机器学习中的优化算法

Lecture01: 优化模型

张立平

清华大学数学科学系

办公室：理科楼#A302, Tel: 62798531

E-mail: lipingzhang@tsinghua.edu.cn

Contents and Acknowledgement

- 教材：最优化：建模、算法与理论

  http://bicmr.pku.edu.cn/ wenzw/bigdata2021.html

- 致谢：北京大学文再文教授

Outline of Lecture01

- 最优化问题
- 稀疏优化
- LASSO问题
- 线性回归与逻辑回归
- 机器学习中的典型问题
- 低秩矩阵恢复

最优化问题

The class of optimization problems considered in this course can all be expressed in the form

$$\textbf{(P)} \quad \begin{array}{ll} \textbf{minimize} & f(x) \\ \textbf{subject to} & x \in \mathcal{X} \end{array} \qquad \textbf{or} \qquad \begin{cases} \min & f(x) \\ \textbf{s.t.} & x \in \mathcal{X} \end{cases}$$

where $\mathcal{X}$ is usually specified by some **约束函数** $c_i(x) : \mathbb{R}^n \to \mathbb{R}$

$$c_i(x) = 0 \quad i \in \mathcal{E}, \quad c_i(x) \le 0 \quad i \in \mathcal{I}.$$

- **决策变量**: $x = (x_1, x_2, \cdots, x_n)^{\mathrm{T}} \in \mathbb{R}^n$; **目标函数**: f: $\mathbb{R}^n \to \mathbb{R}$
- **可行域**: $\mathcal{X} \subseteq \mathbb{R}^n$; **可行解**: $x \in \mathcal{X}$
- 当$\mathcal{X} = \mathbb{R}^n$时, 问题(P)称为**无约束优化**

全局和局部最优解

Definition 1. 对于可行点$\bar{x}$ (即$\bar{x} \in \mathcal{X}$),

(1) 如果$f(\bar{x}) \leq f(x), \forall x \in \mathcal{X}$，那么称$\bar{x}$ 为问题(P)的**全局极小解(点)**，也称为(全局)最优解或最小值点；

(2) 如果存在$\bar{x}$ 的一个ε 邻域$N_\varepsilon(\bar{x})$ 使得

$$f(\bar{x}) \leq f(x), \forall x \in N_\varepsilon(\bar{x}) \cap \mathcal{X},$$

那么称$\bar{x}$ 为问题(P)的**局部极小解(点)**，也称为局部最优解；

(3) 进一步地，如果有$f(\bar{x}) < f(x), \forall x \in N_\varepsilon(\bar{x}) \cap \mathcal{X}$, 且$x \neq \bar{x}$ 成立，则称$\bar{x}$ 为问题(P)的**严格局部极小解(点)**, 也称为严格局部最优解.

优化算法

由于实际问题往往没有办法显式求解，因此常采用迭代算法.

- 迭代算法的基本思想：从一个初始点x^0出发，按照某种给定的规则进行迭代，得到一个序列$\{x^k\}$. 如果迭代在有限步内终止，那么最后一个点就是优化问题的解. 如果迭代点列是无穷集合，那么希望该序列的极限点(或者聚点)为优化问题的解.

