

机器学习中的优化算法



Lecture00: 绪论

清华大学数学科学系 张立平

Email: lipingzhang@tsinghua.edu.cn

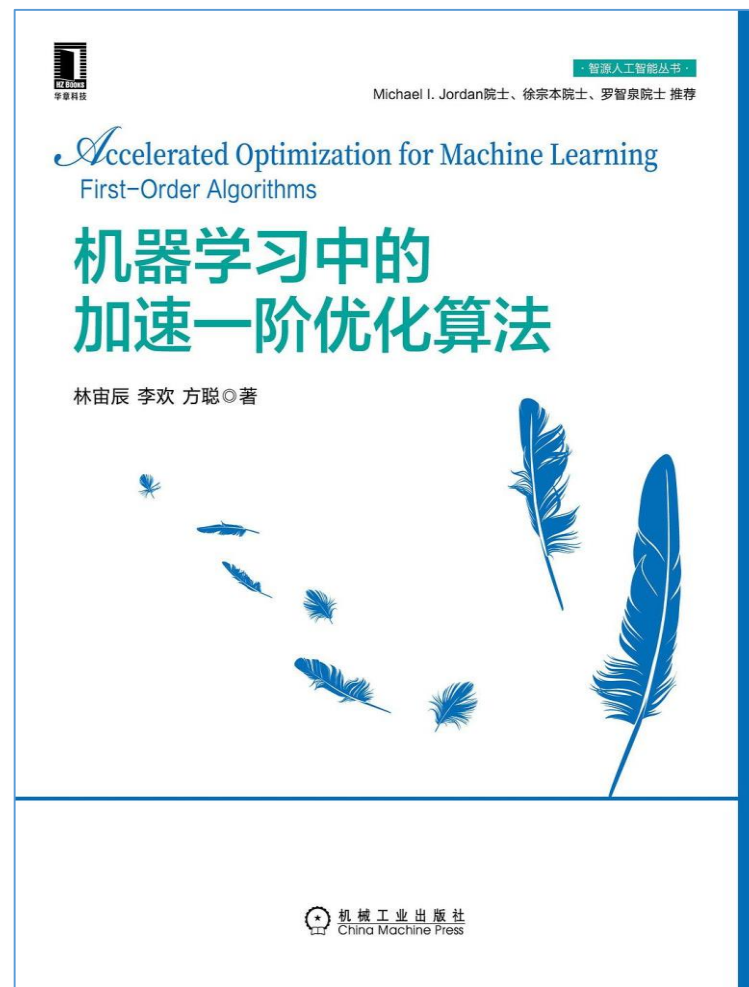
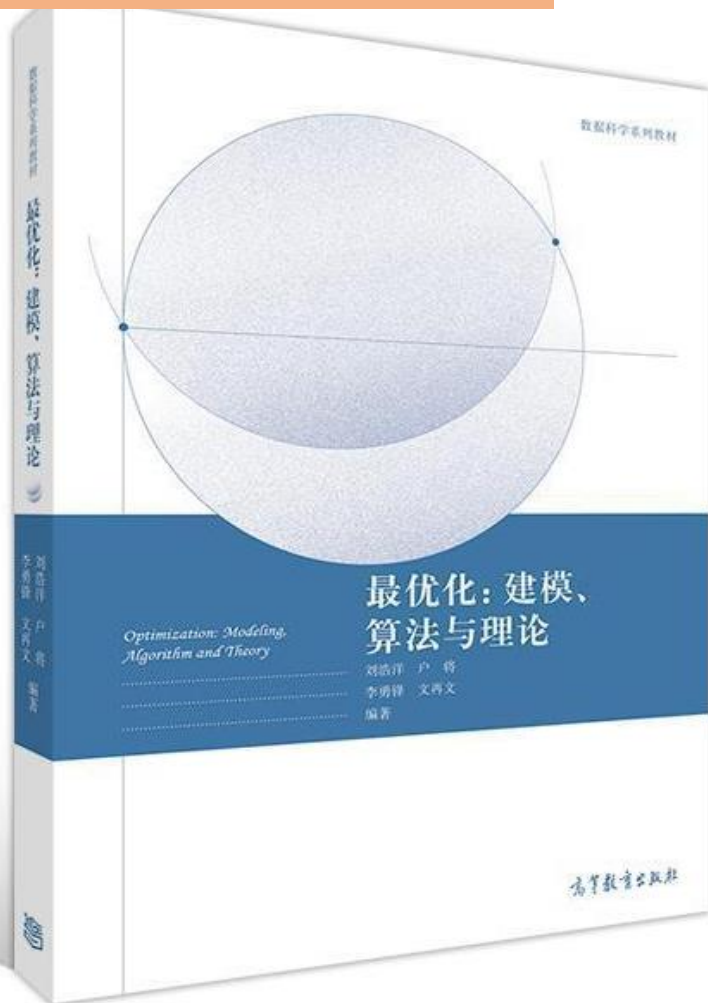
Office: 数学系馆A302

