第十章 优化基础

第 28 讲 优化简介

黄定江

DaSE @ ECNU

djhuang@dase.ecnu.edu.cn

1. 28.1 优化问题的一般形式和重要概念
2. 28.2 优化问题的分类
3. 28.3 数据科学与机器学习中常见的优化问题

引言

- 数据科学、人工智能和机器学习的很多问题都归结为一个优化问题
- 对优化问题的求解已然成为大部分数据分析和机器学习算法的核心组成部分
- 我们将安排三章内容来理清优化问题相关概念、基础理论以及其数值求解算法

引言

在本章中，将主要介绍：

- 优化问题的定义、优化问题的分类、数据科学中常见的优化问题
- 凸集和凸函数的定义和判别方法以及保凸运算
- 凸优化问题的定义和标准形式
- 介绍数据科学中一些典型凸优化问题

28.1.1 优化简介

本讲将介绍：

- 最优化问题的一般形式和一些重要概念
- 最优化问题的分类
- 展示许多应用实例中的最优化问题

最优化问题的一般形式

最优化问题的一般形式表示为：

$$\begin{aligned} \min \quad & f_0(\boldsymbol{x}) \\ \text{subject to}^1 \quad & f_i(\boldsymbol{x}) \le 0, \quad i = 1, \cdots, m \\ & h_j(\boldsymbol{x}) = 0, \quad j = 1, \cdots, p \end{aligned} \tag{1}$$

其中，向量 $\boldsymbol{x} = (x_1, \cdots, x_n)^T$ 称为问题的优化变量，函数 $f_0 : \mathbb{R}^n \to \mathbb{R}$ 称为目标函数，在机器学习中常为损失函数。函数 $f_i : \mathbb{R}^n \to \mathbb{R}$，被称为不等式约束函数，$f_i(\boldsymbol{x}) \le 0,\ i = 1, \cdots, m$ 称为不等式约束，函数 $h_j : \mathbb{R}^n \to \mathbb{R}$，被称为等式约束函数，$h_j(\boldsymbol{x}) = 0,\ j = 1, \cdots, p$ 称为等式约束。

注：如果求最大就是 max，如果最小值最大值不存在，就表示为 inf(sup)。

[1] 下文均简记为 “s.t.”

28.1.2 可行集

定义 1

目标函数和约束函数所有有定义点的集合：

$$\mathcal{D} = \bigcap_{i=0}^{m} \mathbf{dom}\, f_i \cap \bigcap_{j=1}^{p} \mathbf{dom}\, h_j$$

称满足所有约束条件的向量 $\boldsymbol{x} \in \mathcal{D}$ 为可行解或可行点，全体可行点的集合称为可行集，记为 $\mathcal{F}$，其表示为：

$$\mathcal{F} = \{\boldsymbol{x} \in \mathcal{D} | f_i(\boldsymbol{x}) \leq 0,\ i = 1, \cdots, m, \quad h_j(\boldsymbol{x}) = 0,\ j = 1, \cdots, p\}$$

若 $f_i(\boldsymbol{x})$ 和 $h_j(\boldsymbol{x})$ 是连续函数，则 $\mathcal{F}$ 是闭集。

最优值

在可行集中找一点 $\boldsymbol{x}^*$，使目标函数 $f_0(\boldsymbol{x})$ 在该点取最小值，则称 x^* 为问题的最优点或最优解，$f_0(\boldsymbol{x}^*)$ 称为最优值，记为 p^*：

$$p^* = \inf\{f_0(\boldsymbol{x}) | f_i(\boldsymbol{x}) \leq 0, i = 1, \cdots, m,\ h_j(\boldsymbol{x}) = 0, j = 1, \cdots, p\}$$

- $p^* = \infty$，如果问题不可行 (没有 $\boldsymbol{x}$ 满足约束)
- $p^* = -\infty$，问题无下界

全局最优与局部最优

定义 2

整体（全局）最优解： 若 $\boldsymbol{x}^* \in \mathcal{F}$，对于一切 $\boldsymbol{x} \in \mathcal{F}$，恒有 $f_0(\boldsymbol{x}^*) \leq f_0(\boldsymbol{x})$，则称 $\boldsymbol{x}^*$ 是最优化问题(1)的整体最优解。

