{W}} &= \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial w_{11}} & \frac{\partial \mathcal{L}}{\partial w_{12}} & \cdots & \frac{\partial \mathcal{L}}{\partial w_{1n}} \\ \frac{\partial \mathcal{L}}{\partial w_{21}} & \frac{\partial \mathcal{L}}{\partial w_{22}} & \cdots & \frac{\partial \mathcal{L}}{\partial w_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{L}}{\partial w_{m1}} & \frac{\partial \mathcal{L}}{\partial w_{m2}} & \cdots & \frac{\partial \mathcal{L}}{\partial w_{mn}} \end{bmatrix} \\ \frac{\partial \mathcal{L}}{\partial \mathbf{X}} &= \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial x_{11}} & \frac{\partial \mathcal{L}}{\partial x_{12}} & \cdots & \frac{\partial \mathcal{L}}{\partial x_{1p}} \\ \frac{\partial \mathcal{L}}{\partial x_{21}} & \frac{\partial \mathcal{L}}{\partial x_{22}} & \cdots & \frac{\partial \mathcal{L}}{\partial x_{2p}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{L}}{\partial x_{n1}} & \frac{\partial \mathcal{L}}{\partial x_{n2}} & \cdots & \frac{\partial \mathcal{L}}{\partial x_{np}} \end{bmatrix} \\ \frac{\partial \mathcal{L}}{\partial \mathbf{b}} &= \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial b_1} \\ \frac{\partial \mathcal{L}}{\partial b_2} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial b_m} \end{bmatrix}\end{aligned}\tag{2.52}$$

我们通过一个简单的例子就能看出计算过程, 设置 $m = n = p = 2$ 。那么有如下

$$\begin{aligned}\mathbf{W} &= \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \\ \mathbf{X} &= \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \\ \mathbf{b} &= \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}\end{aligned}\tag{2.53}$$

所以有

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} = \begin{bmatrix} w_{11}x_{11} + w_{12}x_{21} + b_1 & w_{11}x_{12} + w_{12}x_{22} + b_1 \\ w_{21}x_{11} + w_{22}x_{21} + b_2 & w_{21}x_{12} + w_{22}x_{22} + b_2 \end{bmatrix}\tag{2.54}$$

那么有如下推导过程

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial w_{11}} &= \frac{\partial \mathcal{L}}{\partial y_{11}} \cdot \frac{\partial y_{11}}{\partial w_{11}} + \frac{\partial \mathcal{L}}{\partial y_{12}} \cdot \frac{\partial y_{12}}{\partial w_{11}} + \cancel{\frac{\partial \mathcal{L}}{\partial y_{21}}} \cancel{\frac{\partial y_{21}}{\partial w_{11}}} + \cancel{\frac{\partial \mathcal{L}}{\partial y_{22}}} \cancel{\frac{\partial y_{22}}{\partial w_{11}}} \\ &= \frac{\partial \mathcal{L}}{\partial y_{11}} \cdot x_{11} + \frac{\partial \mathcal{L}}{\partial y_{12}} \cdot x_{12}\end{aligned}\tag{2.55}$$

同理有

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial w_{12}} &= \frac{\partial \mathcal{L}}{\partial y_{11}} \cdot x_{21} + \frac{\partial \mathcal{L}}{\partial y_{12}} \cdot x_{22} \\ \frac{\partial \mathcal{L}}{\partial w_{21}} &= \frac{\partial \mathcal{L}}{\partial y_{21}} \cdot x_{11} + \frac{\partial \mathcal{L}}{\partial y_{22}} \cdot x_{12} \\ \frac{\partial \mathcal{L}}{\partial w_{22}} &= \frac{\partial \mathcal{L}}{\partial y_{21}} \cdot x_{21} + \frac{\partial \mathcal{L}}{\partial y_{22}} \cdot x_{22}\end{aligned}\tag{2.56}$$

整理成矩阵形式如下：

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{W}} &= \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial w_{11}} & \frac{\partial \mathcal{L}}{\partial w_{12}} \\ \frac{\partial \mathcal{L}}{\partial w_{21}} & \frac{\partial \mathcal{L}}{\partial w_{22}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial y_{11}} & \frac{\partial \mathcal{L}}{\partial y_{12}} \\ \frac{\partial \mathcal{L}}{\partial y_{21}} & \frac{\partial \mathcal{L}}{\partial y_{22}} \end{bmatrix} \cdot \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \end{bmatrix} \\ &= \frac{\partial \mathcal{L}}{\partial \mathbf{Y}} \cdot \mathbf{X}^T \\ &= \mathbf{X} \cdot \left(\frac{\partial \mathcal{L}}{\partial \mathbf{Y}} \right)^T\end{aligned}\tag{2.57}$$

可以看到， $\frac{\partial \mathcal{L}}{\partial \mathbf{W}}$ 具体定义为 \mathbf{X} 还是 \mathbf{X}^T 取决于上游的梯度过来时，是左乘还是右乘。也就是说，矩阵求导无法脱离上游的梯度而单独存在。

重复以上过程，可以得到如下

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial x_{11}} &= \frac{\partial \mathcal{L}}{\partial y_{11}} \cdot w_{11} + \frac{\partial \mathcal{L}}{\partial y_{21}} \cdot w_{21} \\ \frac{\partial \mathcal{L}}{\partial x_{12}} &= \frac{\partial \mathcal{L}}{\partial y_{12}} \cdot w_{11} + \frac{\partial \mathcal{L}}{\partial y_{22}} \cdot w_{21} \\ \frac{\partial \mathcal{L}}{\partial x_{21}} &= \frac{\partial \mathcal{L}}{\partial y_{11}} \cdot w_{12} + \frac{\partial \mathcal{L}}{\partial y_{21}} \cdot w_{22} \\ \frac{\partial \mathcal{L}}{\partial x_{22}} &= \frac{\partial \mathcal{L}}{\partial y_{12}} \cdot w_{12} + \frac{\partial \mathcal{L}}{\partial y_{22}} \cdot w_{22}\end{aligned}\tag{2.58}$$

整理成矩阵形式：

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{X}} &= \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial x_{11}} & \frac{\partial \mathcal{L}}{\partial x_{12}} \\ \frac{\partial \mathcal{L}}{\partial x_{21}} & \frac{\partial \mathcal{L}}{\partial x_{22}} \end{bmatrix} \\ &= \begin{bmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial y_{11}} & \frac{\partial \mathcal{L}}{\partial y_{12}} \\ \frac{\partial \mathcal{L}}{\partial y_{21}} & \frac{\partial \mathcal{L}}{\partial y_{22}} \end{bmatrix} \\ &= \mathbf{W}^T \cdot \frac{\partial \mathcal{L}}{\partial \mathbf{Y}} \\ &= \left(\frac{\partial \mathcal{L}}{\partial \mathbf{Y}} \right)^T \cdot \mathbf{W}\end{aligned}\tag{2.59}$$

由于 \mathbf{b} 在加法运算时，会进行广播。所以我们需要看一下，当然原理还是全微分公式。

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial b_1} &= \frac{\partial \mathcal{L}}{\partial y_{11}} \cdot \frac{\partial y_{11}}{\partial b_1} + \frac{\partial \mathcal{L}}{\partial y_{12}} \cdot \frac{\partial y_{12}}{\partial b_1} \\ &= \frac{\partial \mathcal{L}}{\partial y_{11}} + \frac{\partial \mathcal{L}}{\partial y_{12}} \\ \frac{\partial \mathcal{L}}{\partial b_2} &= \frac{\partial \mathcal{L}}{\partial y_{21}} + \frac{\partial \mathcal{L}}{\partial y_{22}}\end{aligned}\tag{2.60}$$

所以整理成矩阵形式

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial b_1} \\ \frac{\partial \mathcal{L}}{\partial b_2} \end{bmatrix} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial y_{11}} + \frac{\partial \mathcal{L}}{\partial y_{12}} \\ \frac{\partial \mathcal{L}}{\partial y_{21}} + \frac{\partial \mathcal{L}}{\partial y_{22}} \end{bmatrix} \quad (2.61)$$

也就是

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}} = \text{np.sum}\left(\frac{\partial \mathcal{L}}{\partial \mathbf{Y}}, \text{axis}=1, \text{keepdims=True}\right) \quad (2.62)$$

我们来个复杂一点的，也就是线性变换的复合的求导。

$$\mathbf{Y} = \mathbf{W}'(\mathbf{WX} + \mathbf{b}) + \mathbf{b}' \quad (2.63)$$

其中：

- $\mathbf{X} \in \mathbb{R}^{n \times p}$ (输入)
- $\mathbf{W} \in \mathbb{R}^{m \times n}$ (内层线性变换的权重)
- $\mathbf{b} \in \mathbb{R}^{m \times 1}$ (内层线性变换的偏置)
- $\mathbf{W}' \in \mathbb{R}^{k \times m}$ (外层线性变换的权重)
- $\mathbf{b}' \in \mathbb{R}^{k \times 1}$ (外层线性变换的偏置)
- $\mathbf{Y} \in \mathbb{R}^{k \times p}$ (输出)

求 $\frac{\partial \mathcal{L}}{\partial \mathbf{X}} = ?$

我们使用链式求导法则，所以引入中间变量，有如下：

$$\begin{aligned} \mathbf{Z} &= \mathbf{WX} + \mathbf{b} \\ \mathbf{Y} &= \mathbf{W}'\mathbf{Z} + \mathbf{b}' \end{aligned} \quad (2.64)$$

应用链式求导法则

- 步骤 1：

$$\frac{\partial \mathcal{L}}{\partial \mathbf{Z}} = (\mathbf{W}')^T \cdot \frac{\partial \mathcal{L}}{\partial \mathbf{Y}} \quad (2.65)$$

- 步骤 2：

$$\frac{\partial \mathcal{L}}{\partial \mathbf{X}} = \mathbf{W}^T \cdot \frac{\partial \mathcal{L}}{\partial \mathbf{Z}} \quad (2.66)$$

组合一下结果

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{X}} &= \mathbf{W}^T \cdot \left[(\mathbf{W}')^T \cdot \frac{\partial \mathcal{L}}{\partial \mathbf{Y}} \right] \\ &= \mathbf{W}^T (\mathbf{W}')^T \frac{\partial \mathcal{L}}{\partial \mathbf{Y}} \\ &= \left(\frac{\partial \mathcal{L}}{\partial \mathbf{Y}} \right)^T \mathbf{W}' \mathbf{W} \end{aligned} \quad (2.67)$$

求 $\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = ?$

还是应用链式法则

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{Z}} &= (\mathbf{W}')^T \cdot \frac{\partial \mathcal{L}}{\partial \mathbf{Y}} \\ \frac{\partial \mathcal{L}}{\partial \mathbf{W}} &= \frac{\partial \mathcal{L}}{\partial \mathbf{Z}} \cdot \mathbf{X}^T \end{aligned} \quad (2.68)$$

所以结果是

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{W}} &= \left[(\mathbf{W}')^T \cdot \frac{\partial \mathcal{L}}{\partial \mathbf{Y}} \right] \cdot \mathbf{X}^T \\ &= \left(\frac{\partial \mathcal{L}}{\partial \mathbf{Y}} \right)^T \mathbf{W}' \cdot \mathbf{X}\end{aligned}\quad (2.69)$$

求 $\frac{\partial \mathcal{L}}{\partial \mathbf{b}} = ?$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}} = \text{np.sum}\left(\frac{\partial \mathcal{L}}{\partial \mathbf{Z}}, \text{axis}=1, \text{keepdims=True}\right) \quad (2.70)$$

2.3.2 矩阵的逐点运算以及导数

逐点运算的例子

$$\begin{aligned}\mathbf{Y} = \text{ReLU}(\mathbf{X}) &= \text{ReLU}\left(\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}\right) = \begin{bmatrix} \text{ReLU}(x_{11}) & \text{ReLU}(x_{12}) \\ \text{ReLU}(x_{21}) & \text{ReLU}(x_{22}) \end{bmatrix} \\ &= \begin{bmatrix} \max(x_{11}, 0) & \max(x_{12}, 0) \\ \max(x_{21}, 0) & \max(x_{22}, 0) \end{bmatrix}\end{aligned}\quad (2.71)$$

此时 $\frac{\partial \mathbf{Y}}{\partial \mathbf{X}} = ?$

我们还是从标量函数 \mathcal{L} 开始求导数

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial x_{11}} &= \frac{\partial \mathcal{L}}{\partial y_{11}} \cdot \frac{\partial y_{11}}{\partial x_{11}} = \frac{\partial \mathcal{L}}{\partial y_{11}} \cdot \{1 \text{ if } x_{11} > 0 \text{ else } 0\} \\ \frac{\partial \mathcal{L}}{\partial x_{12}} &= \frac{\partial \mathcal{L}}{\partial y_{12}} \cdot \frac{\partial y_{12}}{\partial x_{11}} = \frac{\partial \mathcal{L}}{\partial y_{12}} \cdot \{1 \text{ if } x_{12} > 0 \text{ else } 0\} \\ \frac{\partial \mathcal{L}}{\partial x_{21}} &= \frac{\partial \mathcal{L}}{\partial y_{21}} \cdot \frac{\partial y_{21}}{\partial x_{11}} = \frac{\partial \mathcal{L}}{\partial y_{21}} \cdot \{1 \text{ if } x_{21} > 0 \text{ else } 0\} \\ \frac{\partial \mathcal{L}}{\partial x_{22}} &= \frac{\partial \mathcal{L}}{\partial y_{22}} \cdot \frac{\partial y_{22}}{\partial x_{11}} = \frac{\partial \mathcal{L}}{\partial y_{22}} \cdot \{1 \text{ if } x_{22} > 0 \text{ else } 0\}\end{aligned}\quad (2.72)$$

整理成矩阵形式如下

$$\frac{\partial \mathbf{Y}}{\partial \mathbf{X}} = \frac{\partial \mathcal{L}}{\partial \mathbf{Y}} \odot \underbrace{\{\mathbf{X} > 0\}}_{\substack{\text{numpy 中} \\ \text{的计算方式}}}\quad (2.73)$$

```
1 import numpy as np
2
3 X = np.array([[1, 2], [3, -1]])
4 print(X > 0)
5 print(1 * (X > 0))
```

 Python

2.3.3 矩阵转置的求导

$$\mathbf{Y} = \mathbf{X}^T \quad (2.74)$$

转置也是一个函数，那么我们还是从标量函数开始求导。

$$\begin{aligned}
 \frac{\partial \mathcal{L}}{\partial x_{11}} &= \frac{\partial \mathcal{L}}{\partial y_{11}} \cdot \frac{\partial y_{11}}{\partial x_{11}} = \frac{\partial \mathcal{L}}{\partial y_{11}} \\
 \frac{\partial \mathcal{L}}{\partial x_{12}} &= \frac{\partial \mathcal{L}}{\partial y_{21}} \cdot \frac{\partial y_{21}}{\partial x_{12}} = \frac{\partial \mathcal{L}}{\partial y_{21}} \\
 \frac{\partial \mathcal{L}}{\partial x_{21}} &= \frac{\partial \mathcal{L}}{\partial y_{12}} \cdot \frac{\partial y_{12}}{\partial x_{21}} = \frac{\partial \mathcal{L}}{\partial y_{12}} \\
 \frac{\partial \mathcal{L}}{\partial x_{22}} &= \frac{\partial \mathcal{L}}{\partial y_{22}} \cdot \frac{\partial y_{22}}{\partial x_{22}} = \frac{\partial \mathcal{L}}{\partial y_{22}}
 \end{aligned} \tag{2.75}$$

所以

$$\frac{\partial \mathcal{L}}{\partial \mathbf{X}} = \left(\frac{\partial \mathcal{L}}{\partial \mathbf{Y}} \right)^T \tag{2.76}$$

2.3.4 矩阵 Reshape 的求导

和转置基本一样

$$\mathbf{Y} = \text{reshape}(\mathbf{X}, \text{new_shape}) \tag{2.77}$$

导数如下

$$\frac{\partial \mathcal{L}}{\partial \mathbf{X}} = \text{reshape}\left(\frac{\partial \mathcal{L}}{\partial \mathbf{Y}}, \text{shape}(\mathbf{X})\right) \tag{2.78}$$

举个例子

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \in \mathbb{R}^{4 \times 1} \tag{2.79}$$

$$\mathbf{Y} = \text{reshape}(\mathbf{X}, (2, 2)) = \begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix} \in \mathbb{R}^{2 \times 2}$$

元素对应关系

$$\begin{aligned}
 y_{11} &= x_1 \\
 y_{12} &= x_2 \\
 y_{21} &= x_3 \\
 y_{22} &= x_4
 \end{aligned} \tag{2.80}$$

求导得到

$$\begin{aligned}
 \frac{\partial \mathcal{L}}{\partial x_1} &= \frac{\partial \mathcal{L}}{\partial y_{11}} \\
 \frac{\partial \mathcal{L}}{\partial x_2} &= \frac{\partial \mathcal{L}}{\partial y_{12}} \\
 \frac{\partial \mathcal{L}}{\partial x_3} &= \frac{\partial \mathcal{L}}{\partial y_{21}} \\
 \frac{\partial \mathcal{L}}{\partial x_4} &= \frac{\partial \mathcal{L}}{\partial y_{22}}
 \end{aligned} \tag{2.81}$$

所以有如下

$$\frac{\partial \mathcal{L}}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial y_{11}} \\ \frac{\partial \mathcal{L}}{\partial y_{12}} \\ \frac{\partial \mathcal{L}}{\partial y_{21}} \\ \frac{\partial \mathcal{L}}{\partial y_{22}} \end{bmatrix} = \text{reshape}\left(\frac{\partial \mathcal{L}}{\partial \mathbf{Y}}, (4, 1)\right) \quad (2.82)$$

2.3.5 黑塞矩阵 (Hessian Matrix)

对于二阶可微的标量函数 $f: \mathbb{R}^n \rightarrow \mathbb{R}$, 黑塞矩阵 (Hessian Matrix) 是由所有二阶偏导数组成的 $n \times n$ 方阵:

$$\nabla^2 f(\mathbf{x}) = \mathbf{H}_f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \quad (2.83)$$

简单记作

$$\mathbf{H}_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j} \quad (2.84)$$

2.4 数值优化

2.4.1 数值优化要解决的问题

所谓数值优化基本就是在研究如何得到一个函数的最大值或者最小值, 深度学习在表面上看, 是一个无约束优化问题。

$$\min_{\theta} \mathcal{L}(\theta) \quad (2.85)$$

函数 \mathcal{L} 的参数是 θ , 如何找到参数 θ 使得 \mathcal{L} 最小呢?

这可以说是深度学习要解决的问题。

$$\max_{\theta} J(\theta) \quad \text{等价于} \quad \min_{\theta} \{-J(\theta)\} \quad (2.86)$$

2.4.2 凸函数

先从一元凸函数开始讨论。

函数 $f: \mathbb{R} \rightarrow \mathbb{R}$ 称为凸函数 (convex function), 如果对于定义域内的任意两点 x, y 和任意 $\lambda \in [0, 1]$, 都有:

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) \quad (2.87)$$

几何意义: 连接函数图像上任意两点的线段, 位于函数图像的上方或与之重合。

当 $\lambda \neq 0$ 且 $\lambda \neq 1$, 也就是 $\lambda \in (0, 1)$ 且 $x \neq y$ 时, 不等号严格成立

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y) \quad (2.88)$$

则称 f 为严格凸函数 (strictly convex function)。

如果不等号方向相反, 则 f 为凹函数

$$f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y) \quad (2.89)$$

既不是凸函数也不是凹函数的, 叫做非凸函数。

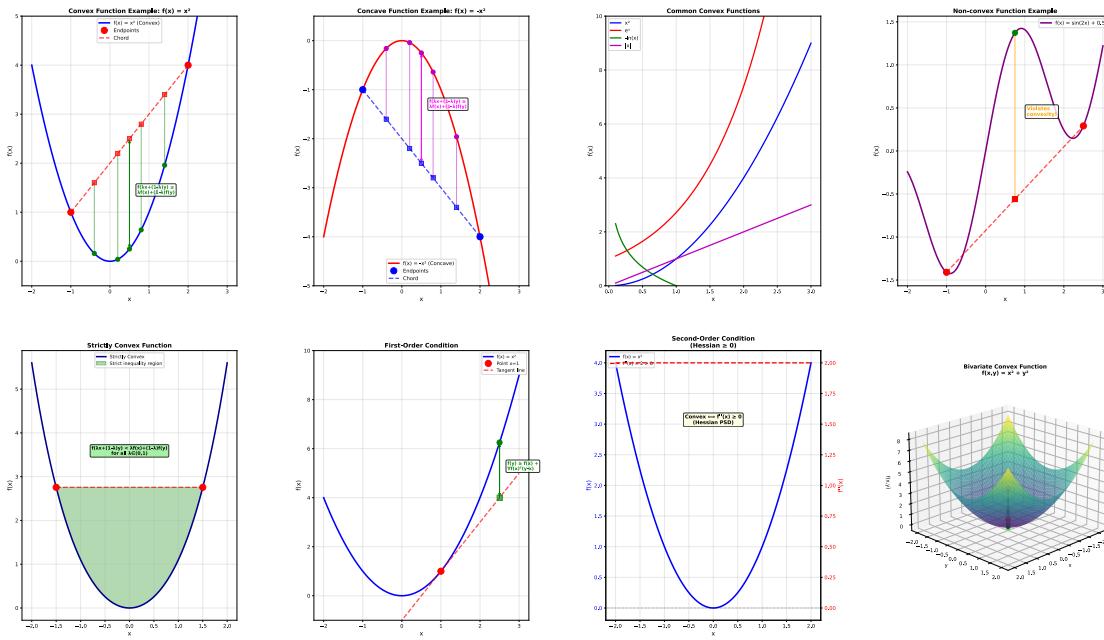


图 2.7 凸函数, 凹函数等

凸函数的重要性质: 局部最小值=全局最小值

判定凸函数的条件:

我们先来看一下一元凸函数

- 一阶条件

对于可微函数, f 是凸函数当且仅当

$$f(y) \geq f(x) + f'(x)(y - x), \forall x, y \in \mathbb{R} \quad (2.90)$$

- 二阶条件

对于二阶可微函数, f 是凸函数 当且仅当:

$$f''(x) \geq 0, \forall x \in \mathbb{R} \quad (2.91)$$

上面两个条件, 满足一个, 就可以判定函数是凸函数。

多元凸函数和一元凸函数很类似

函数 $f: \mathbb{R}^n \rightarrow \mathbb{R}$ 称为凸函数(**convex function**), 如果对于定义域内的任意两点 \mathbf{x}, \mathbf{y} 和任意 $\lambda \in [0, 1]$, 都有:

$$f(\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{y}) \quad (2.92)$$

- 一阶条件

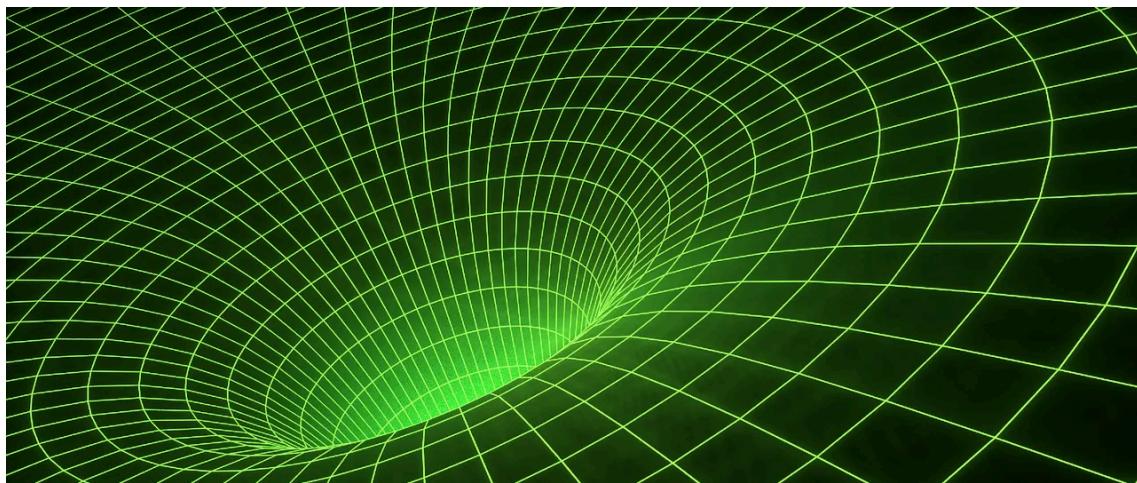
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) \quad (2.93)$$

- 二阶条件

黑塞矩阵 $\nabla^2 f(\mathbf{x})$ 是半正定矩阵。

2.4.3 梯度下降法

梯度下降法奠定了机器学习和深度学习技术的基础。让我们探索它的工作原理、适用场景以及在不同函数中的表现特性。



2.4.3.1 简介

梯度下降是一种迭代式一阶优化算法，用于寻找给定函数的局部最小值/最大值。这种方法在机器学习与深度学习领域广泛用于最小化损失函数（例如线性回归场景）。

我们将会深入探讨一阶梯度下降算法的数学原理、实现方式和行为特性。我们将直接引导自定义的函数来寻找其最小值。

梯度下降法由柯西在 1847 年提出，远早于现代计算机时代。自那时起，计算机科学与数值方法领域取得了长足发展，由此衍生出众多改进版的梯度下降算法。

2.4.3.2 对函数的要求

梯度下降法并不适用于所有函数，有两个特定的要求。函数必须是：

- 可微函数（可以求导的）
- 凸函数

首先，可微是什么意思？如果一个函数是可微的，那么在其定义域内的每个点都有导数——并非所有函数都满足这个标准。首先，我们来看一些满足这个标准的函数示例：

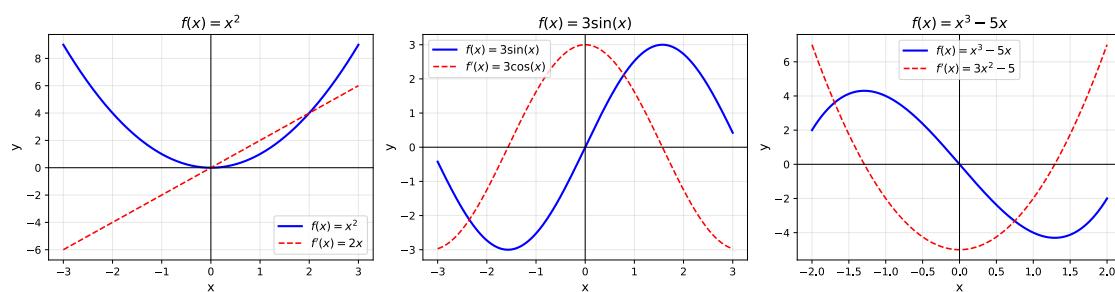


图 2.8 可微函数示例

典型的不可微函数具有阶梯、尖点或不连续点：

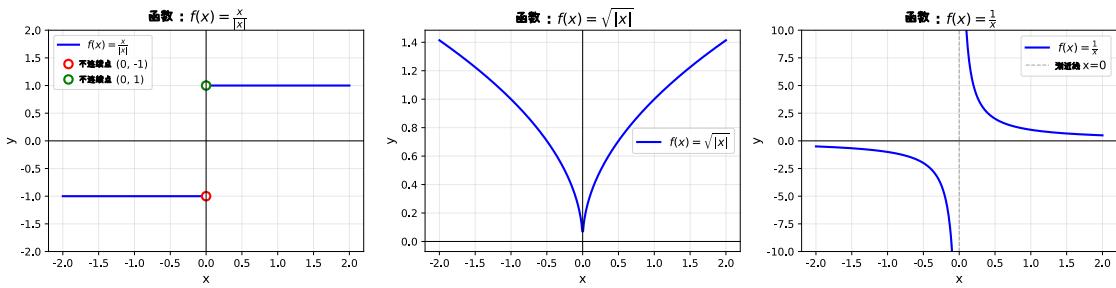


图 2.9 不可微函数示例

下一个要求——函数必须是凸函数。对于一元函数而言，这意味着连接函数上任意两点的线段均位于曲线之上或与之相切(不会穿越曲线)。若线段穿越曲线，则表明函数存在局部最小值而非全局最小值。

数学上，对于位于函数曲线上的两点 x_1 和 x_2 ，一个函数是凸函数的条件可表示为：

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2) \quad (2.94)$$

其中 λ 表示点在截线上的位置，其值必须在0(左侧点)和1(右侧点)之间，例如 $\lambda = 0.5$ 表示中点位置。

以下是两个带有示例截线的函数。

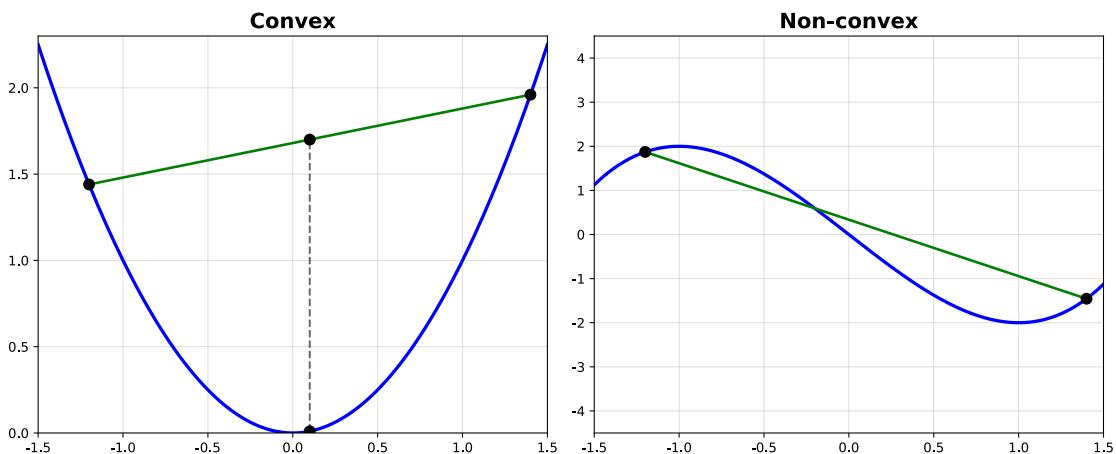


图 2.10 凸函数与非凸函数示例图

判断单变量函数是否为凸函数的另一种数学方法是计算其二阶导数，并检查其值是否始终大于0。

$$\frac{d^2 f(x)}{dx^2} > 0, \text{ 对于定义域内所有的 } x \quad (2.95)$$

我们研究一个由以下公式给出的简单二次函数：

$$f(x) = x^2 - x + 3 \quad (2.96)$$

其一阶导数和二阶导数分别为：

$$\begin{aligned} \frac{df(x)}{dx} &= 2x - 1 \\ \frac{d^2 f(x)}{dx^2} &= 2 \end{aligned} \quad (2.97)$$

由于二阶导数始终大于0，该函数为严格凸函数。

梯度下降法同样可应用于拟凸函数 (**quasi-convex functions**)。然而这类函数常存在所谓鞍点 (**saddle points**)，梯度下降法可能在此陷入停滞。一个拟凸函数的示例如下：

$$\begin{aligned} f(x) &= x^4 - 2x^3 + 2 \\ \frac{df(x)}{dx} &= 4x^3 - 6x^2 = x^2(4x - 6) \\ \frac{d^2f(x)}{dx^2} &= 12x^2 + 12x = 12x(x - 1) \end{aligned} \tag{2.98}$$

我们注意到一阶导数在 $x = 0$ 和 $x = 1.5$ 处为0，这些位置是函数极值点（极小值或极大值）的候选点——该处斜率为零。但首先需要验证二阶导数的情况。

在 $x = 0$ 和 $x = 1.5$ 处，二阶导数的值为0。这些位置被称为拐点——即曲率改变符号的地方——意味着函数从凸变为凹，或反之亦然。通过分析这个方程，我们得出以下结论：

- 当 $x < 0$ 时：函数是凸的
- 当 $0 < x < 1$ 时：函数是凹的（二阶导数 < 0 ）
- 当 $x > 1$ 时：函数再次变为凸的

现在我们看到点 $x = 0$ 处的一阶导数和二阶导数均为0，这表明此处是鞍点，而点 $x = 1.5$ 则是全局最小值点。

我们来看一下这个函数的图像。如前计算，鞍点位于 $x = 0$ 处，最小值点位于 $x = 1.5$ 处。

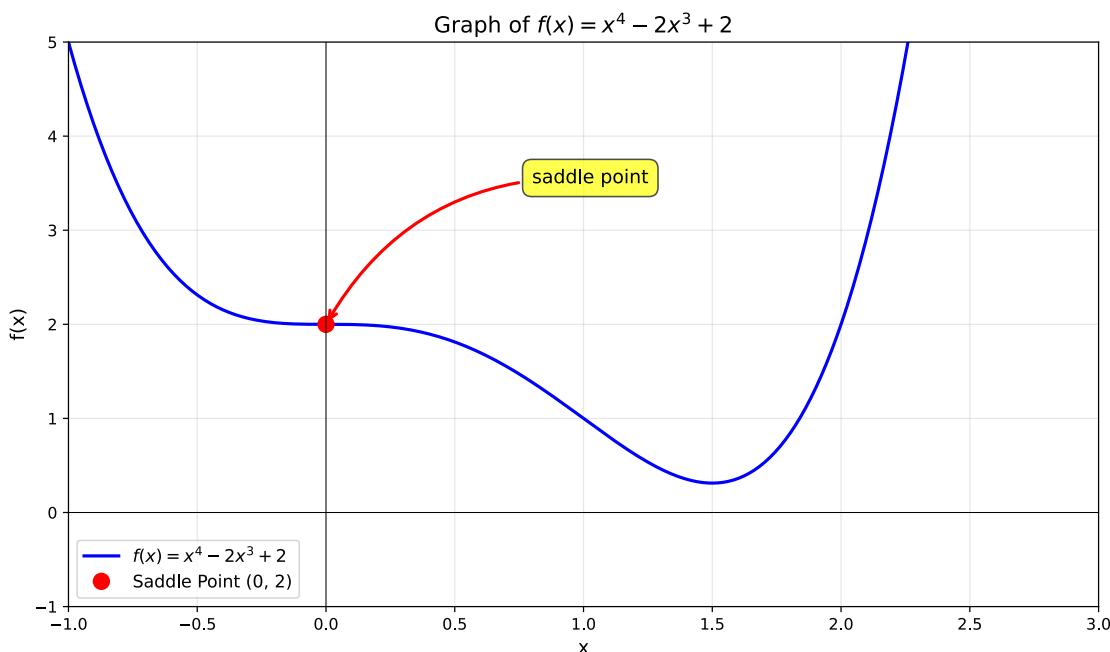
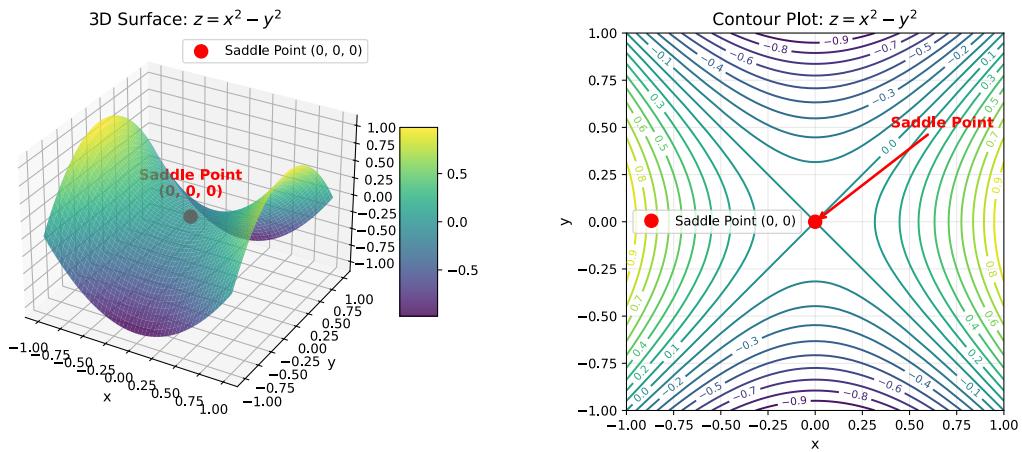


图 2.11 具有鞍点的半凸 (**semi-convex**) 函数

对于多变量函数，判断某点是否为鞍点的最合适方法是计算 **Hessian** 矩阵。

一个双变量函数 $z = x^2 - y^2$ 的鞍点的示例如下图所示。

图 2.12 $z = x^2 - y^2$ 的鞍点示意图

2.4.3.3 梯度

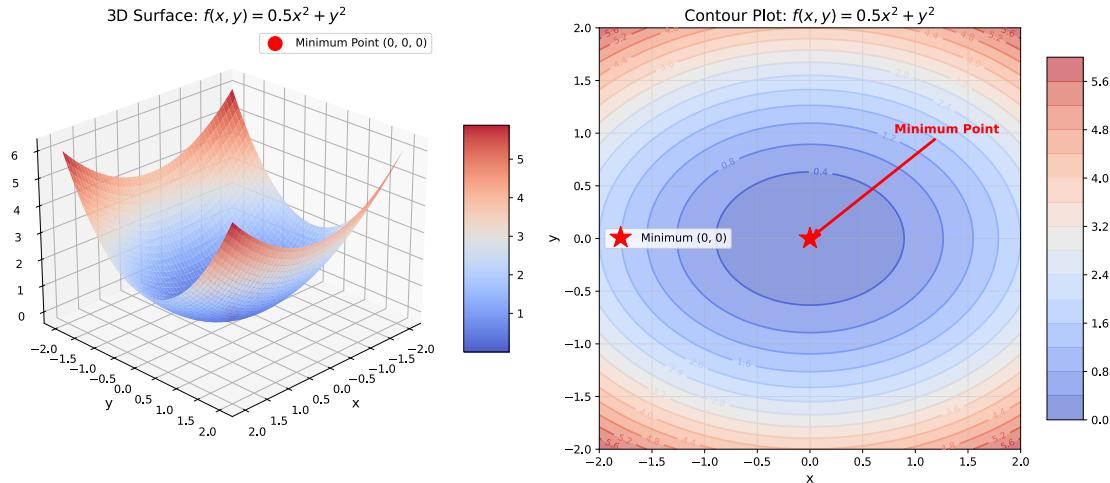
直观地说，梯度代表了在指定方向上某一点处曲线的斜率。

对于单变量函数来说，它就是选定点的一阶导数。对于多变量函数而言，梯度是沿各主方向（顺变量坐标轴）的导数向量。因为我们只关心沿某一坐标轴的斜率，而不在意其他方向的变化，所以这些导数被称为偏导数。

n 维函数 $f(x)$ 在给定点 p 处的梯度定义如下：

$$\nabla f(p) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(p) \\ \vdots \\ \frac{\partial f}{\partial x_n}(p) \end{bmatrix} \quad (2.99)$$

倒三角形 ∇ 就是所谓的“nabla”符号。为了更好地理解如何计算梯度，让我们为下面这个二维示例函数 $f(x) = 0.5x^2 + y^2$ 进行手动计算。

图 2.13 $f(x) = 0.5x^2 + y^2$ 示意图

假设我们关注点 $p(10, 10)$ 处的梯度：

$$\begin{aligned}\frac{\partial f(x, y)}{\partial x} &= x \\ \frac{\partial f(x, y)}{\partial y} &= 2y\end{aligned}\tag{2.100}$$

因此可得：

$$\begin{aligned}\nabla f(x, y) &= \begin{bmatrix} x \\ 2y \end{bmatrix} \\ \nabla f(10, 10) &= \begin{bmatrix} 10 \\ 20 \end{bmatrix}\end{aligned}\tag{2.101}$$

观察这些数值可知，沿 y 轴方向的斜率是 x 轴方向的两倍。

2.4.3.4 梯度下降法

梯度下降法迭代地利用当前位置的梯度计算下一个点，通过学习率进行缩放，并从当前位置减去所得值（即执行一步移动）。之所以减去该值，是因为我们想要最小化函数（若要最大化则应相加）。这一过程可以写作：

$$p_{n+1} = p_n - \eta \nabla f(p_n)\tag{2.102}$$

存在一个关键参数 η ，它通过缩放梯度来控制步长大小。在机器学习中，该参数被称为学习率，对算法性能具有重要影响。

- 学习率越小，梯度下降收敛所需时间越长，甚至可能在达到最优解前触及最大迭代次数限制。
- 若学习率过大，算法可能无法收敛至最优点（持续震荡），甚至可能完全发散。

总而言之，梯度下降法的步骤包括：



图 2.14 梯度下降法步骤

梯度下降法代码实现	
Python	
1 import numpy as np	
2 from typing import Callable	
3	
4 def gradient_descent(
5 start: float, # 起始点	
6 gradient: Callable[[float], float], # 计算梯度的函数	
7 learn_rate: float, # 学习率 η	
8 max_iter: int, # 最大迭代次数	
9 tol: float = 0.01 # 步长的阈值	

```

10  ):
11      x = start
12      steps = [start] # 历史跟踪
13
14      for _ in range(max_iter):
15          diff = learn_rate * gradient(x)
16          if np.abs(diff) < tol:
17              break
18          x = x - diff
19          steps.append(x) # 历史跟踪
20
21      return steps, x

```

这个函数接收 5 个参数：

1. 起始点——我们这里手动定义了起始点，但在实践中，起始点通常是随机初始化的。
2. 梯度函数——计算梯度的函数（需要实现好然后传给上面的函数）
3. 学习率——步长的缩放因子
4. 最大迭代次数——循环次数
5. 阈值——算法停止的一个条件（这里默认是 0.01）

2.4.3.5 示例 1——二次函数

我们的二次函数为

$$f(x) = x^2 - 4x + 1 \quad (2.103)$$

由于是单变量函数，所以梯度函数为

$$\frac{df(x)}{dx} = 2x - 4 \quad (2.104)$$

写成代码如下

```

1 def func1(x: float):
2     return x ** 2 - 4 * x + 1
3
4 def gradient_func1(x: float):
5     return 2 * x - 4

```

Python

当选择起始点 $x = 9$ 以及学习率为 0.1 时，我们可以手动计算一下每一步的过程。例如前三步如下：

$$\begin{aligned}
 x_0 &= 9 \\
 x_1 &= 9 - 0.1 \times (2 \times 9 - 4) = 7.6 \\
 x_2 &= 7.6 - 0.1 \times (2 \times 7.6 - 4) = 6.48 \\
 x_3 &= 6.48 - 0.1 \times (2 \times 6.48 - 4) = 5.584
 \end{aligned} \quad (2.105)$$