Tel: 62798531, 18811783995

机器学习中的优化算法



MOTIVATION



机器学习简介



机器学习定义

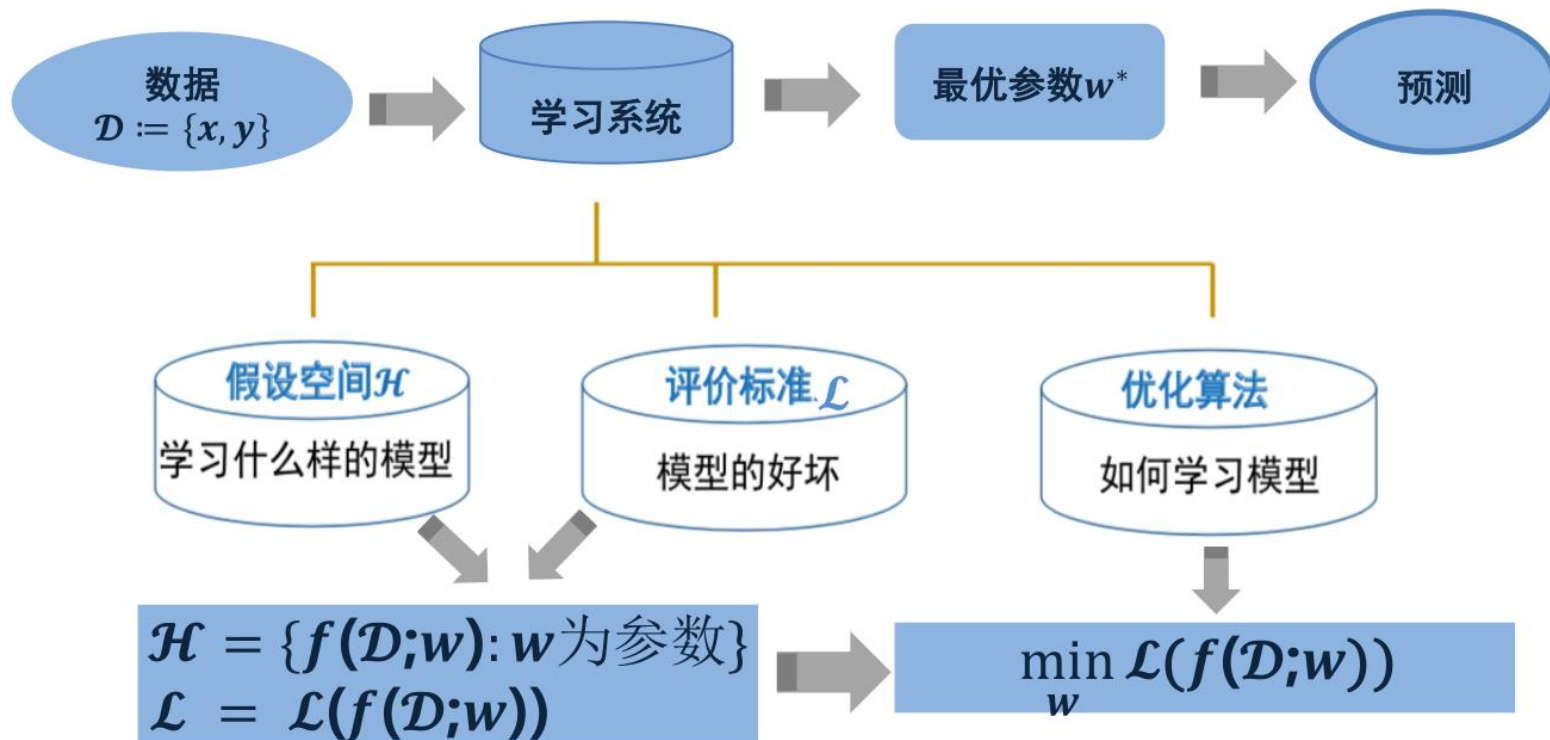
- 机器学习是以数据为研究对象，通过构建数学模型并运用统计与优化等方法，由计算机实现对数据中蕴含内在规律进行挖掘、分析、预测与决策的一门学科。
- 包括：监督学习、无监督学习、强化学习、半监督学习与主动学习等。