定义 3

局部最优解： 若 $\boldsymbol{x}^* \in \mathcal{F}$，存在某个领域 $N_\varepsilon(\boldsymbol{x}^*)$，使得对于一切 $\boldsymbol{x} \in N_\varepsilon(\boldsymbol{x}^*) \cap \mathcal{F}$，恒有 $f_0(\boldsymbol{x}^*) \leq f_0(\boldsymbol{x})$，则称 $\boldsymbol{x}^*$ 是最优化问题(1)的局部最优解。其中 $N_\varepsilon(\boldsymbol{x}^*) = \{\boldsymbol{x} | \|\boldsymbol{x} - \boldsymbol{x}^*\| < \varepsilon, \varepsilon > 0\}$。

最优解与局部最优解

例 1

- $f_0(x) = \frac{1}{x}, \mathbf{dom}\, f_0 = R^{++} : p^* = 0$，无最优解
- $f_0(x) = -\log x, \mathbf{dom}\, f_0 = R^{++} : p^* = -\infty$，无下界
- $f_0(x) = x\log x, \mathbf{dom}\, f_0 = R^{++} : p^* = -\frac{1}{e}, x = \frac{1}{e}$ 是最优解
- $f_0(x) = x^3 - 3x : p^* = -\infty, x = 1$ 是局部最优解

严格最优与非严格最优

定义 4

严格最优解： 当 $\boldsymbol{x} \neq \boldsymbol{x}^*$，有 $f_0(\boldsymbol{x}^*) < f_0(\boldsymbol{x})$ 则称 $\boldsymbol{x}^*$ 为优化问题(1)的严格最优解。

非严格最优解： 若一个点是局部最优解，但不是严格最优解，则称之为非严格最优解。

最优解的相关概念图解：

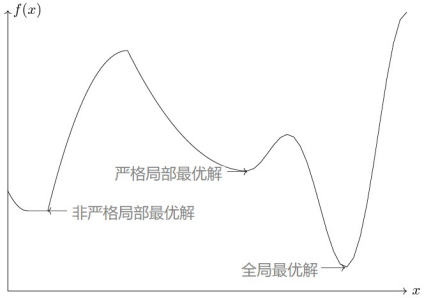

图 1: 函数的全局最优、严格最优和非严格最优解

- 在对优化问题的求解中，我们的目标是求得全局最优解。
- 由于实际问题的复杂性，通常只能求得局部最优解。

28.1.3 最优化算法研究的内容组成

最优化算法研究一般包括三个主要部分：

- 构造最优化模型
- 确定最优化问题的类型和设计算法
- 实现算法或调用优化算法软件包进行求解

优化算法

- 根据优化问题的不同形式，其求解的困难程度可能会有很大差别.
- 对于一个比较简单优化问题，如果我们能用代数表达式给出其最优解，那么这个解称为显式解。
- 然而，对于实际问题往往较为复杂，是没有办法求显式求解的，因此常采用迭代算法。
- 主要思想是寻找一个点列，通过迭代，不断地逼近精确解。我们将在第 12 章详细讨论。

优化问题的分类

优化问题种类繁多，可以按照变量、目标函数、约束函数及其解的性质将其分类。

- 按照变量的性质分类，可以分为连续和离散优化问题。
- 按照约束函数是否存在分类，可以分为无约束和约束优化问题。
- 按照目标函数的性质分类，可以分为凸和非凸优化问题。
- 按照目标函数和约束函数的形式分类，可以分为随机和确定性优化问题，也可以分为线性和非线性规划问题。

28.2.1 连续和离散优化问题

离散优化（Discrete Optimization）问题是指决策变量能够在离散集合上取值，比如离散点集或整数集等。离散优化问题主要有三个分支：

1. **整数规划**（Integer Programming）：输入变量 $\boldsymbol{x} \in \mathbb{Z}^d$ 为整数向量。常见的整数规划问题通常为整数线性规划（Integer Linear Programming，ILP）。
2. **混合整数规划**（Mixed Integer Programming, MIP），即自变量既包含整数也有连续变量。
3. **组合优化**（Combinatorial Optimization）：其目标是从一个有限集合中找出使得目标函数最优的元素。很多机器学习问题都是组合优化问题，比如特征选择、聚类问题、超参数优化问题以及结构化学习（Structured Learning）中标签预测问题等。