代码如下

```
1 history, result = gradient_descent(9, gradient_func1, 0.1, 100)
```

Python

如图所示，对于较小的学习率，随着算法逼近最小值，步长逐渐变小。而较大的学习率则在收敛前在两侧来回跳跃。

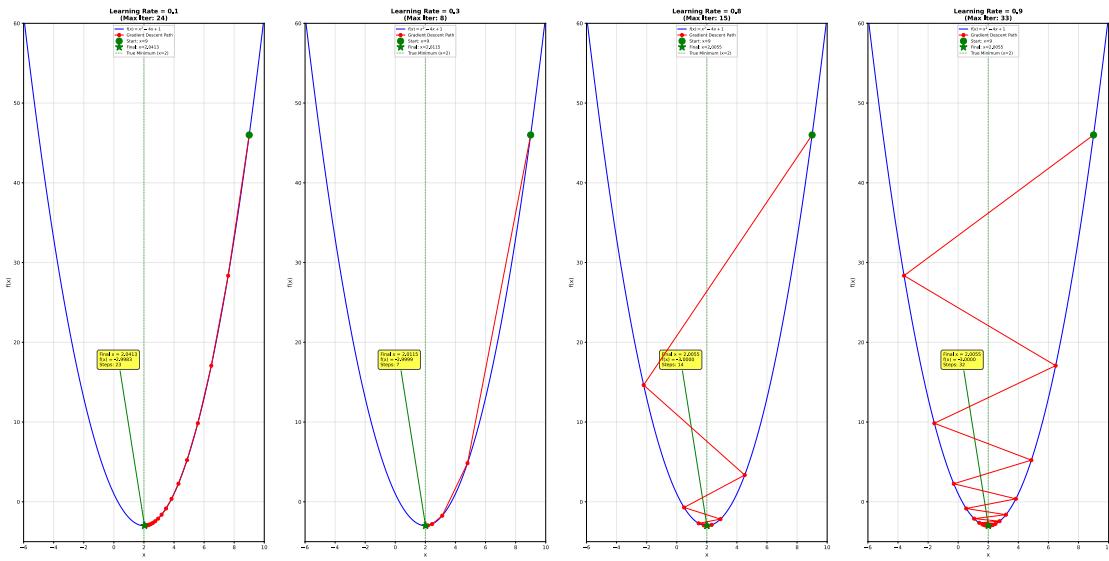


图 2.15 不同学习率的对比

2.4.3.6 示例 2——包含鞍点的函数

现在让我们看看算法将如何处理我们先前进行数学分析的半凸函数。

$$f(x) = x^4 - 2x^3 + 2 \quad (2.106)$$

写成代码如下：

```
1 def func2(x: float):
2     return x ** 4 - 2 * x ** 3 + 2
3
4 def gradient_func2(x: float):
5     return 4 * x ** 3 - 6 * x
```

Python

下方展示了两种学习率与两种不同起始点的运算结果。

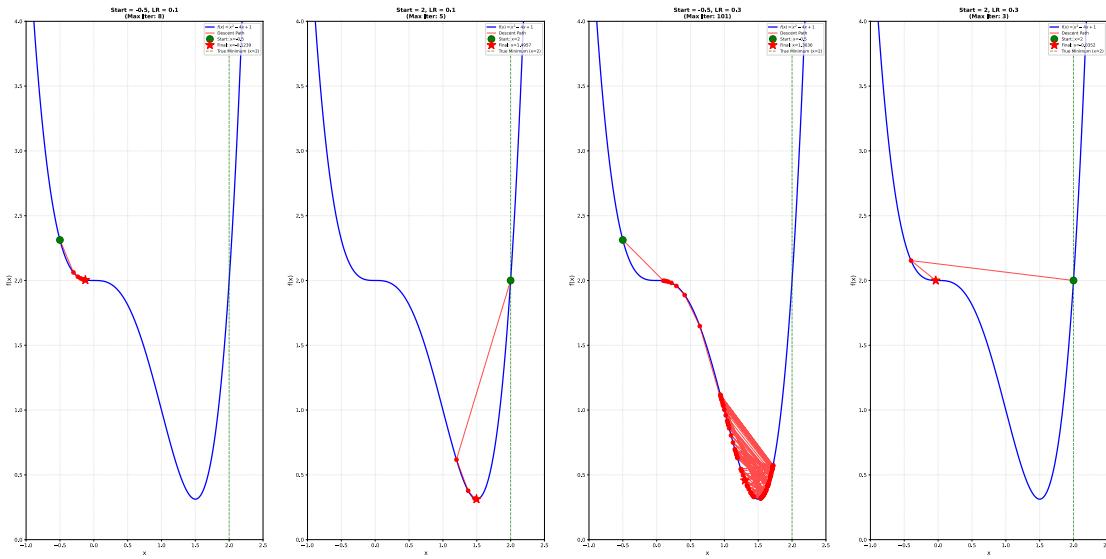


图 2.16 梯度下降法尝试逃离鞍点示意图

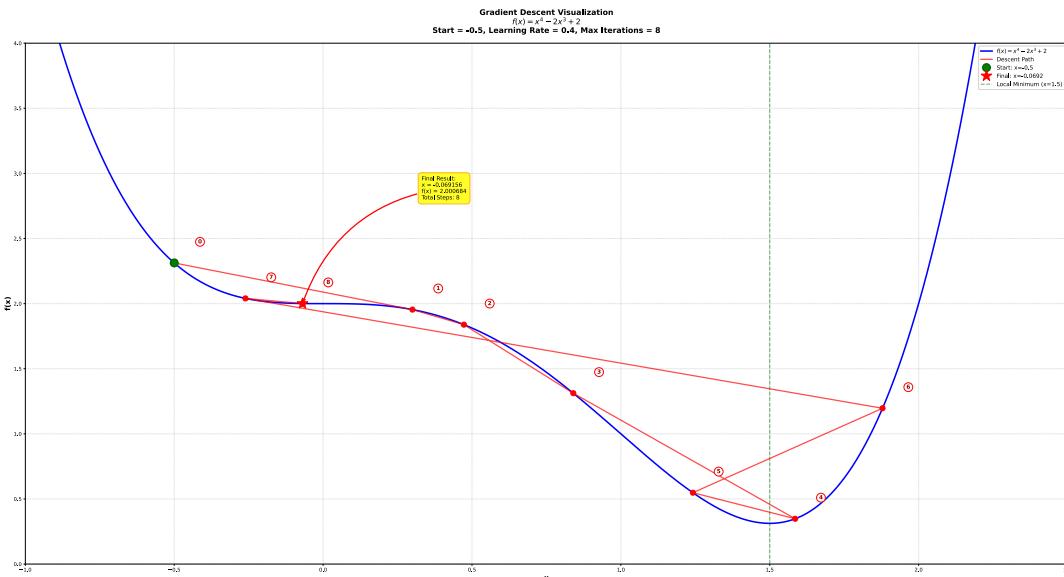


图 2.17 没有逃离鞍点

现在可以看到，鞍点的存在确实给一阶梯度下降法带来了严峻挑战，且无法保证最终能达到全局最小值。二阶优化算法（如牛顿法、拟牛顿法）在此类情况下的表现则更为出色。

我们探讨了梯度下降法的运作机制、适用场景以及使用过程中常见的挑战。后面我们会进一步探索更先进的基于梯度的优化方法，例如动量法、Nesterov 加速梯度下降、RMSprop、Adam，或是牛顿法等二阶优化方法。

2.5 自动微分

2.5.1 数值微分和符号微分的缺点

2.5.1.1 数值微分的缺点

数值微分其实是有优点的，那就是写程序很容易实现。

加入我们有一个 n 元函数 $f(x_0, x_1, \dots, x_{n-1})$ ，要对所有变量求偏导数，也就是梯度

```

1 import copy
2
3 def grad(f, x, i):
4     h = 1e-4
5     delta_x = copy.deepcopy(x)
6     delta_x[i] = x[i] + h
7     delta_y = f(delta_x) - f(x)
8     return delta_y / h
9
10 # f(x_0, x_1) = x_0^2 + x_1^3 + x_0 x_1
11 def f(x):
12     return x[0]**2 + x[1]**3 + x[0] * x[1]
13
14 x = [1.1, 2.2]
15 grad_f = [grad(f, x, 0), grad(f, x, 1)]
16 print(grad_f)
17 # 手动求解
18 print([2 * x[0] + x[1], 3 * x[1]**2 + x[0]])

```

精度问题：

受浮点数精度限制

- 步长 h 太小 → 舍入误差 (roundoff error)
- 步长 h 太大 → 截断误差 (truncation error)
- 难以找到最优步长

计算效率：

- 需要多次函数求值
- 对于 n 维函数，需要 $O(n)$ 次函数调用
- 高阶导数计算代价高

数值不稳定：

- 对病态函数 (ill-conditioned) 敏感
- 可能产生数值噪声

2.5.1.2 符号微分的缺点

例如我们对函数 $f(x) = \frac{x^2 - 1}{x - 1}$ 进行求导

```

1 import sympy as sp
2 x = sp.Symbol("x")
3 f_complex = (x**2 - 1) / (x - 1)
4 f_derivative = sp.diff(f_complex, x)
5 print(f_derivative)

```

使用上面的程序进行求导得到了如下

$$f'(x) = \frac{2x}{x-1} - \frac{x^2-1}{(x-1)^2} \quad (2.107)$$

但是如果我们人类求导的话，会先观察出

$$f(x) = \frac{x^2-1}{x-1} = x+1 \quad (2.108)$$

所以

$$f'(x) = 1 \quad (2.109)$$

符号微分的缺点如下：

- 表达式膨胀 (Expression Swell):
 - 导数表达式可能变得非常复杂
 - 重复子表达式未被优化
 - 内存消耗大
- 计算效率低：
 - 生成的表达式可能包含大量冗余计算
 - 未经优化的符号表达式求值慢
- 不适用于所有函数：
 - 某些函数没有解析形式（如条件语句、循环）
 - 无法处理数值算法（如迭代求解器）
- 实现复杂：
 - 需要复杂的符号操作系统
 - 对于大型程序难以应用

所以我们需要寻找其它自动微分的方法，那就是深度学习的核心算法：大名鼎鼎的反向传播算法。

2.5.2 反向传播算法

反向传播算法是一种求解导数（微分）方法。

我们先来举一个简单的例子，那就是求解 $y = (x+1)^2$ 的导数，在纸上经过辛苦的计算，我们能够知道 $\frac{dy}{dx} = 2(x+1)$ 。

由于手工计算太过于辛苦，所以我们决定使用反向传播算法。

反向传播算法分为两个阶段：

- 前向过程 (forward): 保存中间计算结果
- 反向过程 (backward): 利用链式求导法则和保存的中间计算结果来求解导数。

假设我们要求解在 $x = 2$ 处的导数。通过解析解我们知道导数为 6。

1. 前向过程

$$\begin{aligned} u &= x + 1 = 3 \\ v &= u^2 = 3^2 = 9 \\ y &= v \end{aligned} \quad (2.110)$$

上面的 u, v 是中间计算结果。

2. 反向过程

$$\begin{aligned}
 \frac{dy}{dx} &= \frac{dy}{dv} \frac{dv}{du} \frac{du}{dx} \\
 &= 1 \cdot 2u \cdot 1 \quad (\text{v 在前向过程中保存了}) \\
 &= 2 \times 3 = 6
 \end{aligned} \tag{2.111}$$

正向过程和反向过程的计算图 (computational graph) 如下所示：

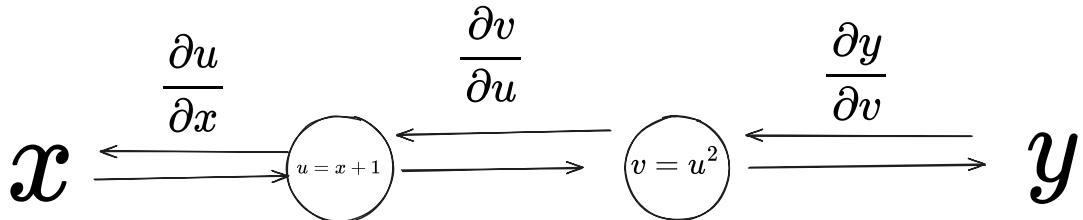


图 2.18 计算图

可以看到，计算图中的反向过程的偏导数相乘就得到了 x 的偏导数。

代码实现如下：

```

1 def forward(x):
2     u = x + 1
3     v = u * u
4     return x, u, v
5
6 def backward(x, u, v):
7     return 2 * u
8
9 x = 2
10 x, u, v = forward(x)
11 grad = backward(x, u, v)
12 print(grad)
  
```

再来看一个例子，我们要求 $y = (x_0 + x_1)(x_1 + 1)$ 在 $(x_0 = 2, x_1 = 1)$ 点的梯度。

还是可以绘制计算图如下：

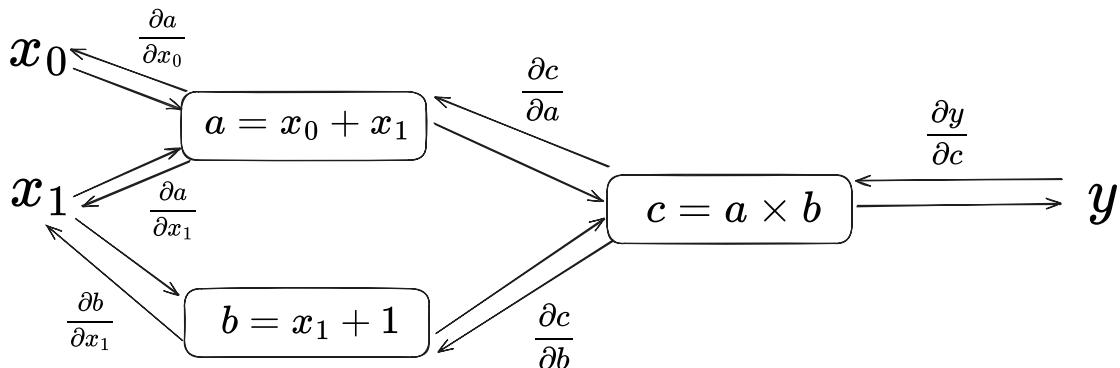


图 2.19 计算图

对应到链式求导公式和全微分公式

$$\begin{aligned}\frac{\partial y}{\partial x_0} &= \frac{\partial y}{\partial c} \cdot \frac{\partial c}{\partial a} \cdot \frac{\partial a}{\partial x_0} \\ \frac{\partial y}{\partial x_1} &= \frac{\partial y}{\partial c} \cdot \frac{\partial c}{\partial a} \cdot \frac{\partial a}{\partial x_1} + \frac{\partial y}{\partial c} \cdot \frac{\partial c}{\partial b} \cdot \frac{\partial b}{\partial x_1}\end{aligned}\tag{2.112}$$

写成代码如下：

```
1 def forward(x0, x1):
2     a = x0 + x1
3     b = x1 + 1
4     c = a * b
5     return a, b, c
6
7 def backward(a, b, c):
8     x0_grad = b
9     x1_grad = b + a
10    return x0_grad, x1_grad
11
12 x0, x1 = 2, 1
13 a, b, c = forward(x0, x1)
14 x0_grad, x1_grad = backward(a, b, c)
15 print(x0_grad, x1_grad)
```

Python

2.6 概率论

2.6.1 概率

2.6.1.1 概率的概念

概率是对事件发生的可能性的度量。通常将事件A的概率写作 $P(A)$ 。

2.6.1.2 概率的计算

事件	概率
A	$P(A) \in [0, 1]$
非A	$P(\bar{A}) = 1 - P(A)$
A和B (联合概率)	$P(A \cap B) = P(A, B) = P(A B)P(B) = P(B A)P(A)$ 当A和B相互独立时, $P(A \cap B) = P(A) \cdot P(B)$
A或B	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ 当A和B互斥时, $P(A \cup B) = P(A) + P(B)$
B情况下A的概率 (条件概率 $P(A B)$)	$P(A B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B A)P(A)}{P(B)}$

表 2.3 概率的计算

例如：现有一个装有10个球的袋子，其中有6个红球和4个蓝球。从中随机抽取两个球。我们定义以下事件：

- 事件A：第一个抽到的是红球。
- 事件B：两个抽到的球都是红球。

1. 计算联合概率 $P(A \cap B)$

第一个球是红球的概率：

$$P(A) = \frac{6}{10} \quad (2.113)$$

在第一个球是红球的情况下，两个球都是红球的概率：

$$P(B|A) = \frac{5}{9} \quad (2.114)$$

联合概率

$$P(A \cap B) = P(B|A)P(A) = \frac{5}{9} \times \frac{6}{10} = \frac{1}{3} \quad (2.115)$$

2. 计算条件概率 $P(A|B)$

条件概率表示在已知两个球都是红球的情况下，第一个球是红球的概率。

两个球都是红球的概率：

$$P(B) = \frac{C_6^2}{C_{10}^2} = \frac{6 \times 5 \div 2}{10 \times 9 \div 2} = \frac{1}{3} \quad (2.116)$$

在两个球都是红球的情况下，第一个球是红球的概率：

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{3}}{\frac{1}{3}} = 1 \quad (2.117)$$

2.6.2 概率分布

概率分布，是指用于表述随机变量取值的概率规律。事件的概率表示了一次试验中某一个结果发生的可能性大小。如果试验结果用变量 X 的取值来表示，则随机试验的概率分布就是随机变量的概率分布，即随机变量的可能取值及取得对应值的概率。

2.6.2.1 期望、方差和标准差

期望

离散型随机变量的期望定义如下

$$\begin{aligned} \mathbb{E}[X] &= \sum_i x_i \cdot p(x_i) \\ \mathbb{E}[f(X)] &= \sum_i f(x_i) \cdot p(x_i) \end{aligned} \quad (2.118)$$

- x_i : 可能的取值
- $p(x_i)$: 事件 x_i 发生的概率

例子：掷骰子

$$\mathbb{E}[X] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5 \quad (2.119)$$

连续型随机变量的期望定义如下

$$\begin{aligned}\mathbb{E}[X] &= \int_{-\infty}^{\infty} x \cdot p(x) dx \\ \mathbb{E}[f(X)] &= \int_{-\infty}^{\infty} f(x) \cdot p(x) dx\end{aligned}\tag{2.120}$$

- $p(x)$: 概率密度函数

方差

方差是衡量数据离散程度的统计量，表示数据与其平均值之间偏离程度的平均值。离散型随机变量 X 的方差是其取值与期望值偏离程度的加权平均。

$$\begin{aligned}\text{Var}[X] &= \sum_{i=1}^n (x_i - \mu)^2 p(x_i) \\ &= \mathbb{E}[(X - \mu)^2] \\ \text{Var}[X] &= \sum_{i=1}^n x_i^2 p(x_i) - \left(\sum_{i=1}^n x_i p(x_i) \right)^2 \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2\end{aligned}\tag{2.121}$$

- μ 是随机变量的期望值 $\mathbb{E}[X]$

标准差的定义

$$\sigma = \sqrt{\text{Var}[X]}\tag{2.122}$$

连续型随机变量的方差定义

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx\tag{2.123}$$

其它性质同离散型随机变量的性质。

2.6.3 均匀分布

离散均匀分布是指所有可能结果出现的概率都相等的离散概率分布。

对于取值范围 $\{a, a+1, a+2, \dots, b\}$ 的离散均匀分布：

- 概率质量函数 (PMF)：

$$P(X = k) = \frac{1}{b-a+1}, k \in \{a, a+1, \dots, b\}\tag{2.124}$$

- 期望值：

$$\mathbb{E}[X] = \frac{a+b}{2}\tag{2.125}$$

- 方差：

$$\text{Var}[X] = \frac{(b-a+1)^2 - 1}{12}\tag{2.126}$$

连续型均匀分布定义如下

- 概率密度函数

$$p(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{else} \end{cases}\tag{2.127}$$

- 期望值:

$$\begin{aligned}\mathbb{E}[X] &= \int_a^b x \cdot \frac{1}{b-a} dx = \frac{a+b}{2} \\ \mathbb{E}[X^2] &= \int_a^b x^2 \cdot \frac{1}{b-a} dx = \frac{a^2 + ab + b^2}{3}\end{aligned}\tag{2.128}$$

- 方差和标准差:

$$\begin{aligned}\text{Var}[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{(b-a)^2}{12} \\ \sigma &= \frac{b-a}{2\sqrt{3}}\end{aligned}\tag{2.129}$$

2.6.4 二项分布（伯努利分布）

2.6.5 多项分布

2.6.6 正态分布（高斯分布）

2.7 信息论

概率论为另一个重要的框架——信息论 (**information theory**) 提供了基础。它量化了数据集中存在的信息，并在机器学习中扮演着重要的角色。

2.7.1 熵

考虑一个离散型随机变量 x ，我们想知道当观察到这个变量的某个特定值时能获得多少信息。信息量可以视为我们在得知 x 的具体值时的“惊讶程度”。如果我们被告知一个极不可能发生的事件发生了，那么我们接收到的信息将多于某个极有可能发生的事件刚刚发生的信息。如果我们知道某个事件确定会发生，那么我们不会接收到任何信息。信息内容的度量将依赖于概率分布 $p(x)$ 。我们需要寻找一个量 $h(x)$ ，它是概率 $p(x)$ 的单调函数，并且表达了信息的内容。注意，如果两个事件 x 和 y 是无关的，那么观察它们所获得的信息增益应该是分别观察它们所获得的信息增益之和，即 $h(x, y) = h(x) + h(y)$ 。两个无关的事件是统计独立的，所以 $p(x, y) = p(x)p(y)$ 。有这两个关系可以知道 $h(x)$ 必须由 $p(x)$ 的对数给出。所以有：

$$h(x) = -\log_2 p(x)\tag{2.130}$$

其中负号确保了信息是正的或者为零。注意低概率事件 x 对应于高信息内容。对数的底数选取是任意的，此刻我们采用信息论中的惯例，即以 2 为底数。在这种情况下，我们可以看到 $h(x)$ 的单位是比特（“二进制位”）。

现在假设发送者希望向接收者传输一个随机变量的值。他们在这个过程中传输的平均信息量是通过对式 (2.130) 关于分布 $p(x)$ 取期望得到的，有下面的式子给出

$$H[x] = -\sum_x p(x) \log_2 p(x)\tag{2.131}$$

这个重要的量称为随机变量 x 的熵 (**entropy**)。注意 $\lim_{\varepsilon \rightarrow 0} (\varepsilon \ln \varepsilon)$ ，所以当遇到 x 在某个取值下使 $p(x) = 0$ 时，令 $p(x) \ln p(x) = 0$ 。

到目前为止，我们已经启发式地给出了信息 [式 (2.130)] 和信息的熵 [式 (2.131)] 的定义。这些定义确实具有一些有用的性质。考虑一个有 8 个可能状态的随机变量 x 。为了将 x 的值传输给接收者，我们需要传输一条长度为 3 比特的信息。注意，这个变量的熵由下面的式子给出：

$$H[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3(\text{比特})\tag{2.132}$$

举个例子，一个变量有 8 个可能的状态 $\{a, b, c, d, e, f, g, h\}$ ，各自的概率分别为 $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$ 。这种情况下的熵由下面的式子给出：

$$H[x] = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{16} \log_2 \frac{1}{16} - 4 \times \frac{1}{64} \log_2 \frac{1}{64} = 2(\text{比特}) \quad (2.133)$$

可以看出，非均匀分布的熵小于均匀分布的熵。考虑如何将变量的状态传输给接收者。我们既可以像之前那样使用一个 3 比特的数字，也可以使用较短编码来表示较大可能发生的事件，或者用较长编码表示较小可能发生的事件，从而利用非均匀分布的特点，实现较短的平均编码长度。这一目的可以通过使用以下一组代码字符串表示状态 $\{a, b, c, d, e, f, g, h\}$ 来实现，例如 0, 10, 110, 1110, 111100, 111101, 111110, 111111。必须传输的代码的平均长度为

$$\text{代码平均长度} = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{16} \times 4 + 4 \times \frac{1}{64} \times 6 = 2(\text{比特}) \quad (2.134)$$

这与随机变量的熵相同。注意，

3. 一元线性回归

机器学习使用数据来解决问题。不是由人来思考问题的解决方案，而是让计算机从收集的数据中找到（学习）问题的解决方案。机器学习的本质就是从数据中寻找解决方案。从现在开始，我们将挑战机器学习问题。本章将实现机器学习中最基本的线性回归。

3.1 玩具数据集

在本步骤，我们将创建一个用于实验的小型数据集。这个小型数据集称为玩具数据集(**toy datasets**)。考虑到重现性，我们用固定的随机种子创建数据，具体代码如下。

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3
4 np.random.seed(0) # 固定随机数种子
5 x = np.random.rand(100, 1) # 形状: 100 × 1
6 #  $y = 5 + 2x + \epsilon$ 
7 # 噪声 $\epsilon$ 的形状是100 × 1
8 #  $y$ 的形状也是100 × 1
9 y = 5 + 2 * x + np.random.rand(100, 1)
10
11 # Plot
12 plt.scatter(x, y, s=10)
13 plt.xlabel("x")
14 plt.ylabel("y")
15 plt.show()
```

上面的代码创建了一个由变量 x 和 y 组成的数据集。这些数据点沿直线分布，是在 y 上增加作为噪声的随机数得到的。下图展示了这些 (x, y) 数据点的分布情况。

可视化如下

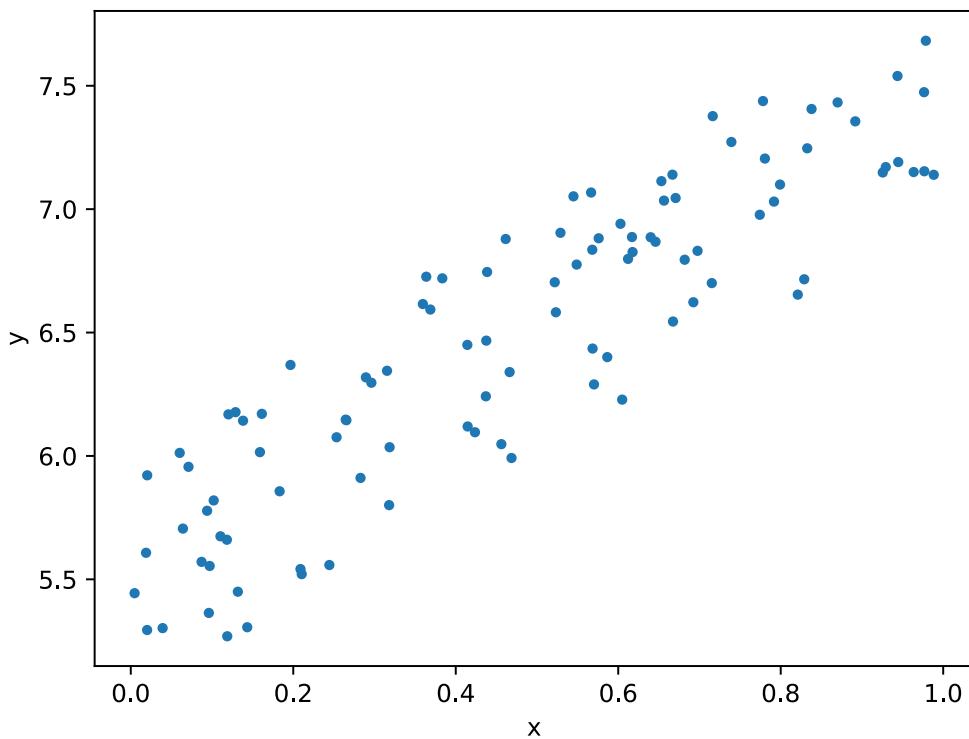


图 3.1 使用的玩具数据集

如图所示，虽然 x 和 y 之间呈线性关系，但数据中存在噪声。我们的目标是创建根据 x 值预测 y 值的模型（式子）。

根据 x 值预测实数值 y 的做法叫作回归（regression）。另外，当预测模型呈线性（直线）时，这种回归分析称为线性回归。

3.2 线性回归的理论知识

接下来的目标是找到拟合给定数据的函数。假设 y 和 x 之间的关系是线性的，函数的式子就可以表示为 $y = Wx + b$ （其中 W 是标量）。 $y = Wx + b$ 这条直线如图所示。

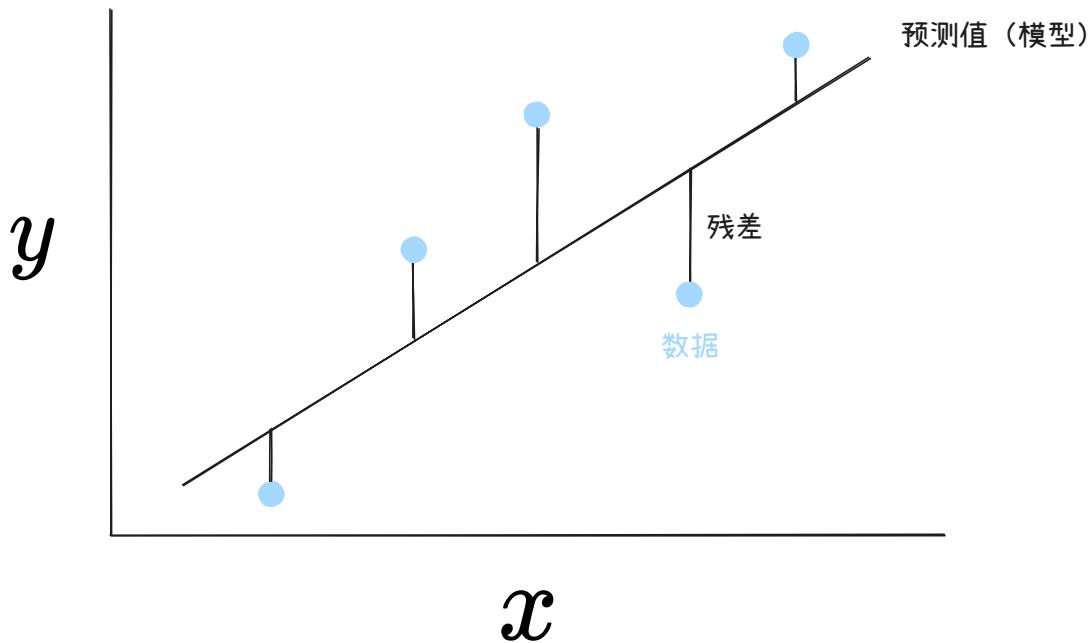


图 3.2 线性回归的示例

如上图所示，我们的目标是找到一条拟合数据的直线 $y = Wx + b$ 。为此，我们需要尽可能地减小数据和预测值之间的差，这个差叫作残差（**residual**）。下面是表示预测值（模型）和数据之间的误差指标的式子。

$$\mathcal{L} = \frac{1}{N} \sum_{i=0}^{N-1} (f(x_i) - y_i)^2 \quad (3.1)$$

在式子中，先求出这 N 个点中的每个点 (x_i, y_i) 的平方误差，然后将它们加起来，之后乘以 $\frac{1}{N}$ 求出平均数。这个式子叫作均方误差（**mean squared error**）。另外，在式子中求平均数时乘的是 $\frac{1}{N}$ ，但在某些情况下，会乘以 $\frac{1}{2N}$ 。但无论哪种情况，在用梯度下降法求解时，都可以通过调整学习率的值来解决同样的问题。

🔥 损失函数

评估模型好坏的函数叫作损失函数（**loss function**）。此时，我们可以说线性回归使用均方误差作为损失函数。

我们的目标是找到使式子表示的损失函数的输出最小的 W 和 b 。这就是函数优化问题。此处使用梯度下降法来找到使式子最小化的参数。

⚡ 梯度下降法

梯度下降法是深度学习中最重要的优化方法，我们会反复讲解！

3.3 线性回归的实现

首先我们来求解损失函数针对参数的梯度 ∇ ，读作“**nabla**”。

$$\nabla = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial W} \\ \frac{\partial \mathcal{L}}{\partial b} \end{bmatrix} \quad (3.2)$$

接下来我们分别求偏导数

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial W} &= \frac{\partial \left\{ \frac{1}{N} \sum_{i=0}^{N-1} (f(x_i) - y_i)^2 \right\}}{\partial W} \\ &= \frac{\partial \left\{ \frac{1}{N} \sum_{i=0}^{N-1} (Wx_i + b - y_i)^2 \right\}}{\partial W} \\ &= \frac{\partial \left\{ \frac{1}{N} \left[(Wx_0 + b - y_0)^2 + (Wx_1 + b - y_1)^2 + \dots + (Wx_{N-1} + b - y_{N-1})^2 \right] \right\}}{\partial W} \\ &= \frac{1}{N} \left[\frac{\partial (Wx_0 + b - y_0)^2}{\partial W} + \dots + \frac{\partial (Wx_{N-1} + b - y_{N-1})^2}{\partial W} \right] \\ &= \frac{1}{N} [2(Wx_0 + b - y_0) \cdot x_0 + \dots + 2(Wx_{N-1} + b - y_{N-1}) \cdot x_{N-1}] \\ &= \frac{2}{N} \sum_{i=0}^{N-1} (Wx_i + b - y_i) \cdot x_i \end{aligned} \quad (3.3)$$

以及

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial b} &= \frac{\partial \left\{ \frac{1}{N} \sum_{i=0}^{N-1} (f(x_i) - y_i)^2 \right\}}{\partial b} \\ &= \frac{\partial \left\{ \frac{1}{N} \sum_{i=0}^{N-1} (Wx_i + b - y_i)^2 \right\}}{\partial b} \\ &= \frac{\partial \left\{ \frac{1}{N} \left[(Wx_0 + b - y_0)^2 + (Wx_1 + b - y_1)^2 + \dots + (Wx_{N-1} + b - y_{N-1})^2 \right] \right\}}{\partial b} \\ &= \frac{1}{N} \left[\frac{\partial (Wx_0 + b - y_0)^2}{\partial b} + \dots + \frac{\partial (Wx_{N-1} + b - y_{N-1})^2}{\partial b} \right] \\ &= \frac{1}{N} [2(Wx_0 + b - y_0) \cdot 1 + \dots + 2(Wx_{N-1} + b - y_{N-1}) \cdot 1] \\ &= \frac{2}{N} \sum_{i=0}^{N-1} (Wx_i + b - y_i) \end{aligned} \quad (3.4)$$

所以损失函数针对参数的梯度为

$$\nabla = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial W} \\ \frac{\partial \mathcal{L}}{\partial b} \end{bmatrix} = \begin{bmatrix} \frac{2}{N} \sum_{i=0}^{N-1} (Wx_i + b - y_i) \cdot x_i \\ \frac{2}{N} \sum_{i=0}^{N-1} (Wx_i + b - y_i) \end{bmatrix} \quad (3.5)$$

使用梯度下降法来寻找损失函数的最小值的算法如下：

$$\begin{aligned} W &\leftarrow W - \alpha \frac{\partial \mathcal{L}}{\partial W} \\ b &\leftarrow b - \alpha \frac{\partial \mathcal{L}}{\partial b} \end{aligned} \quad \begin{array}{l} \text{学习率} \\ \text{学习率} \end{array} \quad (3.6)$$

图 3.3 梯度下降法

在使用梯度下降法时，我们首先需要为参数 W 和 b 选择初始值，我们这里选择 $W = 0$ 以及 $b = 0$ 。

所以有如下代码：

```
1 W = 0
2 b = 0
3
4 def predict(x):
5     """根据输入x做出预测"""
6     y = W * x + b
7     return y
```

 Python

接下来我们使用代码实现损失函数

```
1 def mean_squared_error(x0, x1):
2     diff = x0 - x1
3     return np.sum(diff ** 2) / len(diff)
```

 Python

然后我们使用代码计算参数的梯度

```
1 def gradient(x, y, W, b):
2     N = len(x)
3     W_grad = 2 / N * sum(
4         (W * xi[0] + b - yi[0]) * xi[0] for (xi, yi) in zip(x, y)
5     )
6     b_grad = 2 / N * sum(
7         (W * xi[0] + b - yi[0]) for (xi, yi) in zip(x, y)
8     )
9     return W_grad, b_grad
```

 Python

有了梯度之后，我们使用梯度下降法更新 100 轮参数，学习率设置为 0.1

```
1 lr = 0.1
2 iters = 100
3
4 for i in range(iters):
5     y_pred = predict(x)
6     loss = mean_squared_error(y, y_pred)
7
8     W_grad, b_grad = gradient(x, y, W, b)
9     W = W - lr * W_grad
10    b = b - lr * b_grad
11    print(W, b, loss)
12
13 # 可视化最后拟合出的直线以及训练数据
14 plt.scatter(x, y, s=10)
15 plt.xlabel("x")
16 plt.ylabel("y")
17 y_pred = predict(x)
18 plt.plot(x, y_pred, color="r")
19 plt.show()
```

 Python

可以看到 `loss` 一直在下降，最后得到的参数是

$$\begin{aligned} W &= 2.118073690511974 \\ b &= 5.466089050922982 \end{aligned} \tag{3.7}$$

如果将直线 $y = Wx + b$ 画出来，可视化如下：

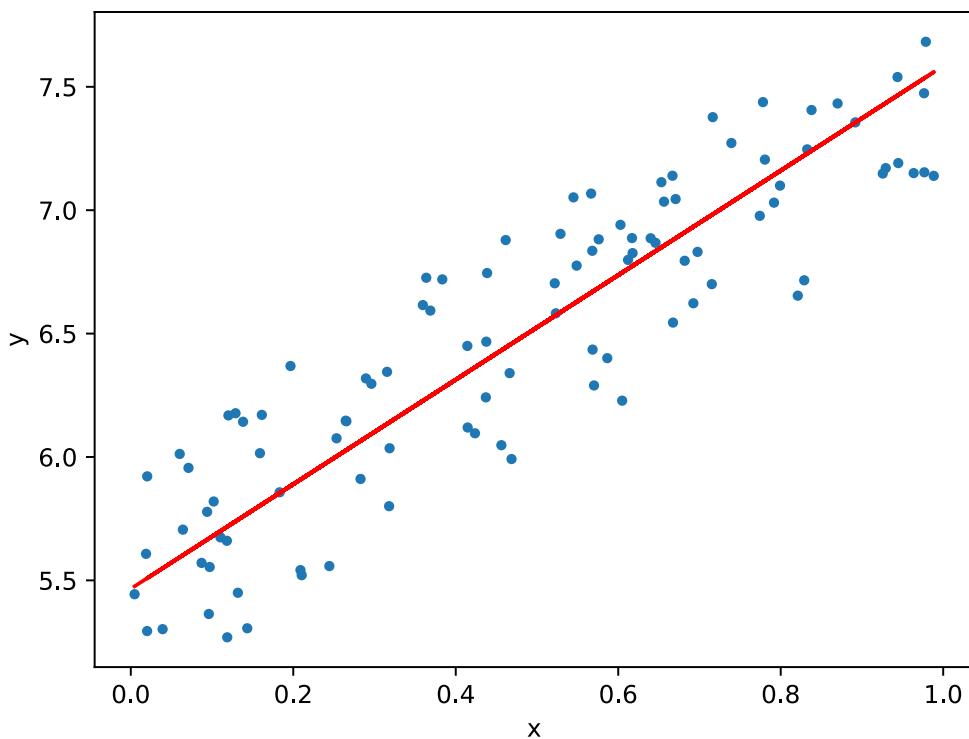


图 3.4 训练后的模型

如图所示，我们已经得到了一个拟合数据的模型。

3.4 梯度下降法为什么可以找到损失函数的最小值呢？

我们都知道损失函数 \mathcal{L} 是参数 W 和 b 的函数。所以我们可以将损失函数可视化出来。

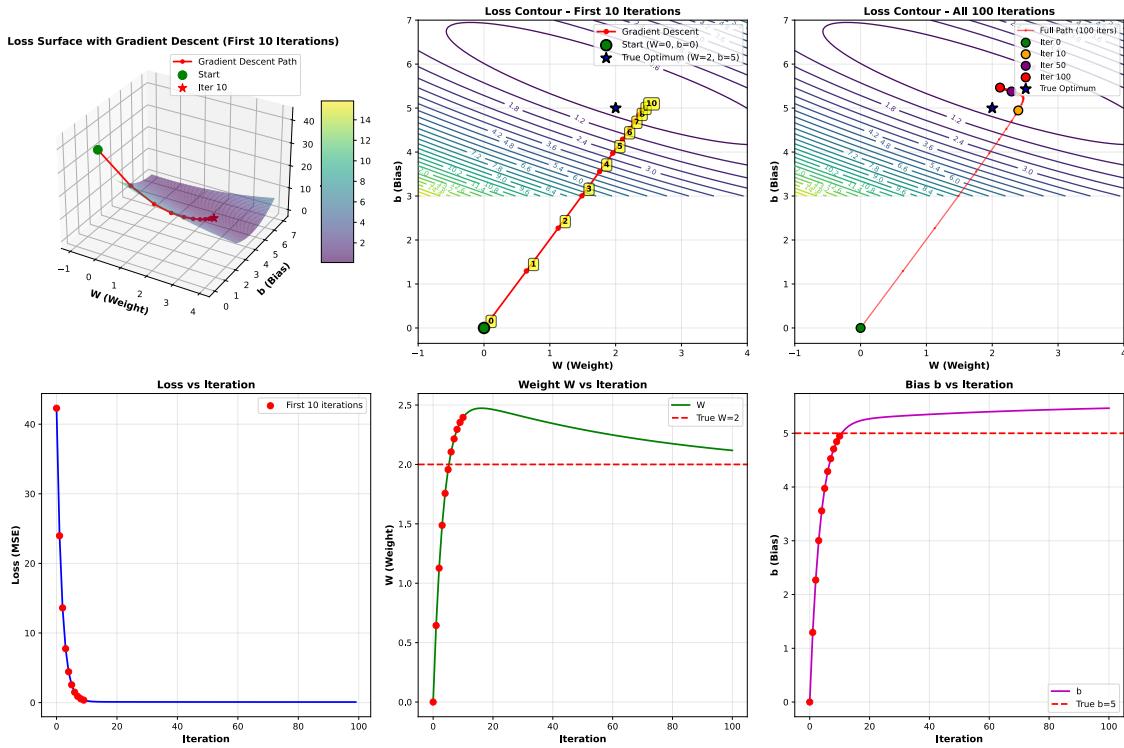


图 3.5 梯度下降法示意图

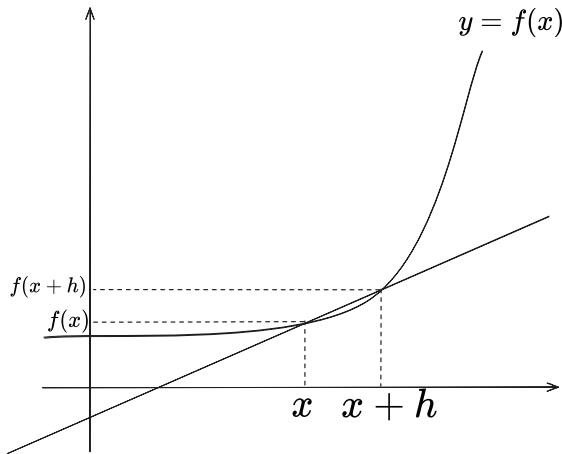
3.5 如何求导? (反向传播算法)

在前面的梯度下降法中，损失函数的梯度是我们手动进行求导得到的。

而在现实中，一个函数的参数是非常多的，每个参数都进行手动求导会非常的繁琐，甚至是不可能的（DeepSeek-V3 的参数数量是 6710 亿）。

所以我们需要寻求能够自动求导（自动微分）的方法，也就是大名鼎鼎的反向传播算法。

我们先来看一下求导。

图 3.6 曲线 $y = f(x)$ 和通过其两点的直线

什么是导数？简单地说，导数是变化率的一种表示方式。比如某个物体的位置相对于时间的变化率就是位置的导数，即速度。速度相对于时间的变化率就是速度的导数，即加速度。像这样，导数表示的是变化率，它被定义为在极短时间内的变化量。函数 $f(x)$ 在 x 处的导数可用下面的式子表示。

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} \quad (3.8)$$

上面的式子中 $\lim_{h \rightarrow 0}$ 表示极限，意思是 h 应该尽可能地接近0。 $\frac{f(x+h)-f(x)}{h}$ 表示通过两点的直线的斜率。

如图所示，函数 $f(x)$ 在 x 和 $x + h$ 两点之间的变化率为 $\frac{f(x+h)-f(x)}{h}$ 。让 h 的值尽可能地接近0，就可以求出 x 处的变化率。这就是 $y = f(x)$ 的导数。另外，在 $y = f(x)$ 的可导区间内，对于该区间的任何 x ，上面的式子都成立。因此，式子中的 $f'(x)$ 也是一个函数，我们称之为 $f(x)$ 的导函数。

3.5.1 数值微分的实现

下面根据导数的定义式来实现求导。需要注意的是，计算机不能处理极限值。因此，这里的 h 表示一个近似值。例如我们可以用 $h = 0.0001 (= 1e-4)$ 这种非常小的值来计算求导公式。利用微小的差值获得函数变化量的方法叫作数值微分。

数值微分使用非常小的值 h 求出的是真的导数的近似值。因此，这个值包含误差。中心差分近似是减小近似值误差的一种方法。中心差分近似计算的不是 $f(x)$ 和 $f(x+h)$ 的差，而是 $f(x-h)$ 和 $f(x+h)$ 的差值。下图中的红线表示的就是中心差分近似。

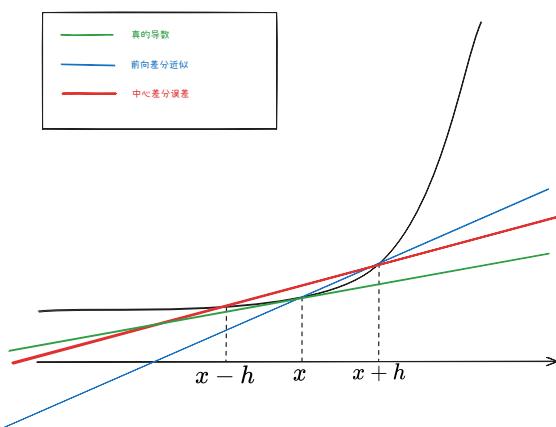


图 3.7 比较真的导数、前向差分近似和中心差分近似

如图所示，计算过 x 和 $x + h$ 这两点的直线的斜率的方法称为前向差分近似，计算 $x - h$ 和 $x + h$ 这两点间斜率的方法称为中心差分近似，中心差分近似实际产生的误差更小。中心差分近似的直线斜率是 $\frac{f(x+h)-f(x-h)}{2h}$ 。

下面我们来实现使用中心差分近似求数值微分的函数，该函数的名称为 `numerical_diff(f, x, eps=1e-4)`。这里的参数 f 是被求导的函数。数值微分可以通过以下代码实现。

