机器学习中的优化基础及案例剖析

刘维湘
深圳大学
wxliu@szu.edu.cn

2019 年 2 月 3 日

常见符号说明

$\mathbb{A}, \mathbb{B}, \mathbb{C}$: 集合 $\mathbb{A}, \mathbb{B}, \mathbb{C}$

$\mathbb{R}$: 实数集

$\mathbb{K}$: 复数集

$\mathbb{N}$: 自然数集

$\mathbb{Z}$: 整数集

$\mathbb{R}^n$: n 维欧式空间

i, j, k, l, m, n, k, r : 整数 i, j, k, l, m, n, k, r

x : 数 x, 比如 $x \in \mathbb{R}$ (零阶张量);

有时在不引起混淆的情况下，我们也用 x_1, x_2 等来表示多个数。

x : 向量 x, 比如 $\mathrm{x} \in \mathbb{R}^n$ (一阶张量), 第 i 个分量 x_i

X : 矩阵 X, 比如 $\mathrm{X} \in \mathbb{R}^{m \times n}$ (二阶张量), 第 i 行 $\mathrm{X}_{i\cdot}$,
第 j 列 $\mathrm{X}_{\cdot j}$, 第 i 行第 j 列的元素 X_{ij}

$\mathcal{X}$: 张量 $\mathcal{X}$, 比如三阶张量 $\mathcal{X} \in \mathbb{R}^{m \times n \times k}$, 四阶张量 $\mathcal{X} \in \mathbb{R}^{m \times n \times k \times t}$

x_* : x 的最优值

$\hat{\cdot}$: 估计值

$\bar{\cdot}$: 算术平均值

$|\cdot|$: 实数的绝对值或复数的模

$||\cdot||$: 向量，矩阵或张量的范数

$.*$: 对应元素乘法

$./$: 对应元素除法

$<\cdot, \cdot>$: 内积

目录

1 简介 **4**
1.1 预备知识 (集合与映射) 4
1.2 机器学习的基本问题 6
1.3 最优化的基本模型 6

2 凸集与凸函数 **7**
2.1 凸集 7
2.2 凸函数 8
2.3 超平面 8
2.4 凸集可分离定理 9
2.5 共轭函数 9

3 对偶理论 **9**
3.1 拉格朗日函数 9
3.2 拉格朗日对偶 10

4 最优性条件 **12**
4.1 无约束最优性条件 12
4.2 约束最优性条件 12

5 最优化方法 **12**
5.1 线搜索与信赖域方法 13
5.2 无约束最优化 13
5.3 约束最优化 13

6 非凸与非光滑优化方法 **14**
6.1 非凸优化方法 14
6.2 非光滑优化方法 14

7 实例 1: Logistic 回归 **14**

8 实例 2: 支持向量机 **14**

9 实例 3: 非负矩阵分解 14
9.1 模型 14
9.2 算法 14
9.3 应用 14

1 简介

优化技术，

智能优化算法，比如遗传算法，蚁群算法等仿生算法。

从机器学习的问题出发（1 章），然后介绍优化相关的基本理论（2 章），再介绍主要的优化方法（2 章），然后举例进行阐述（1 章）。

1.1 预备知识 (集合与映射)

这里主要介绍集合与映射的基本概念，并在此基础上介绍与最优化相关的几个重要概念，比如上下确界以及上镜图等。

1.1.1 集合

集合的概念是现代数学的基础概念之一。我们把具有相同性质的一类事物的全体称为集合。比如，

- 所有的等边三角形构成一个集合
- 所有三维空间的点构成一个集合
- 某种疾病的所有患者构成一个集合

组成集合中的某一个成员称之为集合的元素。集合里所有元素的个数称为集合的基数 (cardinal number)。有的集合具有有限个元素，有的可能具有无限个元素。

描述一个集合通常有两种方法:

- 枚举法: 将集合的所有元素列举出来，比如 0-10 的奇数构成的集合为 {1,3,5,7,9}
- 属性描述法: 二维平面上以坐标原点为中心的一个单位圆上所有的点构成的集合 $\{(x, y)|x^2 + y^2 = 1\}$

对于一个标量集合 S，其上 (下) 界与上 (下) 确界的定义在优化中是比较重要的概念。

如果 α 小于或等于 S 中的任一元素，则称 α 是 S 的一个下界。S 中所有的下界的最大值 (若存在)，则称为其下确界 (inf)，记为 $\alpha = \inf S$。