连续和离散优化问题

连续优化（Continuous Optimization）问题是指决策变量所在的可行集合是连续的。

- 在连续优化问题中，基于决策变量取值空间以及约束和目标函数的连续性，可根据某点领域内的取值信息来判断该点是否最优。
- 离散优化问题不具备该性质。因此通常将离散优化问题转化为一系列连续优化问题来求解。
- 连续优化问题的求解在最优化理论与算法中处于重要地位。一般认为，在深度学习或机器学习中，模型中要学习的参数是连续变量。因此本课程后续内容也将主要围绕讲解连续优化问题展开。

28.2.3 无约束和约束优化问题

根据是否有变量的约束条件，可以将优化问题分为无约束优化问题和约束优化问题。

1. **无约束优化问题**（Unconstrained Optimization）的决策变量没有约束条件限制，即可行域为整个向量空间 $D = \mathbb{R}^d$。在优化问题(1)中，当我们把不等式约束 $f_i(\boldsymbol{x}) \leq 0$ 和等式约束 $h_j(\boldsymbol{x}) = 0$ 去掉时，即退化为无约束优化问题。
2. **约束优化问题**（Constrained Optimization）是指带有约束条件的问题，即变量 $\boldsymbol{x}$ 需要满足一些等式或不等式的约束。在优化问题(1)中，当不等式约束 $f_i(x) \leq 0$ 和等式约束 $h_j(x) = 0$ 只要有一个成立，其即被称为约束优化问题。

28.2.4 随机和确定性优化问题

根据是目标或约束函数中是否涉及随机变量，可以将优化问题分为随机优化问题和确定性优化问题。

1. **随机优化问题**（Stochastic Optimization）是指目标或约束函数中涉及随机变量而带有不确定性的问题。在实际问题中，只能知道参数的某些估计。随机优化在机器学习、深度学习和强化学习中有着重要应用。
2. **确定性优化问题**（Deterministic Optimization）是指目标和约束函数都是确定的优化问题。

许多确定性优化算法都有相应的随机版本，使得在特定问题上具有更低的计算复杂度和更好的收敛性质。在实际问题中，常利用经验分布代替真实分布，或者是将随机变量的优化转化为确定性参数的优化等方式将随机优化转化为确定性优化。因

28.2.5 线性和非线性规划问题

根据函数的线性性质，可以将优化问题分为线性规划（线性优化）和非线性规划（非线性优化）。

1. 在优化问题(1)中，当目标函数和所有的约束函数都为线性函数，则该问题为**线性规划问题**（Linear Programming）。线性规划问题在约束优化问题中具有较为简单的形式，目前求解线性规划问题最流行的两类方法为单纯形法和内点法。
2. 在优化问题(1)中，如果目标函数或任何一个约束函数为非线性函数，则该问题为**非线性规划**问题（Nonlinear Programming）。本课程将要介绍的优化问题主要为非线性优化问题。

28.2.6 凸和非凸优化问题

更进一步，根据目标函数和可行域的凸性，我们还可以把优化问题分为凸优化（Convex Programming）和非凸优化。

1. **凸优化问题**是一种特殊的约束优化问题，需满足目标函数为凸函数，并且等式约束函数为线性函数，不等式约束函数为凸函数。
2. **非凸优化问题**对应于标准形式(1)中的一个或多个目标函数或约束函数不具有凸性的问题。

在凸优化问题中，任意局部最优解都是全局最优解，因此算法设计和理论分析上比非凸优化问题简单很多。

28.2.7 简单和复合优化问题

由于机器学习实际问题的驱动，复合优化问题的求解算法在近年来得到了大量研究。它是根据目标函数进行划分的。

1. **复合优化问题**是一种特殊的无约束优化问题。其目标函数可以分解为两部分函数求和。其中一部分为光滑函数（比如数据拟合项），另一部分可能是非光滑的部分（比如 ℓ_1 正则项、示性函数）。
2. **简单优化问题**是一个相对于复合优化的概念。这里我们更侧重于是指一般目标函数为光滑函数的优化问题。