```
1 def numerical_diff(f, x, eps=1e-4):
2     x0 = x - eps
3     x1 = x + eps
4     y0 = f(x0)
5     y1 = f(x1)
6     return (y1 - y0) / (2 * eps)
```

Python

例如如果 $f(x) = x^2$ ，并在 $x = 2.0$ 处求导，那么有如下代码：

```
1 def f(x):
```

Python

```
2     return x ** 2
3
4 x = 2.0
5 dy = numerical_diff(f, x)
6 print(dy)
```

运行结果如下

```
1 4.000000000004
```

由上面的运行结果可知, $y = x^2$ 在 $x = 2.0$ 时计算得到的导数之为 4.000000000004。不包含误差的导数值为 4.0, 可以说这个结果大体正确。

🔥 解析解求导数

导数也可以通过解析解的方式求解。解析解的方式求解是指只通过式子的变形推导出答案。在上面的例子中, 根据导数的公式可知, $y = x^2$ 的导数为 $\frac{dy}{dx} = 2x$ 。因此, $y = x^2$ 在 $x = 2.0$ 处的导数为 4.0。这个 4.0 是不包含误差的值。前面的数值微分结果虽然不是正好为 4.0, 但我们可以看出误差是相当小的。

当然复合函数也可以使用数值微分的方式进行求导。例如

$$y = (e^{x^2})^2 \quad (3.9)$$

代码如下:

```
1 def f(x):
2     x1 = x ** 2
3     x2 = math.exp(x1)
4     y = x2 ** 2
5     return y
6
7 x = 0.5
8 dy = numerical_diff(f, x)
9 print(dy)
```

Python

运行结果是: 3.2974426293330694。

现在我们已经成功实现了“自动”求导。只要用代码来定义要完成的计算(前面的例子定义了函数 f), 程序就会自动求出导数。使用这种方法, 无论多么复杂的函数组合, 程序都能自动求出导数。今后函数的种类越来越多的话, 不管是什么计算, 只要是可微函数, 程序就能求出它的导数。不过遗憾的是, 数值微分的方法存在一些问题。

⚡ 数值微分存在的问题

1. 数值微分的结果包含误差。在多数情况下，这个误差非常小，但在一些情况下，计算产生的误差可能会很大。数值微分的结果中容易包含误差的主要原因是“精度丢失”。中心差分近似等求差值的方法计算的是相同量级数值之间的差，但由于精度丢失，计算结果中会出现有效位数减少的情况。以有效位数为 4 的情况为例，在计算两个相近的值之间的差时，比如 $1.234 - 1.233$ ，其结果为 0.001，有效位数只有 1 位。本来可能是 $1.234\dots - 1.233\dots = 0.001434\dots$ 之类的结果，但由于精度丢失，结果变成 0.001。同样的情况也会发生在数值微分的差值计算中，精度丢失使结果更容易包含误差。
2. 数值微分更严重的问题是计算成本高。具体来说，在求多个变量的导数时，程序需要计算每个变量的导数。有些神经网络包含几百万个以上的变量（参数），通过数值微分对这么多的变量求导是不现实的。这时，反向传播就派上了用场。

另外，数值微分可以轻松实现，并能计算出大体正确的数值。而反向传播是一种复杂的算法，实现时容易出现 bug。我们可以使用数值微分的结果检查反向传播的实现是否正确。这种做法叫作梯度检验（gradient checking），它是一种将数值微分的结果与反向传播的结果进行比较的方法。

3.5.2 反向传播算法

我们以 $y = (e^{x^2})^2$ 在 $x = 0.5$ 求导为例子来说明一下反向传播算法。反向传播算法分为两个过程，前向过程（forward）和反向过程（backward）。

1. 前向过程：保存中间计算结果

$$\begin{aligned} v_0 &= x = 0.5 \\ v_1 &= v_0^2 = 0.5^2 = 0.25 \\ v_2 &= e^{v_1} = 1.2840254166877414 \\ v_3 &= v_2^2 = 1.648721270700128 \\ y &= v_3 \end{aligned} \tag{3.10}$$

2. 反向过程：利用前向过程保存的中间结果，以及链式求导法则来计算导数。

根据链式求导法则

$$\begin{aligned} \frac{dy}{dx} &= \frac{dy}{dv_3} \cdot \frac{dv_3}{dv_2} \cdot \frac{dv_2}{dv_1} \cdot \frac{dv_1}{dv_0} \cdot \frac{dv_0}{dx} \\ &= 1 \cdot 2v_2 \cdot e^{v_1} \cdot 2v_0 \cdot 1 \\ &= 1 \times (2 \times 1.2840254166877414) \times e^{0.25} \times (2 \times 0.5) \times 1 \\ &= 3.297442541400256 \end{aligned} \tag{3.11}$$

4. 分类问题：以手写数字识别为例

从本章开始，我们开始从代码和数学层面研究一下神经网络。

首先，神经网络是一个有很多个参数的函数。既然是函数，那么就有输入和输出，完成手写数字识别任务的神经网络的函数是什么呢？

$$\text{5} \rightarrow f(\bullet) \rightarrow \begin{bmatrix} 0.01(\text{预测为手写数字 0 的概率}) \\ 0.01(\text{预测为手写数字 1 的概率}) \\ 0.01(\text{预测为手写数字 2 的概率}) \\ 0.01(\text{预测为手写数字 3 的概率}) \\ 0.01(\text{预测为手写数字 4 的概率}) \\ 0.90(\text{预测为手写数字 5 的概率}) \\ 0.01(\text{预测为手写数字 6 的概率}) \\ 0.01(\text{预测为手写数字 7 的概率}) \\ 0.01(\text{预测为手写数字 8 的概率}) \\ 0.02(\text{预测为手写数字 9 的概率}) \end{bmatrix} \quad (4.1)$$

也就是说，我们要构造一个函数，能够接收手写数字图片为输入，而输出是手写数字属于每一个分类的概率所组成的向量！

所以我们需要找到一个函数 $f(\bullet)$ ，使得当函数接收到一张手写数字图片 5 时，输出的概率向量中，索引为 5 的概率是最大的。

这个函数的寻找历经了很多年的演变，函数结构可以是 **MLP**（多层次感知机），可以是 **CNN**（卷积神经网络），还可以是大杀器 **Transformer**。这都是我们未来要仔细学习的重点。

深度学习就是在固定函数结构的情况下，寻找使得预测准确率最高的函数参数。

以线性回归为例，我们固定了函数结构 $y = Wx + b$ ，想要寻找参数 W 和 b ，使得给定某一个输入 x ，输出的 y 能够比较准确。

所以神经网络的训练流程如下：

1. 设计模型结构（固定函数结构）
2. 收集训练数据集
3. 设计损失函数（用于度量预测准确率）
4. 使用最优化算法（例如梯度下降法）寻找使得损失函数最小的参数。

其中最优化算法需要求梯度（求导），所以我们需要类似于 **PyTorch** 这种能够自动求导的框架。

本章使用的数据集为 **MNIST** 数据集。**MNIST** 数据集是机器学习领域的经典基准数据集。它包含 28×28 个像素点（784 个像素点）的手写数字（0-9）灰度图像及其对应标签。我们的目标是构建一个能够根据像素值准确分类这些数字的神经网络。

4.1 先用玩具数据集研究一下分类问题

label	left-top	right-top	left-bottom	right-bottom
0	50	60	55	65
1	120	130	125	135
2	200	210	205	215

4.1.1 准备训练数据集

```
1 import numpy as np
2 import pandas as pd
3 import matplotlib.pyplot as plt
4
5 data = pd.read_csv("toy_dataset.csv")
6 print(data.head())
```

我们数据集中共 3 条数据，每个数据 4 个特征。我们将数据转换成 ndarray 格式。

```
1 data = np.array(data) # 转换成ndarray格式
2 print(data.shape) # 打印形状
3 batch_size, _ = data.shape # 获取一批数据的数量3。
```

可以看到数据的形状是(3,5)。我们将数据进行转置。变成(5,3)的形状。也就是说每一列都是“1个标签+4 个特征”，方便我们后续处理。

```
1 train_data = data.T
2 print(train_data.shape)
```

接下来我们将标签数据和特征数据分开。

```
1 Y = train_data[0] # 提取标签数据
2 X = train_data[1:] # 提取特征数据
3 X = X / 255.0 # 将特征归一化
```

4.1.2 神经网络模型结构的设计

在 式 (4.1) 中，我们知道神经网络其实是一个函数 $f(\bullet)$ 。只不过这个函数可能比较复杂，参数比较多。

一个简单的 ReLU 神经元如下所示：

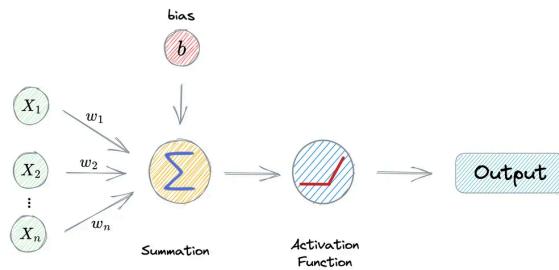


图 4.1 ReLU 神经元

其中 ReLU (Rectified Linear Unit, 整流线型单元) 激活函数的定义如下：

$$\text{ReLU}(x) = \max(0, x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \quad (4.2)$$

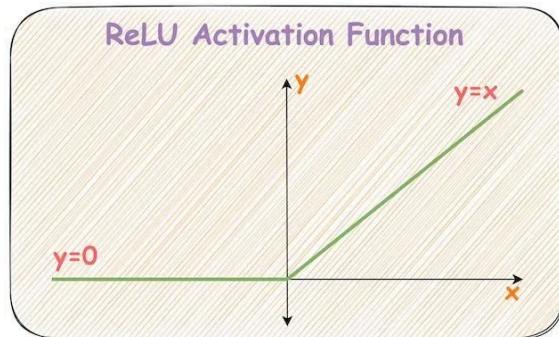


图 4.2 ReLU 激活函数

上面的图像如果写成数学函数，就有

$$\begin{aligned} f(x) &= \text{ReLU}(w_1x_1 + w_2x_2 + \dots + w_nx_n + b) \\ &= \text{ReLU}\left(\begin{bmatrix} w_1 & w_2 & \dots & w_n \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + b\right) \end{aligned} \quad (4.3)$$

这个 ReLU 神经元本身用处不大，但如果有序的组织起来，就能发挥巨大的威力。因为 ReLU 是一个非线性的函数，所以在神经网络中可以引入非线性因素，事实上，ReLU 神经元的叠加可以拟合任意函数，这叫做神经网络的万能逼近定理。

🔥 为什么要引入非线性函数？

考虑一元线性函数的简单复合，例如

$$\begin{aligned} f(x) &= w_2(w_1x + b_1) + b_2 \\ &= \underbrace{w_2w_1}_w x + \underbrace{w_2b_1 + b_2}_b \\ &= wx + b \end{aligned} \quad (4.4)$$

可以看到，线性函数的简单复合还是一个线性函数，所以无法拟合比较复杂的函数。

我们要设计的网络结构如下：

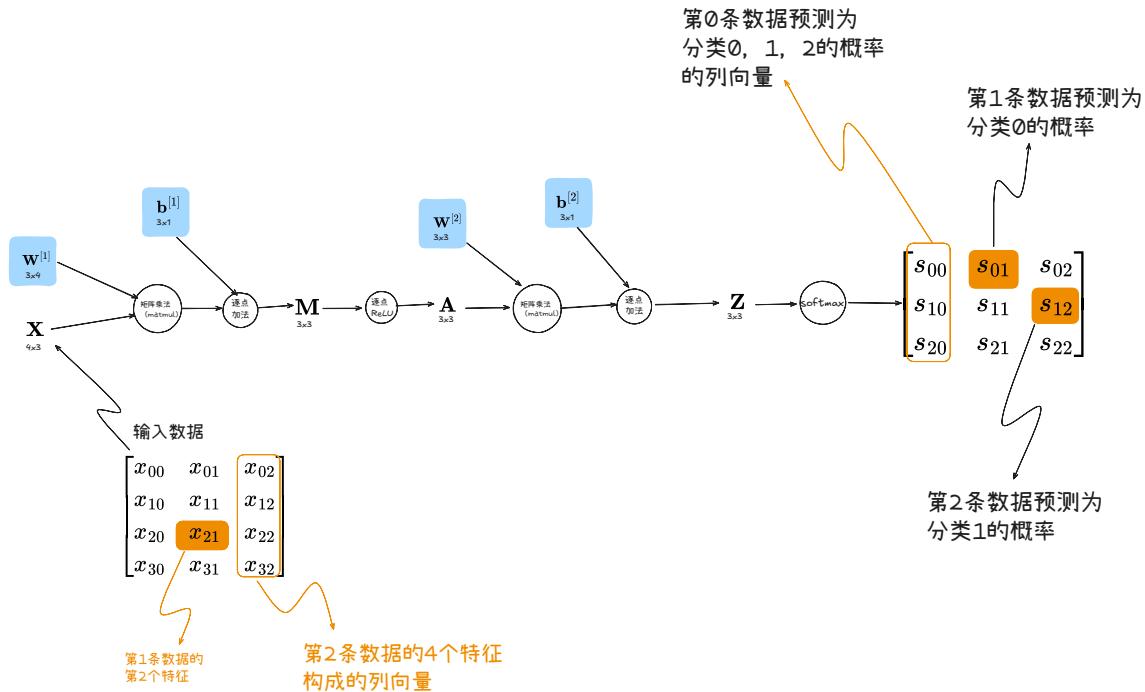


图 4.3

网络对应的函数表达式如下所示：

$$\text{output} = \begin{bmatrix} s_{00} & s_{01} & s_{02} \\ s_{10} & s_{11} & s_{12} \\ s_{20} & s_{21} & s_{22} \end{bmatrix} = \text{softmax}(\mathbf{W}^{[2]} \cdot \text{ReLU}(\mathbf{W}^{[1]} \cdot \mathbf{X} + \mathbf{b}^{[1]}) + \mathbf{b}^{[2]}) \quad (4.5)$$

在上面的网络中，我们的输入数据是 3 条训练数据，每条数据 4 个特征。下面的输入数据中的每一列都是一条数据。

$$\mathbf{X} = \begin{bmatrix} x_{00} & x_{01} & x_{02} \\ x_{10} & x_{11} & x_{12} \\ x_{20} & x_{21} & x_{22} \\ x_{30} & x_{31} & x_{32} \end{bmatrix} \quad (4.6)$$

输出是 3 条数据的每一条属于某一个分类的概率（预测值）。

$$\text{output} = \mathbf{S} = \begin{bmatrix} s_{00} & s_{01} & s_{02} \\ s_{10} & s_{11} & s_{12} \\ s_{20} & s_{21} & s_{22} \end{bmatrix} \quad (4.7)$$

而参数 $\mathbf{W}^{[1]}, \mathbf{b}^{[1]}, \mathbf{W}^{[2]}, \mathbf{b}^{[2]}$ 是需要学习的参数，也就是说在训练过程中不断的改变的参数。我们希望在训练完成以后，给神经网络这个函数输入一条数据，能够得到它的分类的概率。其中概率最大的分类，是这条数据的正确分类。

$$\mathbf{W}^{[1]} = \begin{bmatrix} w_{00}^{[1]} & w_{01}^{[1]} & w_{02}^{[1]} & w_{03}^{[1]} \\ w_{10}^{[1]} & w_{11}^{[1]} & w_{12}^{[1]} & w_{13}^{[1]} \\ w_{20}^{[1]} & w_{21}^{[1]} & w_{22}^{[1]} & w_{23}^{[1]} \end{bmatrix} \quad (4.8)$$

$$\mathbf{b}^{[1]} = \begin{bmatrix} b_0^{[1]} \\ b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \end{bmatrix} \quad (4.9)$$

$$\mathbf{W}^{[2]} = \begin{bmatrix} w_{00}^{[2]} & w_{01}^{[2]} & w_{02}^{[2]} \\ w_{10}^{[2]} & w_{11}^{[2]} & w_{12}^{[2]} \\ w_{20}^{[2]} & w_{21}^{[2]} & w_{22}^{[2]} \end{bmatrix} \quad (4.10)$$

$$\mathbf{b}^{[2]} = \begin{bmatrix} b_0^{[2]} \\ b_1^{[2]} \\ b_2^{[2]} \\ b_3^{[2]} \end{bmatrix} \quad (4.11)$$

其中计算的中间值有

$$\begin{aligned} \mathbf{M} &= \mathbf{W}^{[1]} \mathbf{X} + \mathbf{b}^{[1]} \\ &= \begin{bmatrix} m_{00} & m_{01} & m_{02} \\ m_{10} & m_{11} & m_{12} \\ m_{20} & m_{21} & m_{22} \end{bmatrix} \end{aligned} \quad (4.12)$$

其中

$$\begin{aligned} m_{00} &= w_{00}^{[1]} \cdot x_{00} + w_{01}^{[1]} \cdot x_{10} + w_{02}^{[1]} \cdot x_{20} + w_{03}^{[1]} \cdot x_{30} + b_0^{[1]} \\ m_{01} &= w_{00}^{[1]} \cdot x_{01} + w_{01}^{[1]} \cdot x_{11} + w_{02}^{[1]} \cdot x_{21} + w_{03}^{[1]} \cdot x_{31} + b_0^{[1]} \\ m_{02} &= w_{00}^{[1]} \cdot x_{02} + w_{01}^{[1]} \cdot x_{12} + w_{02}^{[1]} \cdot x_{22} + w_{03}^{[1]} \cdot x_{32} + b_0^{[1]} \\ m_{10} &= w_{10}^{[1]} \cdot x_{00} + w_{11}^{[1]} \cdot x_{10} + w_{12}^{[1]} \cdot x_{20} + w_{13}^{[1]} \cdot x_{30} + b_1^{[1]} \\ m_{11} &= w_{10}^{[1]} \cdot x_{01} + w_{11}^{[1]} \cdot x_{11} + w_{12}^{[1]} \cdot x_{21} + w_{13}^{[1]} \cdot x_{31} + b_1^{[1]} \\ m_{12} &= w_{10}^{[1]} \cdot x_{02} + w_{11}^{[1]} \cdot x_{12} + w_{12}^{[1]} \cdot x_{22} + w_{13}^{[1]} \cdot x_{32} + b_1^{[1]} \\ m_{20} &= w_{20}^{[1]} \cdot x_{00} + w_{21}^{[1]} \cdot x_{10} + w_{22}^{[1]} \cdot x_{20} + w_{23}^{[1]} \cdot x_{30} + b_2^{[1]} \\ m_{21} &= w_{20}^{[1]} \cdot x_{01} + w_{21}^{[1]} \cdot x_{11} + w_{22}^{[1]} \cdot x_{21} + w_{23}^{[1]} \cdot x_{31} + b_2^{[1]} \\ m_{22} &= w_{20}^{[1]} \cdot x_{02} + w_{21}^{[1]} \cdot x_{12} + w_{22}^{[1]} \cdot x_{22} + w_{23}^{[1]} \cdot x_{32} + b_2^{[1]} \end{aligned} \quad (4.13)$$

$$\begin{aligned} \mathbf{A} &= \text{ReLU}(\mathbf{M}) = \begin{bmatrix} a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} \\ a_{20} & a_{21} & a_{22} \end{bmatrix} \\ &= \begin{bmatrix} \max(m_{00}, 0) & \max(m_{01}, 0) & \max(m_{02}, 0) \\ \max(m_{10}, 0) & \max(m_{11}, 0) & \max(m_{12}, 0) \\ \max(m_{20}, 0) & \max(m_{21}, 0) & \max(m_{22}, 0) \end{bmatrix} \end{aligned} \quad (4.14)$$

$$\mathbf{Z} = \mathbf{W}^{[2]} \mathbf{A} + \mathbf{b}^{[2]} = \begin{bmatrix} z_{00} & z_{01} & z_{02} \\ z_{10} & z_{11} & z_{12} \\ z_{20} & z_{21} & z_{22} \end{bmatrix} \quad (4.15)$$

其中

$$\begin{aligned}
 z_{00} &= w_{00}^{[2]} \cdot a_{00} + w_{01}^{[2]} \cdot a_{10} + w_{02}^{[2]} \cdot a_{20} + b_0^{[2]} \\
 z_{01} &= w_{00}^{[2]} \cdot a_{01} + w_{01}^{[2]} \cdot a_{11} + w_{02}^{[2]} \cdot a_{21} + b_0^{[2]} \\
 z_{02} &= w_{00}^{[2]} \cdot a_{02} + w_{01}^{[2]} \cdot a_{12} + w_{02}^{[2]} \cdot a_{22} + b_0^{[2]} \\
 z_{10} &= w_{10}^{[2]} \cdot a_{00} + w_{11}^{[2]} \cdot a_{10} + w_{12}^{[2]} \cdot a_{20} + b_1^{[2]} \\
 z_{11} &= w_{10}^{[2]} \cdot a_{01} + w_{11}^{[2]} \cdot a_{11} + w_{12}^{[2]} \cdot a_{21} + b_1^{[2]} \\
 z_{12} &= w_{10}^{[2]} \cdot a_{02} + w_{11}^{[2]} \cdot a_{12} + w_{12}^{[2]} \cdot a_{22} + b_1^{[2]} \\
 z_{20} &= w_{20}^{[2]} \cdot a_{00} + w_{21}^{[2]} \cdot a_{10} + w_{22}^{[2]} \cdot a_{20} + b_2^{[2]} \\
 z_{21} &= w_{20}^{[2]} \cdot a_{01} + w_{21}^{[2]} \cdot a_{11} + w_{22}^{[2]} \cdot a_{21} + b_2^{[2]} \\
 z_{22} &= w_{20}^{[2]} \cdot a_{02} + w_{21}^{[2]} \cdot a_{12} + w_{22}^{[2]} \cdot a_{22} + b_2^{[2]}
 \end{aligned} \tag{4.16}$$

我们现在可以看到 \mathbf{Z} 是一个 3×3 形状的向量。但是我们都应该知道我们要输出的是数据 \mathbf{X} 属于每个类别概率（共 3 个分类）。

所以我们需要将 \mathbf{Z} 中的每个元素转换成概率值，也就是每一列的 3 个元素的和是 1。

我们需要找到一个函数能够将 \mathbf{Z} 转换成概率值，同时不能改变 \mathbf{Z} 中每个元素的大小顺序。

也即是如果在 \mathbf{Z} 中， $z_{00} < z_{20}$ ，那么这两个元素转换成概率值以后，还得是第 00 个元素小于第 20 个元素。

能够达到这个目的的函数就是大名鼎鼎的“softmax”函数。

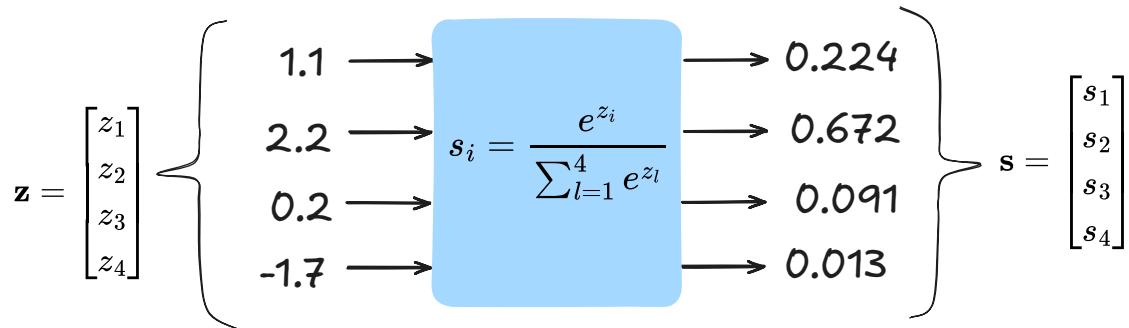


图 4.4 softmax 函数

softmax 函数的主要作用是将任意实数的向量转换为概率。上图公式中的指数函数确保了得到的值是非负的。由于分母中的归一化项，得到的值总和为 1。此外，所有值都介于 0 和 1 之间。softmax 函数的一个重要特性是它能保持其输入值的排序顺序：

$$-1.7 < 0.2 < 1.1 < 2.2 \Rightarrow 0.013 < 0.091 < 0.224 < 0.672 \tag{4.17}$$

所以我们的神经网络的输出是

$$\mathbf{s} = \text{output} = \begin{bmatrix} s_{00} & s_{01} & s_{02} \\ s_{10} & s_{11} & s_{12} \\ s_{20} & s_{21} & s_{22} \end{bmatrix} \tag{4.18}$$

其中

$$\begin{aligned}
 s_{00} &= \frac{e^{z_{00}}}{e^{z_{00}} + e^{z_{10}} + e^{z_{20}}} \\
 s_{01} &= \frac{e^{z_{01}}}{e^{z_{01}} + e^{z_{11}} + e^{z_{21}}} \\
 s_{02} &= \frac{e^{z_{02}}}{e^{z_{02}} + e^{z_{12}} + e^{z_{22}}} \\
 s_{10} &= \frac{e^{z_{10}}}{e^{z_{00}} + e^{z_{10}} + e^{z_{20}}} \\
 s_{11} &= \frac{e^{z_{11}}}{e^{z_{01}} + e^{z_{11}} + e^{z_{21}}} \\
 s_{12} &= \frac{e^{z_{12}}}{e^{z_{02}} + e^{z_{12}} + e^{z_{22}}} \\
 s_{20} &= \frac{e^{z_{20}}}{e^{z_{00}} + e^{z_{10}} + e^{z_{20}}} \\
 s_{21} &= \frac{e^{z_{21}}}{e^{z_{01}} + e^{z_{11}} + e^{z_{21}}} \\
 s_{22} &= \frac{e^{z_{22}}}{e^{z_{02}} + e^{z_{12}} + e^{z_{22}}}
 \end{aligned} \tag{4.19}$$

这样我们就可以将输出解释为属于某个分类的概率了。这里要注意的是我们只是能将输出解释为概率，只有经过神经网络的训练，输出才会慢慢接近真正的概率值。

4.1.3 如何设计损失函数？

由于神经网络在接收 3 条数据作为输入后，输出是属于分类的概率组成的向量。

$$\mathbf{s} = \text{output} = \begin{bmatrix} s_{00} & s_{01} & s_{02} \\ s_{10} & s_{11} & s_{12} \\ s_{20} & s_{21} & s_{22} \end{bmatrix} \tag{4.20}$$

例如我们的训练数据有 3 条，在输入给未经训练的神经网络以后输出的概率如下

$$\begin{bmatrix} 50 & 120 & 200 \\ 60 & 130 & 210 \\ 55 & 125 & 205 \\ 65 & 135 & 215 \end{bmatrix} \Rightarrow f(\bullet) \Rightarrow \begin{bmatrix} 0.01 & 0.5 & 0.49 \\ 0.5 & 0.02 & 0.48 \\ 0.4 & 0.5 & 0.1 \end{bmatrix} \tag{4.21}$$

也就是针对 3 条数据预测为正确分类的概率是 0.01, 0.02, 0.1，错的离谱。而我们的标签数据是 [0, 1, 2]，是没有办法和输出的概率向量做比较的，那么如何做比较呢？那就是将标签数据转换成向量，也就是独热编码（one-hot encoder）。

索引为 0 的元素编码为 1，
其它元素一律编码为 0

$$[0, 1, 2] \Rightarrow \text{独热编码} \Rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.22}$$

也就是我们共有 3 个标签，0, 1, 2，分别独热编码为 3 个列向量。

$$\begin{aligned}
 0 \Rightarrow \text{one-hot} &\Rightarrow \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \\
 1 \Rightarrow \text{one-hot} &\Rightarrow \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\
 2 \Rightarrow \text{one-hot} &\Rightarrow \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}
 \end{aligned} \tag{4.23}$$

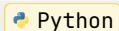
独热编码的直观解释其实就是：第 0 条数据属于分类 0 的概率为 1。

代码如下：

```

1 def one_hot(Y):
2     one_hot_Y = np.zeros((Y.size, Y.max() + 1))
3     one_hot_Y[np.arange(Y.size), Y] = 1
4     one_hot_Y = one_hot_Y.T
5     return one_hot_Y
6
7 print(one_hot(Y)) # 对所有数据进行独热编码
8 print(Y[1]) # 第1条数据的标签
9 print(one_hot(Y)[:, 1]) # 第1条数据的标签的独热编码

```



有了独热编码之后，如何度量预测的概率向量和真实的独热向量之间的损失呢？

$$\mathbf{s} = \begin{bmatrix} s_{00} & s_{01} & s_{02} \\ s_{10} & \mathbf{s}_{11} & s_{12} \\ s_{20} & s_{21} & s_{22} \end{bmatrix} \Leftarrow \text{如何度量两者之间的差异?} \Rightarrow \mathbf{y} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \mathbf{1} & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.24}$$

这就是交叉熵损失函数

$$\mathcal{L}(\mathbf{s}, \mathbf{y}) = - \sum_{j=0}^2 \sum_{i=0}^2 y_{ij} \log(s_{ij}) \tag{4.25}$$

由于 **one-hot** 的性质（只有 $y_{00} = 1, y_{11} = 1, y_{22} = 1$ ），所以化简可以得到

$$\mathcal{L} = -\log(s_{00}) - \log(s_{11}) - \log(s_{22}) \tag{4.26}$$

⚡ 损失函数

到目前为止，我们接触了两种损失函数：

- 均方误差损失 (`mean square error loss, MSE loss`)：用在线性回归中
- 交叉熵损失 (`cross entropy loss, CE loss`)：用在分类任务中

这两种损失函数并不是拍脑袋得来的，而是有着严谨的数学背景，我们后面会讲解。

直觉和例子

交叉熵在做两件事：

- 如果 s 很大（接近 1），则 $-\log(s)$ 很小，损失小
- 如果 s 很小（接近 0），则 $-\log(s)$ 很大，损失大

举个数值感受一下

- 如果 $s = 0.999$ ，那么 $\mathcal{L} = -\log(0.999) \approx 0.001$

- 如果 $s = 0.001$, 那么 $\mathcal{L} = -\log(0.001) \approx 6.907$
- 可见预测的分类越准确, 那么损失越小。可以作为度量。

4.1.4 如何让损失函数最小?

我们使用的算法是梯度下降法, 而梯度下降法需要求导数然后再更新参数, 所以我们分两步走

1. 对参数 $\mathbf{W}^{[1]}, \mathbf{b}^{[1]}, \mathbf{W}^{[2]}, \mathbf{b}^{[2]}$ 求偏导数。
2. 更新参数。

我们还记得神经网络的函数如下所示:

$$\mathbf{S} = \text{output} = \text{softmax}(\mathbf{W}^{[2]} \cdot \text{ReLU}(\mathbf{W}^{[1]} \cdot \mathbf{X} + \mathbf{b}^{[1]}) + \mathbf{b}^{[2]}) \quad (4.27)$$

如果将上面的式子拆分开, 就得到了

$$\begin{aligned} \mathbf{M} &= \mathbf{W}^{[1]} \mathbf{X} + \mathbf{b}^{[1]} \\ \mathbf{A} &= \text{ReLU}(\mathbf{M}) \\ \mathbf{Z} &= \mathbf{W}^{[2]} \mathbf{A} + \mathbf{b}^{[2]} \\ \mathbf{S} &= \text{softmax}(\mathbf{Z}) \end{aligned} \quad (4.28)$$

而这就是反向传播算法的前向传播过程, 代码如下:

```
1 def forward(w1, b1, w2, b2, X):
2     M = w1.dot(X) + b1
3     A = ReLU(M)
4     Z = w2.dot(A) + b2
5     S = softmax(Z)
6     return M, A, Z, S
```

Python

对应的 `ReLU` 函数和 `softmax` 函数如下:

```
1 def ReLU(M):
2     return np.maximum(M, 0)
3
4 def softmax(Z):
5     S = np.exp(Z) / sum(np.exp(Z))
6     return S
```

Python

我们在前向过程中保存了一些中间计算结果: $\mathbf{M}, \mathbf{A}, \mathbf{Z}, \mathbf{S}$ 。用于在反向传播过程中求解损失函数 \mathcal{L} 对参数的导数。

我们的损失函数是交叉熵损失函数 (是一个标量值!)

$$\mathcal{L}(\mathbf{s}, \mathbf{y}) = - \sum_{j=0}^2 \sum_{i=0}^2 y_{ij} \log(s_{ij}) \quad (4.29)$$

那么损失函数如何对参数进行求偏导数呢?

标量对向量或者矩阵求导的形状和向量或者矩阵的形状一样, 例如:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[2]}} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial w_{00}^{[2]}} & \frac{\partial \mathcal{L}}{\partial w_{01}^{[2]}} & \frac{\partial \mathcal{L}}{\partial w_{02}^{[2]}} \\ \frac{\partial \mathcal{L}}{\partial w_{10}^{[2]}} & \frac{\partial \mathcal{L}}{\partial w_{11}^{[2]}} & \frac{\partial \mathcal{L}}{\partial w_{12}^{[2]}} \\ \frac{\partial \mathcal{L}}{\partial w_{20}^{[2]}} & \frac{\partial \mathcal{L}}{\partial w_{21}^{[2]}} & \frac{\partial \mathcal{L}}{\partial w_{22}^{[2]}} \end{bmatrix} \quad (4.30)$$

🔥 标量函数的链式求导

$$\frac{d\mathcal{L}}{dw} = \frac{d\mathcal{L}}{ds} \cdot \frac{ds}{dz} \cdot \frac{dz}{dw} \quad (4.31)$$

可以看到，是比较简单的，但是矩阵或者向量的链式求导就很复杂了！但是依然遵循标量求导的规则。矩阵只是表示法而已！

我们先来求解 $\frac{\partial \mathcal{L}}{\partial \mathbf{Z}}$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{Z}} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial z_{00}} & \frac{\partial \mathcal{L}}{\partial z_{01}} & \frac{\partial \mathcal{L}}{\partial z_{02}} \\ \frac{\partial \mathcal{L}}{\partial z_{10}} & \frac{\partial \mathcal{L}}{\partial z_{11}} & \frac{\partial \mathcal{L}}{\partial z_{12}} \\ \frac{\partial \mathcal{L}}{\partial z_{20}} & \frac{\partial \mathcal{L}}{\partial z_{21}} & \frac{\partial \mathcal{L}}{\partial z_{22}} \end{bmatrix} \quad (4.32)$$

那么偏导数矩阵中的每个元素该如何计算呢？

我们首先知道

$$\begin{aligned} y_{ij} \log(s_{ij}) &= y_{ij} \log \frac{e^{z_{ij}}}{e^{z_{0j}} + e^{z_{1j}} + e^{z_{2j}}} \quad (\text{对数性质}) \\ &= y_{ij} (z_{ij} - \log(e^{z_{0j}} + e^{z_{1j}} + e^{z_{2j}})) \end{aligned} \quad (4.33)$$

又因为独热编码的性质

$$y_{0j} + y_{1j} + y_{2j} = 1 \quad (4.34)$$

所以损失函数可以化简为

$$\begin{aligned} \mathcal{L} &= -\{y_{00}z_{00} + y_{01}z_{01} + y_{02}z_{02} \\ &\quad + y_{10}z_{10} + y_{11}z_{11} + y_{12}z_{12} \\ &\quad + y_{20}z_{20} + y_{21}z_{21} + y_{22}z_{22} \\ &\quad - \log(e^{z_{00}} + e^{z_{10}} + e^{z_{20}}) \\ &\quad - \log(e^{z_{01}} + e^{z_{11}} + e^{z_{21}}) \\ &\quad - \log(e^{z_{02}} + e^{z_{12}} + e^{z_{22}})\} \end{aligned} \quad (4.35)$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial z_{00}} &= -\left\{y_{00} - \frac{e^{z_{00}}}{e^{z_{00}} + e^{z_{10}} + e^{z_{20}}}\right\} \\ &= -\{y_{00} - s_{00}\} = s_{00} - y_{00} \end{aligned} \quad (4.36)$$

最终得到了一个非常漂亮的结果，如下

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{Z}} &= \begin{bmatrix} s_{00} - y_{00} & s_{01} - y_{01} & s_{02} - y_{02} \\ s_{10} - y_{10} & s_{11} - y_{11} & s_{12} - y_{12} \\ s_{20} - y_{20} & s_{21} - y_{21} & s_{22} - y_{22} \end{bmatrix} \\ &= \mathbf{s} - \mathbf{y} \end{aligned} \quad (4.37)$$

那么偏导数矩阵 $\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[2]}}$ 中的每个元素该如何计算呢？例如 $\frac{\partial \mathcal{L}}{\partial w_{00}^{[2]}}$ ？

通过观察得到 \mathcal{L} 中的 z_{00}, z_{01}, z_{02} 依赖于 $w_{00}^{[2]}$ 。根据全微分公式

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial w_{00}^{[2]}} &= \frac{\partial \mathcal{L}}{\partial z_{00}} \frac{\partial z_{00}}{\partial w_{00}^{[2]}} + \frac{\partial \mathcal{L}}{\partial z_{01}} \frac{\partial z_{01}}{\partial w_{00}^{[2]}} + \frac{\partial \mathcal{L}}{\partial z_{02}} \frac{\partial z_{02}}{\partial w_{00}^{[2]}} \\ &= (s_{00} - y_{00})a_{00} + (s_{01} - y_{01})a_{01} + (s_{02} - y_{02})a_{02} \end{aligned} \quad (4.38)$$

所以有如下

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w_{00}^{[2]}} &= (s_{00} - y_{00})a_{00} + (s_{01} - y_{01})a_{01} + (s_{02} - y_{02})a_{02} \\
\frac{\partial \mathcal{L}}{\partial w_{01}^{[2]}} &= (s_{00} - y_{00})a_{10} + (s_{01} - y_{01})a_{11} + (s_{02} - y_{02})a_{12} \\
\frac{\partial \mathcal{L}}{\partial w_{02}^{[2]}} &= (s_{00} - y_{00})a_{20} + (s_{01} - y_{01})a_{21} + (s_{02} - y_{02})a_{22} \\
\frac{\partial \mathcal{L}}{\partial w_{10}^{[2]}} &= (s_{10} - y_{10})a_{00} + (s_{11} - y_{11})a_{01} + (s_{12} - y_{12})a_{02} \\
\frac{\partial \mathcal{L}}{\partial w_{11}^{[2]}} &= (s_{10} - y_{10})a_{10} + (s_{11} - y_{11})a_{11} + (s_{12} - y_{12})a_{12} \\
\frac{\partial \mathcal{L}}{\partial w_{12}^{[2]}} &= (s_{10} - y_{10})a_{20} + (s_{11} - y_{11})a_{21} + (s_{12} - y_{12})a_{22} \\
\frac{\partial \mathcal{L}}{\partial w_{20}^{[2]}} &= (s_{20} - y_{20})a_{00} + (s_{21} - y_{21})a_{01} + (s_{22} - y_{22})a_{02} \\
\frac{\partial \mathcal{L}}{\partial w_{21}^{[2]}} &= (s_{20} - y_{20})a_{10} + (s_{21} - y_{21})a_{11} + (s_{22} - y_{22})a_{12} \\
\frac{\partial \mathcal{L}}{\partial w_{22}^{[2]}} &= (s_{20} - y_{20})a_{20} + (s_{21} - y_{21})a_{21} + (s_{22} - y_{22})a_{22}
\end{aligned} \tag{4.39}$$

通过观察可以知道

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[2]}} &= \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial w_{00}^{[2]}} & \frac{\partial \mathcal{L}}{\partial w_{01}^{[2]}} & \frac{\partial \mathcal{L}}{\partial w_{02}^{[2]}} \\ \frac{\partial \mathcal{L}}{\partial w_{10}^{[2]}} & \frac{\partial \mathcal{L}}{\partial w_{11}^{[2]}} & \frac{\partial \mathcal{L}}{\partial w_{12}^{[2]}} \\ \frac{\partial \mathcal{L}}{\partial w_{20}^{[2]}} & \frac{\partial \mathcal{L}}{\partial w_{21}^{[2]}} & \frac{\partial \mathcal{L}}{\partial w_{22}^{[2]}} \end{bmatrix} \\
&= \begin{bmatrix} s_{00} - a_{00} & s_{01} - a_{01} & s_{02} - a_{02} \\ s_{10} - a_{10} & s_{11} - a_{11} & s_{12} - a_{12} \\ s_{20} - a_{20} & s_{21} - a_{21} & s_{22} - a_{22} \end{bmatrix} \cdot \begin{bmatrix} a_{00} & a_{10} & a_{20} \\ a_{01} & a_{11} & a_{21} \\ a_{02} & a_{12} & a_{22} \end{bmatrix} \\
&= (\mathbf{s} - \mathbf{y}) \cdot \mathbf{A}^T
\end{aligned} \tag{4.40}$$

而我们知道

$$\frac{\partial \mathcal{L}}{\partial \mathbf{Z}} = \mathbf{s} - \mathbf{y} \tag{4.41}$$

所以

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[2]}} &= \frac{\partial \mathcal{L}}{\partial \mathbf{Z}} \frac{\partial \mathbf{Z}}{\partial \mathbf{W}^{[2]}} \\
&= \frac{\partial \mathcal{L}}{\partial \mathbf{Z}} \frac{\partial \mathbf{W}^{[2]} \mathbf{A} + \mathbf{b}^{[2]}}{\partial \mathbf{W}^{[2]}} \\
&= \frac{\partial \mathcal{L}}{\partial \mathbf{Z}} \mathbf{A}^T
\end{aligned} \tag{4.42}$$

这就是矩阵的链式求导公式！

如果你愿意重复上面的过程，会发现

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{[2]}} = \frac{\partial \mathcal{L}}{\partial \mathbf{Z}} \tag{4.43}$$

接下来我们要求解 $\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[1]}}$ ，那么先需要求解 $\frac{\partial \mathcal{L}}{\partial \mathbf{M}}$ 。

还是以 $\frac{\partial \mathcal{L}}{\partial m_{00}}$ 为例子。我们观察到 z_{00}, z_{10}, z_{20} 依赖于 m_{00} 。所以根据全微分公式

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial m_{00}} &= \frac{\partial \mathcal{L}}{\partial z_{00}} \frac{\partial z_{00}}{\partial m_{00}} + \frac{\partial \mathcal{L}}{\partial z_{10}} \frac{\partial z_{10}}{\partial m_{00}} + \frac{\partial \mathcal{L}}{\partial z_{20}} \frac{\partial z_{20}}{\partial m_{00}} \\ &= \frac{\partial \mathcal{L}}{\partial z_{00}} \frac{\partial z_{00}}{\partial a_{00}} \frac{\partial a_{00}}{\partial m_{00}} + \frac{\partial \mathcal{L}}{\partial z_{10}} \frac{\partial z_{10}}{\partial a_{00}} \frac{\partial a_{00}}{\partial m_{00}} + \frac{\partial \mathcal{L}}{\partial z_{20}} \frac{\partial z_{20}}{\partial a_{00}} \frac{\partial a_{00}}{\partial m_{00}} \\ &= \{(s_{00} - y_{00})w_{00}^{[2]} + (s_{10} - y_{10})w_{10}^{[2]} + (s_{20} - y_{20})w_{20}^{[2]}\} \cdot \{1 \text{ if } m_{00} > 0 \text{ else } 0\}\end{aligned}\quad (4.44)$$

那么可以得到矩阵表达形式

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{M}} &= \{\mathbf{W}^{[2]}\}^T \cdot (\mathbf{s} - \mathbf{y}) \odot \{1 \text{ if } m_{ij} > 0 \text{ else } 0\} \\ &= \left\{ \{\mathbf{W}^{[2]}\}^T \cdot \frac{\partial \mathcal{L}}{\partial \mathbf{Z}} \right\} \odot \{1 \text{ if } m_{ij} > 0 \text{ else } 0\}\end{aligned}\quad (4.45)$$

那么根据全微分公式有如下

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial w_{00}^{[1]}} &= \frac{\partial \mathcal{L}}{\partial m_{00}} \frac{\partial m_{00}}{\partial w_{00}^{[1]}} + \frac{\partial \mathcal{L}}{\partial m_{01}} \frac{\partial m_{01}}{\partial w_{00}^{[1]}} + \frac{\partial \mathcal{L}}{\partial m_{02}} \frac{\partial m_{02}}{\partial w_{00}^{[1]}} \\ &= \left[\frac{\partial \mathcal{L}}{\partial m_{00}} \quad \frac{\partial \mathcal{L}}{\partial m_{01}} \quad \frac{\partial \mathcal{L}}{\partial m_{02}} \right] \cdot [x_{00} \ x_{01} \ x_{02}]^T\end{aligned}\quad (4.46)$$

整理成矩阵形式如下

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[1]}} = \frac{\partial \mathcal{L}}{\partial \mathbf{M}} \cdot \mathbf{X}^T \quad (4.47)$$

同理可以得到

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{[1]}} = \frac{\partial \mathcal{L}}{\partial \mathbf{M}} \quad (4.48)$$

写成代码如下