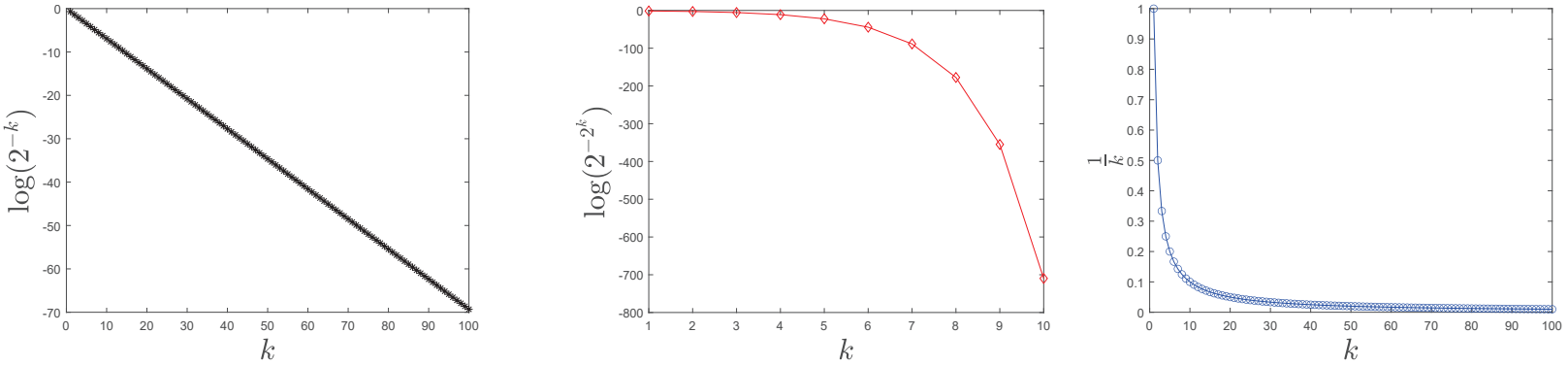
- 算法收敛性：在算法设计中，需考虑算法产生的点列$\{x^k\}$是否收敛到优化问题的解. 如果$\{x^k\}$ 在某种范数$\|\cdot\|$的意义下满足$\lim_{k\to\infty}\|x^k - x^*\| = 0$,且$x^*$ 为一个局部(全局)极小解，那么我们称该点列收敛到局部(全局)极小解，相应的算法称为是**依点列收敛到局部（(全局)极小解**的. 如果从任意初始点x^0 出发，算法都是依点列收敛到局部(全局)极小解的，则称该算法是**全局收敛**的.

算法的渐进收敛速度：Q-收敛速度

设$\{x^k\}$为算法产生的迭代点列且收敛于x^*

- 算法**Q-线性收敛**: 对充分大的k有$\frac{\|x^{k+1}-x^*\|}{\|x^k-x^*\|} \leq a, \quad a \in (0,1)$
- 算法**Q-超线性收敛**: $\lim\limits_{k\to\infty} \frac{\|x^{k+1}-x^*\|}{\|x^k-x^*\|} = 0.$
- 算法**Q-次线性收敛**: $\lim\limits_{k\to\infty} \frac{\|x^{k+1}-x^*\|}{\|x^k-x^*\|} = 1.$
- 算法**Q-二次收敛**: 对充分大的k有$\frac{\|x^{k+1}-x^*\|}{\|x^k-x^*\|^2} \leq a, \quad a > 0.$

- 点列$\{2^{-k}\}$是Q-线性收敛的，点列$\{\frac{1}{k}\}$是Q-次线性收敛的.
- 点列$\{2^{-2^k}\}$是Q-二次收敛的,也是Q-超线性收敛的.



(a) Q-线性收敛 (b) Q-二次收敛 (c) Q-次线性收敛

Figure 1: Q-收敛速度

算法的渐进收敛速度: R-收敛速度

- 算法**R-线性收敛**: 设$\{x^k\}$为算法产生的迭代点且收敛于x^*，若存在Q-线性收敛于0的非负序列t_k使得$\|x^k - x^*\| \leq t_k \quad \forall k.$ 类似地，可定义**R-超线性收敛**和**R-二次收敛**等收敛速度. 当知道t_k的形式时，我们也称算法的收敛速度为$\mathcal{O}(t_k)$.
- **算法复杂度**. 设x^*为全局极小点，某一算法产生的迭代序列$\{x^k\}$满足$f(x^k) - f(x^*) \leq \frac{c}{\sqrt{k}}, \quad \forall\, k > 0,$ 其中$c > 0$为常数. 如果需要计算算法满足精度$f(x^k) - f(x^*) \leq \varepsilon$所需的迭代次数，只需令$\frac{c}{\sqrt{k}} \leq \varepsilon$ 则得到$k \geq \frac{c^2}{\varepsilon^2}$，因此该优化算法对应的(迭代次数)复杂度为$N(\varepsilon) = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$.

优化算法的收敛准则

- 对于无约束优化问题，常用的收敛准则有

$$\frac{f(x^k) - f^*}{\max\{|f^*|, 1\}} \leq \varepsilon_1, \quad \|\nabla f(x^k)\| \leq \varepsilon_2, \tag{1}$$

其中ε_1, ε_2 为给定的很小的正数.