如果 β 大于或等于 S 中的任一元素，则称 β 是 S 的一个上界。S 中所有的上界的最小值 (若存在)，则称为其上确界 (sup)，记为 $\beta = \sup S$。

比如集合 $(0, 1)$，-1 是其一个下界，0 是其下确界，2 是其一个上界，1 是其上确界。

1.1.2 函数与映射

有了集合之后，如何刻画两个集合上的对应关系，这是现代科学和工程实践中的核心任务之一。比如，本文谈到的机器学习其实就是要在训练集上建立一个映射规则到标签集合。

我们首先回顾一下函数的概念，而映射是函数概念的推广。

函数的定义：给定一个数集 A，假设其中的元素为 x。现对 A 中的元素 x 施加对应法则 f，记作 $f(x)$，得到另一数集 B。假设 B 中的元素为 y。则 y 与 x 之间的等量关系可以用 $y = f(x)$ 表示。我们把这个关系式就叫函数关系式，简称函数。函数概念含有三个要素：定义域 A、值域 B 和对应法则 f。其中核心是对应法则 f，它是函数关系的本质特征。函数（function），最早由中国清朝数学家李善兰翻译，出于其著作《代数学》。之所以这么翻译，他给出的原因是“凡此变数中函彼变数者，则此为彼之函数”，也即函数指一个量随着另一个量的变化而变化，或者说一个量中包含另一个量。函数的定义通常分为传统定义和近代定义，函数的两个定义本质是相同的，只是叙述概念的出发点不同，传统定义是从运动变化的观点出发，而近代定义是从集合、映射的观点出发。

两个非空集合 A 与 B 间存在着对应关系 f，而且对于 A 中的每一个元素 x，B 中总有有唯一的一个元素 y 与它对应，就这种对应为从 A 到 B 的映射，记作 f $A \to B$。其中，b 称为元素 a 在映射 f 下的象，记作：$b = f(a)$。a 称为 b 关于映射 f 的原象。集合 A 中所有元素的象的集合称为映射 f 的值域，记作 $f(A)$。

映射在不同的领域有很多的名称，它们的本质是相同的。如函数，算子等等。这里要说明，函数是两个数集之间的映射，其他的映射并非函数。

1.2 机器学习的基本问题

利用集合与映射的关系来对常见的机器学习问题进行描述。

1.2.1 聚类分析

聚类分析指将物理或抽象对象的集合分组为由类似的对象组成的多个类的分析过程。它是一种重要的人类行为。聚类分析的目标就是在相似的基础上收集数据来分类。聚类源于很多领域，包括数学，计算机科学，统计学，生物学和经济学。在不同的应用领域，很多聚类技术都得到了发展，这些技术方法被用作描述数据，衡量不同数据源间的相似性，以及把数据源分类到不同的簇中。

聚类有很多方法

给定一组数据 $x_i \in R^d, i = 1, \ldots, n$

1.2.2 分类与回归

考虑两类分类问题：给定一组数据 $(x_i, y_i), x_i \in R^d, y_i \in -1, 1, i = 1, \ldots, n$

对于回归来说，$y_i \in R$

1.2.3 数据降维

1.2.4 特征选择

1.3 最优化的基本模型

假设 $x \in R^n$

$$\begin{aligned} \underset{\mathrm{x}}{\text{minimize}} \quad & f(\mathrm{x}) && (1)\\ \text{subject to} \quad & g_i(\mathrm{x}) \le 0, \quad i = 1, \ldots, m && (2)\\ & h_j(\mathrm{x}) = 0, \quad j = 1, \ldots, p. && (3) \end{aligned}$$

其中优化模型的三要素：优化变量x,目标函数(1)$f(x)$,约束条件 包括不等式约束 (2) 及等式约束 (3)$g_i(x), h_j(x)$.