```
1 def deriv_ReLU(z):
2     return z > 0
3
4 def backward(M, A, Z, output, w2, Y, X):
5     OneHot_Y = one_hot(Y)
6     dZ = output - OneHot_Y
7     dW2 = 1/m * dZ.dot(A.T)
8     db2 = 1/m * np.sum(dZ)
9     dM = w2.T.dot(dZ) * deriv_ReLU(M)
10    dW1 = 1 / m * dM.dot(X.T)
11    db1 = 1 / m * np.sum(dM)
12    return dW1, db1, dW2, db2
```

Python

4.2 准备训练数据集

```
1 import numpy as np
2 import pandas as pd
3 import matplotlib.pyplot as plt
4
5 data = pd.read_csv("mnist.csv")
```

Python

```
6 print(data.head())
```

可以看到我们将数据集的前 5 行打印了出来。

	label	1x1	1x2	1x3	1x4	1x5	1x6	1x7	1x8	1x9	...	28x19	28x20	28x21	28x22	28x23	28x24	28x25	28x26	28x27	28x28
0	5	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	4	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	9	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 785 columns

图 4.5 mnist 数据集的前 5 行，形状为 5x785

“label”这一列表示的是手写数字的分类，例如第一列的 label 是 5，所以后面的像素点渲染出来是手写数字 5。

从列 1x1 到 28x28 共 784 个像素点，是手写数字图片的每个像素点的值。可以看到很多 0，也就是手写数字图片中的很多像素点都是黑色。我们可以尝试将第一行的 784 个像素点渲染成图片。

```
1 first_row = data.iloc[0]
2 label = first_row.iloc[0]
3 pixels = first_row.iloc[1:].values
4
5 # 将一维数组重塑为28x28的图像
6 image = pixels.reshape(28, 28)
7
8 # 创建图形
9 plt.figure(figsize=(6, 6))
10 plt.imshow(image, cmap='gray')
11 plt.title(f'Label: {int(label)}', fontsize=16)
12 plt.axis('off')
13 plt.colorbar(label='Pixel Intensity')
14 plt.tight_layout()
15 plt.show()
```



可以得到图如下：

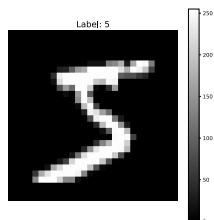
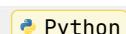


图 4.6 第一行的图片

我们数据集中一共有 60000 张图片，以及对应的 60000 个标签。我们先将数据转换成 numpy 的 ndarray 格式的数据，然后将数据打乱。

```
1 data = np.array(data) # 转换成ndarray格式
2 print(data.shape) # 打印形状
```



```
3 m, n = data.shape
4 np.random.shuffle(data) # 打乱数据
```

可以看到数据的形状是(60000, 785)。

我们接下来要将数据集切分为训练数据集和验证数据集。

🔥 训练数据集和验证数据集

- 训练数据集：用来训练我们的神经网络
- 验证数据集：用来验证训练好的神经网络表现怎么样

我们选择前 1000 行数据作为验证集

```
1 data_dev = data[0: 1000].T # 取前1000行，并转置
2 print(data_dev.shape)
```

Python

可以看到 `data_dev` 的形状是：(785, 1000)。也就是说，数据的每一列都是“1个标签+784个像素点”，方便我们后续处理。

然后我们将验证集数据的标签和像素点数据拆分开，先来提取标签数据

```
1 Y_dev = data_dev[0] # 提取标签数据
2 print(Y_dev.shape) # 形状: (1000,)
```

Python

然后提取标签对应的像素点数据，需要注意的是灰度图的每个像素点的范围是 0 ~ 255，像素点的类型是整型，为了方便处理，我们进行归一化，也就是将所有像素点的值归一化到 0 ~ 1 之间。

```
1 X_dev = data_dev[1 : n] # 提取像素点数据
2 X_dev = X_dev / 255.0 # 将像素点归一化
3 print(X_dev.shape) # 形状: (784, 1000)
```

Python

剩下的 59000 条数据我们作为训练模型用的训练数据集。处理方式和上面的验证集基本相同

```
1 data_train = data[1000 : m].T # 取后面的59000行数据并转置
2 print(data_train.shape) # 形状: (785, 59000)
3 Y_train = data_train[0] # 提取标签，形状为(59000,)
4 X_train = data_train[1 : n] # 提取像素点的数据，形状为(784, 59000)
5 X_train = X_train / 255.0 # 将像素点进行归一化
```

Python

这样我们的数据集就准备好了。

4.3 神经网络模型结构的设计

在式 (4.1) 中，我们知道神经网络其实是一个函数 $f(\bullet)$ 。只不过这个函数可能比较复杂，参数比较多。

一个简单的 ReLU 神经元如下所示：

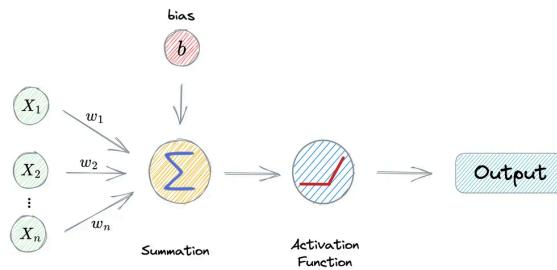


图 4.7 ReLU 神经元

其中 ReLU (Rectified Linear Unit, 整流线型单元) 激活函数的定义如下：

$$\text{ReLU}(x) = \max(0, x) = \begin{cases} x & \text{if } x > 0 \\ 0 & x \leq 0 \end{cases} \quad (4.49)$$

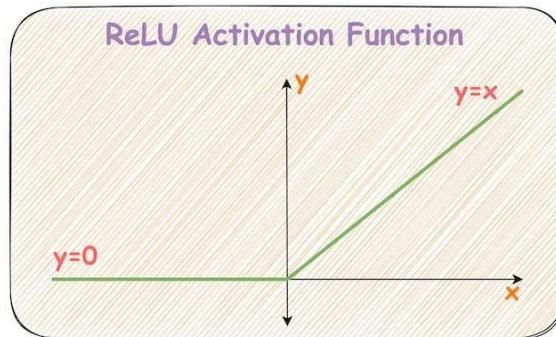


图 4.8 ReLU 激活函数

上面的图像如果写成数学函数，就有

$$\begin{aligned} f(x) &= \text{ReLU}(w_1x_1 + w_2x_2 + \dots + w_nx_n + b) \\ &= \text{ReLU}\left(\begin{bmatrix} w_1 & w_2 & \dots & w_n \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + b\right) \end{aligned} \quad (4.50)$$

这个 ReLU 神经元本身用处不大，但如果有序的组织起来，就能发挥巨大的威力。因为 ReLU 是一个非线性的函数，所以在神经网络中可以引入非线性因素，事实上，ReLU 神经元的叠加可以拟合任意函数，这叫做神经网络的万能逼近定理。

为什么要引入非线性函数？

考虑一元线性函数的简单复合，例如

$$\begin{aligned} f(x) &= w_2(w_1x + b_1) + b_2 \\ &= \underbrace{w_2w_1}_w x + \underbrace{w_2b_1 + b_2}_b \\ &= wx + b \end{aligned} \quad (4.51)$$

可以看到，线性函数的简单复合还是一个线性函数，所以无法拟合比较复杂的函数。

我们要设计的网络结构如下：

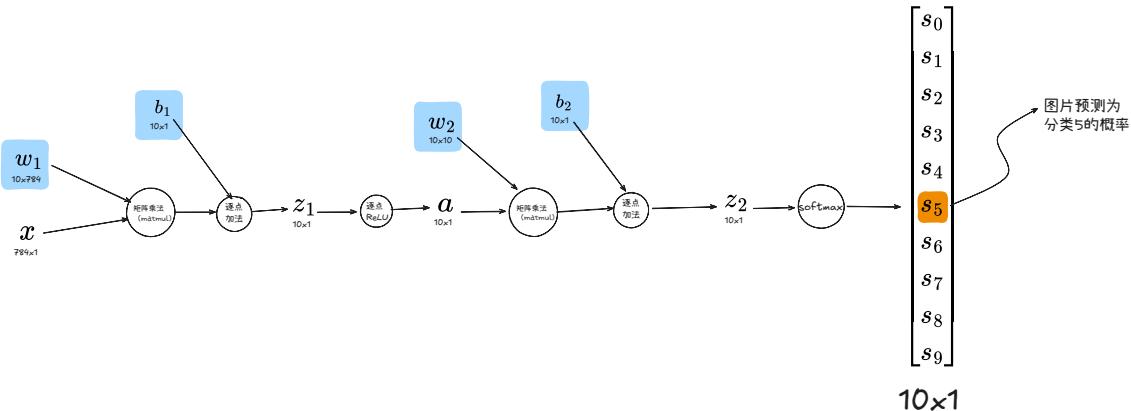


图 4.9 用于解决手写数字分类的神经网络结构，其中蓝色部分为需要学习的参数

网络对应的函数表达式如下所示：

$$\text{output} = \begin{bmatrix} s_0 \\ s_1 \\ s_2 \\ \vdots \\ s_9 \end{bmatrix} = \text{softmax}(\mathbf{W}^{[2]} \cdot \text{ReLU}(\mathbf{W}^{[1]} \cdot \mathbf{X} + \mathbf{b}^{[1]}) + \mathbf{b}^{[2]}) \quad (4.52)$$

在上面的网络中，我们的输入数据是一张 784 像素点的图片数据

$$\mathbf{X} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{783} \end{bmatrix} \quad (4.53)$$

输出是这张图片属于某一个分类的概率（预测值）。

$$\text{output} = \begin{bmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \\ s_7 \\ s_8 \\ s_9 \end{bmatrix} \quad \text{输入图片属于分类5的概率} \quad (4.54)$$

而参数 $\mathbf{W}^{[1]}, \mathbf{b}^{[1]}, \mathbf{W}^{[2]}, \mathbf{b}^{[2]}$ 是需要学习的参数，也就是说在训练过程中不断的改变的参数。我们希望在训练完成以后，给神经网络这个函数输入一张图片数据，能够得到它的分类的概率。其中概率最大的分类，是这张图片的正确分类。

$$\mathbf{W}^{[1]} = \begin{bmatrix} w_{0,0}^{[1]} & w_{0,1}^{[1]} & \cdots & w_{0,783}^{[1]} \\ w_{1,0}^{[1]} & w_{1,1}^{[1]} & \cdots & w_{1,783}^{[1]} \\ \vdots & \vdots & \ddots & \vdots \\ w_{9,0}^{[1]} & w_{9,1}^{[1]} & \cdots & w_{9,783}^{[1]} \end{bmatrix} \quad (4.55)$$

$$\mathbf{b}^{[1]} = \begin{bmatrix} b_0^{[1]} \\ b_1^{[1]} \\ b_2^{[1]} \\ \vdots \\ b_9^{[1]} \end{bmatrix} \quad (4.56)$$

$$\mathbf{W}^{[2]} = \begin{bmatrix} w_{0,0}^{[2]} & w_{0,1}^{[2]} & \cdots & w_{0,9}^{[2]} \\ w_{1,0}^{[2]} & w_{1,1}^{[2]} & \cdots & w_{1,9}^{[2]} \\ \vdots & \vdots & \ddots & \vdots \\ w_{9,0}^{[2]} & w_{9,1}^{[2]} & \cdots & w_{9,9}^{[2]} \end{bmatrix} \quad (4.57)$$

$$\mathbf{b}^{[2]} = \begin{bmatrix} b_0^{[2]} \\ b_1^{[2]} \\ b_2^{[2]} \\ \vdots \\ b_9^{[2]} \end{bmatrix} \quad (4.58)$$

其中计算的中间值有

$$\mathbf{M} = \begin{bmatrix} m_0 \\ m_1 \\ \vdots \\ m_9 \end{bmatrix} = \begin{bmatrix} w_{0,0}^{[1]} \cdot x_0 + w_{0,1}^{[1]} \cdot x_1 + \cdots + w_{0,783}^{[1]} \cdot x_{783} + b_0^{[1]} \\ w_{1,0}^{[1]} \cdot x_0 + w_{1,1}^{[1]} \cdot x_1 + \cdots + w_{1,783}^{[1]} \cdot x_{783} + b_1^{[1]} \\ \vdots \\ w_{9,0}^{[1]} \cdot x_0 + w_{9,1}^{[1]} \cdot x_1 + \cdots + w_{9,783}^{[1]} \cdot x_{783} + b_9^{[1]} \end{bmatrix} \quad (4.59)$$

$$\mathbf{A} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_9 \end{bmatrix} = \begin{bmatrix} \max(m_0, 0) \\ \max(m_1, 0) \\ \vdots \\ \max(m_9, 0) \end{bmatrix} \quad (4.60)$$

$$\mathbf{Z} = \begin{bmatrix} z_0 \\ z_1 \\ \vdots \\ z_9 \end{bmatrix} = \begin{bmatrix} w_{0,0}^{[2]} \cdot a_0 + w_{0,1}^{[2]} \cdot a_1 + \cdots + w_{0,9}^{[2]} \cdot a_9 + b_0^{[2]} \\ w_{1,0}^{[2]} \cdot a_0 + w_{1,1}^{[2]} \cdot a_1 + \cdots + w_{1,9}^{[2]} \cdot a_9 + b_1^{[2]} \\ \vdots \\ w_{9,0}^{[2]} \cdot a_0 + w_{9,1}^{[2]} \cdot a_1 + \cdots + w_{9,9}^{[2]} \cdot a_9 + b_9^{[2]} \end{bmatrix} \quad (4.61)$$

我们现在可以看到 \mathbf{Z} 是一个 10×1 形状的向量。但是我们都知道我们要输出的是图片 \mathbf{X} 属于每个类别的概率（共 10 个分类）。

所以我们需要将 \mathbf{Z} 中的每个元素转换成概率值，也就是 10 个元素的和是 1。

我们需要找到一个函数能够将 \mathbf{Z} 转换成概率值，同时不能改变 \mathbf{Z} 中每个元素的大小顺序。

也即是如果在 \mathbf{Z} 中， $z_0 < z_2$ ，那么这两个元素转换成概率值以后，还得是第 0 个元素小于第 2 个元素。

能够达到这个目的的函数就是大名鼎鼎的“softmax”函数。

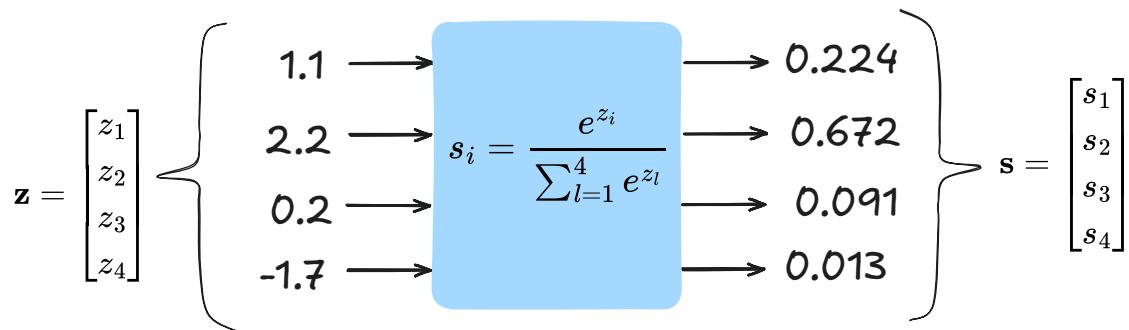


图 4.10 softmax 函数

softmax 函数的主要作用是将任意实数的向量转换为概率。上图公式中的指数函数确保了得到的值是非负的。由于分母中的归一化项，得到的值总和为 1。此外，所有值都介于 0 和 1 之间。**softmax** 函数的一个重要特性是它能保持其输入值的排序顺序：

$$-1.7 < 0.2 < 1.1 < 2.2 \Rightarrow 0.013 < 0.091 < 0.224 < 0.672 \quad (4.62)$$

所以我们的神经网络的输出是

$$\text{output} = \begin{bmatrix} s_0 \\ s_1 \\ \vdots \\ s_9 \end{bmatrix} = \begin{bmatrix} \frac{e^{z_0}}{\sum_{l=0}^9 e^{z_l}} \\ \frac{e^{z_1}}{\sum_{l=0}^9 e^{z_l}} \\ \vdots \\ \frac{e^{z_9}}{\sum_{l=0}^9 e^{z_l}} \end{bmatrix} \quad (4.63)$$

这样我们就可以将输出解释为属于某个分类的概率了。这里要注意的是我们只是能将输出解释为概率，只有经过神经网络的训练，输出才会慢慢接近真正的概率值。

4.4 如何设计损失函数？

由于神经网络在接收一张图片作为输入后，输出是属于分类的概率组成的向量。

$$\text{output} = \begin{bmatrix} s_0 \\ s_1 \\ \vdots \\ s_9 \end{bmatrix} \quad (4.64)$$

例如手写数字 5 在输入给未经训练的神经网络以后输出的概率如下

$$\Sigma \Rightarrow f(\bullet) \Rightarrow \begin{bmatrix} 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ \textcolor{red}{0.001} \\ 0.1 \\ 0.1 \\ 0.199 \\ 0.1 \end{bmatrix} \quad (4.65)$$

也就是针对图片预测为正确分类 5 的概率是 0.001，错的离谱。而我们的标签数据是 5，是没有办法和输出的概率向量做比较的，那么如何做比较呢？那就是将标签数据 5 转换成向量，也就是独热编码（one-hot encoder）。

$$5 \Rightarrow \text{独热编码} \Rightarrow \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad (4.66)$$

索引为 5 的元素编码为 1，其它元素一律编码为 0

独热编码的直观解释其实就是：手写数字 5 图片属于分类 5 的概率为 1。

代码如下：

```
1 def one_hot(Y):
2     one_hot_Y = np.zeros((Y.size, Y.max() + 1))
3     one_hot_Y[np.arange(Y.size), Y] = 1
4     one_hot_Y = one_hot_Y.T
5     return one_hot_Y
6
7 print(one_hot(Y_train)) # 对所有数据进行独热编码
8 print(Y_train[10]) # 第10张图片的标签
9 print(one_hot(Y_train)[:, 10]) # 第10张图片的标签的独热编码
```

Python

有了独热编码之后，如何度量预测的概率向量和真实的独热向量之间的损失呢？

$$\mathbf{s} = \begin{bmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \\ s_7 \\ s_8 \\ s_9 \end{bmatrix} \Leftarrow \text{如何度量两者之间的差异?} \Rightarrow \mathbf{y} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad (4.67)$$

这就是交叉熵损失函数

$$\mathcal{L}(\mathbf{s}, \mathbf{y}) = - \sum_{k=0}^9 y_k \log(s_k) \quad (4.68)$$

由于 one-hot 的性质（只有 $y_5 = 1$ ），所以化简可以得到

$$\mathcal{L} = -\log(s_5) \quad (4.69)$$

⚡ 损失函数

到目前为止，我们接触了两种损失函数：

- 均方误差损失 (`mean square error loss, MSE loss`)：用在线性回归中
- 交叉熵损失 (`cross entropy loss, CE loss`)：用在分类任务中

这两种损失函数并不是拍脑袋得来的，而是有着严谨的数学背景，我们后面会讲解。

直觉和例子

交叉熵在做两件事：

- 如果 s_i 很大（接近 1），则 $-\log(s_i)$ 很小，损失小
- 如果 s_i 很小（接近 0），则 $-\log(s_i)$ 很大，损失大

举个数值感受一下

- 如果 $s_i = 0.999$ ，那么 $\mathcal{L} = -\log(0.999) \approx 0.001$
- 如果 $s_i = 0.001$ ，那么 $\mathcal{L} = -\log(0.001) \approx 6.907$

可见预测的分类越准确，那么损失越小。可以作为度量。

4.5 如何让损失函数最小？

我们使用的算法是梯度下降法，而梯度下降法需要求导数然后再更新参数，所以我们分两步走

1. 对参数 $\mathbf{W}^{[1]}, \mathbf{b}^{[1]}, \mathbf{W}^{[2]}, \mathbf{b}^{[2]}$ 求偏导数。
2. 更新参数。

我们还记得神经网络的函数如下所示：

$$\text{output} = \text{softmax}(\mathbf{W}^{[2]} \cdot \text{ReLU}(\mathbf{W}^{[1]} \cdot \mathbf{X} + \mathbf{b}^{[1]}) + \mathbf{b}^{[2]}) \quad (4.70)$$

如果将上面的式子拆分开，就得到了

$$\begin{aligned} \mathbf{M} &= \mathbf{W}^{[1]} \mathbf{X} + \mathbf{b}^{[1]} \\ \mathbf{A} &= \text{ReLU}(\mathbf{M}) \\ \mathbf{Z} &= \mathbf{W}^{[2]} \mathbf{A} + \mathbf{b}^{[2]} \\ \text{output} &= \text{softmax}(\mathbf{Z}) \end{aligned} \quad (4.71)$$

而这就是反向传播算法的前向传播过程，代码如下：

```
1 def forward(w1, b1, w2, b2, X):
2     M = w1.dot(X) + b1
3     A = ReLU(M)
4     Z = w2.dot(A) + b2
5     output = softmax(Z)
6     return M, A, Z, output
```

Python

对应的 `ReLU` 函数和 `softmax` 函数如下：

```
1 def ReLU(M):
2     return np.maximum(M, 0)
3
4 def softmax(Z):
5     A = np.exp(Z) / sum(np.exp(Z))
6     return A
```

Python

我们在前向过程中保存了一些中间计算结果： \mathbf{M} , \mathbf{A} , \mathbf{Z} , output 。用于在反向传播过程中求解损失函数 \mathcal{L} 对参数的导数。

我们的损失函数是交叉熵损失函数

$$\mathcal{L}(\mathbf{s}, \mathbf{y}) = -\sum_{k=0}^9 y_k \log(s_k) = -\sum_{k=0}^9 y_k \log\left(\frac{e^{z_k}}{\sum_{l=0}^9 e^{z_l}}\right) \quad (4.72)$$

那么损失函数如何对参数进行求偏导数呢？

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[2]}} &=? \\ \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{[2]}} &=? \\ \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[1]}} &=? \\ \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{[1]}} &=? \end{aligned} \quad (4.73)$$

根据链式求导法则

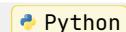
$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[2]}} &= \frac{\partial \mathcal{L}}{\partial \mathbf{Z}} \frac{\partial \mathbf{Z}}{\partial \mathbf{W}^{[2]}} \\ \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{[2]}} &= \frac{\partial \mathcal{L}}{\partial \mathbf{Z}} \frac{\partial \mathbf{Z}}{\partial \mathbf{b}^{[2]}} \\ \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[1]}} &= \frac{\partial \mathcal{L}}{\partial \mathbf{Z}} \frac{\partial \mathbf{Z}}{\partial \mathbf{A}} \frac{\partial \mathbf{A}}{\partial \mathbf{M}} \frac{\partial \mathbf{M}}{\partial \mathbf{W}^{[1]}} \\ \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{[1]}} &= \frac{\partial \mathcal{L}}{\partial \mathbf{Z}} \frac{\partial \mathbf{Z}}{\partial \mathbf{A}} \frac{\partial \mathbf{A}}{\partial \mathbf{M}} \frac{\partial \mathbf{M}}{\partial \mathbf{b}^{[1]}} \end{aligned} \quad (4.74)$$

这里的变量都是矩阵或者向量，所以求导数很复杂，这里我们先给出结论。

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{Z}} &= \text{output} - \mathbf{y} \\ \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[2]}} &= \frac{\partial \mathcal{L}}{\partial \mathbf{Z}} \cdot \mathbf{A}^T \\ \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{[2]}} &= \frac{\partial \mathcal{L}}{\partial \mathbf{Z}} \\ \frac{\partial \mathcal{L}}{\partial \mathbf{M}} &= \mathbf{W}^{[2]} \cdot \frac{\partial \mathcal{L}}{\partial \mathbf{Z}} \odot \{1 \text{ if } \mathbf{M} > 0 \text{ else } 0\} \\ \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[1]}} &= \frac{\partial \mathcal{L}}{\partial \mathbf{M}} \cdot \mathbf{X}^T \\ \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{[1]}} &= \frac{\partial \mathcal{L}}{\partial \mathbf{M}} \end{aligned} \quad (4.75)$$

代码如下：

```
1 def deriv_ReLU(z):
2     return z > 0
3
4 def backward(M, A, Z, output, w2, Y, X):
5     OneHot_Y = one_hot(Y)
6     dZ = output - OneHot_Y
7     dW2 = 1/m * dZ.dot(A.T)
8     db2 = 1/m * np.sum(dZ)
9     dM = w2.T.dot(dZ) * deriv_ReLU(M)
```



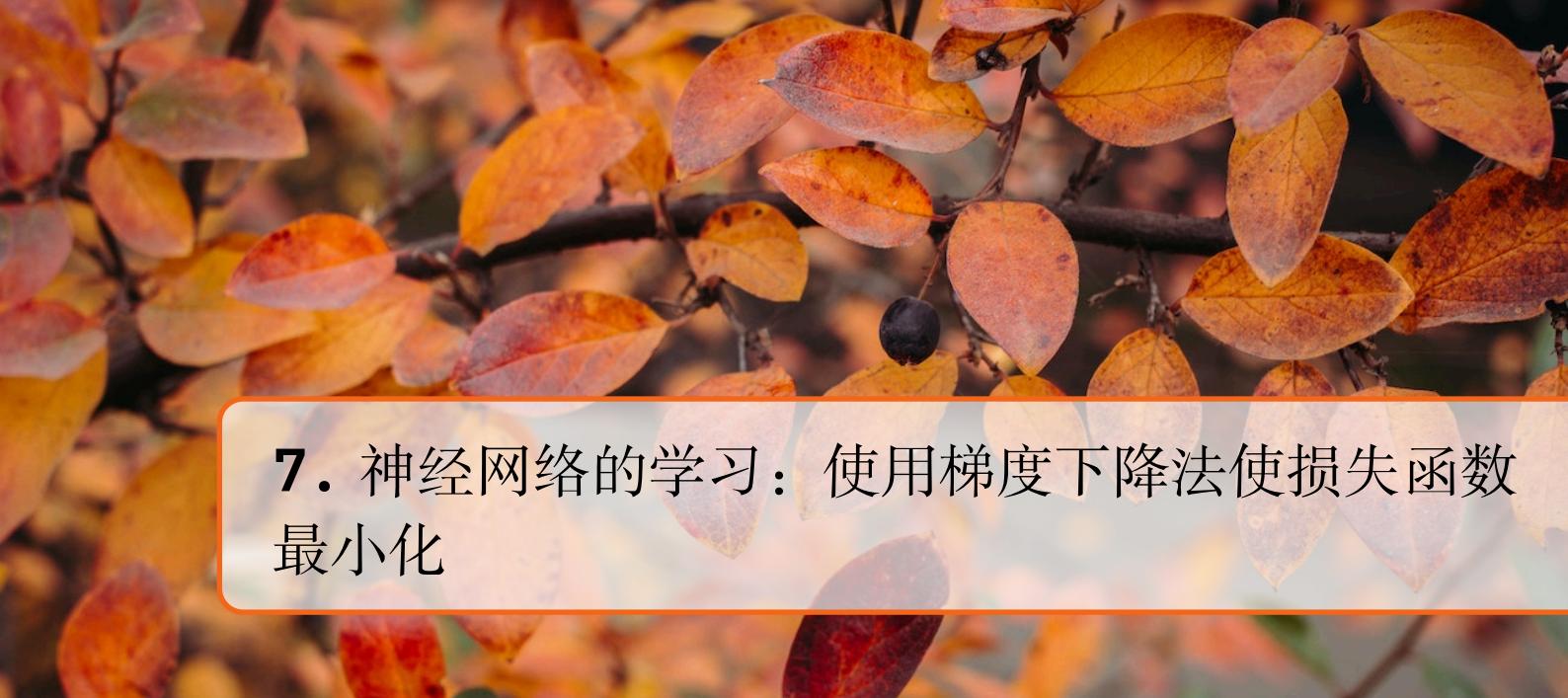
```
10     dW1 = 1 / m * dM.dot(X.T)
11     db1 = 1 / m * np.sum(dM)
12     return dW1, db1, dW2, db2
```



5. 卷积神经网络：将手写数字识别准确率拉满！



6. 损失函数：均方误差损失和交叉熵损失的由来



7. 神经网络的学习：使用梯度下降法使损失函数最小化



8. PyTorch 简介

本附录旨在为你提供必要的技能和知识，以便将深度学习付诸实践并从零开始实现大语言模型。我们将使用 **PyTorch** 作为本书的主要工具，这是一个广泛应用的 **Python** 深度学习库。

首先，我们将指导你搭建一个支持 **PyTorch** 和 **GPU** 的深度学习工作台。然后，我们将介绍张量的基本概念及其在 **PyTorch** 中的用法。接下来，我们将深入探讨 **PyTorch** 的自动微分引擎，这一特性使我们能够方便且高效地使用反向传播，这也是神经网络训练的重要环节。

本附录的目标是为那些刚接触 **PyTorch** 深度学习的读者提供入门资料。虽然我们会从零开始讲解 **PyTorch**，但不会覆盖其所有功能，而是聚焦于实现大语言模型所需的基本概念。如果你对深度学习已有一定了解，那么可以跳过本附录，直接往下阅读。

8.1 什么是 **PyTorch**

PyTorch 是一个开源的基于 **Python** 的深度学习库。根据 **Papers With Code** 这个跟踪和分析研究论文平台的数据，自 2019 年以来，**PyTorch** 已成为研究领域使用最广泛的深度学习库，并且领先优势显著。此外，根据 2022 年 **Kaggle** 数据科学与机器学习调查，大约 40% 的受访者正在使用 **PyTorch**，并且这一比例每年都在增长。

PyTorch 之所以如此受欢迎，原因之一在于其用户友好的界面和高效性。它不仅易于使用，还保留了高度的灵活性，允许专业用户深入修改模型的底层组件，以实现个性化和优化。总之，对许多从业者和研究人员而言，**PyTorch** 在可用性和特性之间提供了恰到好处的平衡。

8.1.1 **PyTorch** 的三大核心组件

PyTorch 是一个相对全面的库，我们可以通过关注其三大核心组件来理解它，如下图所示。

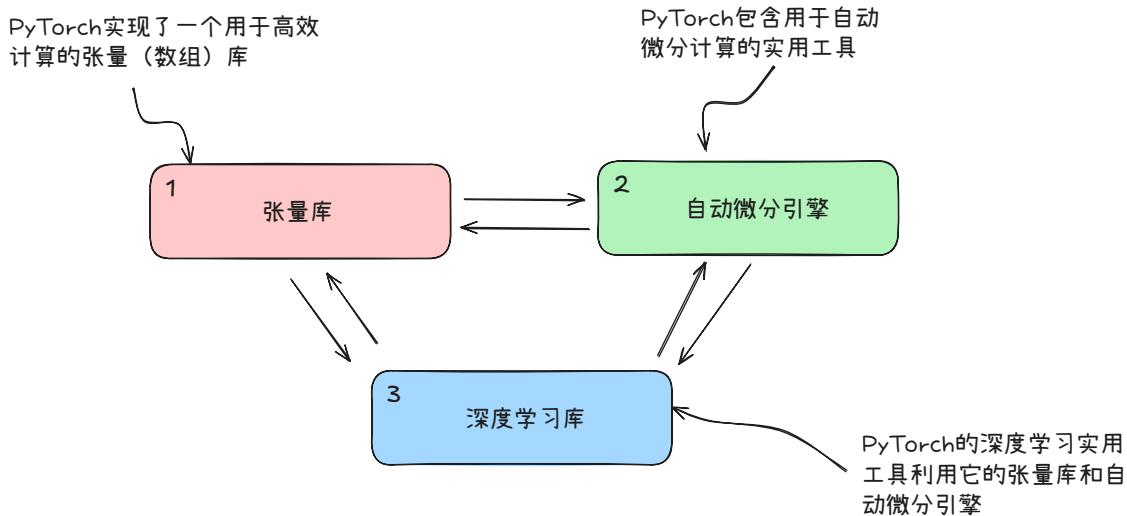


图 8.1 PyTorch 的三大核心组件包括作为计算基础构建块的张量库、用于模型优化的自动微分引擎以及深度学习工具函数，这使得实现和训练深度神经网络模型更加容易

首先，PyTorch 是一个张量库，它扩展了 NumPy 基于数组的编程功能，增加了 GPU 加速特性，从而实现了 CPU 和 GPU 之间的无缝计算切换。其次，PyTorch 是一个自动微分引擎，也称为 `autograd`，它能够自动计算张量操作的梯度，从而简化反向传播和模型优化。最后，PyTorch 是一个深度学习库，它提供了模块化、灵活且高效的构建块（包括预训练模型、损失函数和优化器），能够帮助研究人员和开发人员轻松设计和训练各种深度学习模型。

8.1.2 定义深度学习

在新闻中，大语言模型通常被称为“人工智能模型”。然而，大语言模型实际上也是一种深度神经网络，而 PyTorch 是一个深度学习库。是不是听起来有些困惑？在继续之前，让我们简要总结一下这些术语之间的关系。

人工智能的基本目标是创建能够执行通常需要人类智能水平的任务的计算机系统。这些任务包括自然语言理解、模式识别和决策制定。（尽管取得了显著进展，但人工智能仍远未达到这种通用智能的水平。）

如下图所示，机器学习是人工智能的一个子领域，其专注于学习算法的开发和改进。机器学习背后的主要理念是使计算机能够从数据中学习，并在没有被明确编程的情况下进行预测或决策。这涉及能够识别模式、从历史数据中学习，并随着时间的推移通过更多数据和反馈提升性能的算法。

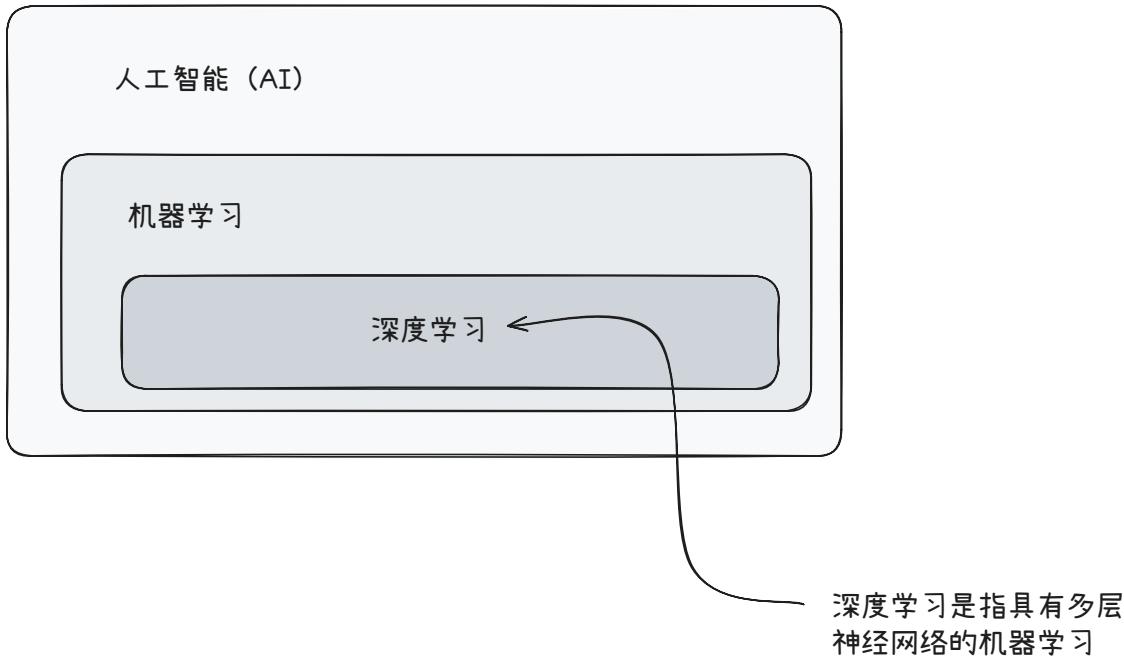


图 8.2 深度学习是机器学习的一个子类别，专注于实现深度神经网络。机器学习是人工智能的一个子类别，涉及从数据中学习的算法。人工智能是一个更广泛的概念，指的是机器能够执行通常需要人类智能水平的任务

机器学习在人工智能的演变中发挥了重要作用，为我们今天所看到的许多进展（包括大语言模型）提供了动力。机器学习还支持在线零售商和流媒体服务使用的推荐系统、垃圾邮件过滤、虚拟助手中的语音识别，甚至自动驾驶汽车等技术。机器学习的引入和发展显著增强了人工智能的能力，使其超越传统的基于规则的系统，并能够适应新的输入或变化的环境。

深度学习是机器学习的一个子类别，专注于深度神经网络的训练和应用。这些深度神经网络最初受到人脑工作原理（特别是许多神经元之间的相互连接）的启发。深度学习中的“深度”指的是人工神经元或节点的多个隐藏层，这些层使它们能够对数据中的复杂非线性关系进行建模。与传统机器学习技术擅长简单模式识别不同，深度学习擅长处理诸如图像、音频、文本之类的非结构化数据，因此特别适合用于大语言模型。

机器学习和深度学习中典型的预测建模工作流程（也称为监督学习）如下图所示。

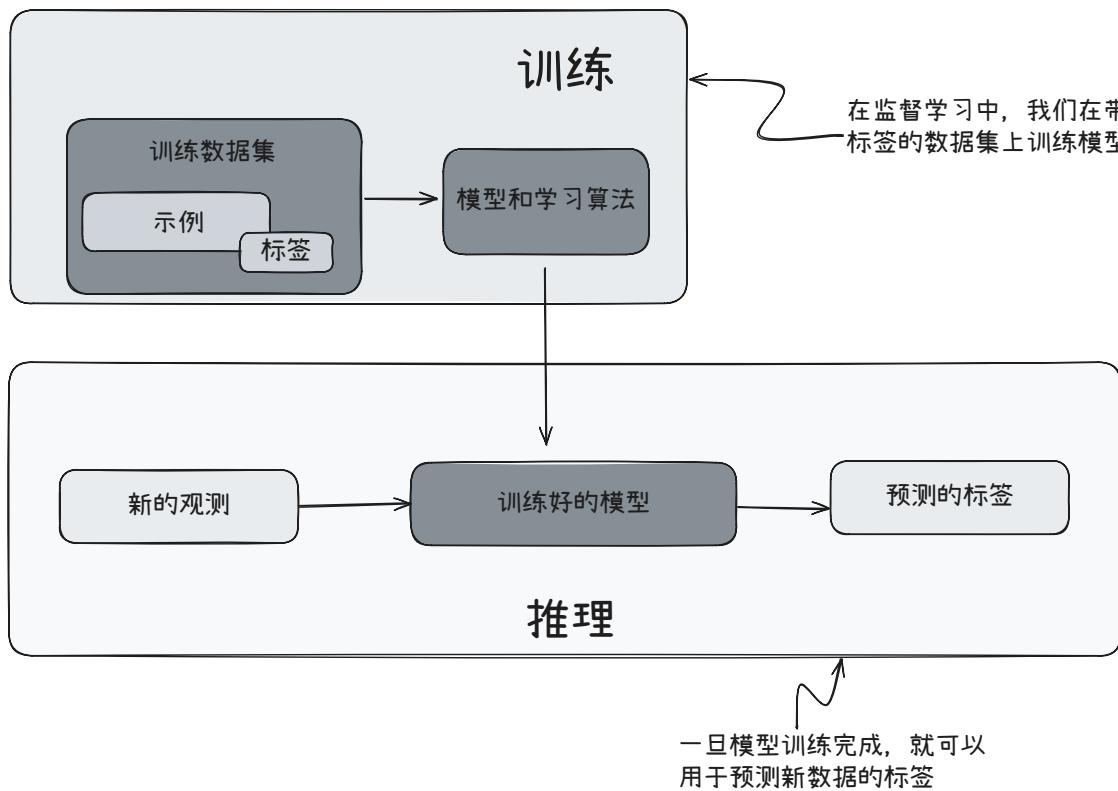


图 8.3 监督学习的预测建模工作流程包括一个训练阶段，在该阶段中，模型在训练数据集中带标签的示例上进行训练。训练好的模型随后可用于预测新观测数据的标签

通过使用学习算法，模型可以在由示例和相应标签组成的训练数据集上进行训练。例如，在垃圾邮件分类器的案例中，训练数据集由电子邮件及其“垃圾消息”和“非垃圾消息”标签组成，这些标签是由人类标注的。然后，训练好的模型可以在新的样本（新的电子邮件）上使用，以预测这些样本的未知标签（“垃圾消息”或“非垃圾消息”）。当然，我们还希望在训练阶段和推理阶段之间添加模型评估，以确保模型在实际应用之前满足性能标准。

如果想要训练大语言模型来对文本进行分类，那么训练和使用大语言模型的工作流程与图 A-3 中描述的类似。即使你关注的是训练大语言模型来生成文本（这也是我们的主要关注点），图 A-3 仍然适用。在这种情况下，预训练期间的标签可以从文本本身获取（第 1 章介绍的下一单词预测任务）。在推理时，大语言模型将在给定输入提示词的情况下生成全新的文本（而不是预测标签）。

8.1.3 安装 PyTorch

PyTorch 可以像其他任何 Python 库或包一样进行安装。然而，由于 PyTorch 是一个包含 CPU 和 GPU 兼容代码的综合性库，安装过程可能需要额外说明。

🔥 Python 版本

许多科学计算库不会立即支持最新版本的 Python。因此，在安装 PyTorch 时，建议使用比最新版本旧一到两个版本的 Python。如果最新的 Python 版本是 Python 3.13，那么推荐使用 Python 3.11 或 Python 3.12。

例如，PyTorch 有两个版本：一个是仅支持 CPU 计算的精简版，另一个是支持 CPU 和 GPU 计算的完整版。如果你的机器有一个兼容 CUDA 的 GPU（理想情况下是 NVIDIA T4、RTX 2080 Ti 或更新的型号），那么推荐安装 GPU 版本。以下是在代码终端中安装 PyTorch 的默认命令：