功能与应用

- 它具有分类、聚类、降维、推断之功能，在工业、农业、军事、工程、国防等领域有广泛应用。
- 美国三院院士Michael Jordan 和机器学习大师Tom Mitchell认为：机器学习是人工智能的核心。

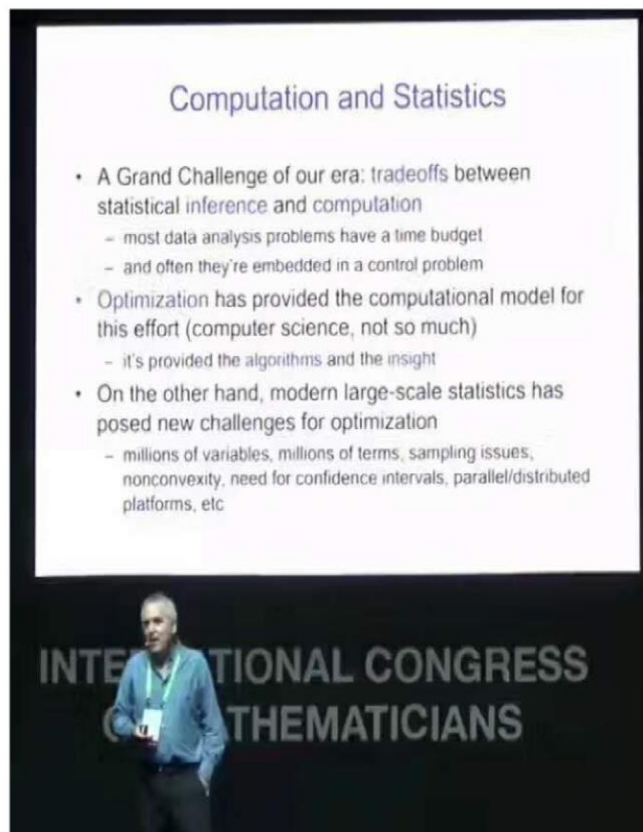
机器学习简介

工作原理：最优化是机器学习的关键技术



美国三院院士Michael Jordan教授国际数学家大会一小时报告

- 我们时代的一个很大挑战是统计推理和计算的平衡。大部分数据分析有时间限制，他们经常被嵌入到某个控制问题里。
- **最优化为这个努力提供了计算模型，给出了算法和深刻的理解。**
- 现代大规模统计给优化带来了新的挑战：百万量级变量/函数项，抽样问题，非凸，置信区间，并行/分布式平台等等



“Statistics and the Oncoming AI Revolution”

- What has made ML so successful? What are the disciplines supporting ML and providing a good basis to understand the challenges, open problems, and limitations of the current techniques?
 - 1) **basic statistical tools**: linear models, generalized linear models, logistic regression, cross validation, overfitting . . .
 - 2) **probability theory and probabilistic modeling**.
- How about engineering disciplines?

Clearly, progress in optimization—particularly in convex optimization—has fueled ML algorithms for the last two decades.

美国科学院院士 Emmanuel Candès



机器学习中的优化问题：0/1损失优化

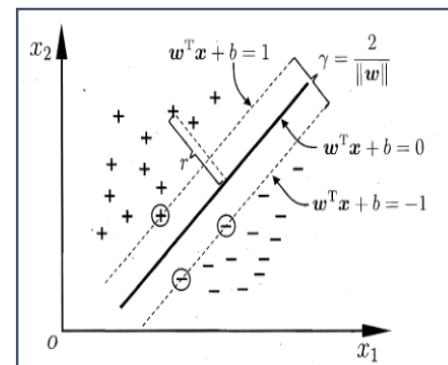


【例】支持向量机(SVM)

硬间隔支持向量机

SVM于1995年首次提出，其基本思想是寻找一个超平面使得它能够尽可能多的将样本点正确分开，同时使分开的样本点距离超平面最远。

$$\min_{w \in R^m, b \in R} \frac{1}{2} \|w\|^2 \quad s.t. \quad y^i (w^T x^i + b) \geq 1.$$

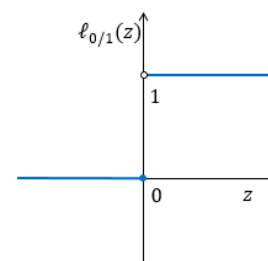
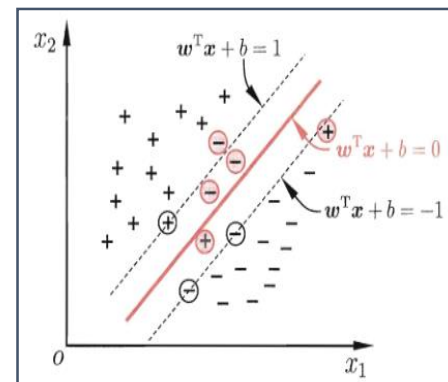


0/1损失优化模型

当训练数据线性不可分时，解决办法是允许存在不满足上述约束的样本，但应尽可能少，于是可得软间隔支持向量机**0/1损失优化**模型：

$$\min_{w \in R^n, b \in R} C \sum_{i=1} l_{0/1}(1 - y^i (w^T x_i + b)) + \frac{1}{2} \|w\|^2$$

其中， $y^i \in \{-1, 1\}$.



