Deep Learning Architectures

A Mathematical Approach

作者：康豪
组织：SMS, UESTC
时间：Aug 31, 2023
版本：1.0
类型：学习笔记

Live long and prosper! ——Vulcans

目录

前言

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 引言

第 2 章 激活函数

第 3 章 损失函数

第 4 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.

4.1 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.

4.1.1 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.

4.1.2 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

$$\frac{\partial f}{\partial x_i}(c)=0,\quad i=1,\ldots,n. \tag{4.1}$$

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4.1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 4.1

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 4.1 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

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

改写一下形式，凑出完全平方后得到：

$$\begin{aligned} u^T H_f u = & \quad f_{xx}\left(a + \frac{f_{xy}}{f_{xx}}b\right)^2 + \frac{b^2}{f_{xx}}(f_{xx}f_{yy} - f_{xy}^2) \\ = & \quad f_{xx}\left(a + \frac{f_{xy}}{f_{xx}}b\right)^2 + \frac{b^2}{f_{xx}} \det H_f. \end{aligned}$$

根据之前的计算，我们可以得到:

- 如果 $f_{xx} > 0$ 和 $detH_f > 0$，那么对于任何 $u \in \mathbb{R}^2$，$u^t H_f u > 0$，即 Hessian 矩阵 H_f 是正定的.
- 如果 $f_{xx} < 0$ 和 $detH_f > 0$，那么对于任何 $u \in \mathbb{R}^2$，$u^t H_f u < 0$，即 Hessian 矩阵 H_f 是负定的.

例题 4.2 二次函数 设 A 是一个对称的、正定的、非退化的 $n \times n$ 矩阵，并考虑以下 n 个变量的实值二次函数

$$f(x) = x^T Ax - 2b^T x + d, \quad x \in \mathbb{R}^n,$$

其中 $b \in \mathbb{R}^n$，$d \in R$，其中 $x^T Ax = \sum_{i,j} a_{ij}x_i x_j$ 和 $b^T x = \sum_k b_k x_k$.。我们有梯度 $\nabla f(x) = 2Ax - 2b$ 和 Hessian 矩阵 $H_f = 2A$，所以这个函数是凸的。$\nabla f(x) = 0$ 的解，即 $c = A^{-1}b$，是 f 的局部最小值。由于该解决方案是唯一的，因此它实际上是一个全局最小值。A 的可逆性是由行列式非零的 Hessian 矩阵的性质得出的.

如果前面的二次函数 $f(x)$ 仅定义在一个紧集 K 上，有时解 $c = a^1 b$ 可能不属于 K，因此我们不能通过求解相关的欧拉系统得到 f 的所有最小值。另一方面，我们知道 f 在紧致域 K 上达到了一个全局最小值。因此，这个最小值必须属于 K 的边界，而找到它需要对最小值进行边界搜索.

例题 4.3 谐函数 如果 $\Delta_x f = 0, \forall x \in D,$，则一个函数 $f(x)$ 在结构域 $D \subset \mathbb{R}^n$ 上被称为谐波，其中 $\Delta_x f = \sum_i \frac{\partial^2 f(x)}{\partial x_i^2}$ 是 f 的拉普拉斯算子. 拉普拉斯的最小（最大）性质表明一个调和函数在域 d 的边界上达到最小（最大值）. 换句话说，一个调和函数的极值总是在域的边界上达到.

由于任何仿射函数 $f(x) = b^T x + db \in R^n d \in R$ 都是调和的，所以 f 的最小值（最大值）在 d 的边界上达到。如果 D 是一个凸多边形的内部，那么在多边形的顶点上达到最小值（极大值）. 寻找解顶点是单纯形算法的基本思想.

我们已经看到，通过求解系统 $\nabla f(x) = 0$，不能总是找到最小值. 即使这是可能的，这种寻找最小值的分析方法在实践中也并不总是可行的，因为涉及到大量的变量. 下面，我们将介绍一些用于使给定函数近似为最小值的鲁棒迭代方法.

4.2 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下一节.

4.2.1 水平集

考虑函数 z = f (x)，x D Rn，与 n≥2，并定义集合 Sc = f−1（c）=x Rn；f（=）=c。假设函数 f 与非零梯度可微， f(x)=0，x Sc。在此条件下，Sc 成为 Rn 中的一个（n−1）维超曲面。族 c 称为函数 f 的水平超曲面。对于 n 个 =2，它们被称为水平曲线。几何上，将 z = f (x) 的图与水平平面 z = c 相交，得到水平超曲面，见图 4.1.

命题 4.2

梯度 ∇f 垂直于水平集 S_c.

证明 ...