- 对于约束优化问题，还需要考虑约束违反度．也即要求最后得到的点满足

$$c_i(x^k) \leq \varepsilon_3, \ i \in \mathcal{I}$$

$$|c_i(x^k)| \leq \varepsilon_4, \ i \in \mathcal{E},$$

其中$\varepsilon_3, \varepsilon_4$ 为很小的正数.

优化算法的停机准则

- 显然(1)不实用，实际中采用的判别准则是点的最优性条件的违反度.
- 常用的停机准则有

$$\frac{\|x^{k+1}-x^k\|}{\max\{\|x^k\|,1\}} \le \varepsilon_5, \quad \frac{|f(x^{k+1})-f(x^k)|}{\max\{|f(x^k)|,1\}} \le \varepsilon_6,$$

- 在算法设计中，这两个条件往往只能反映迭代点列接近收敛，但不能代表收敛到优化问题的最优解.

优化算法的设计技巧

- **泰勒(Taylor)展开：**对于一个非线性的目标或者约束函数，通过其泰勒展开用简单的线性函数或者二次函数来逼近，根据迭代点的更新来重新构造相应的简化问题，多用于DC规划.
- **对偶:**通过求解对偶问题或者同时求解原始问题和对偶问题，可以简化原始问题的求解，从而设计更有效的算法.
- **凸化:**凸优化问题的任何局部最优解都是全局最优解，其相应的算法设计以及理论分析相对非凸优化问题简单很多．对于非凸优化问题，将其转化为一系列凸优化子问题来求解.
- **松弛：**在保留原问题部分性质的条件下，使用简单的项替代目标函数中难以处理的项，进而使得问题更易求解.

优化算法的设计技巧

- **拆分：**对于一个复杂的优化问题，可以将变量进行拆分，比如$\min_x\ h(x)+r(x)$, 可以拆分成

$$\min_{x,y} \quad h(x)+r(y), \quad \textbf{s.t.} \quad x=y$$

通过引入更多的变量，则可以得到每个变量的简单问题（较易求解或者解有显式表达式），从而通过交替求解等方式来得到原问题的解.

- **块坐标下降:**对于一个n维空间（n 很大）的优化问题，可以通过逐步求解分量的方式将其转化为多个低维空间中的优化问题. 比如，对于$n=100$，可以先固定第2—100 个分量，来求解x_1；接着固定下标为$1,3$—100 的分量来求解x_2；依次类推.

稀疏优化

- 稀疏优化：问题(P)的最优解x是稀疏向量(x中只有少量非零分量).
- 背景：在信号传输过程中，希望通过接收到长度为m的数字信号精确地重构原始信号．已知$b \in \mathbb{R}^m$和$A \in \mathbb{R}^{m\times n}$且$m \ll n$, 重构向量$x \in \mathbb{R}^n$使得