机器学习中的优化问题：0/1损失优化



【例】分类学习AUC方法

分类
学习
AUC
方法

AUC(Area Under Curve)是衡量二分类学习器优劣的一种重要性能指标,其基本思想是寻找一个**权重向量**(用 w 表示)使得正样本的预测值尽可能都大于负样本的预测值. 假设正负样本的集合分别是 D 和 H , $|D|$ 和 $|H|$ 分别表示两个集合包含样本的个数,则:

$$\max_{w \in R^n} \frac{1}{|D| \cdot |H|} \sum_{i \in D, j \in H} l_{0/1}(w^T x_i - w^T x_j)$$

0/1
损失
优化
模型

当**权重向量** w 具有**稀疏性**时, 可得AUC的0/1损失优化模型为:

$$\begin{aligned} \min_{w \in R^n} \quad & \frac{1}{|D| \cdot |H|} \sum_{i \in D, j \in H} (1 - l_{0/1}(w^T x_i - w^T x_j)) + \lambda \|w\|_0 \\ \text{s.t.} \quad & \|w\|_2 = 1, w \geq 0. \end{aligned}$$

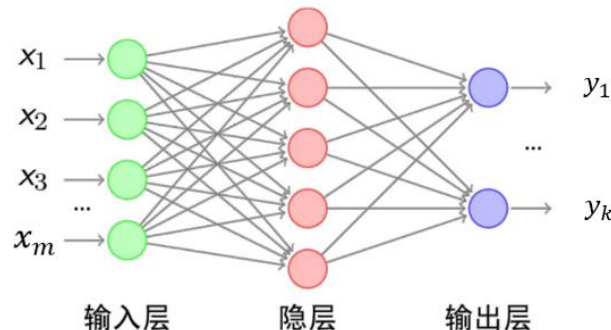
机器学习中的优化问题：0/1损失优化



【例】0/1神经网络-DNN

0/1
神经网络

深度学习首次由G.E. Hinton 等人于 2006 年提出, 它的一个基本模型是0/1-DNN, 即以0/1为激活函数的深度神经网络.



0/1
约束
优化
模型

给定网络结构, 训练集 $T = \{(x^i, y^i)\}_{i=1}^n$, 其中 $x^i \in R^m$ 是输入向量, $y^i \in \{e_1, e_2, \dots, e_k\}$ 是输出标签, 可得0/1损失优化模型为:

$$\begin{aligned} \min_{w, b} \quad & \frac{1}{n} \sum_{i=1}^n \|y^i - \hat{y}^i\|_1 \\ \text{s.t.} \quad & \hat{y}^i = f_{\text{hardmax}}(W^h(\text{Sgn} \dots \text{Sgn}(W^1 x^i + b_1) + \dots + b_h)). \end{aligned}$$

- Goodfellow I., Bengio Y., Courville A., *Deep Learning*, MIT Press, 2016.

■ $L_{0/1}$ 损失函数具有好的统计学性质：稀疏性、无偏性、鲁棒性等；

■ $L_{0/1}$ 损失函数具有可计算函数的三个基本特征：

◆ 次微分：最优性条件分析需要次微分。

◆ 近似点算子：ADMM、APG等算法计算需要近似点算子。

◆ 共轭算子：原始问题推导对偶问题需要函数的共轭。

机器学习中的优化问题：0/1损失优化



优化模型

$$\min_{\mathbf{w} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}} \frac{1}{2} \|\mathbf{w}\|^2 + C \|\mathbf{e} - (\mathbf{A}\mathbf{w} + \mathbf{b}\mathbf{y})\|_0 \quad (1)$$

其中 $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{y} \in \mathbb{R}^m$, $\mathbf{e} = (1, \dots, 1)^T \in \mathbb{R}^m$, $C > 0$ 是已知量.