在等价符号中，对于每个 x Sc，向量 f(x) 正态于 Sc 在 x 处的切面 TxSc，见图 4.2。超曲面的方向的选择使 f 指向向外的方向。切平面 TxSc 作为 Sc 内外 x 左右点的无穷小分隔符，见图 4.3 a。

假设函数 z = f (x) 在 x∗ D 处有一个（局部）<f(x) 最小值，对于所有 x Vx∗，V 邻域为 x∗（我们可以假设 V 是一个以 x∗ 为中心的球）。用 z∗=f（x∗）表示 f 在 x∗ 处的局部最小值。然后有一个 >0，使得任何 c ，z∗，z∗，z∗+)，见图 4.1。对于 c=z∗，超曲面退化到一点，Sc=x∗。对于足够小的，家族 [Sc]c [z∗，z∗+) 是嵌套的，即，如果 c1 < c2，那么 Sc1 Int（Sc2）。

下一个结果说明了任意初始方向曲线的存在，由 x∗ 发出，它们演化垂直于族 Sc 族，见图 4.3 b。回想一下，如果一个函数 φ（−），使得 |φ（−）−φ(y)(y)|≤Kx−y,，则称为 Lipschitz 连续. 下面的两个存在性结果将使用这个假设。

引理 4.1

假设 ∇f 是 Lipschitz 连续的。对于任何向量 $v \in \mathcal{R}^n$，存在 $\delta > 0$ 和一个可微曲线 $\alpha : [0, \delta) \to \mathbb{R}^n$，使：

- $\alpha(0) = x^*$;
- $\dot{\alpha}(0) = v$;
- $\dot{\alpha}(t)$ 垂直于 $\mathcal{S}_{f(\alpha(t))}$, 对所有的 $t \in [0, \delta)$.

♡

证明 ...

为了将来参考，当我们也表示初始方向 v 时，作为上述 ode 系统解的曲线将用 αv (t) 表示。

值得注意的是，在重新参数化时，曲线是唯一的，即曲线可以改变速度，而其几何图像保持不变。

一个显著的参数化是在水平差参数 τ = c−z∗ 上。这是由于对于较小的 t 值，我们有 ˙c(t)>0。如果新参数化中的曲线用 β（τ）表示，则 β(0)=x∗= α(0) 和 β（τ）=˙α（t(τ)）t（τ），那么 β(0) = t(0)v。我们也有方便的关联关系 β(τ) = β（c−z∗） Sc，它也可以写成 f（β（τ））= c。

现在我们陈述并证明以下局部连通性结果，这将在稍后的梯度下降法中使用：

定理 4.1

假设 f 是连续的。对于任何一个足够接近 x∗ 的点 x0，都有一个可微曲线 γ：[0，δ]→Rn，使得

- $\gamma(0) = x^0$;
- $\gamma(\delta) = x^*$;
- $\dot{\gamma}(t)$ 垂直于 $\mathcal{S}_{f(\gamma(t))}$, 对所有的 $t \in [0, \delta)$.

♡

我们将提供一个非建设性的证据。我们先准备好一些符号。设 αv：[0，δv)→Rn 为引理 4.2.2 提供的曲线。假设初始向量子酉，v≤1。由于结束值 δv 相对于 v 是连续的，所以它在酉球上达到其最小值，即 δ=minv≤1δv。现在用 Aδ=αv(t)；t [0，δ)， v Rn，v≤1 表示。集合 Aδ 表示由曲线 αv (t) 交换的实心域，该曲线从 x∗ 发出到所有方向 v，直到时间 δ。δ 值的选择使得 Aδ 的定义有意义。集合 Aδ 不是空的，因为显然是 x∗ Aδ。下一个结果表明，实际上 Aδ 包含一个以 x∗ 为中心的非空球，见图 4.4 a。这实际上是一个等价于定理 4.2.3 的陈述的结果。

定理 4.2

存在 $\epsilon > 0$，使得 $B(x^*, \epsilon) \subset A_\delta$.

♡

证明 我们将首先提供一个经验证明。B（x∗,） Aδ 意味着 x∗ 是 Aδ 的一个内点。如果，由于矛盾，我们假设 x∗ 不是一个内点，那么有一个序列的点（xk）k 收敛到 x∗，使得 xk/ Aδ。这里是经验假设的地方：假设 xk 位于一条光滑的曲线 x (s) 上，从 x(0)=x∗ 开始，满足 x（sk）= xk，sk 为负数递减序列，见图 4.4 b。设 v0=x(0) 为曲线 x (s) 接近于点 x∗ 的方向。引理 4.2.2 产生了一个从这个方向开始的曲线 αv0，它将与一个邻域上的 x (s) 重合。这是由于曲线 x (s) 和 αv0 都具有相同的初始点和速度。因为在这种情况下，我们会有 x(s)=αv0（−） Aδ，对于 −s<δ，这导致了一个矛盾。

我们有两点注记:

- 所有可以通过满足上述性质的曲线 γ 连接到 x∗ 的点都构成了 x∗ 的吸引盆地。这个定理可以等价地说，引力盆地包含一个以 x∗ 为中心的球.
- 方向 v0=x(0) 是一个简并的方向。先前的证明表明不存在退化方向。对这一事实的正式证明可以使用逆函数定理（见附录中的定理 F.1）来完成，如下所示。

证明 ...