```
1 $ pip install torch
2 # 或者指定版本安装
3 $ pip install torch==2.7.0
```

Shell

假设你的计算机支持兼容 CUDA 的 GPU。在这种情况下，如果你正在使用的 Python 环境已安装必要的依赖项（如 pip），那么系统将自动安装支持 CUDA 加速的 PyTorch 版本。

安装 PyTorch 后，可以通过在 Python 中运行以下代码来检查安装是否识别了内置的 NVIDIA GPU：

```
1 import torch
2 print(torch.__version__) # 打印torch版本
3 print(torch.cuda.is_available()) # 打印torch是否支持cuda
```

Python

8.2 理解张量

张量表示一个数学概念，它可以将向量和矩阵推广到潜在的更高维度。换句话说，张量是可以通过其阶数（或秩）来表征的数学对象，其中阶数提供了维度的数量。例如，标量（仅是一个数值）是秩为 0 的张量，向量是秩为 1 的张量，矩阵是秩为 2 的张量，如下图所示。

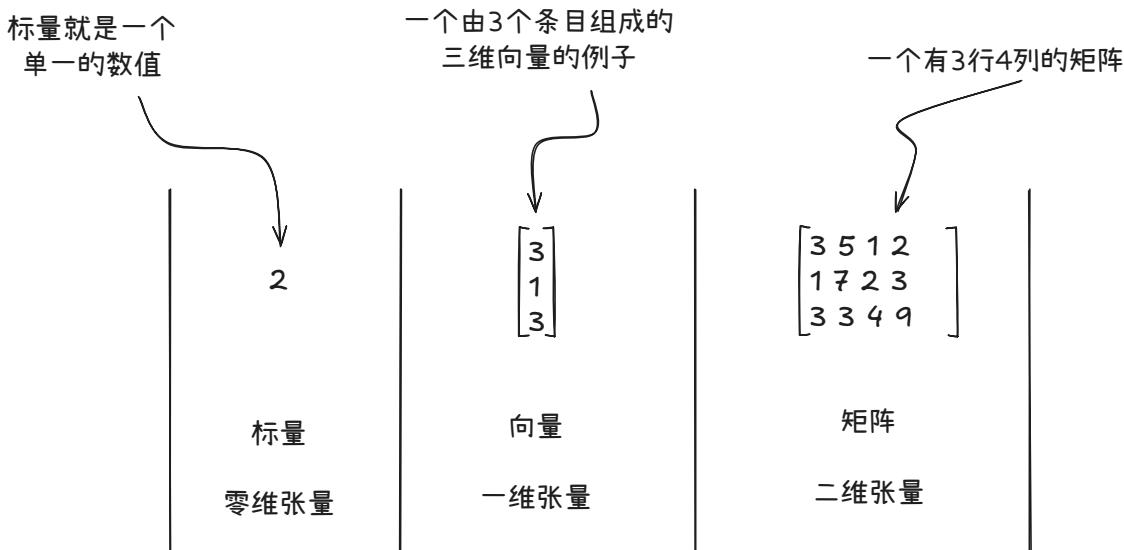


图 8.4 不同秩的张量。这里零维对应于秩 0，一维对应于秩 1，二维对应于秩 2。一个由 3 个元素组成的三维向量仍然是秩为 1 的张量

从计算的角度来看，张量是一种数据容器。举例来说，它们存储多维数据，其中每个维度表示一个不同的特征。像 PyTorch 这样的张量库能够高效地创建、操作和计算这些数组。在这个上下文中，张量库的功能类似于数组库。

PyTorch 张量类似于 NumPy 数组，但具有几个对深度学习至关重要的附加功能。例如，PyTorch 添加了一个自动微分引擎，简化了梯度计算（求导）。PyTorch 张量还支持 GPU 计算，以加速深度神经网络的训练。

8.2.1 标量、向量、矩阵和张量

如前所述，PyTorch 张量是用于与数组类似结构的数据容器。标量是零维张量（例如，仅一个数值），向量是一维张量，矩阵是二维张量。对于更高维的张量没有特定的术语，因此通常将三维张量

称为“3D 张量”，以此类推。可以使用 `torch.tensor()` 函数创建 PyTorch 的 `Tensor` 类对象，如代码所示。

```
创建PyTorch张量
1 import torch
2 # 从Python整数创建一个零维张量（标量）
3 tensor0d = torch.tensor(1)
4 # 从Python列表创建一个一维张量（向量）
5 tensor1d = torch.tensor([1, 2, 3])
6 # 从嵌套的Python列表创建一个二维张量
7 tensor2d = torch.tensor([[1, 2],
8                         [3, 4]])
9 # 从嵌套的Python列表创建一个三维张量
10 tensor3d = torch.tensor([[[1, 2], [3, 4]],
11                          [[5, 6], [7, 8]]])
```

8.2.2 张量数据类型

PyTorch 采用 Python 默认的 64 位整数数据类型。可以通过张量的 `.dtype` 属性来访问张量的数据类型，如下所示：

```
1 tensor1d = torch.tensor([1, 2, 3])
2 print(tensor1d.dtype)
```

输出如下所示：

```
1 torch.int64
```

如果使用 Python 浮点数创建张量，那么 PyTorch 默认会创建具有 32 位精度的张量：

```
1 floatvec = torch.tensor([1.0, 2.0, 3.0])
2 print(floatvec.dtype)
```

输出如下所示：

```
1 torch.float32
```

这种选择主要是为了在精度和计算效率之间取得平衡。32 位浮点数在大多数深度学习任务中提供了足够的精度，同时其消耗的内存和计算资源比 64 位浮点数更少。此外，GPU 架构对 32 位计算进行了优化，使用这种数据类型可以显著加快模型训练和推理速度。还可以使用张量的 `.to` 方法更改精度。以下代码演示了如何将 64 位整数张量更改为 32 位浮点张量：

```
1 floatvec = tensor1d.to(torch.float32)
2 print(floatvec.dtype)
```

这将返回以下内容。

```
1 torch.float32
```

8.2.3 常见的 PyTorch 张量操作

本书无法全面覆盖所有 PyTorch 张量操作和命令。然而，我们将在介绍它们时简要描述相关操作。我们已经介绍了创建新张量的 `torch.tensor()` 函数：

```
1 tensor2d = torch.tensor([[1, 2, 3],
2                           [4, 5, 6]])
```

```
3 print(tensor2d)
```

这将打印以下内容：

```
1 tensor([[1, 2, 3],  
2         [4, 5, 6]])
```

Python

此外，`.shape` 属性允许我们访问张量的形状：

```
1 print(tensor2d.shape)
```

Python

输出如下所示：

```
1 torch.Size([2, 3])
```

Python

如你所见，`.shape` 返回的是`[2, 3]`，这意味着该张量有 2 行 3 列。要将该张量变为 3×2 的形状，可以使用`.reshape` 方法：

```
1 print(tensor2d.reshape(3, 2))
```

Python

这将打印以下内容：

```
1 tensor([[1, 2],  
2         [3, 4],  
3         [5, 6]])
```

Python

然而，请注意，在 PyTorch 中，重塑张量更常用的命令是`.view()`：

```
1 print(tensor2d.view(3, 2))
```

Python

输出如下所示：

```
1 tensor([[1, 2],  
2         [3, 4],  
3         [5, 6]])
```

Python

类似于`.reshape` 和`.view`，在某些情况下，PyTorch 提供了多种语法选项来执行相同的计算。PyTorch 最初遵循了原始 Lua 版本 Torch 的语法约定，但后来应用户的要求，添加了与 NumPy 类似的语法。`(.view()和.reshape())` 的微妙区别在于它们对内存布局的处理方式：`.view()` 要求原始数据是连续的，如果不是，它将无法工作，而`.reshape()` 会工作，如有必要，它会复制数据以确保所需形状。)

接下来，可以使用`.T` 来转置张量，这意味着将其沿对角线翻转。请注意，这与重塑张量类似，你可以从以下结果中看到这一点：

```
1 print(tensor2d.T)
```

Python

输出如下所示：

```
1 tensor([[1, 4],  
2         [2, 5],  
3         [3, 6]])
```

Python

最后，PyTorch 中常用的矩阵相乘方法是`.matmul` 方法：

```
1 print(tensor2d.matmul(tensor2d.T))
```

Python

输出如下所示：

```
1 tensor([[14, 32],  
2         [32, 77]])
```

Python

然而，也可以使用`@`运算符，它能够更简洁地实现相同的功能：

```
1 print(tensor2d @ tensor2d.T)
```

 Python

输出如下所示：

```
1 tensor([[14, 32],  
2         [32, 77]])
```

 Python

如前所述，我们会在需要时介绍额外的操作。

8.3 将模型视为计算图

现在让我们来了解一下 PyTorch 的自动微分引擎，也称为 `autograd`。PyTorch 的 `autograd` 系统能够在动态计算图中自动计算梯度。

计算图是一种有向图，主要用于表达和可视化数学表达式。在深度学习的背景下，计算图列出了计算神经网络输出所需的计算顺序——我们需要用它来计算反向传播所需的梯度，这是神经网络的主要训练算法。

让我们通过一个具体的例子来说明计算图的概念。下面代码实现了一个简单逻辑回归分类器的前向传播（预测步骤），我们可以将其看作一个单层神经网络。它会返回一个介于 0 和 1 之间的分数，当计算损失时，这个分数会与真实的类别标签（0 或 1）进行比较。

逻辑回归的前向传播

```
1 import torch.nn.functional as F  
2  
3 y = torch.tensor([1.0]) # 真实特征  
4 x1 = torch.tensor([1.1]) # 输入特征  
5 w1 = torch.tensor([2.2]) # 权重参数  
6 b = torch.tensor([0.0]) # 偏置单元  
7 z = x1 * w1 + b # 网络输入  
8 a = torch.sigmoid(z) # 激活和输出  
9 loss = F.binary_cross_entropy(a, y)
```

 Python

🔥 代码对应的数学表达式

线性变换

$$z = w_1 x_1 + b \quad (8.1)$$

激活函数（Sigmoid）：

$$a = \sigma(z) = \frac{1}{1 + e^{-z}} \quad (8.2)$$

损失函数（loss）：

$$\text{loss} = -[y \log(a) + (1 - y) \log(1 - a)] \quad (8.3)$$

即使没有完全理解上述代码中的所有部分，也不要担心。这个例子的重点不是实现一个逻辑回归分类器，而是为了说明如何将一系列计算看作一个计算图，如下图所示。

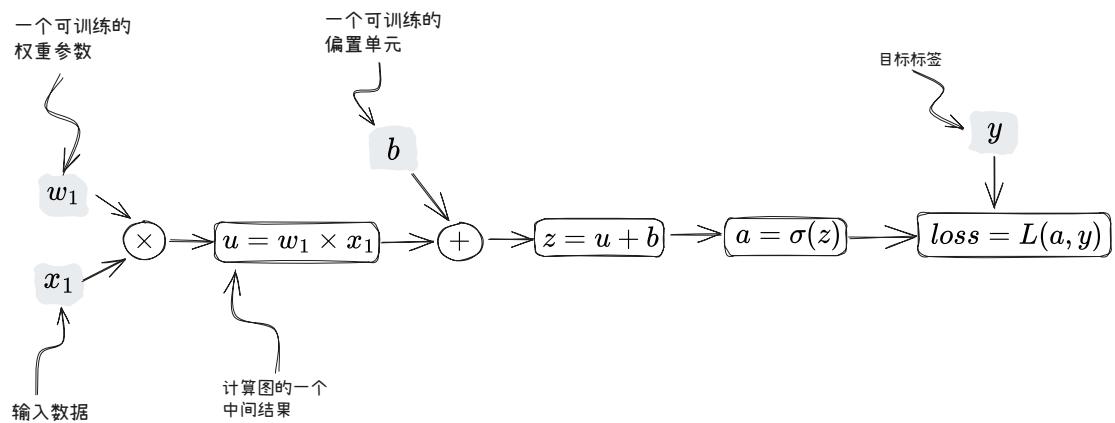


图 8.5 逻辑回归的前向传播作为一个计算图。输入特征 x_1 与模型权重 w_1 相乘，并在加上偏置后通过激活函数 σ 传递。损失是通过比较模型输出 a 与给定标签 y 来计算的

实际上，PyTorch 在后台构建了这样一个计算图，我们可以利用它来计算损失函数相对于模型参数（这里是 w_1 和 b ）的梯度，从而训练模型。

8.4 轻松实现自动微分

如果在 PyTorch 中进行计算，那么只要其终端节点之一的 `requires_grad` 属性被设置为 `True`，PyTorch 默认就会在内部构建一个计算图。这在我们想要计算梯度时非常有用。在训练神经网络时，需要使用反向传播算法计算梯度。反向传播可以被视为微积分中链式法则在神经网络中的应用，如下图所示。

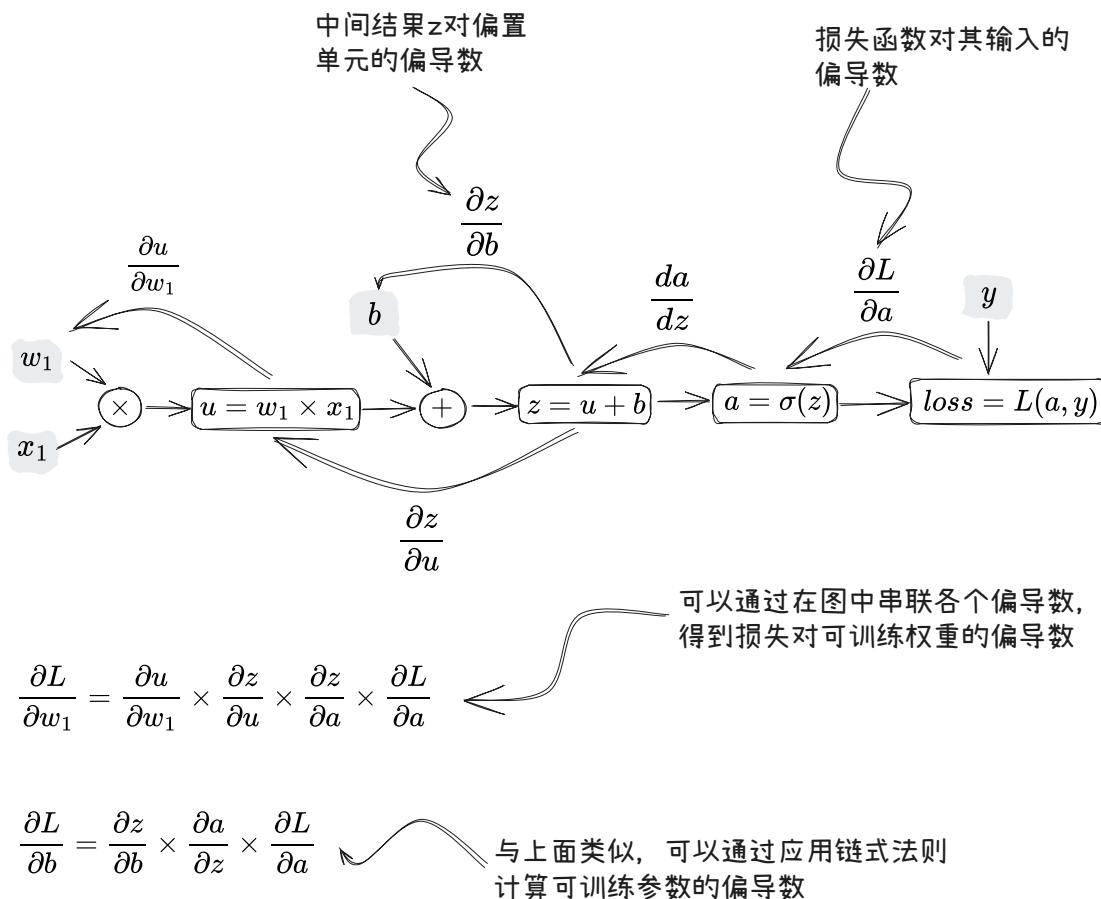


图 8.6 在计算图中计算损失梯度的最常见方法是从右向左应用链式法则，这也称为“反向模型自动求导”或“反向传播”。我们从输出层（或损失本身）开始，向后通过网络一直到输入层。这么做是为了计算损失相对于网络中每个参数（权重和偏置）的梯度，从而为训练过程中如何更新这些参数提供信息

偏导数和梯度

上图展示了偏导数，它测量的是一个函数相对于其中一个变量变化的速率。梯度是一个向量，包含了一个多变量函数（输入变量超过一个的函数）的所有偏导数。

如果你不太熟悉或记不清微积分中的偏导数、梯度或链式法则，不用担心。从高层次来看，你只需要知道链式法则如何在计算图中根据模型参数来计算损失函数的梯度即可。这提供了更新每个参数以最小化损失函数所需的信息，而这个损失函数作为衡量模型性能的代理，可以通过诸如梯度下降之类的方法来实现。我们将在后面重新审视在 PyTorch 中实现这一训练循环的计算过程。

那么，这一切与之前提到的 PyTorch 库的第二个组件——自动微分（`autograd`）引擎有何关联？PyTorch 的 `autograd` 引擎在后台通过跟踪在张量上执行的每个操作来构建计算图。然后，通过调用 `grad` 函数，可以计算损失相对于模型参数 w_1 的梯度，如下面的代码所示。

```
通过autograd计算梯度
1 import torch.nn.functional as F
2 from torch.autograd import grad
3
4 y = torch.tensor([1.0])
5 x1 = torch.tensor([1.1])
6 w1 = torch.tensor([2.2], requires_grad=True)
```

Python

```

7 b = torch.tensor([0.0], requires_grad=True)
8
9 z = x1 * w1 + b
10 a = torch.sigmoid(z)
11
12 loss = F.binary_cross_entropy(a, y)
13
14 # 默认情况下, PyTorch在计算梯度后会销毁计算图以释放内存。然而, 由于我们即将再次使用这个计算图, 因此可以设置`retain_graph=True`, 使其保留在内存中
15 grad_L_w1 = grad(loss, w1, retain_graph=True)
16 grad_L_b = grad(loss, b, retain_graph=True)

```

给定模型参数的损失结果值如下所示:

```

1 print(grad_L_w1)
2 print(grad_L_b)

```

这将打印如下内容:

```

1 (tensor([-0.0898]),)
2 (tensor([-0.0817]),)

```

这里我们手动使用了 `grad` 函数, 这在实验、调试和概念演示中很有用。但是, 在实际操作中, PyTorch 提供了更高级的工具来自动化这个过程。例如, 我们可以对损失函数调用 `.backward` 方法, 随后 PyTorch 将计算计算图中所有叶节点的梯度, 这些梯度将通过张量的 `.grad` 属性进行存储:

```

1 loss.backward()
2 print(w1.grad)
3 print(b.grad)

```

输出如下所示:

```

1 (tensor([-0.0898]),)
2 (tensor([-0.0817]),)

```

我给你提供了很多信息, 你可能会被微积分的概念弄得有些不知所措, 但不用担心。虽然这些微积分术语是为了解释 PyTorch 的 `autograd` 组件, 但你需要记住的仅仅是 PyTorch 通过 `.backward` 方法为我们处理了微积分问题——我们不需要手动计算任何导数或梯度。

手动计算一下梯度验证

- 前向过程, 将中间计算结果都保存下来

$$\begin{aligned}
 z &= w_1 x_1 + b = 2.2 \times 1.1 + 0.0 \\
 a &= \sigma(z) = \sigma(2.42) = \frac{1}{1 + e^{-2.42}} \approx 0.9183 \\
 \mathcal{L} &= -[y \log(a) + (1 - y) \log(1 - a)] \\
 &= -[1.0 \times \log(0.9183) + (1 - 1.0) \times \log(1 - 0.9183)] \\
 &= -[1.0 \times \log(0.9183) + 0] \\
 &= -\log(0.9183) \approx 0.0853
 \end{aligned} \tag{8.4}$$

- 反向过程, 利用前向过程保存的中间结果, 计算 w_1 和 b 的梯度 (偏导数)

🔥 用到的求导公式

对数求导法则

$$\begin{aligned}\frac{d \log x}{dx} &= \frac{1}{x} \\ \frac{d \log(f(x))}{dx} &= \frac{f'(x)}{f(x)}\end{aligned}\tag{8.5}$$

商法则：若 $f(x) = \frac{u(x)}{v(x)}$

$$f'(x) = \frac{u'(x)v(x) - u(x)v'(x)}{(v(x))^2}\tag{8.6}$$

自然对数

$$\frac{de^x}{dx} = e^x\tag{8.7}$$

$$\frac{\partial \mathcal{L}}{\partial w_1} = \frac{\partial \mathcal{L}}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial w_1}\tag{8.8}$$

其中

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial a} &= -\left[\frac{y}{a} + (1-y)\frac{-1}{1-a}\right] \quad (\frac{d \log(x)}{dx} = \frac{1}{x}) \\ &= \frac{1-y}{1-a} - \frac{y}{a} \quad (\text{y}=1.0) \\ &= -\frac{1.0}{0.9183} \quad (\text{利用中间计算结果 } a=0.9183)\end{aligned}\tag{8.9}$$

而

$$\begin{aligned}\frac{\partial a}{\partial z} &= \frac{\partial \sigma(z)}{\partial z} = \frac{\partial \frac{1}{1+e^{-z}}}{\partial z} \\ &= \frac{e^{-z}}{(1+e^{-z})^2} \\ &= \frac{1}{1+e^{-z}} \cdot \left(1 - \frac{1}{1+e^{-z}}\right) \\ &= \sigma(z)(1-\sigma(z)) \\ &= a(1-a) \quad (\text{利用中间计算结果 } a=0.9183) \\ &= 0.9183 \times (1 - 0.9183)\end{aligned}\tag{8.10}$$

显然

$$\frac{\partial z}{\partial w_1} = x_1 = 1.1\tag{8.11}$$

所以

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial w_1} &= \frac{\partial \mathcal{L}}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial w_1} \\ &= \left(-\frac{1.0}{0.9183}\right) \times 0.9183 \times (1 - 0.9183) \times 1.1 \\ &= -0.0898\end{aligned}\tag{8.12}$$

🔥 作业

按照上面的过程手动计算一下 $\frac{\partial \mathcal{L}}{\partial b}$

8.5 实现多层神经网络

接下来，我们将 PyTorch 视为实现深度神经网络的库来进行重点探讨。为了提供一个具体的例子，我们来看看多层感知机（multilayer perceptron），即全连接神经网络，如下图所示。

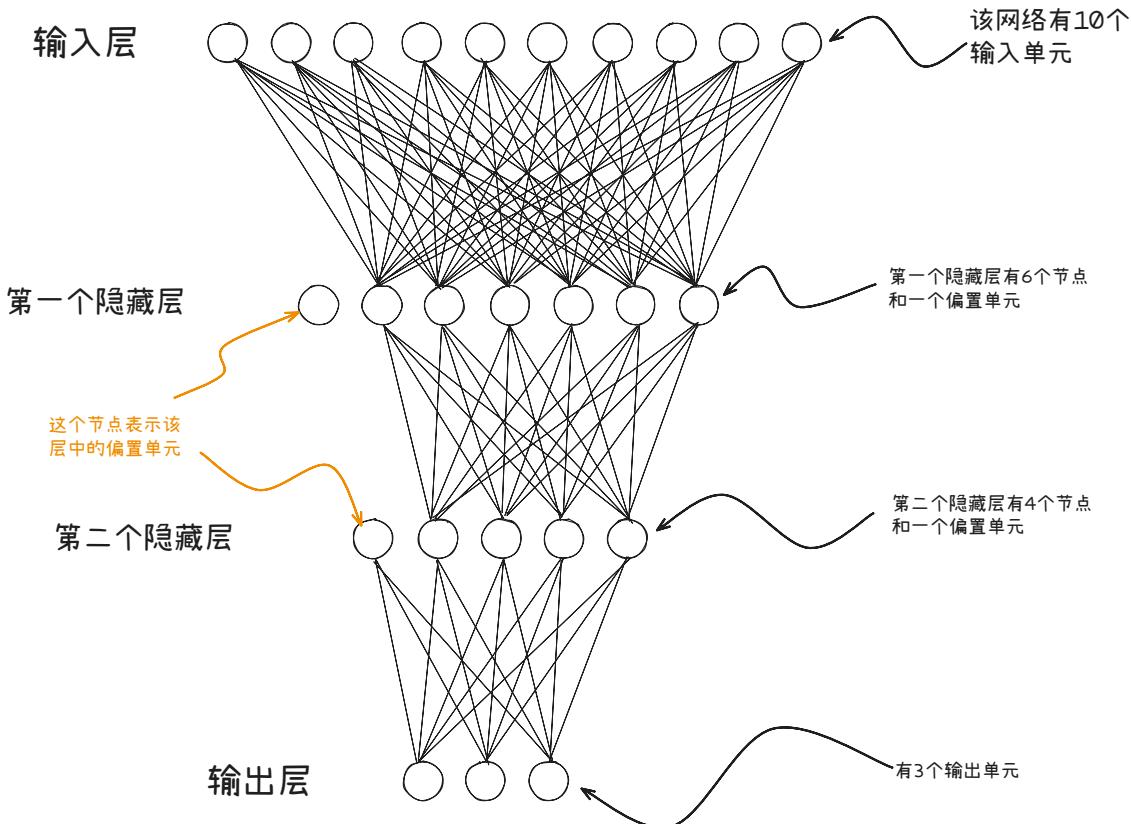


图 8.7 一个具有两个隐藏层的多层感知机。每个节点表示各自层中的一个单元。为了方便展示，这里每层都只有几个节点

在 PyTorch 中实现神经网络时，可以通过子类化 `torch.nn.Module` 类来定义我们自己的自定义网络架构。这个 `Module` 基类提供了很多功能，使得构建和训练模型变得更加容易。例如，它允许我们封装层和操作，并跟踪模型的参数。

在这个子类中，我们在 `__init__` 构造函数中定义网络层，并在 `forward` 方法中指定层与层之间的交互。`forward` 方法描述了输入数据如何通过网络传递，并形成计算图。相比之下，`backward` 方法通常不需要我们自己实现，它在训练期间用于计算给定模型参数的损失函数的梯度。下面的代码通过实现一个具有两个隐藏层的经典的多层感知机展示了 `Module` 类的典型用法。

一个具有两个隐藏层的多层感知机

Python

```

1 class NeuralNetwork(torch.nn.Module):
2     # 将输入和输出的数量编码为变量，使我们可以在具有不同特征数量和类别数量的数据集上重复使用相同的
3     # 代码
4     def __init__(self, num_inputs, num_outputs):
5         super().__init__()

```

```

5
6     self.layers = torch.nn.Sequential(
7         # 第一个隐藏层
8         torch.nn.Linear(num_inputs, 30), # 线性层将输入节点和输出节点的数量作为参数
9         torch.nn.ReLU(), # 非线性激活函数被放置在隐藏层之间
10        # 第二个隐藏层
11        torch.nn.Linear(30, 20), # 一个隐藏层的输出节点数量必须与下一层的输入节点数量相
12        # 匹配
13        torch.nn.ReLU(),
14        # 输出层
15        torch.nn.Linear(20, num_outputs),
16    )
17
18    def forward(self, x):
19        logits = self.layers(x)
20
21    return logits # 最后一层的输出称为logits

```

然后，可以像下面这样实例化一个新的神经网络对象：

```
1 model = NeuralNetwork(50, 3)
```

Python

在使用这个新模型对象之前，可以调用 `print` 函数来查看模型结构的摘要：

```
1 print(model)
```

Python

这将打印以下内容：

```

1 NeuralNetwork(
2     (layers): Sequential(
3         (0): Linear(in_features=50, out_features=30, bias=True)
4         (1): ReLU()
5         (2): Linear(in_features=30, out_features=20, bias=True)
6         (3): ReLU()
7         (4): Linear(in_features=20, out_features=3, bias=True)
8     )
9 )

```

Python

请注意，在实现 `NeuralNetwork` 类时，我们使用了 `Sequential` 类。虽然 `Sequential` 并非必需，但如果有一系列想要按特定顺序执行的层（正如本例中的情况），那么使用它可以让我们的工作更轻松。因此，在 `__init__` 构造函数中实例化 `self.layers = Sequential(...)` 后，只需在 `NeuralNetwork` 的 `forward` 方法中调用 `self.layers`，而无须单独调用每个层。

接下来，检查一下该模型的可训练参数总数：

```

1 num_params = sum(p.numel() for p in model.parameters() if p.requires_grad)
2 print("Total number of trainable model parameters:", num_params)

```

Python

这将打印以下内容：

```
1 Total number of trainable model parameters: 2213
```

Python

每一个 `requires_grad=True` 的参数都会被视为可训练参数，并在训练期间进行更新。

对于前面我们提到的具有两个隐藏层的神经网络模型，这些可训练参数包含在 `torch.nn.Linear` 层中。`Linear` 层会将输入与权重矩阵相乘，并加上一个偏置向量。这有时被称为前馈层或全连接层。

基于这里执行的 `print(model)` 调用，可以看到第一个 `Linear` 层在 `layers` 属性中的索引位置是 0。可以通过以下方式访问对应的权重参数矩阵：

```
1 print(model.layers[0].weight)
```

这将打印以下内容：

```
1 Parameter containing:  
2 tensor([[ 0.1174, -0.1350, -0.1227, ..., 0.0275, -0.0520, -0.0192],  
3         [-0.0169, 0.1265, 0.0255, ..., -0.1247, 0.1191, -0.0698],  
4         [-0.0973, -0.0974, -0.0739, ..., -0.0068, -0.0892, 0.1070],  
5         ...,  
6         [-0.0681, 0.1058, -0.0315, ..., -0.1081, -0.0290, -0.1374],  
7         [-0.0159, 0.0587, -0.0916, ..., -0.1153, 0.0700, 0.0770],  
8         [-0.1019, 0.1345, -0.0176, ..., 0.0114, -0.0559, -0.0088]],  
9         requires_grad=True)
```

由于这个大矩阵未完全显示出来，因此我们使用 `.shape` 属性来查看其维度：

```
1 print(model.layers[0].weight.shape)
```

结果如下所示：

```
1 torch.Size([30, 50])
```

(同样，可以通过 `model.layers[0].bias` 访问偏置向量。)

这里的权重矩阵是一个 30×50 的矩阵，可以看到 `requires_grad` 被设置为 `True`（意味着该矩阵是可训练的）——这是 `torch.nn.Linear` 中权重和偏置的默认设置。

如果你在自己的计算机上执行前面的代码，那么权重矩阵中的数值可能会与本书展示的有所不同。模型权重会用小的随机数进行初始化，每次实例化网络时这些数值都会不同。在深度学习中，使用小的随机数初始化模型权重是为了在训练过程中打破对称性。否则，各节点将执行相同的操作并在反向传播过程中进行相同的更新，导致网络无法学习从输入到输出的复杂映射关系。

然而，虽然我们希望继续使用小的随机数作为层权重的初始值，但可以通过 `manual_seed` 来为 PyTorch 的随机数生成器设定种子，从而使随机数初始化可重复：

```
1 torch.manual_seed(123)  
2 model = NeuralNetwork(50, 3)  
3 print(model.layers[0].weight)
```

结果如下所示：

```
1 Parameter containing:  
2 tensor([[-0.0577, 0.0047, -0.0702, ..., 0.0222, 0.1260, 0.0865],  
3         [ 0.0502, 0.0307, 0.0333, ..., 0.0951, 0.1134, -0.0297],  
4         [ 0.1077, -0.1108, 0.0122, ..., 0.0108, -0.1049, -0.1063],  
5         ...,  
6         [-0.0787, 0.1259, 0.0803, ..., 0.1218, 0.1303, -0.1351],  
7         [ 0.1359, 0.0175, -0.0673, ..., 0.0674, 0.0676, 0.1058],  
8         [ 0.0790, 0.1343, -0.0293, ..., 0.0344, -0.0971, -0.0509]],  
9         requires_grad=True)
```

现在我们已经花了一些时间检查 `NeuralNetwork` 实例，接下来简单看看如何通过前向传播使用它：

```
1 torch.manual_seed(123)  
2 X = torch.rand((1, 50))  
3 out = model(X)
```

```
4 print(out)
```

结果如下所示：

```
1 tensor([[-0.1262, 0.1080, -0.1792]], grad_fn=<AddmmBackward0>)
```

 Python

在上述代码中，我们生成了一个单一的随机训练样本 X 作为示例输入（注意，我们的网络期望接收 50 维的特征向量），并将其输入模型，从而得到了 3 个分数。当我们调用 `model(x)` 时，它会自动执行模型的前向传播。

前向传播是指从输入张量开始到计算获得输出张量的过程。这一过程包括将输入数据从输入层开始，经由隐藏层，最后传递至输出层，贯穿整个神经网络的所有层次。

结果中返回的 3 个数值对应于分配给每个输出节点的分数。注意输出张量还包含了一个 `grad_fn` 值。

这里，`grad_fn=<AddmmBackward0>` 表示计算图中用于计算某个变量的最后一个函数。具体来说，`grad_fn=<AddmmBackward0>` 意味着我们正在查看的张量是通过矩阵乘法和加法操作创建的。`PyTorch` 会在反向传播期间使用这些信息来计算梯度。`grad_fn=<AddmmBackward0>` 中的 `<AddmmBackward0>` 指定了执行的操作。在这种情况下，它执行的是一个 `Addmm` 操作。`Addmm` 代表的是矩阵乘法（`mm`）后接加法（`Add`）的组合运算。

如果只想使用网络进行预测而不进行训练或反向传播（比如在训练之后使用它进行预测），那么为反向传播构建这个计算图可能会浪费资源，因为它会执行不必要的计算并消耗额外的内存。因此，当使用模型进行推理（比如做出预测）而不是训练时，最好的做法是使用 `torch.no_grad()` 上下文管理器。这会告诉 `PyTorch` 无须跟踪梯度，从而可以显著节省内存和计算资源：

```
1 with torch.no_grad():
2     out = model(X)
3     print(out)
```

 Python

结果如下所示：

```
1 tensor([[-0.1262, 0.1080, -0.1792]])
```

 Python

在 `PyTorch` 中，通常的做法是让模型返回最后一层的输出（`logits`），而不将这些输出传递给非线性激活函数。这是因为 `PyTorch` 常用的损失函数会将 `softmax`（或二分类时的 `sigmoid`）操作与负对数似然损失结合在一个类中。这样做是为了提高数值计算的效率和稳定性。因此，如果想为预测结果计算类别成员概率，那么就需要显式调用 `softmax` 函数：

```
1 with torch.no_grad():
2     out = torch.softmax(model(X), dim=1)
3     print(out)
```

 Python

这将打印以下内容：

```
1 tensor([[0.3113, 0.3934, 0.2952]]))
```

 Python

现在这些值可以解释为类别成员的概率，并且它们的总和大约为 1。对于这个随机输入，这些值大致相等，这是未经过训练的随机初始化模型的预期结果。

8.6 设置高效的数据加载器

在我们能够训练模型之前，必须简要讨论如何在 `PyTorch` 中创建高效的数据加载器，这些加载器将在训练过程中被迭代使用。`PyTorch` 中数据加载的整体思路如图所示。

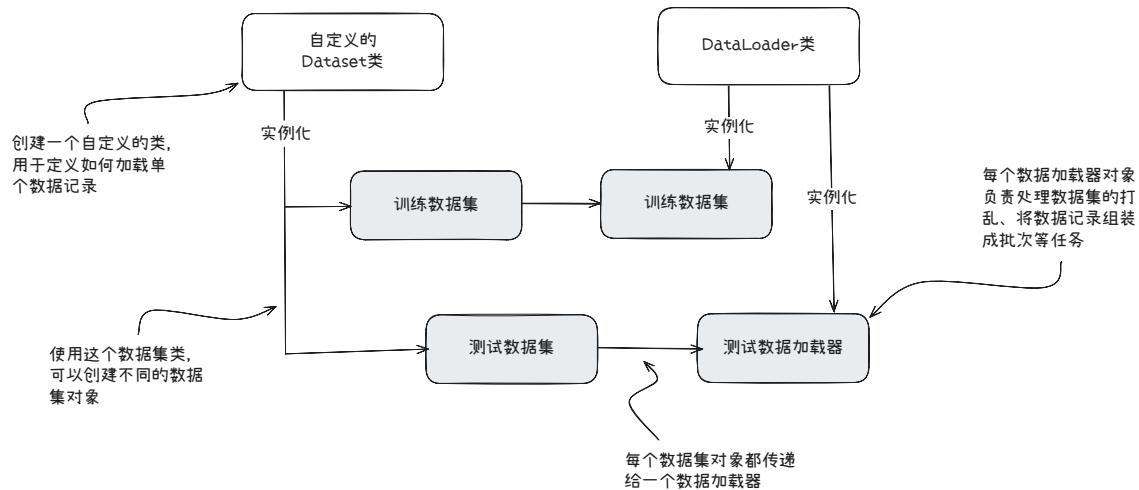


图 8.8 PyTorch 实现了 `Dataset` 类和 `DataLoader` 类。`Dataset` 类用于实例化定义如何加载每条数据记录的对象。`DataLoader` 类负责处理数据的打乱和组装成批次

根据上图，我们将实现一个自定义的 `Dataset` 类，用于创建训练数据集和测试数据集，然后再用这些数据集创建数据加载器。我们首先创建一个简单的示例数据集，其中包含 5 个训练示例，每个示例有两个特征。与训练示例一起，我们还创建了一个包含相应类别标签的张量：3 个示例属于类别标签 0，两个示例属于类别标签 1。此外，我们还构建了一个包含两个样本的测试集。创建此数据集的代码如代码所示。

创建一个小型示例数据集 Python

```

1 X_train = torch.tensor([
2     [-1.2, 3.1],
3     [-0.9, 2.9],
4     [-0.5, 2.6],
5     [2.3, -1.1],
6     [2.7, -1.5]
7 ])
8 y_train = torch.tensor([0, 0, 0, 1, 1])
9
10 X_test = torch.tensor([
11     [-0.8, 2.8],
12     [2.6, -1.6],
13 ])
14 y_test = torch.tensor([0, 1])

```

注意

PyTorch 要求类别标签从标签 0 开始，并且最大的类别标签值不得超过输出节点数减 1（因为 Python 的索引从 0 开始）。因此，如果我们有类别标签 0、1、2、3 和 4，那么神经网络的输出层应包含 5 个节点。

接下来，我们通过继承 PyTorch 的 `Dataset` 父类来创建一个自定义数据集类 `ToyDataset`，如代码所示。

定义一个自定义的Dataset类 Python

```

1 from torch.utils.data import Dataset

```

```

2
3  class ToyDataset(Dataset):
4      def __init__(self, X, y):
5          self.features = X
6          self.labels = y
7
8      def __getitem__(self, index):
9          """检索一条数据记录及其对应标签的说明"""
10         one_x = self.features[index]
11         one_y = self.labels[index]
12         return one_x, one_y
13
14     def __len__(self):
15         """返回数据集总长度的说明"""
16         return self.labels.shape[0]
17
18 train_ds = ToyDataset(X_train, y_train)
19 test_ds = ToyDataset(X_test, y_test)

```

这个自定义的 `ToyDataset` 类的目的是实例化一个 PyTorch `DataLoader`。在进行这一步之前，让我们先简要回顾一下 `ToyDataset` 代码的一般结构。

在 PyTorch 中，自定义的 `Dataset` 类的 3 个主要组成部分是 `__init__` 方法、`__getitem__` 方法和 `__len__` 方法。在 `__init__` 方法中，我们设置一些可以在 `__getitem__` 方法和 `__len__` 方法中访问的属性。这些属性可以是文件路径、文件对象、数据库连接器等。由于我们创建了一个位于内存中的张量数据集，因此只需将 `X` 和 `y` 分配给这些代表张量对象的占位符属性即可。

在 `__getitem__` 方法中，我们定义了通过索引返回数据集中单个项目的具体指令。这指的是与单个训练示例或测试实例对应的特征和类别标签。（数据加载器将提供这个索引，稍后我们会介绍。）

最后，`__len__` 方法包含了检索数据集长度的指令。在这里，我们使用张量的 `.shape` 属性来返回特征数组中的行数。就训练数据集而言，我们有 5 行数据，下面可以再次确认一下：