我们将在最优性理论和优化算法部分，着重讨论复合优化问题。包括近年来被大量研究的近似梯度法和交替方向法。

28.2.8 参数和超参数优化

在机器学习中，优化问题又可以分为参数优化和超参数优化。

- 模型 $f(\boldsymbol{x};\boldsymbol{\theta})$ 中的 $\boldsymbol{\theta}$ 称为模型的参数，可以通过优化算法进行学习，除了可学习的参数 $\boldsymbol{\theta}$ 之外，还有一类参数是用来定义模型结构或优化策略的，这类参数叫做超参数（Hyper-Parameter）。
- 常见的超参数包括：聚类算法中的类别个数、梯度下降法的步长、正则项的系数、神经网络的层数、支持向量机中的核函数等。
- 超参数的选取一般都是组合优化问题，很难通过优化算法来自动学习。
- 因此，超参数优化是机器学习的一个经验性很强的技术，通常通过搜索的方法对一组超参数组合进行不断试错调整。

28.2.9 其他分类方式

除了上述分类，还有按目标函数的个数分类：单目标最优化问题，多目标最优化问题；以及按约束条件和目标函数是否是时间的函数分类：静态最优化问题和动态最优化问题（动态规划）。

接下来我们将通过一些数据科学与机器学习中常见的应用场景和实例来加深对不同类型优化问题的理解：

- 所熟知的最小二乘相关优化问题
- 在自然语言处理下的优化问题
- 在推荐系统应用场景中的优化问题
- 目标分类或预测任务中的回归模型和支持向量机模型
- 无监督学习的相关模型

28.3.1 最小二乘问题相关优化问题

- 最小二乘问题

$$\min_{\boldsymbol{x}} \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2 \tag{2}$$

- 加权最小二乘

$$\min_{\boldsymbol{x}} \|\boldsymbol{A}_w\boldsymbol{x}-\boldsymbol{y}_w\|_2^2$$

- 约束最小二乘

$$\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2$$
$$\text{s.t.} \quad \boldsymbol{B}\boldsymbol{x}=\boldsymbol{f}$$

- 总体最小二乘

$$\min_{\Delta\boldsymbol{A},\Delta\boldsymbol{b},\boldsymbol{x}} \quad \|\Delta\boldsymbol{A}\|_{\mathrm{F}}^2+\|\Delta\boldsymbol{b}\|_2^2$$
$$\text{s.t.} \quad (\boldsymbol{A}+\Delta\boldsymbol{A})\boldsymbol{x}=\boldsymbol{b}+\Delta\boldsymbol{b}$$

28.3.2 自然语言处理下的优化问题

- 词向量模型:

$$\max_{\boldsymbol{w},b} \prod_{i=1}^{m} (h_{\boldsymbol{w},b}(\boldsymbol{x}_i)^{y_i} * (1 - h_{\boldsymbol{w},b}(\boldsymbol{x}_i)^{1-y_i}))$$

或

$$\min_{\boldsymbol{w},b} -\sum_{i=1}^{m} (y_i \log h_{\boldsymbol{w},b}(\boldsymbol{x}_i)) + (1 - y_i) \log(1 - h_{\boldsymbol{w},b}(\boldsymbol{x}_i))$$

- 连续词袋模型

$$\min_{\boldsymbol{u},\boldsymbol{v}} -\boldsymbol{u}_c^T \hat{\boldsymbol{v}} + \log \sum_{j=1}^{|V|} \exp\left(\boldsymbol{u}_j^T \hat{\boldsymbol{v}}\right)$$

- 跳格模型

$$\min_{\boldsymbol{u},\boldsymbol{v}} -\sum_{i=0, i\neq m}^{2m} \boldsymbol{u}_{c-m+j}^T \boldsymbol{v}_c + 2m \log \sum_{k=1}^{|V|} \exp\left(\boldsymbol{u}_k^T \boldsymbol{v}_c\right)$$