P-稳定点定义

□ 给定 $C > 0$, 我们称 $(\bar{\mathbf{w}}, \bar{\mathbf{b}}, \bar{\mathbf{u}})$ 是问题(1)的邻近点算子稳定点(**P-稳定点**), 如果存在 $\bar{\boldsymbol{\lambda}} \in \mathbb{R}^m$ 和 $\gamma > 0$ 满足

$$\begin{cases} \bar{\mathbf{w}} + \mathbf{A}^T \bar{\boldsymbol{\lambda}} = \mathbf{0}; & (i) \\ \mathbf{y}^T \bar{\boldsymbol{\lambda}} = 0; & (ii) \\ \bar{\mathbf{u}} \in \text{prox}_{\gamma C \|\cdot\|_0}(\bar{\mathbf{u}} - \gamma \bar{\boldsymbol{\lambda}}); & (iii) \\ \bar{\mathbf{u}} = \mathbf{e} - (\mathbf{A}\bar{\mathbf{w}} + \bar{\mathbf{b}}\mathbf{y}). & (iv) \end{cases}$$

P-稳定点与最优解关系

全局最优解



P-稳定点

+ 矩阵 (\mathbf{A}, \mathbf{y}) 列满秩

局部最优解



P-稳定点

- H.J Wang, Y.H Shao, S.L .Zhou, C. Zhang and N.H. Xiu, Support vector machine classifier via L0/1 soft-margin loss, **IEEE TPAMI**, online, 2021. 北京交通大学修乃华教授团队

机器学习中的优化问题：0/1损失优化



模型
等价
转化

$$\min_{w \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{2} \|w\|^2 + C \| (e - (Aw + by))_+ \|_0$$

$$\begin{aligned} \min_{w \in \mathbb{R}^n, b \in \mathbb{R}, u \in \mathbb{R}^m} \quad & \frac{1}{2} \|w\|^2 + C \|u_+\|_0 \\ \text{s.t.} \quad & u = e - (Aw + by). \end{aligned}$$

$L_{0/1}$ -
ADMM
算法
框架

步1: (支持集选取) $T_k := \{i \in N_m | u_i^{k+1} = 0\}$.

步2: (子空间参数更新) 子空间更新 $u^{k+1}, w^{k+1}, b^{k+1}$.

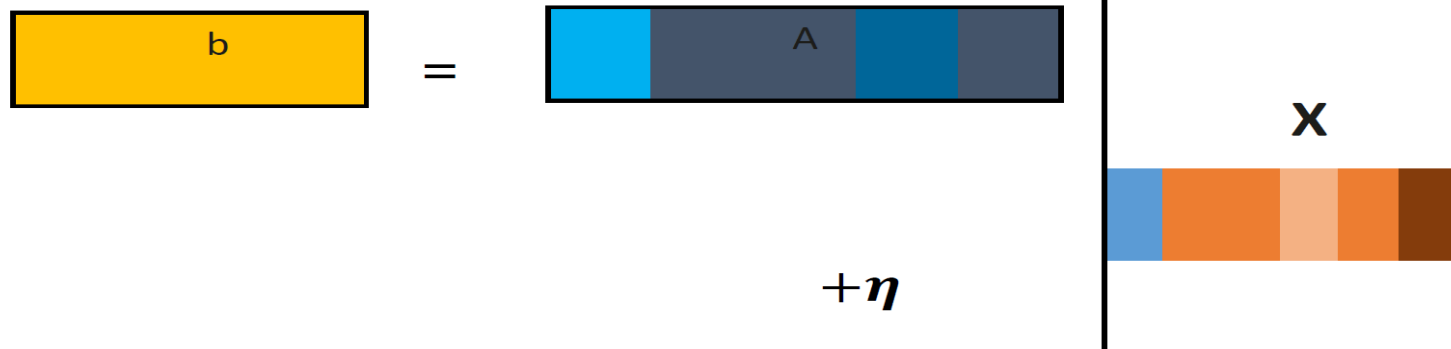
$$\text{特点: } (I + A^T A)^{-1} \quad (I + A_{T_k}^T A_{T_k})^{-1}$$

步3: (乘子更新) $\lambda^{k+1} = \lambda^k + \tau \sigma(u^{k+1} - e + Aw^{k+1} + b^{k+1}y)$.

- H.J Wang, Y.H Shao, S.L .Zhou, C. Zhang and N.H. Xiu, Support vector machine classifier via L0/1 soft-margin loss, **IEEE TPAMI**, online, 2021. 特点: 支持集选取, 降维。

机器学习中的优化问题：稀疏优化

挑战：在工程领域，已知观测信号在某组字典下有稀疏表示。由于噪声的干扰，希望能利用此先验信息去稳定、快速的去噪，且能精确做出预测。


$$b = AX + \eta$$

字典 A 是通过理论模型离散化得到的矩阵，观测矩阵 b 在字典 A 下有稀疏表示，且不同观测下的**非零位置相同，但是系数不同**。信号受到噪音干扰，但是可以估计出噪音的协方差矩阵。需要稳定快速恢复出系数矩阵 X 。

机器学习中的优化问题：稀疏优化



考虑下面的带 ℓ_1 范数正则的优化问题

LASSO问题

$$\min_x \mu \|x\|_1 + \frac{1}{2} \|Ax - b\|^2.$$

- ℓ_1 范数不可微， $f(x)$ 的一个次梯度为 $A^T(Ax - b) + \mu \text{sign}(x)$.