```
1 print(len(train_ds))
```

结果如下所示。

```
1 5
```

现在我们已经定义了一个可用于示例数据集的 PyTorch `Dataset` 类，我们可以使用 PyTorch 的 `DataLoader` 类从中进行采样，如代码所示。

实例化数据加载器

```

1 from torch.utils.data import DataLoader
2
3 torch.manual_seed(123)
4
5 # 之前创建的示例数据集实例作为数据加载器的输入
6 train_loader = DataLoader(
7     dataset=train_ds,
8     batch_size=2,
9     shuffle=True, # 是否打乱数据
10    num_workers=0 # 后台进程的数量
11 )
12
13 test_loader = DataLoader(

```

```
14     dataset=test_ds,
15     batch_size=2,
16     shuffle=False, # 测试数据集无须打乱顺序
17     num_workers=0
18 )
```

在实例化训练数据加载器后，可以对其进行迭代。对 `test_loader` 的迭代与之类似，但为简洁起见，这里省略了具体细节：

```
1 for idx, (x, y) in enumerate(train_loader):
2     print(f"Batch {idx+1}:", x, y)
```

Python

结果如下所示：

```
1 Batch 1: tensor([[-1.2000, 3.1000],
2                   [-0.5000, 2.6000]]) tensor([0, 0])
3 Batch 2: tensor([[ 2.3000, -1.1000],
4                   [-0.9000, 2.9000]]) tensor([1, 0])
5 Batch 3: tensor([[ 2.7000, -1.5000]]) tensor([1])
```

Python

根据前面的输出，可以看到 `train_loader` 迭代了训练数据集，每个训练示例正好访问一次。这被称为一个训练轮次。由于我们使用 `torch.manual_seed(123)` 设置了随机数生成器，因此你应该得到完全相同的训练示例打乱顺序。然而，当你再次迭代数据集时，你会发现打乱的顺序已经发生变化。这是为了防止深度神经网络在训练过程中陷入重复更新循环。

我们在这里指定的批次大小为 2，但第三批次仅包含一个示例。这是因为我们有 5 个训练示例，而 5 不能被 2 整除。

在实践中，如果一个训练轮次的最后一个批次显著小于其他批次，那么可能会影响训练过程中的收敛。为此，可以设置 `drop_last=True`，这将在每轮中丢弃最后一个批次，如代码所示。

```
1 train_loader = DataLoader(
2     dataset=train_ds,
3     batch_size=2,
4     shuffle=True,
5     num_workers=0,
6     drop_last=True
7 )
```

一个丢弃最后一个批次的训练加载器

Python

现在，迭代训练加载器，可以看到最后一个批次被省略了：

```
1 for idx, (x, y) in enumerate(train_loader):
2     print(f"Batch {idx+1}:", x, y)
```

Python

结果如下所示。

```
1 Batch 1: tensor([[-0.9000, 2.9000],
2                   [ 2.3000, -1.1000]]) tensor([0, 1])
3 Batch 2: tensor([[ 2.7000, -1.5000],
4                   [-0.5000, 2.6000]]) tensor([1, 0])
```

Python

最后，我们来讨论 `DataLoader` 中的 `num_workers=0` 设置。这个参数在 PyTorch 的 `DataLoader` 函数中对于并行加载和预处理数据至关重要。当 `num_workers` 设置为 0 时，数据加载将在主进程中而不是单独的工作进程中进行。这看起来似乎没有问题，但在使用 GPU 训练较大的网络时，这可能会导致模型训练显著减慢。这是因为 CPU 不仅要处理深度学习模型，还要花时间加载和预处理数据。

因此，GPU 在等待 CPU 完成这些任务时可能会闲置。相反，当 `num_workers` 设置为大于 0 的数值时，会启动多个工作进程并行加载数据，从而释放主进程专注于训练模型，并更好地利用系统资源。

然而，如果你处理的是非常小的数据集，那么可能并不需要将 `num_workers` 设置为 1 或更大的数值，因为总训练时间只需几秒。因此，如果你使用的是小型数据集或交互式环境（如 Jupyter Notebook），那么增加 `num_workers` 可能不会显著提高速度，反而会导致一些问题。一个潜在的问题是，启动多个工作进程的开销可能会比实际数据加载所需的时间更长，尤其是数据集很小的时候。

此外，对于 Jupyter Notebook，将 `num_workers` 设置为大于 0 有时可能会导致不同进程之间资源共享的问题，从而引发错误或导致笔记本崩溃。因此，理解这种权衡并对 `num_workers` 参数进行合理设置是非常重要的。如果使用得当，这可以成为一个有益的工具，但应根据你的特定数据集大小和计算环境进行调整，以获得最佳效果。

根据我的经验，设置 `num_workers=4` 通常会在许多真实世界数据集上获得最佳性能，但最佳设置取决于你的硬件和用于加载 `Dataset` 类中训练示例的代码。

8.7 典型的训练循环

现在让我们在示例数据集上训练一个神经网络。代码展示了训练过程。

```

    在PyTorch中进行神经网络训练
Python

1 import torch.nn.functional as F
2
3 torch.manual_seed(123)
4 model = NeuralNetwork(num_inputs=2, num_outputs=2)
5 optimizer = torch.optim.SGD(
6     model.parameters(), lr=0.5
7 )
8
9 num_epochs = 3
10 for epoch in range(num_epochs):
11     model.train()
12     for batch_idx, (features, labels) in enumerate(train_loader):
13         logits = model(features)
14         loss = F.cross_entropy(logits, labels)
15         optimizer.zero_grad()
16         loss.backward()
17         optimizer.step()
18
19         ### LOGGING
20         print(f"Epoch: {epoch+1:03d}/{num_epochs:03d}"
21             f" | Batch {batch_idx:03d}/{len(train_loader):03d}"
22             f" | Train Loss: {loss:.2f}")
23
24 model.eval()
25 # 插入可选的模型评估代码

```

运行上述代码会产生以下输出：

```

1 Epoch: 001/003 | Batch 000/002 | Train Loss: 0.75
2 Epoch: 001/003 | Batch 001/002 | Train Loss: 0.65
3 Epoch: 002/003 | Batch 000/002 | Train Loss: 0.44
4 Epoch: 002/003 | Batch 001/002 | Train Loss: 0.13
5 Epoch: 003/003 | Batch 000/002 | Train Loss: 0.03
6 Epoch: 003/003 | Batch 001/002 | Train Loss: 0.00

```

如你所见，损失在 3 轮后降至 0，这表明模型已经在训练集上收敛。这里初始化了一个具有两个输入和两个输出的模型，因为我们的示例数据集有两个输入特征和两个类别标签需要预测。我们使用了一个学习率 (`lr`) 为 0.5 的随机梯度下降 (SGD) 优化器。学习率是一个超参数，意味着这是可调的设置，我们必须根据观察到的损失进行实验。理想情况下，我们希望选择一个学习率，使得损失在一定轮数后收敛——轮数是另一个需要选择的超参数。

🔥 练习

代码中介绍的神经网络有多少个参数？

在实际操作中，我们通常会使用第三个数据集，即所谓“验证数据集”，来找到最优的超参数设置。验证集类似于测试集。然而，虽然我们只想精确地使用一次测试集以避免评估偏差，但通常会多次使用验证集来调整模型设置。

我们还引入了新的设置：`model.train()` 和 `model.eval()`。顾名思义，这些设置用于将模型置于训练模式或评估模式。这对于在训练和推断过程中具有不同行为的组件（如 `dropout` 或批归一化层）是必要的。由于我们的 `NeuralNetwork` 类中没有受到这些设置影响的 `dropout` 或其他组件，因此在之前的代码中并未使用 `model.train()` 和 `model.eval()`。然而，最好还是包含这些设置，以避免在更改模型架构或重用代码训练其他模型时出现意外行为。

正如之前讨论的那样，我们直接将 `logits` 传递给 `cross_entropy` 损失函数，后者会在内部应用 `softmax` 函数，以提高效率并增强数值稳定性。接下来，调用 `loss.backward()` 会计算由 PyTorch 在后台构建的计算图中的梯度。`optimizer.step()` 方法会利用这些梯度来更新模型参数以最小化损失。对 SGD 优化器而言，这意味着将梯度与学习率相乘，然后将缩放后的负梯度加到参数上。

⚡ 危险

为了避免不必要的梯度累积，确保在每次更新中调用 `optimizer.zero_grad()` 来将梯度重置为 0，这很重要。否则，梯度会逐渐累积起来，这往往是我们不愿意见到的。

在训练好模型后，可以使用它进行预测：

```
1 model.eval()  
2 with torch.no_grad():  
3     outputs = model(X_train)  
4 print(outputs)
```

Python

结果如下所示：

```
1 tensor([[ 2.8569, -4.1618],  
2         [ 2.5382, -3.7548],  
3         [ 2.0944, -3.1820],  
4         [-1.4814,  1.4816],  
5         [-1.7176,  1.7342]])
```

Python

为了获得类别成员概率，可以使用 PyTorch 的 `softmax` 函数：

```
1 torch.set_printoptions(sci_mode=False)  
2 probas = torch.softmax(outputs, dim=1)  
3 print(probas)
```

Python

输出如下所示：

```
1 tensor([[ 0.9991,    0.0009],  
2         [ 0.9982,    0.0018],  
3         [ 0.9949,    0.0051],
```

Python

```
4      [    0.0491,    0.9509],
5      [    0.0307,    0.9693]])
```

来看一下上面代码输出的第 1 行。在这里，第一个值（列）表示该训练示例属于类别标签 0 的概率为 99.91%，属于类别标签 1 的概率为 0.09%。（这里使用 `set_printoptions` 是为了让输出更加易读。）

可以使用 PyTorch 的 `argmax` 函数将这些概率值转换为类别标签预测。如果设置 `dim=1`，它将返回每行中最大值的索引位置（设置 `dim=0` 则返回每列中最大值的索引位置）：

```
1 predictions = torch.argmax(probas, dim=1)
2 print(predictions)
```

这将打印如下内容：

```
1 tensor([0, 0, 0, 1, 1])
```

请注意，为了获得类别标签，计算 `softmax` 概率并非必需步骤，也可以直接对 `logits`（输出）应用 `argmax` 函数：

```
1 predictions = torch.argmax(outputs, dim=1)
2 print(predictions)
```

输出如下所示：

```
1 tensor([0, 0, 0, 1, 1])
```

在这里，我们计算了训练数据集的预测标签。鉴于训练数据集相对较小，我们可以通过肉眼将其与真实的训练标签进行比较，结果显示模型预测的准确率为 100%。可以使用比较运算符 `=` 来再次确认这一点：

```
1 predictions = y_train
```

结果如下所示：

```
1 tensor([True, True, True, True, True])
```

使用 `torch.sum` 可以计算正确预测的数量：

```
1 torch.sum(predictions == y_train)
```

输出如下所示：

```
1 5
```

由于数据集由 5 个训练示例组成，我们的 5 个预测全部正确，准确率为 $\frac{5}{5} \times 100\% = 100\%$ 。

为了使预测准确率的计算更加通用，让我们实现一个 `compute_accuracy` 函数，如代码所示。

一个计算预测准确率的函数

```
1 def compute_accuracy(model, dataloader):
2     model = model.eval()
3     correct = 0.0
4     total_examples = 0
5
6     for idx, (features, labels) in enumerate(dataloader):
7         with torch.no_grad():
8             logits = model(features)
9
10            predictions = torch.argmax(logits, dim=1)
11            compare = labels == predictions # 根据标签是否匹配，返回一个True/False值的张量
```

```
12     correct += torch.sum(compare) # 求和操作计算True值的数量
13     total_examples += len(compare)
14     # 正确预测的比例是一个介于0和1之间的值。调用`.item()`会将张量的值以Python浮点数的形式返回
15     return (correct / total_examples).item()
```

这段代码通过迭代数据加载器来计算正确预测的数量和比例。在处理大规模数据集时，由于内存限制，通常我们只能对数据集的一小部分调用模型。这里的 `compute_accuracy` 函数是一种通用方法，适用于任意大小的数据集，因为在每次迭代中，模型所接收的数据集块的大小与训练期间的批次大小相同。`compute_accuracy` 函数的内部逻辑类似于我们之前将 `logits` 转换为类别标签时使用的方法。

接下来，可以将该函数应用于训练数据：

```
1 print(compute_accuracy(model, train_loader))
```

结果如下所示：

```
1 1.0
```

类似地，可以在测试集上应用这个函数：

```
1 print(compute_accuracy(model, test_loader))
```

这将打印如下内容。

```
1 1.0
```

8.8 保存和加载模型

现在我们的模型已经训练好了，接下来看看如何保存它，以便以后可以重用。下面是在 PyTorch 中保存和加载模型的推荐方法：

```
1 torch.save(model.state_dict(), "model.pth")
```

模型的 `state_dict` 是一个 Python 字典对象，它可以将模型中的每一层映射到其可训练参数（权重和偏置）。`model.pth` 是保存到磁盘的模型文件的任意文件名，我们可以使用任何名称和文件后缀，不过 `.pth` 和 `.pt` 是最常见的约定。

保存模型后，可以从磁盘中恢复它：

```
1 model = NeuralNetwork(2, 2)
2 model.load_state_dict(torch.load("model.pth"))
```

`torch.load("model.pth")` 函数读取文件 `model.pth`，并重建包含模型参数的 Python 字典对象，`model.load_state_dict()` 则将这些参数应用到模型中，有效地恢复了我们保存模型时模型的学习状态。

在同一会话中执行此代码时，`model = NeuralNetwork(2, 2)` 这一行并不是严格必需的。然而，这里包含它是为了说明我们需要在内存中拥有一个模型的实例，这样才能应用保存的参数。此外，`NeuralNetwork(2, 2)` 的架构必须与最初保存的模型完全匹配。

8.9 使用 GPU 优化训练性能

接下来，让我们探讨如何利用 GPU 来加速深度神经网络的训练。（相较于普通 CPU，GPU 能够显著提升训练速度。）首先，我们将了解 PyTorch 中 GPU 计算的主要概念。然后，我们将在单个 GPU 上训练模型。最后，我们将讨论如何使用多个 GPU 进行分布式训练。

8.9.1 在 GPU 设备上运行 PyTorch

修改训练循环使其便于在 GPU 上运行相对简单，只需更改 3 行代码即可。在进行这些修改之前，理解 PyTorch 中 GPU 计算的主要概念非常重要。在 PyTorch 中，设备是执行计算和存储数据的地方。CPU 和 GPU 是设备的示例。如果一个 PyTorch 张量存放在某个设备上，那么其操作也会在同一个设备上执行。

来看一下这一过程是如何进行的。假设你已经安装了兼容 GPU 的 PyTorch 版本，可以使用以下代码再次检查一下你的运行环境是否真的支持 GPU 计算：

```
1 print(torch.cuda.is_available())
```

结果如下所示：

```
1 True
```

现在，假设我们有两个张量可以相加。默认情况下，这个计算将在 CPU 上执行：

```
1 tensor_1 = torch.tensor([1., 2., 3.])
2 tensor_2 = torch.tensor([4., 5., 6.])
3 print(tensor_1 + tensor_2)
```

输出如下所示：

```
1 tensor([5., 7., 9.])
```

现在可以使用`.to()`方法。这个方法与我们用来更改张量数据类型的方法相同，它能够将这些张量转移到 GPU 上并在那里执行加法操作：

```
1 tensor_1 = tensor_1.to("cuda")
2 tensor_2 = tensor_2.to("cuda")
3 print(tensor_1 + tensor_2)
```

输出如下所示：

```
1 tensor([5., 7., 9.], device='cuda:0')
```

生成的张量现在包含了设备信息`device='cuda:0'`，这意味着这些张量位于第一个 GPU 上。如果你的机器有多个 GPU，那么可以指定要将张量转移到哪个 GPU 上。这可以通过在传输命令中指定设备 ID 来实现，比如使用`.to("cuda:0")`、`.to("cuda:1")`等命令。

然而，所有的张量必须位于同一个设备上。否则，如果一个张量位于 CPU，另一个张量位于 GPU，计算就会失败：

```
1 tensor_1 = tensor_1.to("cpu")
2 print(tensor_1 + tensor_2)
```

结果如下所示：

```
1 RuntimeError      Traceback (most recent call last)
2 <ipython-input-7-4ff3c4d20fc3> in <cell line: 2>()
3     1 tensor_1 = tensor_1.to("cpu")
4 ----> 2 print(tensor_1 + tensor_2)
5 RuntimeError: Expected all tensors to be on the same device, but found at
6 least two devices, cuda:0 and cpu!
```

总之，只需要将张量传输到同一个 GPU 设备上，PyTorch 会处理其余的工作。

8.9.2 单个 GPU 训练

我们已经熟悉了将张量传输到 GPU 的过程，现在可以修改训练循环以在 GPU 上运行。这一步仅需要更改 3 行代码，如代码所示。

```

GPU 上的训练循环
Python

1 torch.manual_seed(123)
2 model = NeuralNetwork(num_inputs=2, num_outputs=2)
3
4 device = torch.device("cuda") # 定义一个默认使用GPU的设备变量
5 model = model.to(device) # 将模型转移到GPU上
6
7 optimizer = torch.optim.SGD(model.parameters(), lr=0.5)
8
9 num_epochs = 3
10
11 for epoch in range(num_epochs):
12     model.train()
13     for batch_idx, (features, labels) in enumerate(train_loader):
14         # 将数据转移到GPU上
15         features, labels = features.to(device), labels.to(device)
16         logits = model(features)
17         loss = F.cross_entropy(logits, labels) # Loss function
18
19         optimizer.zero_grad()
20         loss.backward()
21         optimizer.step()
22
23     ### LOGGING
24     print(f"Epoch: {epoch+1:03d}/{num_epochs:03d}"
25           f" | Batch {batch_idx:03d}/{len(train_loader):03d}"
26           f" | Train/Val Loss: {loss:.2f}")
27
28 model.eval()
29 # 插入可选的模型评估代码

```

运行上述代码将输出以下内容，类似于在 CPU 上获得的结果：

```

1 Epoch: 001/003 | Batch 000/002 | Train/Val Loss: 0.75
2 Epoch: 001/003 | Batch 001/002 | Train/Val Loss: 0.65
3 Epoch: 002/003 | Batch 000/002 | Train/Val Loss: 0.44
4 Epoch: 002/003 | Batch 001/002 | Train/Val Loss: 0.13
5 Epoch: 003/003 | Batch 000/002 | Train/Val Loss: 0.03
6 Epoch: 003/003 | Batch 001/002 | Train/Val Loss: 0.00

```

可以使用`.to("cuda")`来代替`device = torch.device("cuda")`。将张量传输到 "cuda" 而不是`torch.device("cuda")`也可以工作，并且更简洁。还可以修改该语句，这样即使没有 GPU，代码也能在 CPU 上执行。这被认为是分享 PyTorch 代码时的最佳实践：

```

1 device = torch.device("cuda" if torch.cuda.is_available() else "cpu")

```

在当前修改后的训练循环中，由于从 CPU 转移到 GPU 的内存传输成本，我们可能不会看到速度的提升。然而，我们可以期待在训练深度神经网络，尤其是大语言模型时，会有显著的速度提升。

🔥 练习

比较矩阵乘法在 CPU 和 GPU 上的运行时间。在多大尺寸的矩阵上，你开始看到 GPU 上的矩阵乘法比 CPU 上的矩阵乘法更快？（提示：在 Jupyter Notebook 中使用 `%timeit` 命令来比较运行时间。例如，对于矩阵 `a` 和 `b`，在新的笔记本单元中运行命令 `%timeit a @ b。`）

8.9.3 使用多个 GPU 训练

分布式训练的概念是将模型训练分配到多个 GPU 和机器上。为什么要这样做？虽然在单个 GPU 或机器上训练模型是可行的，但这个过程可能会非常耗时。通过将训练过程分布到多台机器上（每台机器可能有多个 GPU），可以显著减少训练时间。这在模型开发的实验阶段尤为重要，因为可能需要进行大量训练迭代来微调模型参数和架构。

让我们从分布式训练最基础的案例开始：`PyTorch` 的分布式数据并行（`DistributedDataParallel`, `DDP`）策略。`DDP` 通过将输入数据分割到可用设备上并同时处理这些数据子集来实现并行化。

这是如何工作的呢？`PyTorch` 会在每个 GPU 上启动一个独立的进程，每个进程都会接收并保存一份模型副本，这些副本在训练过程中会进行同步。假设有两个 GPU，我们想要用它们来训练一个神经网络，如图所示。

每个 GPU 都会接收到一份模型副本。然后，在每次训练迭代中，每个模型都会从数据加载器中接收一个小批次（或简称“批次”）数据。可以使用 `DistributedSampler` 来确保在使用 `DDP` 时，每个 GPU 接收到的批次不同且不重叠。

由于每个模型副本会看到不同的训练数据样本，因此模型副本在反向传播时将返回不同的 `logits` 并计算出不同的梯度。然后，这些梯度在训练过程中会被平均和同步，以便更新模型。通过这种方式，可以确保模型不会出现分歧，如图所示。

使用 `DDP` 的好处在于，与单个 GPU 相比，它能够更快地处理数据集。除去设备之间由于使用 `DDP` 而产生的少量通信开销，理论上使用两个 GPU 可以将训练一轮的时间缩短一半。时间效率会随着 GPU 数量的增加而提高，如果有 8 个 GPU，那么可以将一轮的处理速度提高 8 倍，以此类推。

现在让我们看看这在实践中是如何工作的。为简洁起见，我们将专注于需要为 `DDP` 训练调整的核心代码部分。

首先，导入一些用于分布式训练的 `PyTorch` 附加子模块、类和函数，如代码所示。

用于分布式训练的 PyTorch 工具	Python
1 <code>import torch.multiprocessing as mp</code>	
2 <code>from torch.utils.data.distributed import DistributedSampler</code>	
3 <code>from torch.nn.parallel import DistributedDataParallel as DDP</code>	
4 <code>from torch.distributed import init_process_group, destroy_process_group</code>	

在深入讨论使训练与 `DDP` 兼容的更改之前，先简要回顾一下与 `DistributedDataParallel` 类一起使用的这些新导入的工具的原理和用途。

`PyTorch` 的 `multiprocessing` 子模块包含诸如 `multiprocessing.spawn` 之类的函数，我们将使用这些函数来生成多个进程，然后再并行地将一个函数应用于多个输入。我们将为每个 GPU 生成一个训练进程。如果想要为训练生成多个进程，则需要用一种方法将数据集划分给这些不同的进程。为此，可以使用 `DistributedSampler`。

`init_process_group` 和 `destroy_process_group` 用于初始化和退出分布式训练模式。`init_process_group` 函数应在训练脚本开始时调用，以初始化分布式设置中每个进程的进程组，而 `destroy_process_group` 应在训练脚本结束时调用，以销毁给定的进程组并释放其资源。代码展示了这些新组件如何用于实现我们之前实现的 `NeuralNetwork` 模型的 `DDP` 训练。

使用 <code>DistributedDataParallel</code> 策略进行模型训练	Python
1 <code>def ddp_setup(rank, world_size):</code>	