等式约束可以转换成不等式约束进行。$x = y :: x \ge y, x \le y$

需要说明的是，优化变量有可能来自其他空间，比如在独立成分分析 [] 或非负矩阵分解 [] 中，我们的优化变量是一个矩阵。尽管可以将一个矩阵拉直成向量形式，但有时候矩阵形式能较好的保持数据原始的空间结构信息，比如在支持矩阵机 (SMM)[] 中。

下面讨论典型的几种规划问题：

1. 线性规划
2. 非线性规划
3. 凸规划
4. 非凸规划

2 凸集与凸函数

上 (下) 界与上 (下) 确界 sup inf
两个集合上的对应关系。

2.1 凸集

定义 给定一个集合 C，如果对任意的 $x, y \in C, \alpha, \beta \in [0, 1]$, 且 $\alpha+\beta = 1$, 有 $\alpha x + \beta y \in C$ 成立，则 C 为凸集。也就是说，凸集中任意两点之间的线段仍然属于集合 C, 这正是凸集的几何意义所在. 图?? 给出了凸集和非凸集的实例。

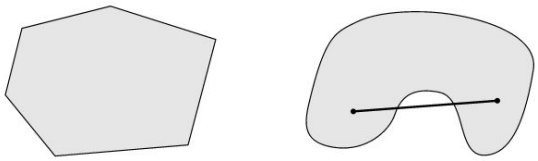

Figure 1: Examples of a convex set (a) and a non-convex set (b).

图 1: 凸集 (左) 和非凸集 (右) 的示例

凸集的相关运算的结果，比如凸集关于加法，数乘和交运算是封闭的，参见如下引理。

2.2 凸函数

定义 给定非空凸集 $C \subset R^n$, 有函数 $f : C \leftarrow R$. 如果 $\forall x, y \in C, \theta \in [0, 1]$, 有

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) \tag{4}$$

那么称 f 为 C 上的凸函数。若上式 (4) 中的不等式变为等式，则 f 为线性函数。

凸函数的几何意义

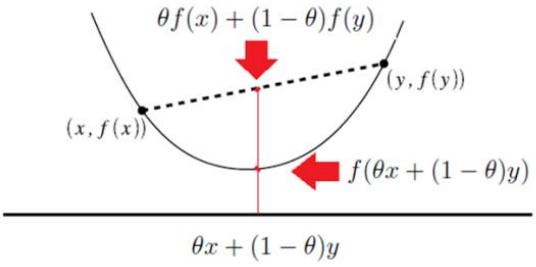

图 2: 凸函数的几何意义

凸函数的判定条件

一阶条件，

二阶条件

给定一个函数，判定其凸性，在优化模型与算法设计中，具有重要的意义。

下面给出几个凸函数的例子。

上图的定义

上图的几何意义

上图 (epi-graph) 将凸集和凸函数联系起来。

2.3 超平面

超平面的概念是凸分析与凸优化中的重要概念之一。超平面在机器学习中的支持向量机方法中就是一个例子。

定义 设 $x \in R^n$, 由任意非零向量 $a \neq 0$ 及标量 b 所限制构成的集合 $\{x | a^T x = b, x \in R^n, a \neq 0\}$ 就是 超平面。

当 $n = 1$ 时，超平面是直线上的一个点

当 $n = 2$ 时，超平面是平面上的一条直线。

当 $n = 3$ 时，超平面是三维空间上的一个平面

以 $n = 2$ 为例，如图所示

2.4 凸集可分离定理

简单的说，两个不相交的凸集，可以被一个超平面分离。这在优化理论中是一个很重要的基础原则。

在机器学习算法中，感知器以及支持向量机都和超平面有关。

2.5 共轭函数

共轭函数的几何意义

3 对偶理论

以常见的凸优化模型来进行，比如线性规划，二次规划等。

对偶理论的共轭性框架 (借用共轭函数) 与 MC-MC 几何框架 (几何性质)，前者适合算法设计与分析，后者适合从几何角度来理解对偶理论，特别是对初学者来说。一般的优化理论书籍重于前者。

3.1 拉格朗日函数

再次考虑最优化的通用基本模型

假设 $\mathrm{x} \in \mathbb{R}^n$

$$\begin{aligned} \underset{\mathrm{x}}{\text{minimize}} \quad & f(\mathrm{x}) && (5)\\ \text{subject to} \quad & g_i(\mathrm{x}) \le 0, \quad i = 1, \dots, m && (6)\\ & h_j(\mathrm{x}) = 0, \quad j = 1, \dots, p. && (7) \end{aligned}$$

我们定义其拉格朗日函数为：

$$L(\mathrm{x}, \lambda, \mathrm{v}) = f(\mathrm{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathrm{x}) + \sum_{j=1}^{p} \mathrm{v}_j g_j(\mathrm{x}) \tag{8}$$