28.3.3 推荐系统中的优化问题

- 推荐系统的优化问题可以转为如下低秩矩阵恢复的优化问题

$$\begin{aligned} \min_{\boldsymbol{X}} \quad & \operatorname{rank}(\boldsymbol{X}) \\ s.t. \quad & \boldsymbol{X}_{ij} = \boldsymbol{M}_{ij} \quad \forall i,j \in \mathbb{E} \end{aligned}$$

或者转化为限定在秩为 r 的条件下，求矩阵使得观测到的评分与预测的评分最接近：

$$\begin{aligned} \min_{\boldsymbol{X}} \quad & \sum_{ij} (\boldsymbol{X}_{ij} - \boldsymbol{M}_{ij})^2 \quad \forall i,j \in \mathbb{E} \\ s.t. \quad & \operatorname{rank}(\boldsymbol{X}) = r \end{aligned}$$

28.3.4 低秩矩阵相关优化问题

- 鲁棒 PCA

$$\min_{\boldsymbol{A},\boldsymbol{E}} \|\boldsymbol{A}\|_* + \lambda\|\boldsymbol{E}\|_1 \quad s.t.\ \boldsymbol{X} = \boldsymbol{A} + \boldsymbol{E}$$

- 低秩矩阵补全

$$\min_{\boldsymbol{A}} \|\boldsymbol{A}\|_* \quad s.t.\ P_\Omega(\boldsymbol{A}) = P_\Omega(\boldsymbol{D})$$

- 低秩矩阵表示

$$\min_{\boldsymbol{Z}} \|\boldsymbol{Z}\|_* \quad s.t.\ \boldsymbol{D} = \boldsymbol{B}\boldsymbol{Z}$$

以及

$$\min_{\boldsymbol{Z},\boldsymbol{E}} \|\boldsymbol{Z}\|_* + \lambda\|\boldsymbol{E}\|_{2,1} \quad s.t.\ \boldsymbol{D} = \boldsymbol{D}\boldsymbol{Z} + \boldsymbol{E}$$

28.3.5 目标分类或预测中的优化问题

- 线性回归模型

$$\min_{\boldsymbol{w},b} \sum_{i=1}^{m} (\boldsymbol{w}^T \boldsymbol{x}_i + b - y_i)^2$$

- 逻辑回归：

$$\min_{\boldsymbol{w}} \sum_{i=1}^{N} [y_i(\boldsymbol{w}^T \boldsymbol{x}_i) - \log(1 + \exp(\boldsymbol{w}^T \boldsymbol{x}_i))]$$

- 感知机

$$\min_{\boldsymbol{w},b} - \sum_{\boldsymbol{x}_i \in M} y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b)$$

目标分类或预测中的优化问题

- 支持向量机

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N} L_{0/1}(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1),$$

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N} \max(0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b))$$

$$\begin{aligned} \min_{\boldsymbol{w},b,\xi} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N} \xi_i \\ s.t. \quad & y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i, i = 1, 2, \ldots, N \\ & \xi_i \geq 0, i = 1, 2, \ldots, N \end{aligned}$$

目标分类或预测中的优化问题

- 非线性支持向量机

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) - \sum_{i=1}^{N}\alpha_i$$

$$s.t. \quad \sum_{i=1}^{N}\alpha_i y_i = 0, 0 \le \alpha_i \le C, i = 1, 2, \ldots, N$$

28.3.6 无监督学习的相关模型

- PCA

$$\min_{\boldsymbol{W}} \mathrm{Tr}(-\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{W})$$
$$s.t.\ \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}$$

- k 均值聚类

$$\min_{C} \sum_{l=1}^{k} \sum_{C(i)=l} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_l\|^2$$

- 谱聚类

$$\min_{\boldsymbol{x}} \boldsymbol{x}^T\boldsymbol{L}\boldsymbol{x}$$
$$\text{s.t. } \boldsymbol{x}^T\mathbf{1} = 0$$