```
2     os.environ["MASTER_ADDR"] = "localhost"
3     os.environ["MASTER_PORT"] = "12345"
4     init_process_group(
5         backend="nccl",
6         rank=rank,
7         world_size=world_size
8     )
9     torch.cuda.set_device(rank)
10
11 def prepare_dataset():
12     # 插入数据集准备代码
13     train_loader = DataLoader(
14         dataset=train_ds,
15         batch_size=2,
16         shuffle=False,
17         pin_memory=True,
18         drop_last=True,
19         sampler=DistributedSampler(train_ds)
20     )
21     return train_loader, test_loader
22
23 def main(rank, world_size, num_epochs):
24     ddp_setup(rank, world_size)
25     train_loader, test_loader = prepare_dataset()
26     model = NeuralNetwork(num_inputs=2, num_outputs=2)
27     model.to(rank)
28     optimizer = torch.optim.SGD(model.parameters(), lr=0.5)
29     model = DDP(model, device_ids=[rank])
30     for epoch in range(num_epochs):
```

8.10 小结

- PyTorch 是一个开源库，包含 3 个核心组件：张量库、自动微分函数和深度学习工具。
- PyTorch 的张量库类似于 NumPy 等数组库。
- 在 PyTorch 中，张量是表示标量、向量、矩阵和更高维数组的类数组数据结构。
- PyTorch 张量可以在 CPU 上执行，但 PyTorch 张量格式的一个主要优势是它支持 GPU 加速计算。
- PyTorch 中的自动微分 (autograd) 功能使我们能够方便地使用反向传播训练神经网络，而无须手动推导梯度。
- PyTorch 的深度学习工具提供了创建自定义深度神经网络的构建块。
- PyTorch 提供了 Dataset 类和 DataLoader 类来建立高效的数据加载流水线。
- 在 CPU 或单个 GPU 上训练模型是最简单的。
- 如果有多个 GPU 可用，那么使用 DistributedDataParallel 是 PyTorch 中加速训练的最简单方式。



9. Transformer: 从零实现大语言模型



10. 监督学习

监督学习模型定义了从一个或多个输入到一个或多个输出的映射。例如，输入可以是二手丰田普锐斯的车龄和里程数，输出可以是汽车的估价（美元）。

这个模型只是一个数学函数；当输入通过这个函数时，它会计算输出，这称为推理。模型函数也包含参数。不同的参数值会改变计算结果；模型函数描述了输入和输出之间的一组可能的关系，而模型参数则指定了特定的关系。

当训练模型时，目标是寻找能描述输入和输出之间真实关系的参数。学习算法会取一个输入/输出对作为训练集，并调整参数，直到输入尽可能准确地预测其对应的输出。如果模型在这些训练对的数据上表现良好，那么我们希望它能对新输入数据（即真实输出未知的情况）做出良好的预测。

本章的目标是扩展这些概念。首先，将更正式地描述这个框架并引入一些符号。然后，通过一个简单示例演示如何用一条直线来描述输入与输出之间的关系。这个线性模型既为人熟知又易于可视化，但仍能很好地阐明监督学习的所有主要思想。

10.1 监督学习概述

在监督学习中，我们的目标是构建一个模型，它可以接收一个输入 x 并输出一个预测 y 。为简单起见，假设输入 x 和输出 y 都是内容已预先确定且大小固定的向量，并且每个向量的元素始终按照相同的方式排列；在前述的普锐斯示例中，输入 x 总会按顺序包含汽车的车龄，然后是里程数。这类数据被称为结构化数据或表格数据。

为了做出预测，我们需要一个模型 $f[\bullet]$ ，它接收输入 x 并返回输出 y ，所以：

$$y = f(x) \tag{10.1}$$

我们把用输入 x 计算预测结果 y 的过程称为推理。

模型只是一个固定形式的数学函数，代表了输入和输出之间一系列不同的关系。模型也包含参数。参数的选择决定了输入和输出之间的具体关系，所以应该更准确地写成：

$$y = f[x, \Phi] \tag{10.2}$$

当谈论学习或训练模型时，我们的目的是试图找到能够通过输入做出合理输出预测的参数。我们使用一个包含对输入和输出示例的训练数据集 $\{(x_i, y_i)\}$ 来学习获得这些参数。目标是选择参数，尽可能准确地将每个训练输入数据映射到其相应的输出数据。用损失函数 \mathcal{L} 来量化这种映射的不匹配程度。损失是一个标量值，概括了在给定参数的情况下，模型从对应的输入中预测训练输出的误差。

可将损失 \mathcal{L} 视为参数的函数 $\mathcal{L}[\Phi]$ 。当训练模型时，我们就是在寻找能使这个损失函数值最小的参数 $\hat{\Phi}$ ：

$$\hat{\Phi} = \underset{\Phi}{\operatorname{argmin}} [\mathcal{L}[\Phi]] \quad (10.3)$$

提示

更准确地说，损失函数还依赖于训练数据 $\{x_i, y_i\}$ ，所以应该写作 $\mathcal{L}[\{x_i, y_i\}, \Phi]$ ，但这样表示会显得相当繁杂。

如果在完成这一最小化过程后损失函数的值很小，我们就已经找到了能准确预测输入 x_i 到输出 y_i 的模型参数。

训练完一个模型后，必须评估其性能；在独立的测试数据上运行模型，观测它在训练期间未观察到的样本上的泛化能力。如果性能令人满意，就可以准备部署模型了。

10.2 线性回归示例

现在，让我们通过一个简单例子来具体化这些概念。设想一个模型 $y = f[x, \Phi]$ ，它通过一个输入 x 预测一个输出 y 。接下来，我们会设计一个损失数，最后讨论模型的训练过程。

10.2.1 一维（1D）线性回归模型

一维线性回归模型用一条直线描述输入 x 和输出 y 之间的关系：

$$\begin{aligned} y &= f[x, \Phi] \\ &= \phi_0 + \phi_1 x \end{aligned} \quad (10.4)$$

该模型有两个参数 $\Phi = [\phi_0, \phi_1]^T$ ，其中 ϕ_0 是直线的 y 轴截距， ϕ_1 是斜率。不同的 y 轴截距和斜率选择会决定输入和输出之间的不同关系。因此，上面的式子定义了一组可能的输入/输出关系（所有可能的直线），参数的选择决定了这一组中的具体成员（特定的直线）。

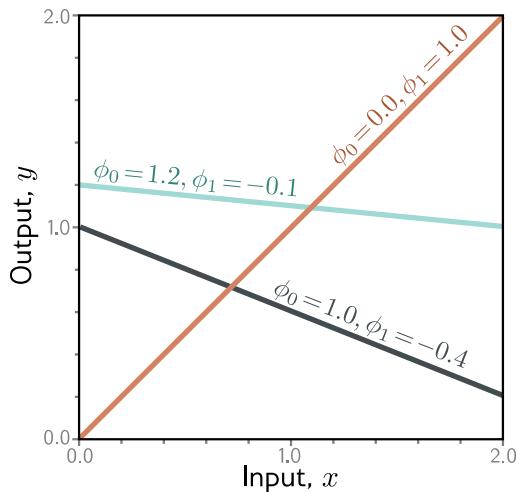


图 10.1 线性回归模型。对于给定的参数 $\Phi = [\phi_0, \phi_1]^T$ ，模型根据输入（ x 轴）对输出（ y 轴）进行预测。不同的截距 ϕ_0 和斜率 ϕ_1 的选择会改变这些预测结果（青色、橙色和灰色直线）。线性回归模型（式（10.4））定义了一组输入/输出关系（直线），而参数决定了该组中的具体成员（特定的直线）

10.2.2 损失

对于这个模型，训练数据集（图 10.2(a)）由 I 个输入/输出对 $\{x_i, y_i\}$ 构成。图 10.2(b) ~ (d) 显示了由三组参数定义的三条直线。图 10.2(d) 中的绿色直线比其他两条线更准确地描述了数据，

因为它更接近数据点。然而，我们需要一个有依据的方法来决定哪个参数比其他参数更好。为此，给每个参数分配一个数值，用该数值来量化模型与数据之间的不匹配程度。将这个值称为损失；更低的损失意味着更好的拟合效果。

这种不匹配由模型的预测 $f[x, \Phi]$ （在 x 处线上方的高度）和实际输出 y 之间的偏差所捕获。这些偏差在图 10.2(b) ~ (d) 中以橙色虚线表示。将所有 I 个训练对的偏差的平方和称为总的不匹配，训练误差或损失：

$$\begin{aligned}\mathcal{L}[\Phi] &= \sum_{i=1}^I (f[x_i, \Phi] - y_i)^2 \\ &= \sum_{i=1}^I (\phi_0 + \phi_1 x_i - y_i)^2\end{aligned}\quad (10.5)$$

由于最佳参数会最小化这个式子的值，所以我们称之为最小二乘损失。平方计算意味着偏差的方向（即直线是在数据上方还是下方）无关重要。后面我们会讲解这样做的理论依据。

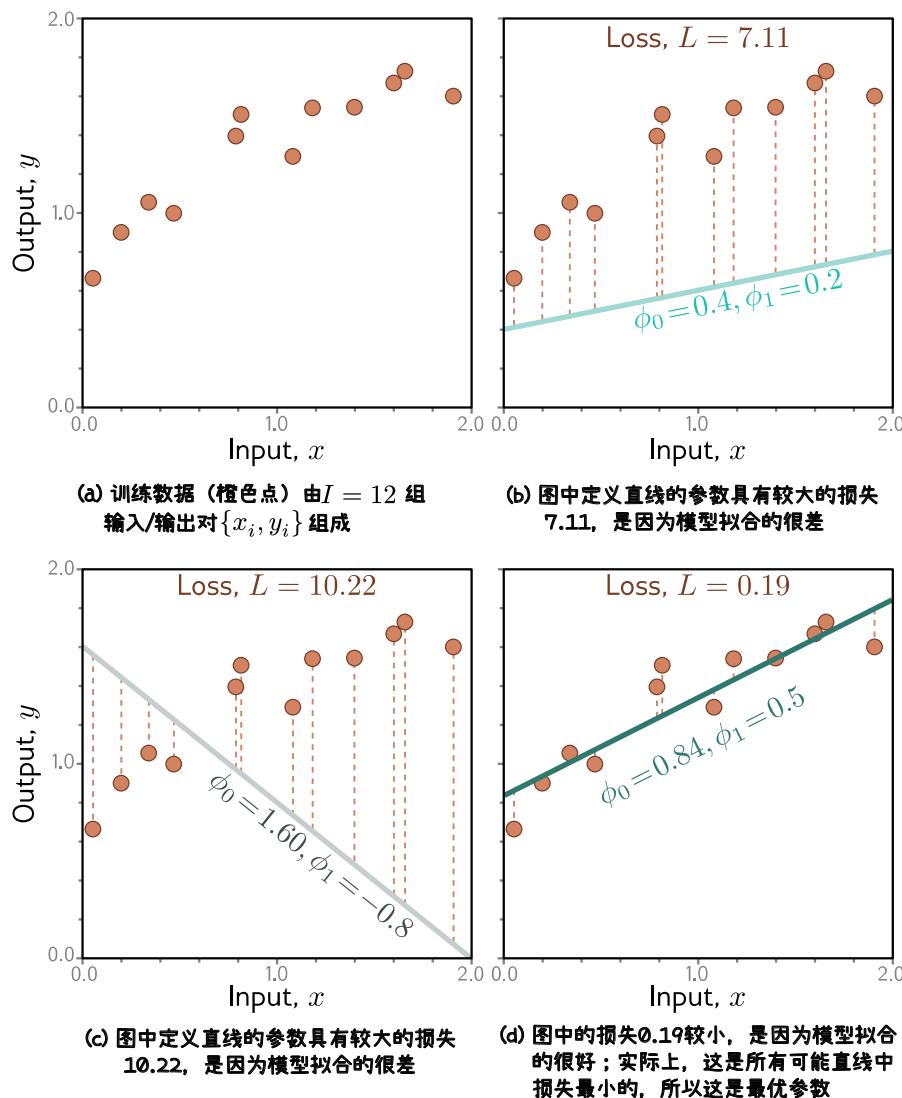


图 10.2 线性回归的训练数据、模型和损失。图(b) ~ (d) 分别展示了具有不同参数的线性回归模型。根据截距和斜率参数 $\Phi = [\phi_0, \phi_1]^T$ 的选择，模型误差（橙色虚线）可能更大或者更小。损失 \mathcal{L} 是这些误差平方的总和

损失 \mathcal{L} 是参数 Φ 的函数；当模型拟合较差时（图 10.2(b) 和图 10.2(c)）损失更大，当拟合良好时（图 10.2(d)）损失更小。从这个角度看，将 $\mathcal{L}[\Phi]$ 称为损失函数或者代价函数。训练模型的目标是找到最小化这个值的参数 $\hat{\Phi}$ ：

$$\begin{aligned}\hat{\Phi} &= \operatorname{argmin}_{\Phi} [\mathcal{L}[\Phi]] \\ &= \operatorname{argmin}_{\Phi} \left[\sum_{i=1}^I (f[x_i, \Phi] - y_i)^2 \right] \\ &= \operatorname{argmin}_{\Phi} \left[\sum_{i=1}^I (\phi_0 + \phi_1 x_i - y_i)^2 \right]\end{aligned}\quad (10.6)$$

只有两个参数（ y 轴截距 ϕ_0 和斜率 ϕ_1 ），因此我们可以计算每种参数值组合的损失，并将损失函数可视化为一个曲面。“最佳”参数位于该曲面的最低点。

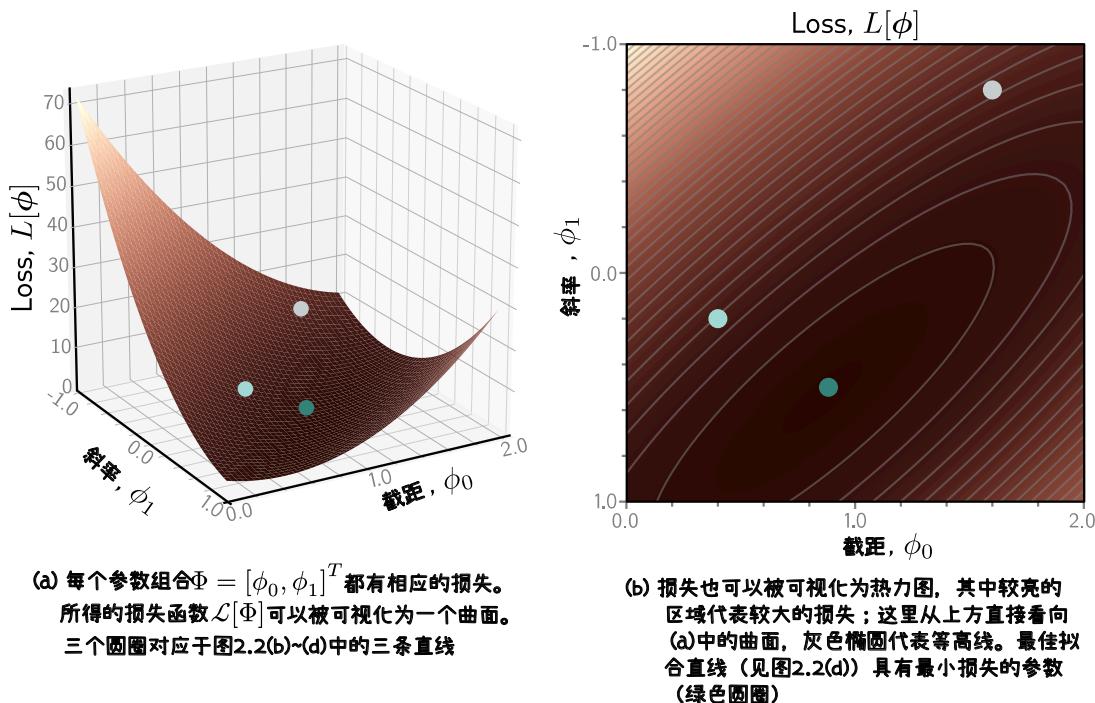


图 10.3 针对图 10.2(a)的数据集的线性回归模型的损失函数

10.2.3 训练

寻找使损失最小化的参数的过程称为模型拟合、训练或者学习。基本方法是随机选择初始参数，然后“沿着”损失函数“下降”的方向改进它们，直至到达最低点。一种方法是测量当前点所在曲面的梯度，并向最陡峭的下坡方向迈出一步。此后重复这个过程，直到梯度变得平坦，无法进一步改进为止。

⚡ 危险

对于线性回归模型来说，这种迭代方法实际上并不是必要的。在这里，我们可以找到参数的解析解表达式。然而，梯度下降法适用于更复杂的模型，这些模型中没有解析解，并且参数特别多，无法评估每种参数组合的损失。

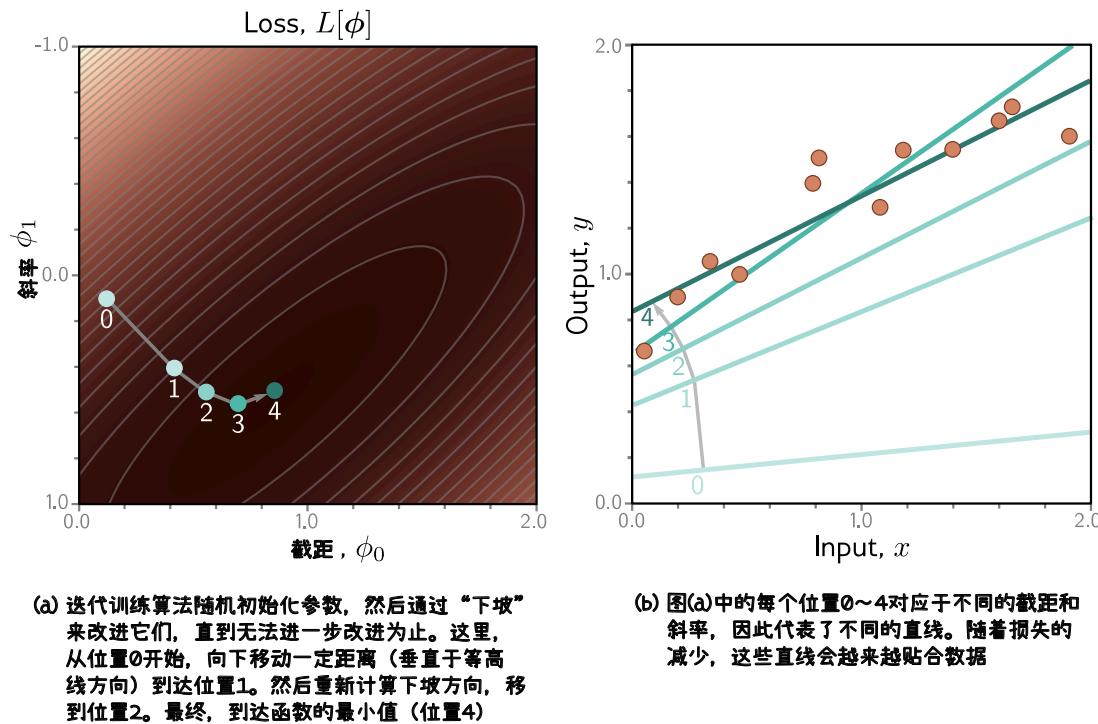


图 10.4 线性回归训练。训练目标是找到对应于最小损失的截距和斜率参数

10.2.4 测试

完成模型训练之后，我们需要知道它在现实世界中的表现如何。通过在单独的测试数据集上面计算损失来评估这一点。预测准确性会在何种程度上泛化到测试数据部分取决于训练数据的代表性和完整性，也取决于模型的表达能力。一个简单模型（如一条直线）可能无法捕捉输入和输出之间的真正关系，这称为欠拟合。相反，一个表达力很强的模型可能描述训练数据的一些非典型的统计特性，同时会引起异常的预测。这被称为过拟合。

10.2.5 最小二乘法

我们还记得一元线性回归的损失函数是

$$\mathcal{L}[\Phi] = \sum_{i=1}^I (\phi_0 + \phi_1 x_i - y_i)^2 \quad (10.7)$$

而我们的目标是想让损失函数 \mathcal{L} 最小化。而通过观察损失函数的图像，我们可以得知这个损失函数是存在最小值的。那么怎么找出这个最小值呢？由于我们有两个参数：截距 ϕ_0 和斜率 ϕ_1 。根据微积分的知识，损失函数对这两个参数求导，导数为 0 的点就是最小值。

所以损失函数对 ϕ_0 的导数如下推导：

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \phi_0} &= \frac{\partial \sum_{i=1}^I (\phi_0 + \phi_1 x_i - y_i)^2}{\partial \phi_0} \quad (\text{求导的线性法则}) \\
&= \sum_{i=1}^I \frac{\partial (\phi_0 + \phi_1 x_i - y_i)^2}{\partial \phi_0} \quad (\text{链式法则}) \\
&= \sum_{i=1}^I \frac{\partial (\phi_0 + \phi_1 x_i - y_i)^2}{\partial (\phi_0 + \phi_1 x_i - y_i)} \cdot \frac{\partial (\phi_0 + \phi_1 x_i - y_i)}{\partial \phi_0} \\
&= \sum_{i=1}^I \{2(\phi_0 + \phi_1 x_i - y_i) \cdot 1\} \\
&= 2 \sum_{i=1}^I \{\phi_0 + \phi_1 x_i - y_i\} \\
&= 2 \cdot I \cdot \phi_0 + 2(x_1 + x_2 + \dots + x_I) \phi_1 - 2(y_1 + y_2 + \dots + y_I) \\
&= 0
\end{aligned} \tag{10.8}$$

化简可以得到如下

$$I \cdot \phi_0 + \left[\sum_{i=1}^I x_i \right] \phi_1 = \sum_{i=1}^I y_i \tag{10.9}$$

所以损失函数对 ϕ_1 的导数如下推导：

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \phi_1} &= \frac{\partial \sum_{i=1}^I (\phi_0 + \phi_1 x_i - y_i)^2}{\partial \phi_1} \quad (\text{求导的线性法则}) \\
&= \sum_{i=1}^I \frac{\partial (\phi_0 + \phi_1 x_i - y_i)^2}{\partial \phi_1} \quad (\text{链式法则}) \\
&= \sum_{i=1}^I \frac{\partial (\phi_0 + \phi_1 x_i - y_i)^2}{\partial (\phi_0 + \phi_1 x_i - y_i)} \cdot \frac{\partial (\phi_0 + \phi_1 x_i - y_i)}{\partial \phi_1} \\
&= \sum_{i=1}^I \{2(\phi_0 + \phi_1 x_i - y_i) \cdot x_i\} \\
&= 2 \sum_{i=1}^I \{(\phi_0 + \phi_1 x_i - y_i) \cdot x_i\} \\
&= 2(x_1 + x_2 + \dots + x_I) \phi_0 + 2(x_1^2 + x_2^2 + \dots + x_I^2) \phi_1 - 2(x_1 y_1 + x_2 y_2 + \dots + x_I y_I) \\
&= 0
\end{aligned} \tag{10.10}$$

化简可以得到如下

$$\left[\sum_{i=1}^I x_i \right] \phi_0 + \left[\sum_{i=1}^I x_i^2 \right] \phi_1 = \sum_{i=1}^I x_i y_i \tag{10.11}$$

也就是说我们要求解的

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \phi_0} = 0 \\ \frac{\partial \mathcal{L}}{\partial \phi_1} = 0 \end{cases} \tag{10.12}$$

最终化简得到了一个二元一次方程组

$$\begin{cases} I \cdot \phi_0 + \left[\sum_{i=1}^I x_i \right] \phi_1 = \sum_{i=1}^I y_i \\ \left[\sum_{i=1}^I x_i \right] \phi_0 + \left[\sum_{i=1}^I x_i^2 \right] \phi_1 = \sum_{i=1}^I x_i y_i \end{cases} \tag{10.13}$$

这样就可以求解得到 ϕ_0 和 ϕ_1 的数值，也就是最优参数就可以被求解出来了。

如果写成矩阵的表示法，可以看到如下

$$\begin{bmatrix} I & \sum_{i=1}^I x_i \\ \sum_{i=1}^I x_i & \sum_{i=1}^I x_i^2 \end{bmatrix} \cdot \begin{bmatrix} \phi_0 \\ \phi_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^I y_i \\ \sum_{i=1}^I x_i y_i \end{bmatrix} \quad (10.14)$$

根据矩阵的性质

$$A \cdot B = C \Rightarrow A^{-1} \cdot A \cdot B = A^{-1}C \Rightarrow B = A^{-1}C \quad (10.15)$$

可以得到如下

$$\begin{bmatrix} \phi_0 \\ \phi_1 \end{bmatrix} = \begin{bmatrix} I & \sum_{i=1}^I x_i \\ \sum_{i=1}^I x_i & \sum_{i=1}^I x_i^2 \end{bmatrix}^{-1} \cdot \begin{bmatrix} \sum_{i=1}^I y_i \\ \sum_{i=1}^I x_i y_i \end{bmatrix} \quad (10.16)$$

2×2 逆矩阵的求解对于计算机来说是很容易的，但如果参数很多呢？那么求解逆矩阵可能就是一件根本无法做到的事情了。

例如，如果我们的训练数据是 $\{(x_{i,1}, x_{i,2}, \dots, x_{i,N}), y_i\}$ 数据对呢？如果我们还是用线性回归的话，我们需要用多元线性回归来解决这个问题。此时，我们的损失函数变成了

$$\begin{aligned} \mathcal{L}[\Phi] &= \sum_{i=1}^I (\phi_0 + \phi_1 x_{i,1} + \phi_2 x_{i,2} + \dots + \phi_N x_{i,N} - y_i)^2 \\ &= \sum_{i=1}^I \left(\phi_0 + \sum_{j=1}^N \phi_j x_{i,j} - y_i \right)^2 \end{aligned} \quad (10.17)$$

那么损失函数需要对 $\{\phi_0, \phi_1, \phi_2, \dots, \phi_N\}$ 共 $N+1$ 个参数求导数为0的点。

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \phi_0} &= 2 \sum_{i=1}^I \left[\left(\phi_0 + \sum_{j=1}^N \phi_j x_{i,j} - y_i \right) \cdot 1 \right] \\ \frac{\partial \mathcal{L}}{\partial \phi_1} &= 2 \sum_{i=1}^I \left[\left(\phi_0 + \sum_{j=1}^N \phi_j x_{i,j} - y_i \right) \cdot x_{i,1} \right] \\ \frac{\partial \mathcal{L}}{\partial \phi_2} &= 2 \sum_{i=1}^I \left[\left(\phi_0 + \sum_{j=1}^N \phi_j x_{i,j} - y_i \right) \cdot x_{i,2} \right] \\ &\vdots \\ \frac{\partial \mathcal{L}}{\partial \phi_N} &= 2 \sum_{i=1}^I \left[\left(\phi_0 + \sum_{j=1}^N \phi_j x_{i,j} - y_i \right) \cdot x_{i,N} \right] \end{aligned} \quad (10.18)$$

如果 N 很大，例如 6710 亿参数（DeepSeek-V3 的参数数量），求解上面这个方程组是做不到的。

所以我们才会使用梯度下降法来去寻找损失函数的极小值点。

我们可以将上面的式子整理成矩阵表示法。

设计矩阵

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,N} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{I,1} & x_{I,2} & \cdots & x_{I,N} \end{bmatrix} \quad (10.19)$$

参数向量

$$\Phi = \begin{bmatrix} \phi_0 \\ \phi_1 \\ \phi_2 \\ \vdots \\ \phi_N \end{bmatrix} \in \mathbb{R}^{N+1} \quad (10.20)$$

目标向量

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^I \quad (10.21)$$

预测值

$$\hat{\mathbf{y}} = \mathbf{X}\Phi = \begin{bmatrix} \phi_0 + \sum_{j=1}^N \phi_j x_{1,j} \\ \phi_0 + \sum_{j=1}^N \phi_j x_{2,j} \\ \vdots \\ \phi_0 + \sum_{j=1}^N \phi_j x_{I,j} \end{bmatrix} \quad (10.22)$$

残差（误差）向量

$$\mathbf{e} = \hat{\mathbf{y}} - \mathbf{y} = \mathbf{X}\Phi - \mathbf{y} = \begin{bmatrix} \phi_0 + \sum_{j=1}^N \phi_j x_{1,j} - y_1 \\ \phi_0 + \sum_{j=1}^N \phi_j x_{2,j} - y_2 \\ \vdots \\ \phi_0 + \sum_{j=1}^N \phi_j x_{I,j} - y_I \end{bmatrix} \quad (10.23)$$

梯度的矩阵形式

观察偏导数的结构

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \phi_2} &= 2 \sum_{i=1}^I \left[\left(\phi_0 + \sum_{j=1}^N \phi_j x_{i,j} - y_i \right) \cdot x_{i,2} \right] \\ &= 2 \sum_{i=1}^I (\mathbf{e}_i \cdot x_{i,2}) \\ &= 2 [\mathbf{e}_1 \ \mathbf{e}_2 \ \cdots \ \mathbf{e}_I] \cdot \begin{bmatrix} x_{1,2} \\ x_{2,2} \\ x_{3,2} \\ \vdots \\ x_{I,2} \end{bmatrix} \\ &= 2 [x_{1,2} \ x_{2,2} \ x_{3,2} \ \cdots \ x_{I,2}] \cdot \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_I \end{bmatrix} \end{aligned} \quad (10.24)$$

这实际上是残差向量与设计矩阵第 2 列的内积！

所以

$$\nabla_{\Phi} \mathcal{L} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial \phi_0} \\ \frac{\partial \mathcal{L}}{\partial \phi_1} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial \phi_N} \end{bmatrix} = 2 \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_{1,1} & x_{2,1} & x_{3,1} & \cdots & x_{I,1} \\ x_{1,2} & x_{2,2} & x_{3,2} & \cdots & x_{I,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{1,N} & x_{2,N} & x_{3,N} & \cdots & x_{I,N} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_I \end{bmatrix} = 2\mathbf{X}^T \mathbf{e} \quad (10.25)$$

所以损失函数的梯度是

$$\nabla_{\Phi} \mathcal{L} = 2\mathbf{X}^T \mathbf{e} = 2\mathbf{X}^T(\mathbf{X}\Phi - \mathbf{y}) \quad (10.26)$$

如果我们要求解梯度为 $\mathbf{0}$ 的解，也就是

$$\nabla_{\Phi} \mathcal{L} = 2\mathbf{X}^T \mathbf{e} = 2\mathbf{X}^T(\mathbf{X}\Phi - \mathbf{y}) = 0 \quad (10.27)$$

先将上面的式子把 2 消去。然后分配率展开得到

$$\begin{aligned} & \mathbf{X}^T(\mathbf{X}\Phi - \mathbf{y}) = 0 \\ & \Downarrow \\ & \mathbf{X}^T\mathbf{X}\Phi - \mathbf{X}^T\mathbf{y} = 0 \\ & \Downarrow \\ & \mathbf{X}^T\mathbf{X}\Phi = \mathbf{X}^T\mathbf{y} \quad (\text{正规方程}) \\ & \Downarrow \\ & \Phi = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \quad (\text{计算机算不动}) \end{aligned} \quad (10.28)$$

主要是上面的矩阵求逆运算计算机无法完成。

IT

大语言模型

11	理解大语言模型	147
11.1	什么是大语言模型	147
11.2	大语言模型的应用	148
11.3	构建和使用大语言模型的各个阶段	149
11.4	Transformer 架构介绍	149
11.5	利用大型数据集	150
11.6	深入剖析 GPT 架构	151
11.7	构建大语言模型	152
11.8	小结	152
12	大语言模型的架构	153
13	处理文本数据	155
13.1	对文本分词	155
13.2	创建词嵌入查找表	171



11. 理解大语言模型

近年来，OpenAI 推出的 ChatGPT 等大语言模型作为深度神经网络模型的代表，为自然语言处理 (*natural language processing, NLP*) 领域带来了革命性的变化。在大语言模型出现之前，传统方法（如手工规则或简单模型）在垃圾邮件检测、简单模式识别等分类任务中表现优异。然而，这些传统方法在需要具备复杂的理解和生成能力的语言任务（比如解析详细的指令、进行语境分析或创作连贯且符合语境的原创文本）中通常表现不佳。举例来说，早期的语言模型无法根据关键词列表来编写电子邮件，而现今的大语言模型能轻松完成这一任务。

大语言模型在理解、生成和解释人类语言方面拥有出色的能力。但需要澄清的是，当我们谈论语言模型的“理解”能力时，实际上是指它们能够处理和生成看似连贯且符合语境的文本，而这并不意味着它们真的拥有像人类一样的意识或理解能力。

深度学习 (*deep learning*) 是机器学习 (*machine learning*) 和人工智能 (*artificial intelligence, AI*) 领域的一个重要分支，主要聚焦于神经网络的研究。深度学习的发展使得大语言模型能够利用海量的文本数据进行训练，从而相比于以往的方法能够捕获更深层次的上下文信息和人类语言的细微之处。因此，大语言模型在文本翻译、情感分析、问答等各类自然语言处理任务中都有显著的性能提升。

现代大语言模型与早期自然语言处理模型之间的另一个重要区别在于，早期自然语言处理模型通常是为特定任务（如文本分类，语言翻译等）而设计的。尽管这些早期自然语言处理模型在其特定应用中表现卓越，但大语言模型在各种自然语言处理任务中展现了更广泛的能力。

大语言模型的成功，一方面得益于为其提供支撑的 **Transformer** 架构，另一方面得益于用于训练这些模型的海量数据。这使得它们能够捕捉到语言中的各类细微差别、上下文信息和模式规律，而这些都是手动编码难以实现的。

这一转变，即以 **Transformer** 架构为核心，使用大型数据集来训练大语言模型，已经从根本上变革了自然语言处理领域，为机器理解并与人类语言互动提供了更强大的工具。

本章接下来所讨论的内容为实现本书的主要目标奠定了基础：通过代码逐步实现一个基于 **Transformer** 架构的类 ChatGPT 大语言模型，以此深入理解大语言模型。

11.1 什么是大语言模型

大语言模型是一种用于理解、生成和响应类似人类语言文本的神经网络。这类模型属于深度神经网络 (*deep neural network*)，通过大规模文本数据训练而成，其训练资料甚至可能涵盖了互联网上大部分公开的文本。

“大语言模型”这一名称中的“大”字，既体现了模型训练时所依赖的庞大数据集，也反映了模型本身庞大的参数规模。这类模型通常拥有数百亿甚至数千亿个参数 (*parameter*)。这些参数是神经网络中的可调整权重，在训练过程中不断被优化，以预测文本序列中的下一个词。下一单词预

测 (**next-word prediction**) 任务合理地利用了语言本身具有顺序这一特性来训练模型，使得模型能够理解文本中的上下文、结构和各种关系。然而，由于这项任务本身非常简单，因此许多研究人员对其能够孕育出如此强大的模型深感惊讶。在后续章节中，我们将逐步讨论并实现下一单词预测的训练过程。

大语言模型采用了一种名为 **Transformer** 的架构，这种架构允许模型在进行预测时有选择地关注输入文本的不同部分，从而使得它们特别擅长应对人类语言的细微差别和复杂性。

由于大语言模型能够生成文本，因此它们通常也被归类为生成式人工智能 (**generative artificial intelligence**，简称 **generative AI** 或 **GenAI**)。如图所示，人工智能是一个囊括机器学习、深度学习等众多分支的领域，旨在开发能够执行需要人类智能水平的任务（包括语言理解、模式识别、决策制定等）的机器。

实现人工智能的算法是机器学习领域的重点研究内容。具体而言，机器学习涉及开发能够从数据中学习的算法。无须明确编程，这些算法就能基于数据做出预测或决策。举个例子，垃圾邮件过滤器是机器学习技术的一个典型应用。与手动编写规则来识别垃圾邮件不同，机器学习算法会接收标记为垃圾邮件和正常邮件的示例。通过在训练数据集上最小化预测误差，模型能够学习到如何识别垃圾邮件的模式和特征，进而将新的邮件分类为垃圾邮件或非垃圾邮件。

如图所示，深度学习是机器学习的一个分支，它主要利用 3 层及以上的神经网络（深度神经网络）来建模数据中的复杂模式和抽象特征。与深度学习不同，传统的机器学习往往需要人工进行特征提取。这意味着人类专家需要为模型识别和挑选出最相关的特征。

尽管人工智能领域现在由机器学习和深度学习所主导，但该领域也涉及其他方法，比如基于规则的系统、遗传算法、专家系统、模糊逻辑或符号推理。

仍以垃圾邮件分类为例，在传统的机器学习方法中，人类专家需要手动从电子邮件文本中提取诸如特定触发词（“**prize**”、“**win**”、“**free**”）的出现频率、感叹号的数量、全大写单词的使用情况或可疑链接的存在等特征。这些基于专家定义的特征所构造的数据集将被用来训练模型。相比之下，深度学习并不依赖人工提取的特征，这意味着不再需要由人类专家为模型识别和选择最相关的特征。然而，无论是传统的机器学习还是用于垃圾邮件分类任务的深度学习，仍然需要收集标签（比如垃圾邮件或非垃圾邮件，这些标签通常由专家或用户提供）。

接下来我们将介绍大语言模型目前能够解决的一些问题、它们所面临的挑战，以及本书中将要实现的大语言模型的通用架构。

11.2 大语言模型的应用

大语言模型在解析和理解非结构化文本数据方面的能力非常强，因此它们在许多领域得到了广泛应用。如今，大语言模型已被应用于机器翻译、文本生成、情感分析、文本摘要等多种任务。最近，它们还被用于进行内容创作，包括撰写小说和文章，甚至编写计算机代码。

此外，大语言模型还可以为复杂的聊天机器人和虚拟助手提供支持，包括 OpenAI 的 ChatGPT、谷歌的 Gemini（前称为 Bard）等。这些系统可以回答用户的问题，并增强谷歌搜索、微软必应等传统搜索引擎的能力。

在医学、法律等专业领域中，大语言模型还被用于从大量文本中有效地提取知识，包括筛选文献、总结长篇段落和回答技术性问题。

简而言之，大语言模型在几乎所有需要解析和生成文本的任务的自动化处理中都具有重要价值。它们的应用领域极为广阔，并且显而易见的是，随着我们不断创新和探索这些模型的使用方法，它们有潜力重塑我们与科技的关系，使其变得更具互动性、更为直观且更易使用。

在本书中，我们将致力于从零开始理解大语言模型的工作原理，并实现一个可以生成文本的大语言模型。此外，你还将学习使大语言模型能够执行各类任务（包括回答问题、文本总结、多语言翻译等）的技术。换言之，在本书中，你将通过逐步构建一个像 ChatGPT 这样复杂的大语言模型助手，来学习其工作原理。

11.3 构建和使用大语言模型的各个阶段

为什么要自己构建大语言模型?从零开始构建大语言模型不仅是一次深入了解模型机制和局限性的绝佳机会,还为我们提供了预训练和微调开源大语言模型,使其适应特定领域的数据集或任务的必要知识。

研究表明,针对特定领域或任务量身打造的大语言模型在性能上往往优于 ChatGPT 等为多种应用场景而设计的通用大语言模型。这样的例子包括专用于金融领域的模型 BloombergGPT 和专用于医学问答的大语言模型。

使用定制的大语言模型具有多个优势,尤其是在数据隐私方面。例如,出于机密性考虑,公司可能不愿将敏感数据共享给像 OpenAI 这样的第三方大语言模型提供商。此外,如果开发较小的定制的大语言模型,那么就可以将其直接部署到客户设备(笔记本电脑和智能手机)上。这也是苹果公司等企业正在探索的方向。本地部署可以显著减少延迟并降低与服务器相关的成本。此外,定制的大语言模型使开发者拥有完全的自主权,能够根据需要控制模型的更新和修改。

大语言模型的构建通常包括预训练(**pre-training**)和微调(**fine-tuning**)两个阶段。“预训练”中的“预”表明它是模型训练的初始阶段,此时模型会在大规模、多样化的数据集上进行训练,以形成全面的语言理解能力。以预训练模型为基础,微调阶段会在规模较小的特定任务或领域数据集上对模型进行针对性训练,以进一步提升其特定能力。图 1-3 展示了由预训练和微调组成的两阶段训练方法。

创建大语言模型的第一步是在大量文本数据上进行训练,这些数据也被称作原始文本(**raw text**)。“原始”指的是这些数据只是普通的文本,没有附加任何标注信息。(在这一步中,我们通常会进行数据过滤,比如删除格式字符或未知语言的文档。)

预训练是大语言模型的第一个训练阶段,预训练后的大语言模型通常称为基础模型(**foundation model**)。一个典型例子是 ChatGPT 的前身——GPT-3,这个模型能够完成文本补全任务,即根据用户的前半句话将句子补全。此外,它还展现了有限的少样本学习能力,这意味着它可以在没有大量训练数据的情况下,基于少量示例来学习并执行新任务。

通过在无标注数据集上训练获得预训练的大语言模型后,我们可以在带标注的数据集上进一步训练这个模型,这一步称为微调。

微调大语言模型最流行的方法是指令微调和分类任务微调。在指令微调(**instruction fine-tuning**)中,标注数据集由“指令-答案”对(比如翻译任务中的“原文-正确翻译文本”)组成。在分类任务微调(**classification fine-tuning**)中,标注数据集由文本及其类别标签(比如已被标记为“垃圾邮件”或“非垃圾邮件”的电子邮件文本)组成。

在本书中,我们将介绍预训练和微调大语言模型的代码实现,并且在预训练基础模型之后,我们将深入探讨指令微调和分类任务微调的具体技术细节。

11.4 Transformer 架构介绍

大部分的现代大语言模型基于 Transformer 架构,这是一种深度神经网络架构,该架构是在谷歌于 2017 年发表的论文“Attention Is All You Need”中首次提出的。为了理解大语言模型,我们需要简单回顾一下最初的 Transformer。Transformer 最初是为机器翻译任务(比如将英文翻译成德语和法语)开发的。Transformer 架构的一个简化版本如图 1-4 所示。

Transformer 架构由两个子模块构成:编码器和解码器。编码器(**encoder**)模块负责处理输入文本,将其编码为一系列数值表示或向量,以捕捉输入的上下文信息。然后,解码器(**decoder**)模块接收这些编码向量,并据此生成输出文本。以翻译任务为例,编码器将源语言的文本编码成向量,解码器则解码这些向量以生成目标语言的文本。编码器和解码器都是由多层组成,这些层通过自注意力机制连接。关于如何对输入进行预处理和编码,我们将在后续章节中逐步解答。

Transformer 和大语言模型的一大关键组件是自注意力机制(**self-attention mechanism**),它允许模型衡量序列中不同单词或词元之间的相对重要性。这一机制使得模型能够捕捉到输入数据中长距离的依赖和上下文关系,从而提升其生成连贯且上下文相关的输出的能力。然而,由于自注意力机制较为复杂,我们将在第 3 章中详细解释并逐步实现。

为了适应不同类型的下游任务，Transformer 的后续变体，如 BERT (Bidirectional Encoder Representations from Transformer, 双向编码预训练 Transformer) 和各种 GPT (Generative Pretrained Transformer, 生成式预训练 Transformer) 模型，都基于这一理念构建。

BERT 基于原始 Transformer 的编码器模块构建，其训练方法与 GPT 不同。GPT 主要用于生成任务，而 BERT 及其变体专注于掩码预测 (masked word prediction)，即预测给定句子中被掩码的词，如图 1-5 所示。这种独特的训练策略使 BERT 在情感预测、文档分类等文本分类任务中具有优势。例如，截至本书撰写时，X (以前的 Twitter) 在检测有害内容时使用的是 BERT。

GPT 则侧重于原始 Transformer 架构的解码器部分，主要用于处理生成文本的任务，包括机器翻译、文本摘要、小说写作、代码编写等。

GPT 模型主要被设计和训练用于文本补全 (text completion) 任务，但它们表现出了出色的可扩展性。这些模型擅长执行零样本学习任务和少样本学习任务。零样本学习 (zero-shot learning) 是指在没有任何特定示例的情况下，泛化到从未见过的任务，而少样本学习 (few-shot learning) 是指从用户提供的少量示例中进行学习，如图 1-6 所示。



Transformer 与大语言模型

当今的大语言模型大多基于前文介绍的 Transformer 架构，因此，Transformer 和大语言模型在文献中常常被作为同义词使用。然而，并非所有的 Transformer 都是大语言模型，因为 Transformer 也可用于计算机视觉领域。同样，并非所有的大语言模型都基于 Transformer 架构，因为还存在基于循环和卷积架构的大语言模型。推动这些新架构发展的主要动机在于提高大语言模型的计算效率。然而，这些非 Transformer 架构的大语言模型是否能够与基于 Transformer 的大语言模型相媲美，以及它们是否会在实践中被广泛应用，还有待观察。为简便起见，本书中使用“大语言模型”一词来指代类似于 GPT 的基于 Transformer 的大语言模型。

11.5 利用大型数据集

主流的 GPT、BERT 等模型所使用的训练数据集涵盖了多样而全面的文本语料库。这些语料库包含数十亿词汇，涉及广泛的主题，囊括自然语言与计算机语言。表 1-1 通过一个具体的例子总结了用于预训练 GPT-3 的数据集。GPT-3 被视作第一代 ChatGPT 的基础模型。

表 1-1 展示了各种数据集的词元数量。词元 (token) 是模型读取文本的基本单位。数据集中的词元数量大致等同于文本中的单词和标点符号的数量。我们将在第 2 章中更详细地介绍分词，即将文本转换为词元的过程。

我们能得到的主要启示是，训练数据集的庞大規模和丰富多样性使得这些模型在包括语言语法、语义、上下文，甚至一些需要通用知识的任务上都拥有了良好表现。

🔥 GPT-3 数据集的细节

表 1-1 显示了用于训练 GPT-3 的数据集。表中的“训练数据中的比例”一列总计为 100%，根据舍入误差进行调整。尽管“词元数量”一列总计为 4990 亿，但该模型仅在 3000 亿个词元上进行了训练。GPT-3 论文的作者并没有具体说明为什么该模型没有对所有 4990 亿个词元进行训练。

为了更好地理解，以 CommonCrawl 数据集为例，它包含 4100 亿个词元，需要约 570GB 的存储空间。相比之下，GPT-3 等模型的后续版本（如 Meta 的 Llama），已经扩展了它们的训练范围，涵盖了包括 Arxiv 研究论文（92GB）和 StackExchange 上的代码问答（78GB）在内的更多数据源。

GPT-3 论文的作者并未公开其训练数据集，但我们可以参考一个公开可用的类似数据集——Dolma：这是一个用于大语言模型预训练的 3 万亿兆词元大小的开放语料库。然而，该数据集可能包含受版权保护的内容，具体使用条款可能取决于预期的使用情境和国家。

这些模型的预训练特性使它们在针对下游任务进行微调时表现出了极高的灵活性，因此它们也被称为“基础模型”。预训练大语言模型需要大量资源，成本极其高昂。例如，预训练 GPT-3 的云计算费用成本估计高达 460 万美元。

好消息是，许多预训练的大语言模型是开源模型，可以作为通用工具，用于写作、摘要和编辑那些未包含在训练数据中的文本。同时，这些大语言模型可以使用相对较小的数据集对特定任务进行微调，这不仅减少了模型所需的计算资源，还提升了它们在特定任务上的性能。

在本书中，我们将实现预训练代码，并将其应用于预训练一个用于教育目的的大语言模型，其中的所有计算都可以在消费级硬件上执行。在实现预训练代码之后，我们将学习如何复用公开可用的模型权重，并将它们加载到要实现的架构中。这样后续在微调大语言模型时，我们就可以跳过昂贵的预训练阶段。

11.6 深入剖析 GPT 架构

GPT 最初是由 OpenAI 的 Radford 等人在论文“*Improving Language Understanding by Generative Pre-Training*”中提出的。GPT-3 是该模型的扩展版本，它拥有更多的参数，并在更大的数据集上进行了训练。此外，ChatGPT 中提供的原始模型是通过使用 OpenAI 的 InstructGPT 文中的方法，在一个大型指令数据集上微调 GPT-3 而创建的。正如我们在图 1-6 中所见，这些模型不仅是强大的文本补全模型，还可以胜任拼写校正、分类或语言翻译等任务。考虑到 GPT 模型仅在相对简单的下一单词预测任务（参见图 1-7）上进行了预训练，它们能有如此强大而全面的能力实在令人惊叹。

下一单词预测任务采用的是自监督学习（self-supervised learning）模式，这是一种自我标记的方法。这意味着我们不需要专门为训练数据收集标签，而是可以利用数据本身的结构。也就是说，我们可以使用句子或文档中的下一个词作为模型的预测标签。由于该任务允许“动态”创建标签，因此我们可以利用大量的无标注文本数据集来训练大语言模型。

与 1.4 节中讨论的原始 Transformer 架构相比，GPT 的通用架构更为简洁。如图 1-8 所示，本质上，它只包含解码器部分，并不包含编码器。由于像 GPT 这样的解码器模型是通过逐词预测生成文本，因此它们被认为是一种自回归模型（autoregressive model）。自回归模型将之前的输出作为未来预测的输入。因此，在 GPT 中，每个新单词都是根据它之前的序列来选择的，这提高了最终文本的一致性。

GPT-3 等架构的规模远超原始 Transformer 模型。例如，原始的 Transformer 模型将编码器模块和解码器模块重复了 6 次，而 GPT-3 总共有 96 层 Transformer 和 1750 亿个参数。GPT-3 发布于 2020 年，按照深度学习和大语言模型的迅猛发展速度来衡量，这已是非常久远的事情了。然而，像 Meta 的 Llama 模型这样更近期的架构仍然基于相同的基本理念，仅进行了些许调整。因此，理解 GPT 仍然非常重要。本书侧重于实现 GPT 背后的核心架构，同时会介绍其他大语言模型采用的特别调整。

虽然原始的 **Transformer** 模型（包含编码器模块和解码器模块）专门为语言翻译而设计，但 **GPT** 模型采用了更大且更简单的纯解码器架构，旨在预测下一个词，并且它们也能执行翻译任务。这种能力起初让研究人员颇为意外，因为其来自一个主要在下一单词预测任务上训练的模型，而这项任务并没有特别针对翻译。

模型能够完成未经明确训练的任务的能力称为涌现（**emergence**）。这种能力并非模型在训练期间被明确教授所得，而是其广泛接触大量多语言数据和各种上下文的自然结果。即使没有经过专门的翻译任务训练，**GPT** 模型也能够“学会”不同语言间的翻译模式并执行翻译任务。这充分体现了这类大规模生成式语言模型的优势和能力。因此，无须针对不同的任务使用不同的模型，我们便可执行多种任务。

11.7 构建大语言模型

在本章，我们为理解大语言模型打下了基础。在本书的后续章节里，我们将从零开始，一步步构建自己的模型。我们将以 **GPT** 的核心原理为指导，按照图 1-9 所示的路线图，分 3 个阶段来逐步实现这一目标。

在第一阶段，我们将学习数据预处理的基本流程，并着手实现大语言模型的核心组件——注意力机制。

在第二阶段，我们将学习如何编写代码并预训练一个能够生成新文本的类 **GPT** 大语言模型。同时，我们还将探讨评估大语言模型的基础知识，这对于开发高效的自然语言处理系统至关重要。

需要指出的是，从头开始预训练大语言模型是一项艰巨的任务。训练类 **GPT** 模型所需的计算成本可能高达数千到数百万美元。鉴于本书的目的是教学，因此我们将使用较小的数据集进行训练。此外，本书也提供了用于展示如何加载那些公开可用的模型参数的示例代码。

最后，在第三阶段，我们将对一个预训练后的大语言模型进行微调，使其能够执行回答查询、文本分类等任务——这是许多真实应用程序和研究中常见的需求。

希望你已经做好准备。快与我们一起踏上这段激动人心的探索之旅吧！

11.8 小结

- 大语言模型彻底革新了自然语言处理领域。在此之前，自然语言处理领域主要采用基于明确规则的系统和较为简单的统计方法。而如今，大语言模型的兴起为这一领域引入了基于深度学习的新方法，在理解、生成和翻译人类语言方面取得了显著的进步。
- 现代大语言模型的训练主要包含两个步骤。
 - 首先，在海量的无标注文本上进行预训练，将预测的句子中的下一个词作为“标签”。
 - 随后，在更小规模且经过标注的目标数据集上进行微调，以遵循指令和执行分类任务。
- 大语言模型采用的是基于 **Transformer** 的架构。这一架构的核心组件是注意力机制，它使得大语言模型在逐词生成输出时，能够根据需要选择性地关注输入序列中的各个部分。
- 原始的 **Transformer** 架构由两部分组成：一个是用于解析文本的编码器，另一个是用于生成文本的解码器。
- 专注于生成文本和执行指令的大语言模型（如 **GPT-3** 和 **ChatGPT**）只实现了解码器部分，从而简化了整个架构。
- 由数以亿计的语料构成的大型数据集是预训练大语言模型的关键。
- 尽管类 **GPT** 大语言模型的常规预训练任务是预测句子中的下一个词，但它们展现出了能够完成分类、翻译或总结文本等任务的“涌现”特性。
- 当一个大语言模型完成预训练后，该模型便能作为基础模型，通过高效的微调来适应各类下游任务。
- 在自定义数据集上进行微调的大语言模型能够在特定任务上超越通用的大语言模型。

12. 大语言模型的架构

大语言模型要解决的问题是根据上文 (prompt, 提示词) 来“预测下一个 token”(next token prediction, ntp)。

例如，如果 prompt 是“君不见黄河”，那么大语言模型的工作流程如下：

君不见黄河 → LLM → 之
君不见黄河之 → LLM → 水
君不见黄河之水 → LLM → 天
君不见黄河之水天 → LLM → 上
⋮

(12.1)

而大语言模型如何进行训练呢？

首先我们需要有训练数据，也就是输入输出对。而大语言模型的预训练无需人工标注，可以直接由文本数据生成，所以有时也叫做“自监督学习”。

例如如果我们有一条文本数据“君不见黄河之水天上来”，那么训练数据如下：

输入	预测目标
君	不
君不	见
君不见	黄
君不见黄	河
君不见黄河	之
君不见黄河之	天
君不见黄河之水	上
君不见黄河之水上	来

表 12.1 训练数据：输入 - 预测目标对

我们现在知道了大语言模型要解决的问题是什么（预测下一个 token），以及对应的训练数据长什么样子，那么我们接下来就需要设计一个模型结构，然后使用训练数据训练这个模型结构，训练出来之后应该能够很好的完成预测下一个 token 的任务。

这个模型结构经过多年的研究，目前收敛到的架构是：Decoder-Only Transformer（纯解码器的 Transformer 架构）。

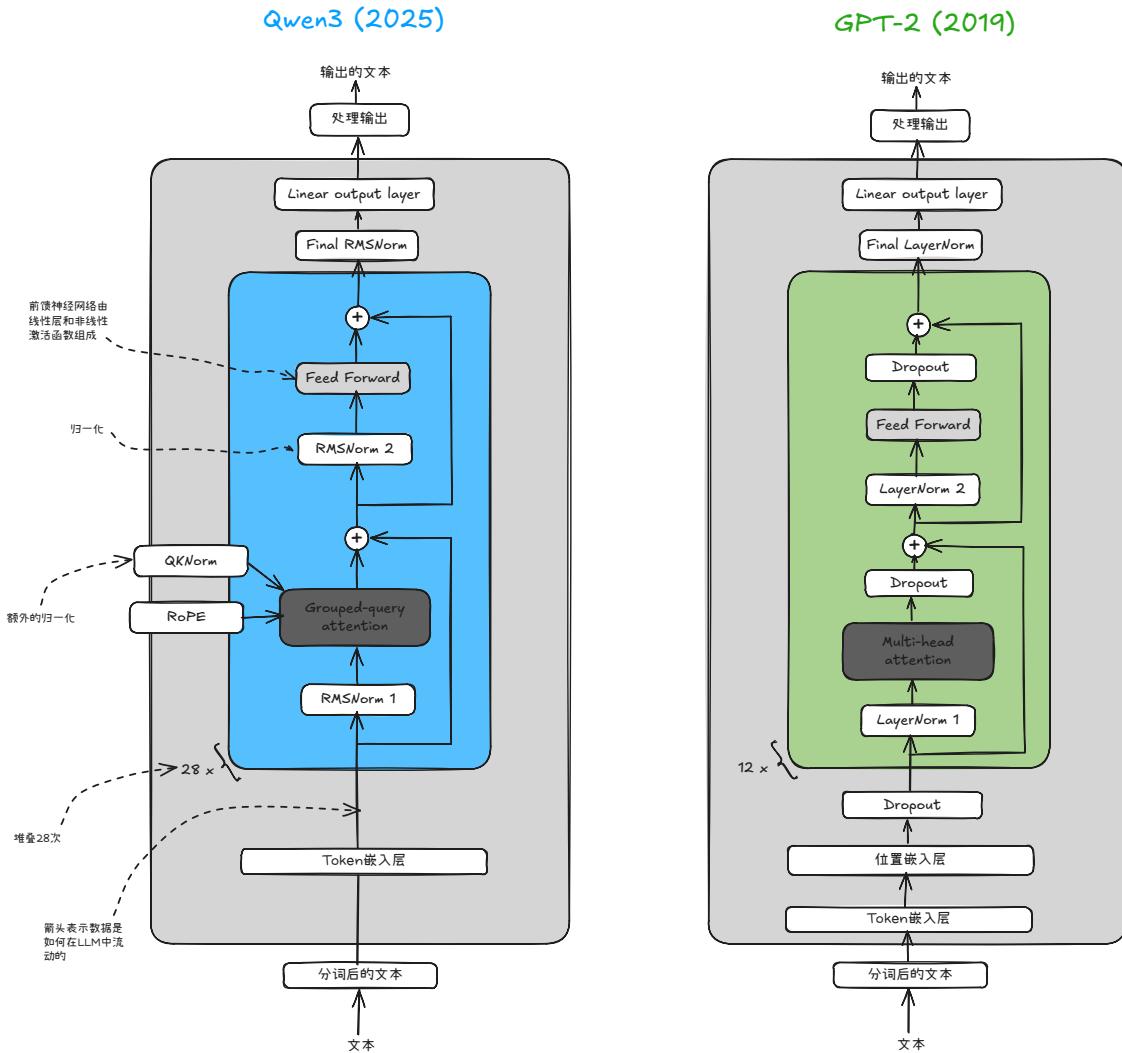


图 12.1 GPT-2 和 Qwen3 的架构图



13. 处理文本数据

文本数据的处理分两步：

1. 将文本分词转换成整数 (`input_ids`)
2. 将每个 `input_id` 对应一个数值向量 (`token 嵌入, word embeddings`)

13.1 对文本分词

这里我们实现一个 BPE (Byte Pair Encoding) 字节对编码分词器。

13.1.1 BPE 背后的核心理念

BPE 的核心思想是将文本转换为整数表示 (`token ID`), 以便用于 LLM 训练。

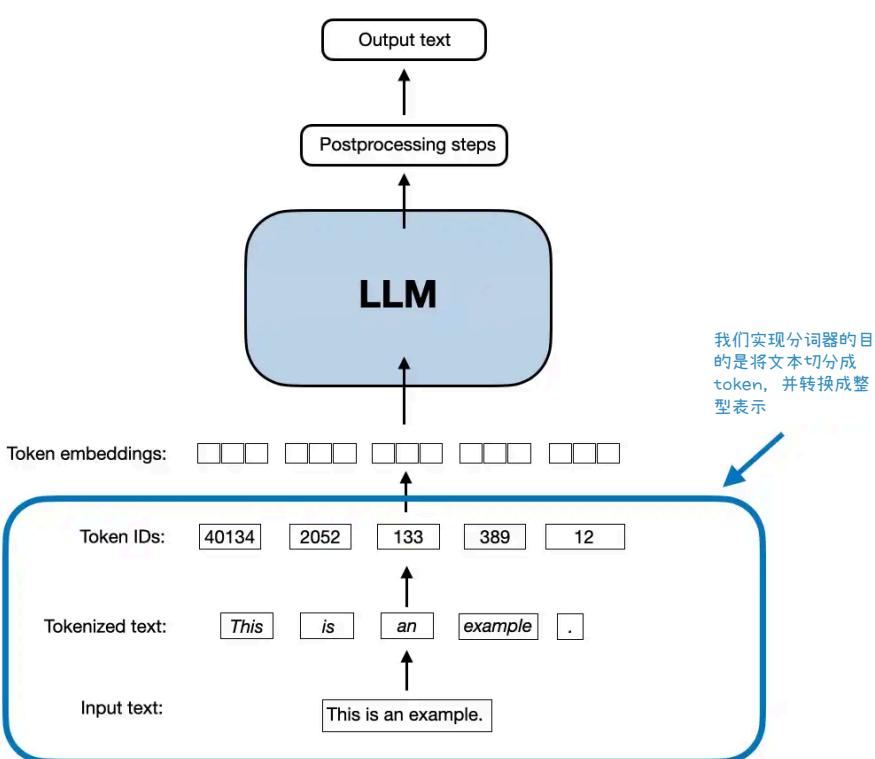


图 13.1 分词器

13.1.1.1 位和字节 (Bits and Bytes)

考虑将文本转换为字节数组 (毕竟 BPE 代表“字节”对编码):

```
1 text = "This is some text"
2 byte_ary = bytearray(text, "utf-8")
3 print(byte_ary)
4 # 输出: bytearray(b'This is some text')
```

Python

当我们对 `bytearray` 对象调用 `list()` 时，会将每个字节视为独立元素，结果将是一个与字节值相对应的整数列表：

```
1 ids = list(byte_ary)
2 print(ids)
3 # 输出: [84, 104, 105, 115, 32, 105, 115, 32, 115, 111, 109, 101, 32, 116, 101, 120, 116]
```

Python

这是一种将文本转换为 `token ID` 表示形式的有效方法，这是 LLM 嵌入层所需要的。

但这种方法的一个缺点是，它为每个字符创建一个 ID (对于一段简短文本来说，需要创建大量 ID！)

也就是说，这意味着对于 17 个字符的输入文本，我们必须使用 17 个 `token ID` 作为 LLM 的输入：