几个实例

3.2 拉格朗日对偶

$$
\begin{aligned}
L_d(\lambda, \mathrm{v}) &= \inf_{\mathbf{x}} L(\mathrm{x}, \lambda, \mathrm{v}) && (9)\\
&= \inf_{\mathbf{x}} \left\{ f(\mathrm{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathrm{x}) + \sum_{j=1}^{p} \mathrm{v}_j g_j(\mathrm{x}) \right\} && (10)
\end{aligned}
$$

几个实例

实例 1：$x \in \mathbb{R}$，且只有不等式约束

$$
\begin{aligned}
\underset{x}{\text{minimize}} \quad & f(x) = 3x^2 + 4x && (11)\\
\text{subject to} \quad & (x-1)(x-5) \leq 0, && (12)
\end{aligned}
$$

可行域为 $x \in [1, 5]$, 最优变量 $x_* = 1$, 最优值为 $f_* = 7$.

我们定义其约束函数为 $g(x) = (x-1)(x-5) \leq 0$, 其拉格朗日函数为：

$$
\begin{aligned}
L(x, \lambda) &= f(x) + \lambda g(x) && (13)\\
&= 3x^2 + 4x + \lambda(x-1)(x-5) && (14)\\
&= (3+\lambda)x^2 + (4-6\lambda)x + 5\lambda && (15)
\end{aligned}
$$

我们画出 $f(x)$, 以及不同 λ 值的情况下 $L(x, \lambda)$ 的图形, 如图**??**所示。

下面来求其对偶函数。若 $3 + \lambda \leq 0$ 时，拉格朗日函数 $L(x, \lambda)$ 关于 x 无下界。当 $3 + \lambda > 0$, 拉格朗日函数 $L(x, \lambda)$ 是关于 x 的凸函数。此时，令

$$
\frac{dL(x, \lambda)}{dx} = 2(3+\lambda)x + (4-6\lambda) = 0 \tag{16}
$$

所以，当 $x = \frac{-2+3\lambda}{3+\lambda}$，拉格朗日函数 $L(x, \lambda)$ 取得最小值

$$
L(\frac{-2+3\lambda}{3+\lambda}, \lambda) = \frac{-4+27\lambda-4\lambda^2}{3+\lambda} \tag{17}
$$