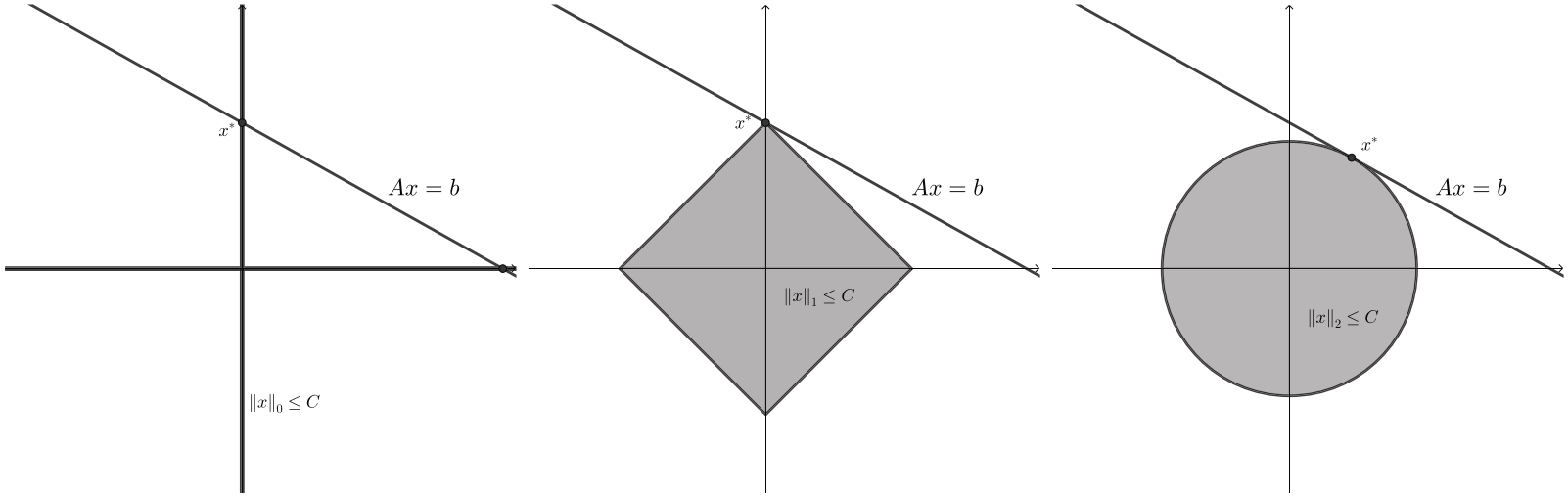
$$Ax = b. \tag{2}$$

- 方程组(2)是欠定的($m \ll n$)，因此存在无穷多个解，重构出原始信号**难!** 原始信号中有较多的零元素, 真正有用的是“稀疏解”u.
- 利用稀疏性这一先验信息和矩阵A以及原问题的解u满足的某些条件，可以通过求解稀疏优化问题把u与方程组(2) 的其他解区别开．这类技术广泛应用于压缩感知(compressive sensing: 通过部分信息恢复全部信息的解决方案).

Sparse Optimization of Compressive Sensing

$$(\ell_0)\begin{cases}\min & \|x\|_0,\\ \text{s.t.} & Ax=b.\end{cases}\quad (\ell_2)\begin{cases}\min & \|x\|_2,\\ \text{s.t.} & Ax=b.\end{cases}\quad (\ell_1)\begin{cases}\min & \|x\|_1,\\ \text{s.t.} & Ax=b.\end{cases}$$

- $\|x\|_0$: x 中非零元素的个数. (ℓ_0)是NP难的，求解起来非常困难.
- (ℓ_1)又称基追踪问题(BP), 是一个线性规划. 理论上可以证明：
  若A, b满足一定的条件，向量u也是(ℓ_1)的唯一最优解.
- (ℓ_2)有唯一最优解$x^* = A^T(AA^T)^+b$, 但不具有稀疏性且u不是(ℓ_2)的解.



(a) ℓ_0范数 (b) ℓ_1范数 (c) ℓ_2范数

- 基追踪问题(ℓ_1)的理论和算法研究在2006年左右带来了革命性的影响．这是一个非光滑优化问题，虽可等价于线性规划，但是数据矩阵A通常是稠密矩阵，甚至A的元素未知或者不能直接存储，只能提供Ax或$A^{\mathrm{T}}y$ 等运算结果．在这些特殊情况下，线性规划经典的单纯形法和内点法通常不太适用于求解大规模的基追踪问题．

LASSO问题

考虑带ℓ_1范数正则项的优化问题

$$\min_{x\in\mathbb{R}^n} \quad \mu\|x\|_1 + \frac{1}{2}\|Ax-b\|_2^2, \tag{3}$$

其中$\mu > 0$是给定的正则化参数.

- 问题(3) 又称为LASSO(Least Absolute Shrinkage and Selection Operator), 可看作(ℓ_0)的二次罚函数形式.

- 注意： 课件关注的大部分数值算法都将针对问题(ℓ_0)或(3) 给出具体形式.

回归分析

- 回归模型将响应变量$b \in \mathbb{R}$与自变量$a \in \mathbb{R}^d$通过函数f联系在一起，形如：

$$b = f(a) + \varepsilon, \tag{4}$$

  $\varepsilon \in \mathbb{R}$是模型误差或噪声.

