Deep Learning Architectures

A Mathematical Approach

作者：康豪
组织：SMS, UESTC
时间：Sept 1, 2023
版本：1.0
类型：学习笔记

Live long and prosper! ——Vulcans

目录

前言

矩阵理论是数学的一个重要分支，在多种工程学科中有极其重要的应用。本门课程以电子科技大学《矩阵理论》为主要参考教材，同时参考了 R.A.Horn, C.R.Johnson 等学者所著的经典教材 *Matrix Analysis* 以及其他书籍，本笔记是由作者根据教学内容编写。

笔记的 LaTeX 模板来自 *ElegantLatex* 团队编写的作品 *elegantbook*，该系列风格优雅、功能齐全，被各类讲义、笔记编著者广泛采纳，实为佳作。

封面图片为船底座大星云，由韦伯望远镜拍摄，图片来自 NASA 官网。

第 1 章 优化理论与算法

监督学习中的学习过程包括调整网络参数（权值和偏差），直到一定的代价函数被最小化。由于参数的数量相当大（它们很容易成数千），需要一个鲁棒的最小化算法. 本章介绍了许多不同口味的最小化算法，并强调了它们的优缺点.

1.1 最小值的一般性质

本节回顾关于具有一个实变量或几个实变量的最小函数值的基本概念. 这些理论上可行的技术只有在变量的数量不是太大的情况下，在实践中才是有效的。然而，在机器学习中，变量的数量是成千上万或更多的，所以这些经典的寻找最小值的理论方法对于这些应用来说并不是有利可图的。我们在这里包含它们只是为了完整性，并有一个构建未来更复杂的方法的基础.

1.1.1 单变量实值函数

设 $f:[a,b]\to\mathbb{R}$, 是定义在紧区间 $[a,b]$ 上的一个连续函数，在这里紧区间就是有界闭区间。根据微积分的知识，有界闭区间上的连续函数必有界，且一定存在最大最小值. 也就是说，存在 $c\in[a,b]$，使得 $f(c)=\min_{x\in[a,b]}f(x)$. 这个点被称为 f 的全局最小值点，当然 f 可能也有一些局部极小值点. 此外，根据 Fermat 引理，若 c 是函数 f 的极小值点，且 f 在 c 处可导，那么 $f'(c)=0$. 注意，这是一个必要但不充分条件，当补充条件 $f''(x)\geq 0)$ 时，上述条件就变成了充要条件.

1.1.2 多变量实值函数

设 K 是 $\mathbb{R}^n$ 上的紧集，即 K 是一个有界闭集，与单变量实值函数类似，当 K 中有满足条件的全局或局部极值点 c 时，必然满足

$$\frac{\partial f}{\partial x_i}(c)=0,\quad i=1,\ldots,n. \tag{1.1}$$

上式可以等价地描述为 $\nabla f(c)=0$，其中 $\nabla f=\sum_{i=1}^{n}\frac{\partial f}{\partial x_i}\tilde{e}_i$. 系统1.1等价于曲面 z = f (x) 处（c，f (c)）的切平面是水平的，即平行于 x 超平面.

在 $x=c$ 的邻域中，$f(x)$ 的二阶泰勒近似给出如下:

$$\begin{aligned}f(x)&=f(c)+\sum_i\frac{\partial f}{\partial x_i}(c)(x_i-c_i)\\&\quad+\frac{1}{2}\sum_{i,j}\frac{\partial^2 f}{\partial x_j\partial x_k}(c)(x_j-c_j)(x_k-c_k)+o(\|x-c\|^2)\\&=f(c)+(x-c)^T\nabla f(c)+\frac{1}{2}(x-c)^T H_f(c)(x-c)+o(\|x-c\|^2),\end{aligned}$$

其中，$H_f=\frac{\partial^2 f}{\partial x_j\partial x_k}$，是 f 的 Hessian 矩阵.

命题 1.1

设 c 是 $\nabla f(c)=0$ 的解，假设 H_f 在 c 的邻域上是正定的。那么 c 是 f 的局部最小值点.

证明 设 $x(t)$ 为具有 $x(0)=c$ 的 f 的结构域上的任意固定曲线，并考虑复合函数 $g(t)=f(x(t))$. 为了证明 c 是 $f(x)$ 的局部最小值，等价于证明 $t=0$ 是 $g(t)$ 的局部最小值，对于任何曲线 $x(t)$ 的局部最小值。

例题 1.1 二维空间中的正定 Hessian 矩阵 考虑一个在 $\mathcal{R}^2$ 上具有连续导数的二次可微函数 f, 它的 Hessian 矩阵

是一个 2×2 矩阵:

$$H_f(x,y) = \begin{pmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{pmatrix},$$

其中，我们使用符号 $f_{xx} = \frac{\partial^2 f}{\partial x^2}$ 来表示 x 中的二次偏导数. 对于任何向量，$u = (ab)$ 提供了二次形式

$$u^T H_f u = f_{xx}a^2 + 2f_{xy}ab + f_{yy}b^2,$$

例题 1.2 二次函数

例题 1.3 谐函数

1.2 梯度下降法

梯度下降算法是一种通过通过相关的水平集进入最大成本降低的方向来找到函数的最小值的过程。我们将首先介绍水平集的成分。对令人生畏的细节不感兴趣的读者可以直接跳过到第 4.2.2 节.

附录 A