28.3.7 概率模型

- 最大熵模型

$$\min_{P\in C} \; -H(P) = \sum_{x,y} \tilde{P}(x)P(y|x)\log P(y|x)$$

$$s.t.\; E_p(f_i) - E_{\tilde{P}}(f_i) = 0, i = 1, 2, ..., n$$

$$\sum_y P(y|x) = 1$$

- 线性链条件随机场

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu}} \quad \sum_{j=1}^{N}\left(\sum_{i,k}\lambda_k t_k\left(y_{i-1}, y_i, x, i\right) + \sum_{i,l}\mu_l s_l\left(y_i, x, i\right)\right) - \sum_{j=1}^{N}\log Z_{\boldsymbol{\lambda},\boldsymbol{\mu}}\left(x_j\right).$$

概率模型

- 深度信念网络

$$\max \quad \frac{1}{N}\sum_{n=1}^{N}\log p\left(\hat{\boldsymbol{v}}^{(n)};\, \boldsymbol{W}^{(1)},\cdots,\boldsymbol{W}^{(L)},\boldsymbol{a}^{(1)},\cdots,\boldsymbol{a}^{(L)}\right).$$

- 变分自编码器

$$\max_{\theta,\phi}\mathrm{ELBO}(q,\boldsymbol{x};\theta,\phi)=\max_{\theta,\phi}\mathbb{E}_{\boldsymbol{z}\sim q(z;\phi)}\left[\log\frac{p(\boldsymbol{x}\mid\boldsymbol{z};\theta)p(\boldsymbol{z};\theta)}{q(\boldsymbol{z};\phi)}\right]$$

28.3.8 神经网络模型

- 多层感知机模型（全连接神经网络）：

$$\min_{\boldsymbol{A}_i,\boldsymbol{b}_i,i=1,\cdots,L} \|\sigma\left(\boldsymbol{A}_L(\cdots(\sigma(\boldsymbol{A}_1\boldsymbol{x}+\boldsymbol{b}_1))\cdots)+\boldsymbol{b}_L\right)-\boldsymbol{y}\|,$$

其中 $\boldsymbol{x},\boldsymbol{y}$ 分别表示输入特征和对应的标签，L 代表神经网络的层数，$\boldsymbol{A}_i,\boldsymbol{b}_i$ 分别表示连接参数和偏置项，$\sigma(\cdot)$ 代表激活函数，如 sigmoid、ReLU 函数等。

- 卷积神经网络：

$$\min_{\boldsymbol{K}_i,\boldsymbol{B}_i,i=1,\cdots,L} \|\sigma\left(\boldsymbol{K}_L*(\cdots(\sigma(\boldsymbol{K}_1*\boldsymbol{X}+\boldsymbol{B}_1))\cdots)+\boldsymbol{B}_L\right)-\boldsymbol{y}\|,$$

其中 $\boldsymbol{K}_i,\boldsymbol{B}_i$ 分别表示卷积和和对应的偏置项，这里我们用 $*$ 表示卷积运算，其他变量及符号同上。注意，在数学上，卷积神经网络可以看作是全连接网络的连接剪枝和参数共享的网络。另外，在实际计算中，可能还需要添加一些池化层或全连接层。

28.3.9 强化学习模型

强化学习即为求解如下**优化问题**：

$$\max_{\pi} E_{\tau\sim\pi}[R(\tau)].$$

其中 $R(\tau)$ 为累计奖励（带折扣因子 γ）可表示为：

$$R(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t.$$

这里 τ 为智能体与环境交互产生的一条轨迹。

本讲小结

优化问题的一般概念

- 一般形式
- 重要概念：全局最优、局部最优、（非）严格最优
- 优化问题分类：连续与离散、线性与非线性、凸与非凸 ···

数据科学与机器学习中常见的优化问题

- 自然语言处理：词向量模型、词袋模型
- 推荐系统：低雉矩阵恢复
- 目标分类或预测：logistic 回归、支持向量机和感知机模型
- ···

在这些优化问题中，我们将主要讨论凸优化的问题。因此，接下来我们将了解凸优化相关的基础知识。