如何定义海瑟矩阵？

- 经典牛顿法的更新格式为：

$$x^{k+1} = x^k - \nabla^2 f(x^k)^{-1} \nabla f(x^k).$$

不可微情形如何定义牛顿法？

机器学习中的优化问题：低秩矩阵优化



- 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。电影评级由用户打分1星到5星表示，记为取值1~5的整数。我们将电影评级放在一个矩阵 M 中，矩阵 M 的每一行表示不同用户，每一列表示不同电影。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	电影4	...	电影n
用户1	4	?	?	3	...	?
用户2	?	2	4	?	...	?
用户3	3	?	?	?	...	?
用户4	2	?	5	?	...	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮
用户m	?	3	?	4	...	?

机器学习中的优化问题：低秩矩阵优化



- 令 Ω 是矩阵 M 中所有已知评级元素的下标的集合，则该问题可以初步描述为构造一个矩阵 X ，使得在给定位置的元素等于已知评级元素，即满足 $X_{ij} = M_{ij}, (i, j) \in \Omega$.
- 低秩矩阵恢复 (low rank matrix completion)

$$\begin{aligned} \min_{X \in \mathbb{R}^{m \times n}} \quad & \text{rank}(X), \\ \text{s.t.} \quad & X_{ij} = M_{ij}, (i, j) \in \Omega. \end{aligned}$$

$\text{rank}(X)$ 正好是矩阵 X 所有非零奇异值的个数

机器学习中的优化问题：低秩矩阵优化



清华大学
Tsinghua University

网络流量恢复

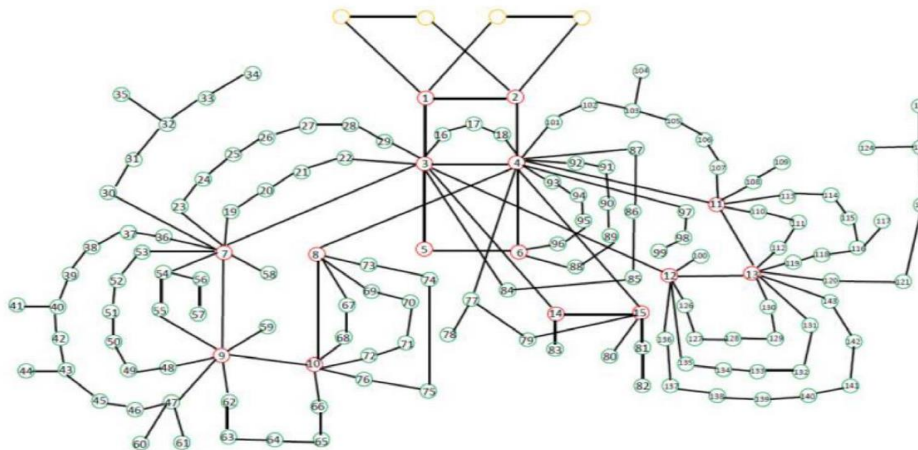


Figure 1: The topology of a network system

- Suppose the IPRAN network comprises S routers (nodes) and M links.
- Sparsity of topology: $M \ll S^2$.
- A node can be regarded as an origin (O) as well as a destination (D) of traffic flows \implies there are S^2 OD pairs.

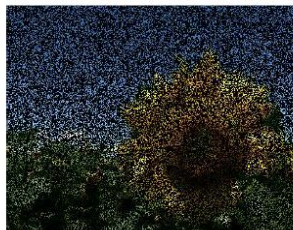
$$L = R \cdot X, \quad L \in \mathbb{R}^{M \times T}, \quad R \in \mathbb{R}^{M \times S^2}, \quad X \in \mathbb{R}^{S^2 \times T}.$$

