优化方法 期末论文题目

日期：2021 年 12 月 22 日

本次作业请于 **2022年1月9日前** 提交，需包含以下内容：

- 纸质版论文：需包括摘要、引言、文献综述、正文、结论以及参考文献等部分。
- 补充材料：包括电子版论文（`.pdf` 格式）及相关补充材料（补充数据及**带有注释说明**的源代码等）。请将补充材料打包并命名为**"期末论文-题目1/题目2-学号1-姓名1-学号2-姓名2-……"** 后发送至助教邮箱：meetziyang@outlook.com。

可以以小组形式进行（最多 5 人）；两道题目任选其一即可；可使用任意编程语言，如无特殊要求，可使用现成软件包。

1 拟合广义可加模型

广义可加模型具有如下形式：

$$f(\mathbf{x}) = \alpha + \sum_{j=1}^{p} f_j(x_j) \tag{1}$$

其中，$\mathbf{x} \in \mathbb{R}^p$ 是自变量；$\alpha \in \mathbb{R}$ 是截距项，$f_j : \mathbb{R} \to \mathbb{R}$ 是未定义的（非参）光滑函数且 $f_j(0) = 0$。

为简化问题，我们限制 $f_j(\cdot), j = 1, \ldots, p$ 为分段线性函数，并给定结点 $-\infty < p_1 < \ldots < p_K < \infty$，也就是说函数 $f_j(\cdot)$ 在区间 $(-\infty, p_1], [p_1, p_2], \ldots, [p_K, \infty)$ 上为线性函数并且在结点处连续（如图 1 所示）。对于该广义可加模型 $f(\mathbf{x})$，定义 C 为所有分段线性函数 $f_j(\cdot)$ 在所有结点附近斜率的变化程度。该指标 C 度量所有分段线性函数 $f_j(\cdot)$ 的非线性程度，当 $C = 0$ 时，该广义可加模型退化为线性回归模型。

现给定一组数据 $(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)}) \in \mathbb{R}^p \times \mathbb{R}$，我们希望该广义可加模型能够尽可能地拟合给定数据，即

$$\min_{\alpha, f_j} \quad \sum_{i=1}^{n} \left[y^{(i)} - \alpha - \sum_{j=1}^{p} f_j(\mathbf{x}_j^{(i)}) \right]^2 + \lambda C \tag{2}$$

其中，$\lambda > 0$ 为正则化参数。

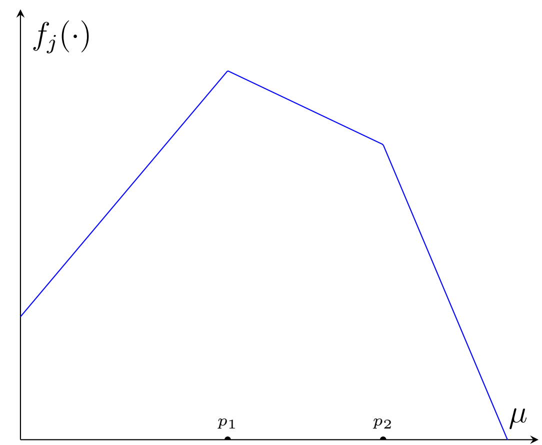

图 1: 分段线性函数 $f_j(\cdot)$ 示意图

1. 考虑如何表达该分段线性函数 $f_j(\cdot)$ 并定义 $f_j(\cdot)$ 的非线性程度，从而可以使用本学期内学过的优化方法求解该问题。
2. 基于第（1）问中提出的方法，选择正则化参数 λ 对 `problem1.py` 中生成的数据进行拟合。该文件包括一个 $n \times p$ 矩阵 $\mathbf{X}$（矩阵中每一行是 $(\mathbf{x}^i)^{\mathrm{T}}$），一个列向量 $\mathbf{y}$ （向量中每一行是 $y^{(i)}$）以及一个向量 $\mathbf{p}$ 为给定的结点。给出求解得出的均方误差，并对比估计得到的函数 $\hat{f}_j(\cdot)$ 与真实的函数 $f_j(\cdot)$。（真实函数以及绘图程序在文件中已给出）
3. 基于实现的方法（需自己确定分段线性函数结点 $\mathbf{p}$ 及正则化参数 λ），选择任意一真实数据[1]进行拟合。
4. 查阅相关资料，考虑其他类型的光滑函数（如样条函数等），重新求解第（1）—（3）问，并与给定的分段线性函数进行对比和讨论。

注. 分段线性函数 $f_j(\cdot)$ 可以表示为一系列基函数的线性组合

$$f_j(\mu) = \sum_{k=0}^{K} \beta_{jk} b_k(\mu) \tag{3}$$

其中，

$$b_0(\mu) = \mu, \quad b_k(\mu) = (\mu - p_k)_+ - (-p_k)_+, \, k = 1, \dots, K, \quad (\mu)_+ = \max(\mu, 0) \tag{4}$$

2 投资组合优化问题

现有上证 50 的 50 支股票在过去一年内的价格记录，利用该数据可计算出第 i 只股票的历史平均收益 μ_i 和上证 50 的相关系数矩阵 $\mathbf{C}$。我们可以依靠这些信息构建一个投资

[1]参考 https://archive.ics.uci.edu/ml/index.php。

组合，我们用 x_i 代表第 i 支股票在投资组合中所占的比例，令 $\mathbf{x} = (x_1, \dots, x_{50})^{\mathrm{T}}$，其中 $\sum_{i=1}^{50} x_i = 1$。我们可以的计算得到投资组合的平均收益和方差分别为

$$\mu_x = \mu^{\mathrm{T}}\mathbf{x}, \quad \mathbf{V}_{\mathbf{x}} = \mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x} \tag{5}$$

则我们可以构建一个最大化收益，同时最小化方差的投资组合：

$$\begin{aligned} \min_{\mathbf{x}} \quad & q(\mathbf{x}) = -\mu^{\mathrm{T}}\mathbf{x} + \lambda\mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x} \\ \text{s.t.} \quad & \sum_{i=1}^{50} x_i = 1 \\ & x_i \geq 0, \quad i = 1, \dots, 50 \end{aligned} \tag{6}$$

1. 写出优化问题的拉格朗日乘子形式与 KKT 条件，并解释 KKT 条件的含义。
2. 使用已有的优化算法包，编写代码求解上述问题，并比较不同 λ 下的最优解 q^* 的变化。
3. 尝试自己编写优化算法求解问题，并验证结果。