```
1 print("Number of characters:", len(text))
2 # 输出: Number of characters: 17
3 print("Number of token IDs:", len(ids))
4 # 输出: Number of token IDs: 17
```

Python

BPE 分词器使用的是词汇表机制——其工作原理并非为每个字符分配一个 token ID，而是针对完整的单词或子词来建立这种映射关系。

例如，GPT-2 分词器会将相同的文本（“This is some text”）分词为 4 个 token，而非 17 个：1212, 318, 617, 2420。

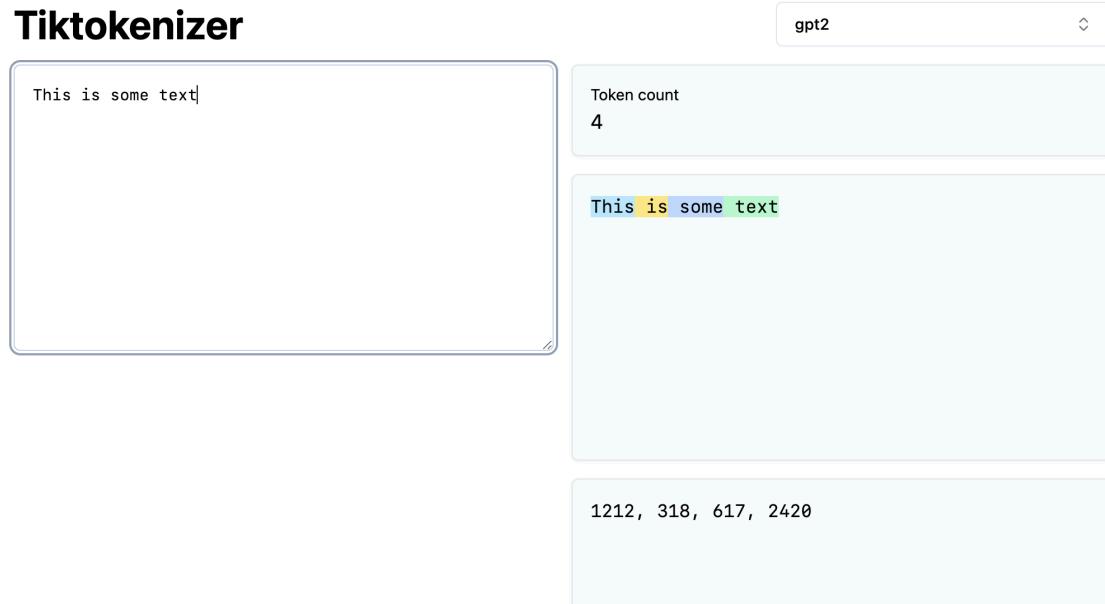


图 13.2 tiktoken 分词器

程序如下

```
1 import tiktoken
2
3 gpt2_tokenizer = tiktoken.get_encoding("gpt2")
4 gpt2_tokenizer.encode("This is some text")
5 # prints [1212, 318, 617, 2420]
```

一个字节由 8 位组成，因此单个字节能够表示从 0 到 255 的 256 个可能取值。

一个 BPE 分词器通常将这 256 个值作为其最初的 256 个单字符 token；你可以通过运行以下代码来直观地检查：

```
1 import tiktoken
2 gpt2_tokenizer = tiktoken.get_encoding("gpt2")
3
4 for i in range(300):
5     decoded = gpt2_tokenizer.decode([i])
6     print(f"{i}: {decoded}")
7 """
8 prints:
9 0: !
10 1: "
11 2: #
12 ...
13 255: ♦ # ----- single character tokens up to here
14 256: t
```

```

15 257: a
16 ...
17 298: ent
18 299: n
19 """

```

请注意，256 和 257 并非单字符值，而是双字符值（一个空格加一个字母），这是原始 GPT-2 BPE 分词器的一个小缺陷（该问题已在 GPT-4 分词器中得到改进）。

13.1.1.2 构建词汇表

BPE 分词算法的目标是构建一个常见子词的词汇表，例如像 298: ent 这样的子词（可以在 entangle、entertain、enter、entrance、entity 等词中找到），甚至包括像

```

1 318: is
2 617: some
3 1212: This
4 2420: text

```

BPE 算法最初在 1994 年由菲利普·盖奇（Philip Gage）所著的《一种新的数据压缩算法》（A New Algorithm for Data Compression）中描述。

13.1.1.3 BPE 算法概述

1. 识别常见配对

在每一次迭代中，扫描文本以找出最频繁出现的字节（或字符）对。

2. 替换并记录

将字节对替换为一个新的占位符 ID（使用尚未被占用的数字，例如：若起始编号范围为 0 至 255，则首个新占位符应为 256）。

将此映射关系记录在查找表中。

查找表的大小是一个超参数，亦称为“词汇表大小”（对于 GPT-2 模型，该值为 50,257）。

3. 重复直到无法进一步压缩

持续重复步骤 1 和步骤 2，不断地合并出现频率最高的配对

当无法进一步压缩时（例如，没有配对出现超过一次）即停止

解码

为了恢复原始文本，需反转此过程，使用查找表将每个 ID 替换为其对应的配对

13.1.1.4 BPE 算法举例

编码部分的具体示例（对应步骤 1 和 2）

假设我们有一段文本（训练数据集）the cat in the hat，希望基于此构建一个 BPE 分词器的词表

- 第 1 轮迭代

1. 识别高频对

在这段文字中，“th” 出现了两次（分别位于开头和第二个“e”之前）

2. 进行替换并记录

将“th” 替换为尚未使用的新 token ID，例如 256

新文本为：<256>e cat in <256>e hat

新词汇表为

```
1 0: ...
2 ...
3 256: "th"
```

- 第 2 轮迭代

1. 识别高频对

在文本<256>e cat in <256>e hat 中，组合<256>e 出现了两次

2. 进行替换并记录

使用未在使用的、新的 token ID (如 257) 替换<256>e。

新文本是：<257> cat in <257> hat

更新后的词汇表是：

```
1 0: ...
2 ...
3 256: "th"
4 257: "<256>e"
```

- 第 3 轮迭代

1. 识别高频对

在文本<257> cat in <257> hat 中，词语对<257>（注意这里有空格）出现了两次（一次在开头，一次在“hat”之前）。

2. 进行替换并记录

使用未在使用的、新的 token ID (如 258) 替换<257>。

新文本为：<258>cat in <258>hat

更新后的词汇表是：

```
1 0: ...
2 ...
3 256: "th"
4 257: "<256>e"
5 258: "<257> "
```

- 继续迭代

解码部分具体例子

要恢复原始文本，我们通过按引入顺序的逆序将每个 token ID 替换为其对应的字符对来反转这一过程

从最终的压缩文本开始：<258>cat in <258>hat

将<258>替换为<257>：<257> cat in <257> hat

将<257>替换为<256>e：<256>e cat in <256>e hat

将<256>替换为“th”：the cat in the hat

13.1.2 一个简单的 BPE 实现

1. 将输入文本分割为单独的字节
2. 反复查找并替换（合并）相邻的 token（对），当它们匹配到学习到的 BPE 合并中的任何一对时（按照从高到低的“等级”，即它们被学习的顺序进行）。
3. 继续合并操作直至无法再进行任何合并
4. 最终的 token ID 列表即为编码输出

代码如下：

```
1  from collections import Counter, deque
2  from functools import lru_cache
3  import re
4  import json
5
6
7  class BPETokenizerSimple:
8      def __init__(self):
9          # 词汇表: {token_id: token_str} (e.g., {11246: "some"})
10         self.vocab = {}
11
12         # 反向词汇表: {token_str: token_id} (e.g., {"some": 11246})
13         self.inverse_vocab = {}
14
15         # BPE的合并字典: {(token_id1, token_id2): merged_token_id}
16         self.bpe_merges = {}
17
18         # OpenAI官方的GPT-2合并字典使用了一个排序字典
19         # 格式为: {(string_A, string_B): rank}, rank越小, 优先级越高
20         self.bpe_ranks = {}
21
22
23     Args:
24         text (str): 训练文本
25         vocab_size (int): 词汇表大小
26         allowed_special (set): 特殊token集合
27
28         """
29
30         # 预处理: 将空格替换为"\u202f"
31         # 注意! \u202f仅在GPT-2 BPE实现中使用
32         # E.g., "Hello world" 会被分词为 ["Hello", "\u202fworl\u202f"]
33         # (GPT-4 BPE 会分词为 ["Hello", " world"])
34         processed_text = []
35         for i, char in enumerate(text):
36             if char == " " and i != 0:
37                 processed_text.append("\u202f")
38             if char != " ":
39                 processed_text.append(char)
40         processed_text = "\u202f".join(processed_text)
41
42         # 使用单个字符初始化词汇表, 包括"\u202f" (如果文本有空格的话)
43         # 词汇表从256个ascii字符开始
44         unique_chars = [chr(i) for i in range(256)]
45         unique_chars.extend(
46             char for char in sorted(set(processed_text))
47             if char not in unique_chars
48         )
```

```
49         if "\u00d7" not in unique_chars:
50             unique_chars.append("\u00d7")
51
52         self.vocab = {i: char for i, char in enumerate(unique_chars)}
53         self.inverse_vocab = {char: i for i, char in self.vocab.items()}
54
55     # 添加特殊token
56     if allowed_special:
57         for token in allowed_special:
58             if token not in self.inverse_vocab:
59                 new_id = len(self.vocab)
60                 self.vocab[new_id] = token
61                 self.inverse_vocab[token] = new_id
62
63     # 将处理后的文本分词为token IDs
64     token_ids = [self.inverse_vocab[char] for char in processed_text]
65
66     # BPE 1-3 步: 重复的发现和替换高频对
67     for new_id in range(len(self.vocab), vocab_size):
68         pair_id = self.find_freq_pair(token_ids, mode="most")
69         if pair_id is None:
70             break
71         token_ids = self.replace_pair(token_ids, pair_id, new_id)
72         self.bpe_merges[pair_id] = new_id
73
74     # 使用合并后的token构建词汇表
75     for (p0, p1), new_id in self.bpe_merges.items():
76         merged_token = self.vocab[p0] + self.vocab[p1]
77         self.vocab[new_id] = merged_token
78         self.inverse_vocab[merged_token] = new_id
79
80     def load_vocab_and_merges_from_openai(self, vocab_path, bpe_merges_path):
81         """
82             加载预训练的词汇表和OpenAI GPT-2的BPE合并文件
83
84         Args:
85             vocab_path (str): 词汇表文件路径(GPT-2叫做'encoder.json').
86             bpe_merges_path (str): bpe_merges文件路径(GPT-2叫做'vocab.bpe').
87         """
88         # 加载词汇表
89         with open(vocab_path, "r", encoding="utf-8") as file:
90             loaded_vocab = json.load(file)
91             # encoder.json字典格式为: {token_str: id}; 我们需要id→str和str→id
92             self.vocab = {int(v): k for k, v in loaded_vocab.items()}
93             self.inverse_vocab = {k: int(v) for k, v in loaded_vocab.items()}
94
95         # 必须包含 GPT-2的可打印换行字符 '\u2028' (U+010A) , 且 id 必须是 198
96         if "\u2028" not in self.inverse_vocab or self.inverse_vocab["\u2028"] != 198:
97             raise KeyError("Vocabulary missing GPT-2 newline glyph '\u2028' at id 198.")
```

```
98
99      # 必须包含 <endoftext> 且 id 为 50256
100     if "<endoftext>" not in self.inverse_vocab or self.inverse_vocab["<endoftext>"] != 50256:
101         raise KeyError("Vocabulary missing <endoftext> at id 50256.")
102
103     # 为换行符 '\n' 起别名, id 为 198
104     # 将可打印字符' '保存在词汇表中, 这样BPE的合并机制才能起作用
105     if "\n" not in self.inverse_vocab:
106         self.inverse_vocab["\n"] = self.inverse_vocab[" "]
107
108     if "\r" not in self.inverse_vocab:
109         if 201 in self.vocab:
110             self.inverse_vocab["\r"] = 201
111         else:
112             raise KeyError("Vocabulary missing carriage return token at id 201.")
113
114     # 加载GPT-2的合并然后存储优先级rank
115     self.bpe_ranks = {}
116     with open(bpe_merges_path, "r", encoding="utf-8") as file:
117         lines = file.readlines()
118         if lines and lines[0].startswith("#"):
119             lines = lines[1:]
120
121         rank = 0
122         for line in lines:
123             token1, *rest = line.strip().split()
124             if len(rest) != 1:
125                 continue
126             token2 = rest[0]
127             if token1 in self.inverse_vocab and token2 in self.inverse_vocab:
128                 self.bpe_ranks[(token1, token2)] = rank
129                 rank += 1
130             else:
131                 # 如果对的符号都不在词汇表中, 则直接跳过
132                 pass
133
134
135     def encode(self, text, allowed_special=None):
136         """
137             将输入文本编码为token ID列表, 使用tiktoken处理特殊token的风格
138
139         Args:
140             text (str): 输入文本
141             allowed_special (set or None): 是否允许特殊token。如果为None, 则禁用特殊
142             token处理机制
143
144         Returns:
145             token ID列表
```

```
145      """
146
147      # ---- 下面的代码模仿tiktoken处理特殊token的实现 ----
148      specials_in_vocab = [
149          tok for tok in self.inverse_vocab
150          if tok.startswith("｟") and tok.endswith("｠")
151      ]
152      if allowed_special is None:
153          # 如果文本中出现了特殊token, 则抛出异常
154          disallowed = [tok for tok in specials_in_vocab if tok in text]
155          if disallowed:
156              raise ValueError(f"Disallowed special tokens encountered in text: {disallowed}")
157      else:
158          # 允许一些特殊token (e.g., 例如｟endoftext｠)
159          disallowed = [tok for tok in specials_in_vocab if tok in text and tok not in allowed_special]
160          if disallowed:
161              raise ValueError(f"Disallowed special tokens encountered in text: {disallowed}")
162      #
163      -----
164      token_ids = []
165      # If some specials are allowed, split around them and passthrough those
166      # ids
167      if allowed_special is not None and len(allowed_special) > 0:
168          special_pattern = "(" + "|".join(
169              re.escape(tok) for tok in sorted(allowed_special, key=len,
170                                          reverse=True)
171          ) + ")"
172
173          last_index = 0
174          for match in re.finditer(special_pattern, text):
175              prefix = text[last_index:match.start()]
176              token_ids.extend(self.encode(prefix, allowed_special=None))  # encode prefix normally
177
178              special_token = match.group(0)
179              if special_token in self.inverse_vocab:
180                  token_ids.append(self.inverse_vocab[special_token])
181              else:
182                  raise ValueError(f"Special token {special_token} not found in vocabulary.")
183              last_index = match.end()
184
185              text = text[last_index:]  # remainder to process normally
186
187              # Extra guard for any other special literals left over
188              disallowed = [
189                  tok for tok in self.inverse_vocab
```

```

188             if tok.startswith("｟") and tok.endswith("｠") and tok in text and
189                 tok not in allowed_special
190         ]
190     if disallowed:
191         raise ValueError(f"Disallowed special tokens encountered in text:
191                         {disallowed}")
192
193
194     # ---- Newline and carriage return handling ----
195     tokens = []
196     parts = re.split(r'(\r\n|\r|\n)', text)
197     for part in parts:
198         if part == "":
199             continue
200         if part == "\r\n":
201             tokens.append("\r")
202             tokens.append("\n")
203             continue
204         if part == "\r":
205             tokens.append("\r")
206             continue
207         if part == "\n":
208             tokens.append("\n")
209             continue
210
211     # Normal chunk without line breaks:
212     # - If spaces precede a word, prefix the first word with 'Ġ' and
213     #   add standalone 'Ġ' for additional spaces
214     # - If spaces trail the chunk (e.g., before a newline) add
215     #   standalone 'Ġ' tokens (tiktoken produces id 220 for 'Ġ')
216     pending_spaces = 0
217     for m in re.finditer(r'( +)|(\\S+)', part):
218         if m.group(1) is not None:
219             pending_spaces += len(m.group(1))
220         else:
221             word = m.group(2)
222             if pending_spaces > 0:
223                 for _ in range(pending_spaces - 1):
224                     tokens.append("Ġ") # remaining spaces as standalone
225                     tokens.append("Ġ" + word) # one leading space
226                     pending_spaces = 0
227             else:
228                 tokens.append(word)
229             # Trailing spaces (no following word): add standalone 'Ġ' tokens
230             for _ in range(pending_spaces):
231                 tokens.append("Ġ")
232             # -----
233
234     # Map tokens → ids (BPE if needed)
235     for tok in tokens:
236         if tok in self.inverse_vocab:

```

```
237         token_ids.append(self.inverse_vocab[tok])
238     else:
239         token_ids.extend(self.tokenize_with_bpe(tok))
240
241     return token_ids
242
243 def tokenize_with_bpe(self, token):
244     """
245     Tokenize a single token using BPE merges.
246
247     Args:
248         token (str): The token to tokenize.
249
250     Returns:
251         List[int]: The list of token IDs after applying BPE.
252     """
253     # Tokenize the token into individual characters (as initial token IDs)
254     token_ids = [self.inverse_vocab.get(char, None) for char in token]
255     if None in token_ids:
256         missing_chars = [char for char, tid in zip(token, token_ids) if tid is
257                         None]
258         raise ValueError(f"Characters not found in vocab: {missing_chars}")
259
260     # If we haven't loaded OpenAI's GPT-2 merges, use my approach
261     if not self.bpe_ranks:
262         can_merge = True
263         while can_merge and len(token_ids) > 1:
264             can_merge = False
265             new_tokens = []
266             i = 0
267             while i < len(token_ids) - 1:
268                 pair = (token_ids[i], token_ids[i + 1])
269                 if pair in self.bpe_merges:
270                     merged_token_id = self.bpe_merges[pair]
271                     new_tokens.append(merged_token_id)
272                     # Uncomment for educational purposes:
273                     # print(f"Merged pair {pair} → {merged_token_id}"
274                     # ('{self.vocab[merged_token_id]}')")
275                     i += 2 # Skip the next token as it's merged
276                     can_merge = True
277                 else:
278                     new_tokens.append(token_ids[i])
279                     i += 1
280             if i < len(token_ids):
281                 new_tokens.append(token_ids[i])
282             token_ids = new_tokens
283
284     # Otherwise, do GPT-2-style merging with the ranks:
285     # 1) Convert token_ids back to string "symbols" for each ID
286     symbols = [self.vocab[id_num] for id_num in token_ids]
```

```
286
287     # Repeatedly merge all occurrences of the lowest-rank pair
288     while True:
289         # Collect all adjacent pairs
290         pairs = set(zip(symbols, symbols[1:]))
291         if not pairs:
292             break
293
294         # Find the pair with the best (lowest) rank
295         min_rank = float("inf")
296         bigram = None
297         for p in pairs:
298             r = self.bpe_ranks.get(p, float("inf"))
299             if r < min_rank:
300                 min_rank = r
301                 bigram = p
302
303         # If no valid ranked pair is present, we're done
304         if bigram is None or bigram not in self.bpe_ranks:
305             break
306
307         # Merge all occurrences of that pair
308         first, second = bigram
309         new_symbols = []
310         i = 0
311         while i < len(symbols):
312             # If we see (first, second) at position i, merge them
313             if i < len(symbols) - 1 and symbols[i] == first and symbols[i+1] == second:
314                 new_symbols.append(first + second) # merged symbol
315                 i += 2
316             else:
317                 new_symbols.append(symbols[i])
318                 i += 1
319         symbols = new_symbols
320
321         if len(symbols) == 1:
322             break
323
324     # Finally, convert merged symbols back to IDs
325     merged_ids = [self.inverse_vocab[sym] for sym in symbols]
326     return merged_ids
327
328     def decode(self, token_ids):
329         """
330         将token ID列表解码为文本字符串
331
332         Args:
333             token_ids (List[int]): The list of token IDs to decode.
334
```

```
335     Returns:  
336         str: The decoded string.  
337     """  
338     out = []  
339     for tid in token_ids:  
340         if tid not in self.vocab:  
341             raise ValueError(f"Token ID {tid} not found in vocab.")  
342         tok = self.vocab[tid]  
343  
344         # Map GPT-2 special chars back to real chars  
345         if tid == 198 or tok == "\n":  
346             out.append("\n")  
347         elif tid == 201 or tok == "\r":  
348             out.append("\r")  
349         elif tok.startswith("Ġ"):  
350             out.append(" " + tok[1:])  
351         else:  
352             out.append(tok)  
353     return "".join(out)  
354  
355     def save_vocab_and_merges(self, vocab_path, bpe_merges_path):  
356     """  
357         将词汇表和BPE合并保存到JSON文件中  
358  
359         Args:  
360             vocab_path (str): Path to save the vocabulary.  
361             bpe_merges_path (str): Path to save the BPE merges.  
362         """  
363         # Save vocabulary  
364         with open(vocab_path, "w", encoding="utf-8") as file:  
365             json.dump(self.vocab, file, ensure_ascii=False, indent=2)  
366  
367         # Save BPE merges as a list of dictionaries  
368         with open(bpe_merges_path, "w", encoding="utf-8") as file:  
369             merges_list = [{"pair": list(pair), "new_id": new_id}  
370                         for pair, new_id in self.bpe_merges.items()]  
371             json.dump(merges_list, file, ensure_ascii=False, indent=2)  
372  
373     def load_vocab_and_merges(self, vocab_path, bpe_merges_path):  
374     """  
375         Load the vocabulary and BPE merges from JSON files.  
376  
377         Args:  
378             vocab_path (str): Path to the vocabulary file.  
379             bpe_merges_path (str): Path to the BPE merges file.  
380         """  
381         # Load vocabulary  
382         with open(vocab_path, "r", encoding="utf-8") as file:  
383             loaded_vocab = json.load(file)  
384             self.vocab = {int(k): v for k, v in loaded_vocab.items()}
```

```

385         self.inverse_vocab = {v: int(k) for k, v in loaded_vocab.items()}
386
387     # Load BPE merges
388     with open(bpe_merges_path, "r", encoding="utf-8") as file:
389         merges_list = json.load(file)
390         for merge in merges_list:
391             pair = tuple(merge["pair"])
392             new_id = merge["new_id"]
393             self.bpe_merges[pair] = new_id
394
395     @lru_cache(maxsize=None)
396     def get_special_token_id(self, token):
397         return self.inverse_vocab.get(token, None)
398
399     @staticmethod
400     def find_freq_pair(token_ids, mode="most"):
401         pairs = Counter(zip(token_ids, token_ids[1:]))
402
403         if not pairs:
404             return None
405
406         if mode == "most":
407             return max(pairs.items(), key=lambda x: x[1])[0]
408         elif mode == "least":
409             return min(pairs.items(), key=lambda x: x[1])[0]
410         else:
411             raise ValueError("Invalid mode. Choose 'most' or 'least'.")
412
413     @staticmethod
414     def replace_pair(token_ids, pair_id, new_id):
415         dq = deque(token_ids)
416         replaced = []
417
418         while dq:
419             current = dq.popleft()
420             if dq and (current, dq[0]) == pair_id:
421                 replaced.append(new_id)
422                 # Remove the 2nd token of the pair, 1st was already removed
423                 dq.popleft()
424             else:
425                 replaced.append(current)
426
427         return replaced

```

13.1.3 BPE 实现逐步讲解

13.1.3.1 训练、编码与解码

首先，让我们考虑一些样本文本作为训练数据集：

```
1 text_path = "data.txt"
```

 Python

```

2
3 with open(text_path, "r", encoding="utf-8") as f:
4     text = f.read()

```

接下来，让我们初始化和训练 BPE 分词器，设置词汇表大小为 1000

注意，由于先前讨论的字节值，词汇表大小默认已是 256，因此我们只需“学习” 744 个词汇（若将 `<endoftext>` 特殊 token 和 6 空白 token 计算在内；确切而言，实际为 742 个）

作为对比，GPT-2 的词汇表包含 50,257 个 token，GPT-4 的词汇表有 100,256 个 token (`tiktoken` 中的 `cl100k_base`)，而 GPT-4o 则使用 199,997 个 token (`tiktoken` 中的 `o200k_base`)；相比我们上述的简单示例文本，它们的训练集规模都要大得多

```

1 tokenizer = BPETokenizerSimple()
2 tokenizer.train(text, vocab_size=1000, allowed_special={"<endoftext>"})
3 # print(tokenizer.vocab)
4 print(len(tokenizer.vocab))

```

本词汇表通过合并 742 次 (= 1000 - len(range(0, 256)) - len(special_tokens) - "G" = 1000 - 256 - 1 - 1 = 742) 创建而成

```
1 print(len(tokenizer.bpe_merges))
```

这意味着前 256 个条目是单字符 token

接下来，我们使用通过 `encode` 方法创建的合并操作来编码一些文本：

```

1 input_text = "Jack embraced beauty through art and life."
2 token_ids = tokenizer.encode(input_text)
3 print(token_ids)
4
5 input_text = "Jack embraced beauty through art and life.<endoftext> "
6 token_ids = tokenizer.encode(input_text, allowed_special={"<endoftext>"})
7 print(token_ids)
8
9 print("Number of characters:", len(input_text))
10 print("Number of token IDs:", len(token_ids))
11
12 print(tokenizer.decode(token_ids))

```

从上述长度可以看出，一个包含 42 个字符的句子被编码成了 20 个 token ID，与基于逐字符字节的编码方式相比，有效缩短了输入长度约一半

请注意，词汇表本身在 `decode()` 方法中被使用，这使得我们能够将 token ID 映射回文本。

迭代每个 token ID 可以让我们更好地理解这些 token ID 是如何通过词汇表解码的：

```

1 for token_id in token_ids:
2     print(f"{token_id} → {tokenizer.decode([token_id])}")

```

输出

```

1 424 → Jack
2 256 →
3 654 → em
4 531 → br
5 302 → ac
6 311 → ed
7 256 →
8 296 → be

```

```

9  97 -> a
10 465 -> ut
11 121 -> y
12 595 -> through
13 841 -> ar
14 116 -> t
15 287 -> a
16 466 -> nd
17 256 ->
18 326 -> li
19 972 -> fe
20 46 -> .
21 257 -> <|endoftext|>
22 256 ->

```

由此可见，大多数 token ID 都代表 2 字符的子词；这是因为训练数据的文本非常简短，没有那么多重复的单词，同时因为我们使用了相对较小的词汇量

总而言之，调用 `decode(encode())` 应能还原任意输入文本：

```

1 tokenizer.decode(
2     tokenizer.encode("This is some text."))
3 )
4 tokenizer.decode(
5     tokenizer.encode("This is some text with \n newline characters."))
6 )

```

13.1.3.2 保存和加载分词器

接下来，让我们看看如何保存训练好的分词器，以便后续重复使用：

```

1 # Save trained tokenizer
2 tokenizer.save_vocab_and_merges(vocab_path="vocab.json",
3                                 bpe_merges_path="bpe_merges.txt")
4 # Load tokenizer
5 tokenizer2 = BPETokenizerSimple()
6 tokenizer2.load_vocab_and_merges(vocab_path="vocab.json",
7                                 bpe_merges_path="bpe_merges.txt")

```

加载的分词器应能产生与之前相同的结果：

```

1 print(tokenizer2.decode(token_ids))
2 tokenizer2.decode(
3     tokenizer2.encode("This is some text with \n newline characters."))
4 )

```

13.1.3.3 从 OpenAI 加载原始的 GPT-2 BPE 分词器

```

1 files_to_download = {
2     "https://openaipublic.blob.core.windows.net/gpt-2/models/124M/vocab.bpe":
3         "vocab.bpe",
4     "https://openaipublic.blob.core.windows.net/gpt-2/models/124M/encoder.json":
5         "encoder.json"
6 }
7
8 tokenizer_gpt2 = BPETokenizerSimple()

```

```

7  tokenizer_gpt2.load_vocab_and_merges_from_openai(
8      vocab_path="encoder.json", bpe_merges_path="vocab.bpe"
9  )
10
11 print(len(tokenizer_gpt2.vocab))

```

词汇表大小应为 50257。

我们现在可以通过 `BPETokenizerSimple` 对象使用 GPT-2 分词器。

```

1 input_text = "This is some text"
2 token_ids = tokenizer_gpt2.encode(input_text)
3 print(token_ids)
4 print(tokenizer_gpt2.decode(token_ids))

```

使用官方的 `tiktoken` 库验证一下

```

1 import tiktoken
2
3 gpt2_tokenizer = tiktoken.get_encoding("gpt2")
4 gpt2_tokenizer.encode("This is some text")
5 # prints [1212, 318, 617, 2420]

```

13.2 创建词嵌入查找表

13.2.1 词嵌入深入讨论

在 PyTorch 中，嵌入层实现的功能与执行矩阵乘法的线性层相同；我们使用嵌入层主要是出于计算效率的考虑。

```

1 # 假设我们有以下3个训练样本,
2 # 它们可能代表LLM上下文中的token ID
3 idx = torch.tensor([2, 3, 1])
4
5 # 嵌入矩阵的行数可以通过
6 # 获得最大的token ID + 1来确定。
7 # 如果最大的token ID是3, 那么我们需要4行, 用于可能的
8 # token ID: 0, 1, 2, 3
9 num_idx = max(idx)+1
10
11 # 期望的嵌入维度是一个超参数
12 out_dim = 5

```

让我们实现一个简单的嵌入层：

```

1 torch.manual_seed(123)
2 embedding = torch.nn.Embedding(num_idx, out_dim)
3 print(embedding.weight) # 查看嵌入权重
4 # 获得训练示例ID为1的向量表示
5 print(embedding(torch.tensor([1])))
6 # 获得训练示例ID为2的向量表示
7 print(embedding(torch.tensor([2])))
8 # 查找一批ID的向量表示
9 idx = torch.tensor([2, 3, 1])

```

```
10 print(embedding(idx))
```

1) Index of training example [1]

2) Embedding matrix

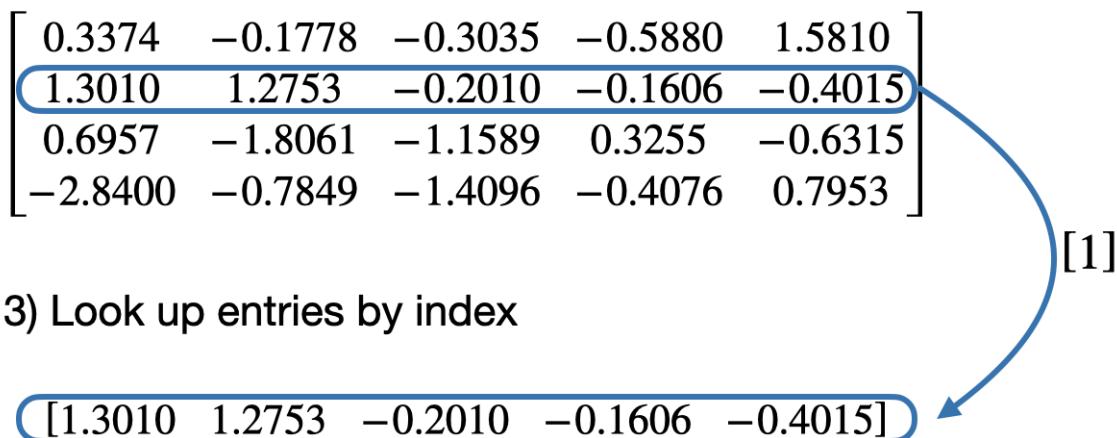


图 13.3 训练示例 ID 为 1 的向量表示

1) Index of training example [2]

2) Embedding matrix

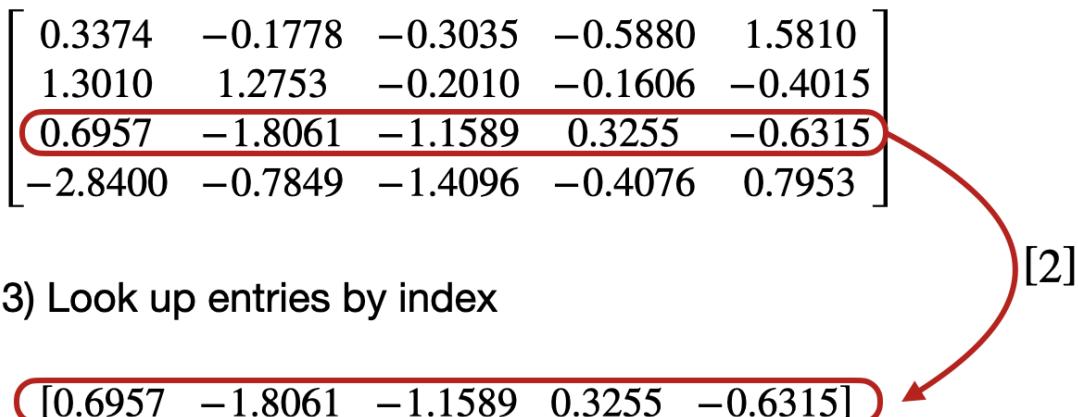


图 13.4 训练示例 ID 为 2 的向量表示

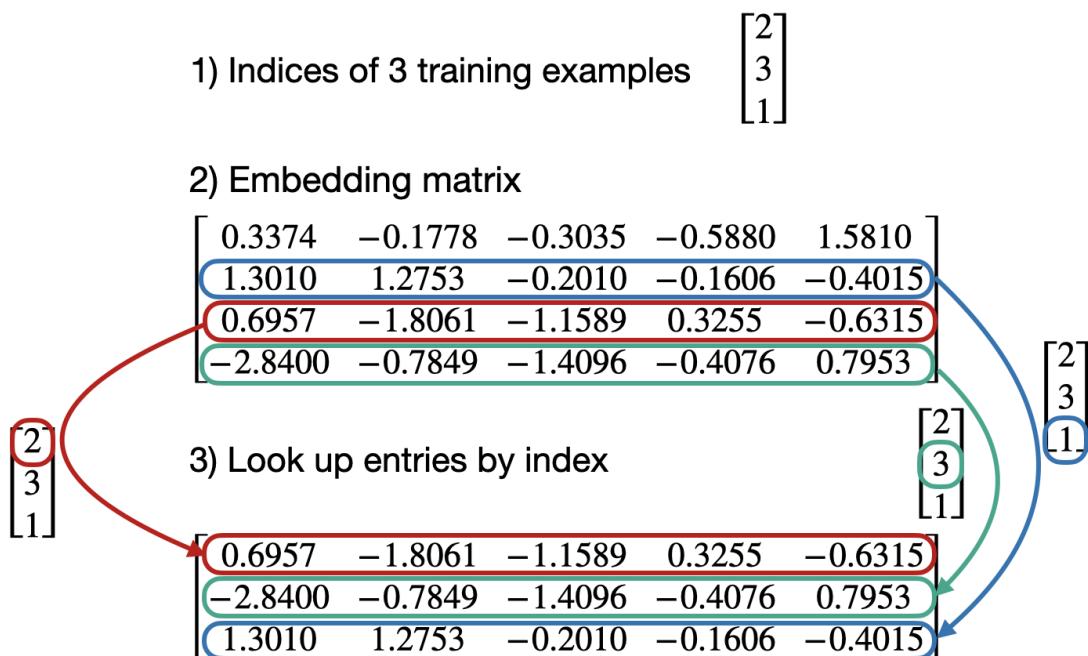


图 13.5 查找一批 ID 的向量表示

现在我们用独热编码和 `nn.Linear` 实现和上面的嵌入层一样的功能。

首先将 `token ID` 转换为独热编码

```
1 onehot = torch.nn.functional.one_hot(idx)
2 print(onehot)
```

Python

接下来，我们初始化一个 `Linear` 层，它会执行一个矩阵乘法 XW^T

```
1 torch.manual_seed(123)
2 linear = torch.nn.Linear(num_idx, out_dim, bias=False)
3 print(linear.weight)
```

Python

请注意，PyTorch 中的线性层同样也是用小的随机权重初始化的；为了与上面的 `Embedding` 层直接比较，我们必须使用相同的小随机权重，这就是我们在此处重新赋值它们的原因：

```
1 linear.weight = torch.nn.Parameter(embedding.weight.T)
```

Python

现在我们可以将线性层应用于输入数据的独热编码表示

```
1 print(linear(onehot.float()))
2 print(embedding(idx))
```

Python

正如我们所见，这与使用嵌入层时得到的结果完全相同。

底层执行的是对第一个训练示例的 `token ID` 进行的如下计算：

1) Convert indices of 3 training examples to one-hot encoding

$$\begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

2) Multiply one-hot encoded inputs with weight matrix

$$\begin{aligned}
 & \begin{array}{l}
 0 \times 0.3374 \\
 +0 \times 1.3010 \\
 +1 \times 0.6957 \\
 +0 \times -2.8400 \\
 = 0.6957
 \end{array} \\
 & \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0.3374 & -0.1778 & -0.3035 & -0.5880 & 1.5810 \\ 1.3010 & 1.2753 & -0.2010 & -0.1606 & -0.4015 \\ 0.6957 & -1.8061 & -1.1589 & 0.3255 & -0.6315 \\ -2.8400 & -0.7849 & -1.4096 & -0.4076 & 0.7953 \end{bmatrix} \\
 & = \begin{bmatrix} 0.6957 & -1.8061 & -1.1589 & 0.3255 & -0.6315 \\ -2.8400 & -0.7849 & -1.4096 & -0.4076 & 0.7953 \\ 1.3010 & 1.2753 & -0.2010 & -0.1606 & -0.4015 \end{bmatrix}
 \end{aligned}$$

图 13.6 对第一个训练示例的底层计算过程

第二个训练样本对应的 token ID 是：

1) Convert indices of 3 training examples to one-hot encoding

$$\begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

2) Multiply one-hot encoded inputs with weight matrix

$$\begin{aligned}
 & \begin{array}{l}
 0 \times 0.3374 \\
 +0 \times 1.3010 \\
 +0 \times 0.6957 \\
 +1 \times -2.8400 \\
 = -2.8400
 \end{array} \\
 & \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0.3374 & -0.1778 & -0.3035 & -0.5880 & 1.5810 \\ 1.3010 & 1.2753 & -0.2010 & -0.1606 & -0.4015 \\ 0.6957 & -1.8061 & -1.1589 & 0.3255 & -0.6315 \\ -2.8400 & -0.7849 & -1.4096 & -0.4076 & 0.7953 \end{bmatrix} \\
 & = \begin{bmatrix} 0.6957 & -1.8061 & -1.1589 & 0.3255 & -0.6315 \\ -2.8400 & -0.7849 & -1.4096 & -0.4076 & 0.7953 \\ 1.3010 & 1.2753 & -0.2010 & -0.1606 & -0.4015 \end{bmatrix}
 \end{aligned}$$

图 13.7 对第二个训练示例的底层计算过程

由于每行独热编码中除一个索引外全为 0 (这是设计的必然结果), 该矩阵乘法实质上等同于对独热元素的查表操作。

这种在独热编码上使用矩阵乘法的做法等价于嵌入层查找操作, 但当处理大型嵌入矩阵时会效率低下, 因为存在大量与零相乘的无效计算。

13.2.2 词嵌入的反向传播

1. 问题设定

参数设置

- 词汇表大小 (`vocab_size`): 2
- 嵌入维度 (`embedding_dim`): 2
- 输入序列 (`input_ids`): [0, 1, 0]
- 序列长度: 3

嵌入矩阵

$$W \in \mathbb{R}^{2 \times 2} = \begin{bmatrix} w_{00} & w_{01} \\ w_{10} & w_{11} \end{bmatrix} \quad (13.1)$$

其中:

- 第 0 行: `token ID = 0` 的嵌入向量
- 第 1 行: `token ID = 1` 的嵌入向量

2. 将 `token id` 列表转换成独热编码

$$\text{OneHot}([0, 1, 0]) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \in \mathbb{R}^{3 \times 2} \quad (13.2)$$

3. 前向传播

$$\begin{aligned} E &= \text{OneHot} \cdot W \\ &= \begin{bmatrix} e_{00} & e_{01} \\ e_{10} & e_{11} \\ e_{20} & e_{21} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} w_{00} & w_{01} \\ w_{10} & w_{11} \end{bmatrix} = \begin{bmatrix} w_{00} & w_{01} \\ w_{10} & w_{11} \\ w_{00} & w_{01} \end{bmatrix} \end{aligned} \quad (13.3)$$

4. 反向传播 (求导)

从标量函数 \mathcal{L} 开始推导, 得到

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial W} &= \text{OneHot}^T \cdot \frac{\partial \mathcal{L}}{\partial E} \\ &= \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial e_{00}} & \frac{\partial \mathcal{L}}{\partial e_{01}} \\ \frac{\partial \mathcal{L}}{\partial e_{10}} & \frac{\partial \mathcal{L}}{\partial e_{11}} \\ \frac{\partial \mathcal{L}}{\partial e_{20}} & \frac{\partial \mathcal{L}}{\partial e_{21}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial e_{00}} + \frac{\partial \mathcal{L}}{\partial e_{20}} & \frac{\partial \mathcal{L}}{\partial e_{01}} + \frac{\partial \mathcal{L}}{\partial e_{21}} \\ \frac{\partial \mathcal{L}}{\partial e_{10}} & \frac{\partial \mathcal{L}}{\partial e_{11}} \end{bmatrix} \end{aligned} \quad (13.4)$$

有两个发现

- `Token ID = 0` 出现了 2 次 (位置 0 和位置 2)
 - 第 0 行梯度 = 两个位置的梯度之和
- `Token ID = 1` 出现了 1 次 (位置 1)
 - 第 1 行梯度 = 该位置的梯度

即：某个 `token` 的嵌入向量的梯度 = 所有使用该 `token` 位置的梯度之和。

Python

```
1 import torch
2 import torch.nn as nn
3
4 # 设置随机种子以便复现
5 torch.manual_seed(42)
6
7 # 定义参数
8 vocab_size = 2
9 embedding_dim = 2
10 input_ids = torch.tensor([0, 1, 0])
11
12 # 创建嵌入层
13 embedding = nn.Embedding(vocab_size, embedding_dim)
14
15 # 查看初始权重
16 print("初始嵌入矩阵 W:")
17 print(embedding.weight)
18 print()
19
20 # 前向传播
21 output = embedding(input_ids)
22 print("嵌入输出 E:")
23 print(output)
24 print()
25
26 # 假设一个简单的损失：所有元素的平方和
27 loss = (output ** 2).sum()
28 print(f"损失 L = {loss.item():.4f}")
29 print()
30
31 # 反向传播
32 loss.backward()
33
34 # 查看梯度
35 print("嵌入矩阵的梯度 ∂L/∂W:")
36 print(embedding.weight.grad)
37 print()
38
39 # 手动计算验证
40 print("≡≡ 手动验证 ≡≡")
41 # ∂L/∂E = 2 * E (因为 L = sum(E2) )
42 grad_E = 2 * output
43 print("∂L/∂E (梯度矩阵 G):")
44 print(grad_E)
45 print()
46
47 # 手动计算 ∂L/∂W
48 grad_W_manual = torch.zeros(vocab_size, embedding_dim)
```

```
49 for i, idx in enumerate(input_ids):
50     grad_W_manual[idx] += grad_E[i]
51
52 print("手动计算的 ∂L/∂W:")
53 print(grad_W_manual)
54 print()
55
56 # 验证是否一致
57 print("PyTorch 梯度与手动计算是否一致:")
58 print(torch.allclose(embedding.weight.grad, grad_W_manual))
```