机器学习中的优化问题：低秩矩阵优化

Original frame



Observation



QRTC



MQRTC



(a) AN119T

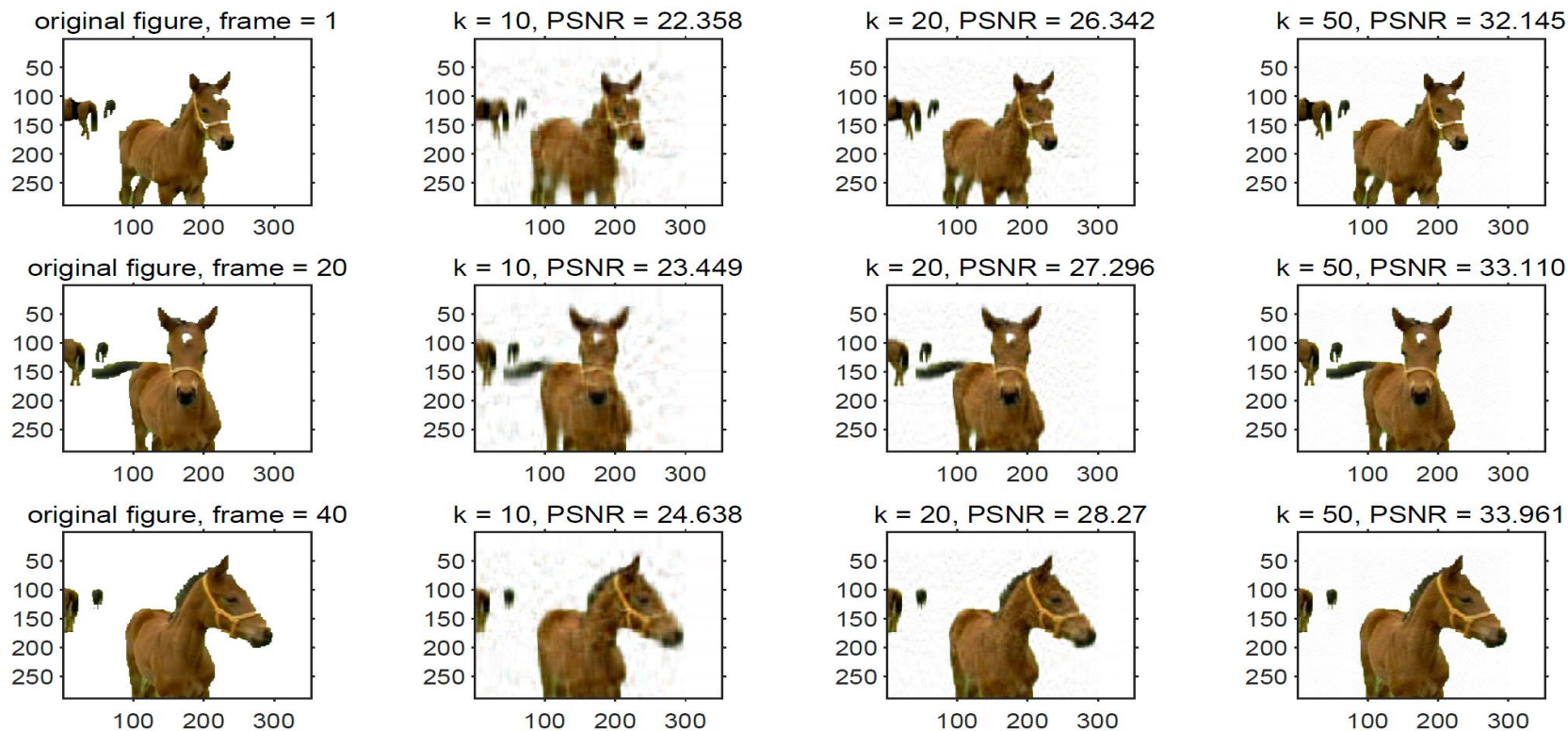
(b) BR128T

(c) DO01-013

(d) DO01-030

(e) M07-058

机器学习中的优化问题：低秩矩阵优化



$$\min_{\mathcal{C} \in \mathbb{H}^{n_1 \times n_2 \times n_3}} \text{rank}_{Q_t}(\mathcal{C}), \text{ s.t. } P_{\Omega}(\mathcal{C} - \mathcal{M}) = 0, \Re(\mathcal{C}) = 0.$$

机器学习中的优化算法：教学内容



第 1 讲 机器学习中的优化模型

- 1.1 最优化问题基本概念
- 1.2 稀疏优化模型
- 1.3 LASSO 问题
- 1.4 线性回归模型
- 1.5 逻辑回归模型
- 1.6 机器学习中的典型问题
- 1.7 低秩矩阵恢复模型

第 2 讲 凸分析

- 2.1 向量范数和矩阵范数
- 2.2 凸集和凸函数
- 2.3 共轭函数
- 2.4 次梯度

第 3 讲 最优性理论

- 3.1 最优化问题解的存在性
- 3.2 不可微无约束优化的最优性理论
- 3.3 对偶理论
- 3.4 凸优化的最优性理论
- 3.5 约束优化的最优性理论
- 3.6 复合优化的最优性理论