因此，当 $\lambda > -3$, 所得拉格朗日对偶函数为

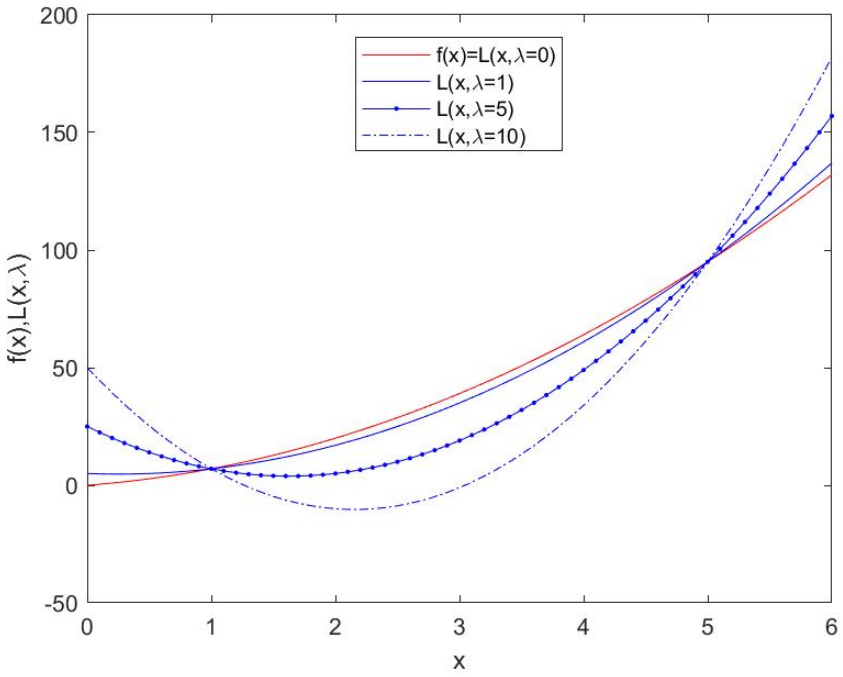

图 3: 目标函数 $f(x)$ 及拉格朗日函数 $L(x,\lambda)$

$$
\begin{aligned}
L_d(\lambda) &= \inf_x L(x,\lambda) && (18)\\
&= \inf_x\{f(x)+\lambda g(x)\} && (19)\\
&= \inf_x\{3x^2+4x+\lambda(x-1)(x-5)\} && (20)\\
&= \inf_x\{(3+\lambda)x^2+(4-6\lambda)x+5\lambda\} && (21)\\
&= \inf_x\{3x^2+4x+\lambda(x^2-6x+5)\} && (22)\\
&= \frac{-4+27\lambda-4\lambda^2}{3+\lambda} \quad (\lambda>-3) && (23)
\end{aligned}
$$

如图??所示

一个优化问题的原问题和对偶问题

通过求解对偶问题来实现原问题的求解。

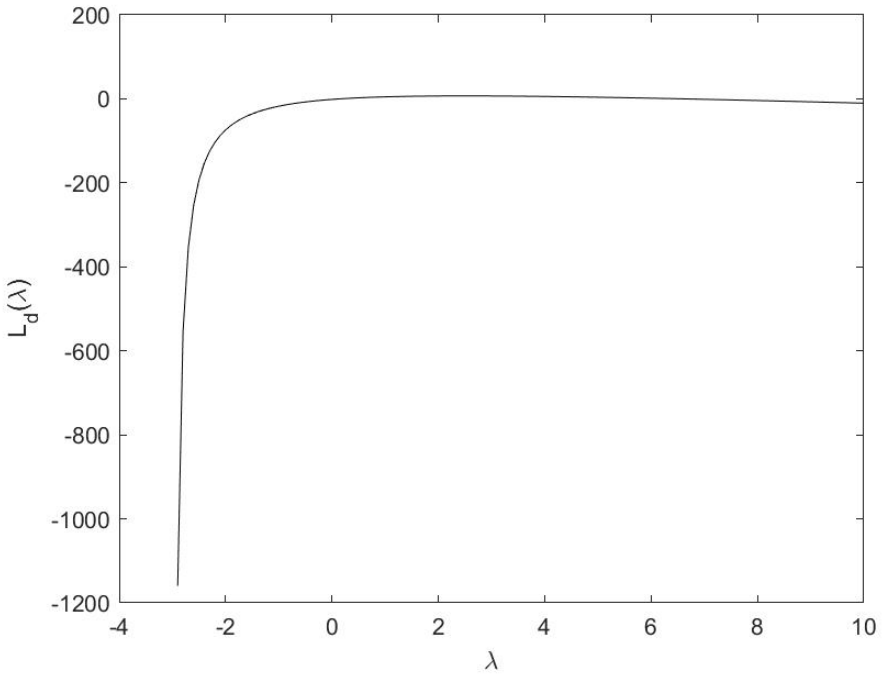

图 4: 拉格朗日对偶函数 $L_d(\lambda)$

4 最优性条件

4.1 无约束最优性条件

4.2 约束最优性条件

5 最优化方法

在现代优化方法的数值实现中，各种优化技术都是采用迭代的范式进行，即从一个初始点进行，按照一定的参数更新规则进行迭代，直到满足一定的条件。可以用如下的格式表示：

$$= \tag{24}$$

在以上的迭代过程中，优化参数与两个因素有关：迭代步长也叫学习速率，和搜索前进方向。线搜索方法与信赖域方法就是在处理步长和前进方向上的策略不一样而形成的两大类方法。

5.1 线搜索与信赖域方法

5.1.1 线搜索

5.1.2 信赖域方法

5.2 无约束最优化

Logistic regression

5.2.1 最速下降与梯度法

最速下降法 (负梯度法)，共轭梯度法

5.2.2 拟牛顿法

牛顿法，拟牛顿法

5.2.3 其他方法

5.3 约束最优化

SVM

5.3.1 可行方向法

5.3.2 罚函数法

5.3.3 序列二次规划法

6 非凸与非光滑优化方法

6.1 非凸优化方法

6.2 非光滑优化方法

7 实例 1: Logistic 回归

无约束优化

8 实例 2: 支持向量机

不等式约束优化

9 实例 3: 非负矩阵分解

不等式约束优化
参考 SO-NMF []

9.1 模型

9.2 算法

9.3 应用