- 一般只知道a和b的观测值，而误差ε是未知的．建立回归模型的最终任务是利用m 个观测值(a_i, b_i)来求解出f的具体形式，然后可以利用新观测的自变量对响应变量做出预测.

- 函数f取值于函数空间中，一般将其进行参数化，即模型(4) 为

$$b = f(a; x) + \varepsilon,$$

  其中$f(a; x)$的含义是f以$x \in \mathbb{R}^n$为参数，通过选取不同的x得到不同的f．参数化的重要意义在于其将f选取的范围缩小到了有限维空间$\mathbb{R}^n$中，求解f的过程实际上就是求解参数x的过程.

线性回归

- 考虑线性回归：$b = Ax + \varepsilon$.
- 假设ε_i是高斯白噪声，即$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$，则

$$p(b_i \mid a_i; x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(b_i - a_i^{\mathrm{T}} x)^2}{2\sigma^2}\right),$$

对数似然函数为

$$\ell(x) = \ln \prod_{i=1}^{m} p(b_i \mid a_i; x) = -\frac{m}{2}\ln(2\pi) - m\ln\sigma - \sum_{i=1}^{m} \frac{(b_i - a_i^{\mathrm{T}} x)^2}{2\sigma^2}.$$

- 最大似然估计是极大化对数似然函数得到最小二乘问题：

$$\min_{x \in \mathbb{R}^n} \quad \frac{1}{2}\|Ax - b\|_2^2. \tag{5}$$

当误差是高斯白噪声时，最小二乘解就是线性回归的最大似然解.
当ε_i不是高斯白噪声时，二者并不等价.

岭回归(ridge regression)

- 为了平衡模型(5)的拟合性质和解的光滑性，Tikhonov 正则化: 添加ℓ_2范数平方正则项.

- 假设ε_i 是高斯白噪声，则带ℓ_2范数平方正则项的线性回归模型为岭回归:

$$\min_{x\in\mathbb{R}^n} \quad \frac{1}{2}\|Ax-b\|_2^2+\mu\|x\|_2^2.$$

由于正则项的存在，该问题的目标函数是强凸函数，解的性质得到改善.

- 另一种常见的变形是给定参数$\sigma>0$，求解:

$$\min_{x\in\mathbb{R}^n} \quad \frac{1}{2}\|Ax-b\|_2^2, \quad \textbf{s.t.} \quad \|x\|_2\le\sigma. \tag{6}$$

稀疏线性回归：LASSO

- 如果希望解x是稀疏的，可以考虑添加ℓ_1范数为正则项得到LASSO 问题：

$$\min_{x\in\mathbb{R}^n} \quad \frac{1}{2}\|Ax-b\|_2^2+\mu\|x\|_1,$$

或

$$\min_{x\in\mathbb{R}^n} \quad \frac{1}{2}\|Ax-b\|_2^2, \quad \textbf{s.t.} \quad \|x\|_1\le\sigma. \tag{7}$$

其中$\mu>0,\sigma>0$ 为给定常数，x 是待估计的参数．LASSO 问题通过惩罚参数的ℓ_1 范数来控制解的稀疏性.

- 考虑到噪声ε的存在，通常考虑模型(给定$\nu>0$)：

$$\min_{x\in\mathbb{R}^n} \quad \|x\|_1, \quad \textbf{s.t.} \quad \|Ax-b\|_2\le\nu. \tag{8}$$

模型(7)与(8)本质思想都是“在控制误差的条件下使得x的ℓ_1 范数尽量小”，也就是说尽量得到稀疏解.

LASSO变形

- 如果ε不是高斯白噪声，则需要根据具体类型选择损失函数，例：

$$\min_{x\in\mathbb{R}^n} \quad \|Ax-b\|_2+\mu\|x\|_1, \tag{9}$$

$$\min_{x\in\mathbb{R}^n} \quad \|Ax-b\|_1+\mu\|x\|_1. \tag{10}$$

上述两个模型和LASSO问题的差别在于对损失函数选择的范数不同，它们的性能可能很不一样.

- 损失函数项还有很多变化形式，如同时考虑ℓ_2范数和ℓ_1范数的组合，或选择分位数等.