第 4 讲 无约束优化算法

- 4.1 引言
- 4.2 次梯度算法
- 4.3 牛顿类算法
- 4.4 拟牛顿类算法
- 4.5 信赖域算法
- 4.6 非线性最小二乘问题算法
- 4.7 应用举例

第 5 讲 约束优化算法

- 5.1 引言
- 5.2 二次罚函数法及应用举例
- 5.3 约束优化的增广拉格朗日函数法
- 5.4 凸优化的增广拉格朗日函数法
- 5.5 基追踪问题的增广拉格朗日函数法
- 5.6 半定规划问题的增广拉格朗日函数法

第 6 讲 复合优化算法

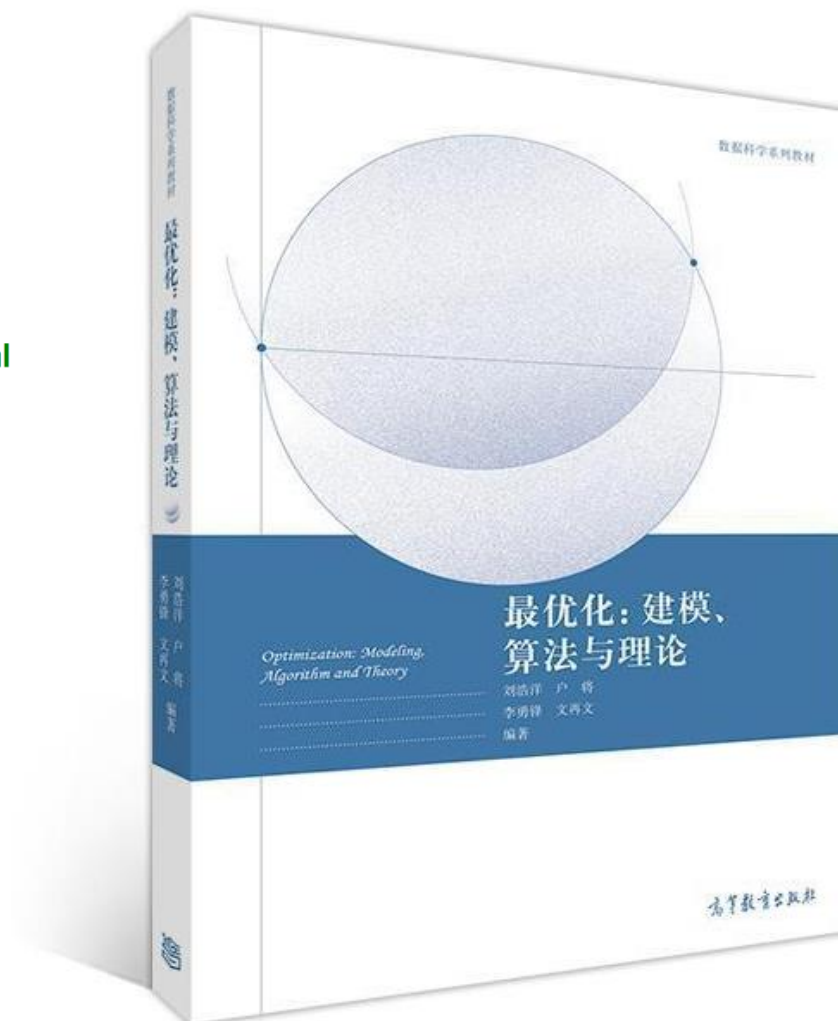
- 6.1 近似点梯度法及应用举例
- 6.2 Nesterov 加速算法及应用举例
- 6.3 近似点算法及应用举例
- 6.4 对偶近似点梯度法及应用举例
- 6.5 交替方向乘子法及应用举例
- 6.6 随机梯度算法及应用举例

机器学习中的优化算法：教材



- 教材：最优化：建模、算法与理论

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



机器学习中的优化算法：助教、考核



➤ 助教

李 妍: 16622723023, li-yan20@mails.tsinghua.edu.cn

➤ 考核：平时作业+期末大作业+课外拓展

总评=4次平时作业60%+期末大作业40%+课外拓展5%

注意：①4次平时作业，在网络学堂布置，按时提交（有5分上课随机点名）
②期末大作业是开放式作业，期末按要求提交
③课外拓展：鼓励按自己的专业阅读相关的论文，提交阅读报告和相应的算法实现结果

➤ 答疑

请联系助教答疑