关于 η 的大小，我们有一些评论：

- 如果 η 较大，算法停止过早，然后达到点 x∗ 的良好近似，见图 4.6 a；
- 如果 η 太小，停止阶数 m 很大，在计算机实现的情况下可能没有时间有效，见图 4.6 b。

在实际应用中，步长 η 的大小是一个正在运行的应用程序的误差幅度和时间有效性之间的权衡。我们将在第 4.2.4 节中进一步正式阐述这一想法.

4.2.2 方向导数

后来使用的另一个概念是方向导数，它测量一个函数在给定方向上的某一点上的瞬间变化率。更准确地说，假设 v 是 Rn 中的一个酉向量，并考虑可微函数 f：U Rn→R。f 在第 x0 U 点处的方向导数定义为

$$\frac{\partial f}{\partial v}(x^0) = \lim_{t \searrow 0} \frac{f(x^0 + tv) - f(x^0)}{t}.$$

注意，关于坐标的偏导数，∂x∂fk，是关于坐标向量 v =（0，...，1，...，0）T 的方向导数。链规则的应用提供了作为标量积的方向导数的计算：

$$\begin{aligned}\frac{\partial f}{\partial v}(x^0) = \frac{d}{dt} f(x^0 + tv)\big|_{t=0+} &= \sum_{k+1}^{n} \frac{\partial f}{\partial x_k}(x^0 + tv)v^k \Big|_{t=0} \\ &= v^T \nabla f(x^0) = \langle \nabla f(x^0), v \rangle.\end{aligned}$$

4.2.3 最速下降法

最速下降法（或梯度下降法）是一种基于贪婪算法的数值方法，该算法通过将给定的一步指向函数值减少最大的方向来搜索函数的最小值。人们可以想象一个被蒙着眼睛的游客想要以最快的方式下山的方法。在每一点上，游客都在检查距离，以找到下降最陡的方向，然后朝那个方向走一步。然后重复这个过程，直到游客最终到达山谷的底部（或者，如果他的脚步太小，就会被困在当地的最小值）.

为了应用这种方法，我们感兴趣的是找到酉方向 v，其中函数 f 在给定的小步长 η 内尽可能地减小。函数 f 在初始点 x0 处的值和在方向 v 上大小为 η 的一步后的值之间的变化，用线性近似表示为

$$\begin{aligned}f(x^0 + \eta v) - f(x^0) &= \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(x^0)\eta v^k + o(\eta^2) \\ &= \eta \langle \nabla f(x^0), v \rangle + o(\eta^2).\end{aligned}$$

下面，由于 η 很小，我们将忽略二次项 o（η2）的影响。因此，为了得到 v，使函数的变化具有最大的负值，我们将使用标量积的柯西不等式

$$-\left\|\nabla f(x^0)\right\| \|v\| \le \left\langle \nabla f(x^0), v \right\rangle \le \left\|\nabla f(x^0)\right\| \|v\|.$$

已知道，对于负比例的向量达到左边的不等式。3 由于 v=1，最小值为

$$v = -\frac{\nabla f(x^0)}{\|\nabla f(x^0)\|}.$$

那么这个函数中的最大变化近似等于

$$f(x^0 + \eta v) - f(x^0) = \eta \langle \nabla f(x^0), v \rangle = -\eta \|\nabla f(x^0)\|.$$

常数 η 被称为学习速率。从前面的关系来看，每一步后函数的变化与梯度的大小和学习速率成正比.

该算法由以下迭代组成，它构造了以下序列（xn）：

- 在全局最小值 x∗ 的吸引盆地中选择一个初始点 x0,
- 使用迭代来构造序列 $(xn)n$.

$$x^{n+1} = x^n - \eta \frac{\nabla f(x^n)}{\|\nabla f(x^n)\|}. \tag{4.2}$$

这保证了目标函数的负变化，这是由 f（xn+1）−f（xn）=−η f（xn）<0 给出的。

我们注意到线 xnxn+1 是正常的水平超表面 Sf（xn）。因此，我们从前一节中得到了多边形线 Pm = [x0，...，xm]，它是由定理 4.2.3 提供的曲线 γ 的近似值。

然而，这种结构有一个缺点，它很快就会被修复。由于 xn+1−xn= η > 0，近似序列（xn）n 不收敛，因此很容易错过最小点 x∗，见图 4.7 a。为了克服这个问题，我们将假设学习速率 η 是可调的，即随着函数变化的较慢（当梯度较小时），它会变得更小，见图 4.7 b。我们假设现在有一个正的常数 δ > 0，这样在第 n 次迭代中的学习率与梯度，ηn = δ f（xn）. 成正比然后，迭代（4.2.3）变为

$$x^{n+1} = x^n - \delta \nabla f(x^n). \tag{4.3}$$

命题 4.3

当且仅当梯度序列收敛于零， f（xn）→0 时，n→∞ 时，由（4.2.4）定义的序列（xn）n 为收敛

证明 ...

例题 4.4

4.2.4 线搜索法

4.2.5 运动学解释

4.3 动量法

附录 A