LASSO变形

- 当特征x本身不稀疏但在某种变换下是稀疏的，则需调整正则项

$$\min_{x\in\mathbb{R}^n} \quad \frac{1}{2}\|Ax-b\|_2^2+\mu\|Fx\|_1,$$

假设x在某线性变换$F\in\mathbb{R}^{p\times n}$下是稀疏的．如果要求$x$相邻点之间的变化是稀疏的，取$F$为

$$F=\begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \\ & & & & 1 & -1 \end{pmatrix}.$$

- 实际上$\|Fx\|_1$还可以与$\|x\|_1$结合起来，这表示同时对Fx和x提出稀疏性的要求．例如fused-LASSO:

$$\min_{x\in\mathbb{R}^n} \quad \frac{1}{2}\|Ax-b\|_2^2+\mu_1\|x\|_1+\mu_2\sum_{i=2}^{n}|x_i-x_{i-1}|,$$

其中$\mu_2\sum_{i=2}^{n}|x_i-x_{i-1}|$用来控制相邻系数之间的平稳度．

逻辑回归: logistic regression

- 对于二分类问题，预测变量只有两个取值，即$-1, 1$.
- 给定特征a，逻辑回归假设这个样本属于类别1的概率

$$p(1|a; x) = P(t = 1 \mid a; x) = \theta(a^{\mathrm{T}} x),$$

其中$\theta(\cdot)$为Sigmoid 函数：

$$\theta(z) = \frac{1}{1 + \exp(-z)}, \tag{11}$$

那么属于类别-1的概率

$$p(-1|a; x) = 1 - p(1 \mid a; x) = \theta(-a^{\mathrm{T}} x).$$

因此对于$b \in \{-1, 1\}$，有$p(b \mid a; x) = \theta(b \cdot a^{\mathrm{T}} x)$.

逻辑回归

- 假设数据对$\{a_i, b_i\}, i = 1, 2, \cdots, m$之间独立同分布，则在给定$a_1, a_2, \cdots, a_m$情况下，$b_1, b_2, \cdots, b_m$的联合概率密度是

$$\begin{aligned} p(b_1, b_2, \cdots, b_m \mid a_1, a_2, \cdots, a_m; x) &= \prod_{i=1}^{m} p(b_i \mid a_i; x) \\ &= \frac{1}{\prod_{i=1}^{m}(1 + \exp(-b_i \cdot a_i^{\mathrm{T}} x))}. \end{aligned}$$

- 最大似然估计是求解如下最优化问题：

$$\min_{x \in \mathbb{R}^n} \quad \sum_{i=1}^{m} \ln(1 + \exp(-b_i \cdot a_i^{\mathrm{T}} x)). \tag{12}$$

- 加上正则项，如Tikhonov和ℓ_1范数正则化模型：

$$\min_{x\in\mathbb{R}^n} \quad \sum_{i=1}^{m} \ln(1+\exp(-b_i\cdot a_i^{\mathrm{T}}x)) + \lambda\|x\|_2^2, \tag{13}$$

$$\min_{x\in\mathbb{R}^n} \quad \sum_{i=1}^{m} \ln(1+\exp(-b_i\cdot a_i^{\mathrm{T}}x)) + \lambda\|x\|_1. \tag{14}$$

机器学习中典型问题

很多机器学习中的问题可以写为：

$$\min \quad \sum_{i=1}^{N} \tfrac{1}{2}\|a_i^\top x - b_i\|_2^2 + \mu\varphi(x) \qquad \text{线性回归}$$

$$\min \quad \tfrac{1}{N}\sum_{i=1}^{N} \log(1+\exp(-b_i a_i^\top x)) + \mu\varphi(x) \qquad \text{逻辑回归}$$

$$\min \quad \tfrac{1}{N}\sum_{i=1}^{N} \ell(f(a_i,x),b_i) + \mu\varphi(x) \qquad \text{一般形式}$$

- (a_i, b_i) 是给定的数据对，b_i 是数据a_i对应的标签
- $\ell(\cdot)$: 损失函数，度量模型拟合数据点i的程度(避免拟合不足)
- $\varphi(x)$: 避免过拟合的正则项，$\|x\|_2^2$ 或$\|x\|_1$等，与噪声类型有关
- $f(a,x)$: 线性函数或者由深度神经网络构造的模型

低秩矩阵恢复问题的背景

- 某视频网站提供了约48万用户对1万7千多部电影的上亿条评级数据，希望对用户的电影评级进行预测，从而改进用户电影推荐系统，为每个用户更有针对性地推荐影片.
- 显然每一个用户不可能看过所有的电影，每一部电影也不可能收集到全部用户的评级. 由于用户只对看过的电影给出自己的评价，矩阵M中很多元素是未知的.

	电影1	电影2	电影3	$\cdots$	电影n
用户1	4	?	?	$\cdots$	?
用户2	?	2	4	$\cdots$	?
用户3	3	?	?	$\cdots$	?
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
用户m	?	3	?	$\cdots$	?

低秩矩阵恢复问题的性质

该问题在推荐系统、图像处理等方面有着广泛的应用.

- 由于用户对电影的偏好可进行分类，按年龄可分为：年轻人，中年人，老年人；电影也能分为不同的题材：战争片，悬疑片，言情片等。故这类问题隐含的假设为补全后的矩阵应为低秩的.
- 由于低秩矩阵可分解为两个低秩矩阵的乘积，所以低秩限制下的矩阵补全问题是比较实用的，这样利于储存且有更好的诠释性。
- 有些用户的打分可能不为自身真实情况，对评分矩阵有影响，所以原矩阵是可能有噪声的.

低秩矩阵恢复

- 令Ω是矩阵M中所有已知元素的下标的集合，构造一个矩阵$X \in \mathbb{R}^{m\times n}$，使得$X_{ij} = M_{ij},\ (i,j) \in \Omega$.
- Low Rank Matrix Completion是指根据部分观察数据恢复全部数据:

$$\begin{aligned} \min \quad & \mathrm{rank}(X), \\ \textbf{s.t.} \quad & X_{ij} = M_{ij},\ (i,j) \in \Omega. \end{aligned} \tag{15}$$

- $\mathrm{rank}(X)$正好是矩阵X所有非零奇异值的个数, 矩阵X的核范数(nuclear norm)为X所有奇异值的和: $\|X\|_* = \sum_i \sigma_i(X)$.

$$\begin{aligned} \min \quad & \|X\|_*, \\ \textbf{s.t.} \quad & X_{ij} = M_{ij},\ (i,j) \in \Omega. \end{aligned} \tag{16}$$

- 可以证明问题(16) 是一个凸优化，在一定条件下与问题(15) 等价.

- 考虑到观测可能出现误差，对于给定的参数$\mu > 0$，给出(16)的二次罚函数形式：

$$\min \quad \mu\|X\|_* + \frac{1}{2}\sum_{(i,j)\in\Omega}(X_{ij} - M_{ij})^2. \tag{17}$$

- 秩r情形: $X = LR^T$,其中$L \in \mathbb{R}^{m\times r}$, $R \in \mathbb{R}^{n\times r}$, $r \ll \min(m,n)$. 问题(16)可写为

$$\min_{L,R} \sum_{(i,j)\in\Omega}\left(\left[LR^T\right]_{ij} - M_{ij}\right)^2 + \alpha\left\|L\right\|_F^2 + \beta\left\|R\right\|_F^2.$$

α, β为正则化参数，作用是消除解L, R在放缩意义下的不唯一性. 该问题非凸，避免了SVD分解.

矩阵分离问题: 稀疏和低秩矩阵

- 给定矩阵M，我们想找到一个低秩矩阵$W \in \mathbb{R}^{m\times n}$和稀疏矩阵$E \in \mathbb{R}^{m\times n}$，使得

$$W + E = M.$$

- 非凸模型：

$$\begin{aligned} \min_{W,E} \quad & \mathrm{rank}(W) + \mu\|E\|_0, \\ \textbf{s.t.} \quad & W + E = M. \end{aligned}$$

- 凸松弛：

$$\min_{W,E} \ \|W\|_* + \mu\|E\|_1, \quad \textbf{s.t.} \quad W + E = M.$$