arXiv:2303.04301v1 [stat.ML] 8 Mar 2023

使用决策树桩进行最优稀疏恢复

Kiarash Banihashem、MohammadTaghi Hajiaghayi、Max Springer

马里兰大学 kiarash@umd.edu、
hajiagha@cs.umd.edu、mss423@umd.edu

摘要

决策树因计算成本低、预测性能良好且能够评估特征的重要性而得到广泛应用。尽管决策树在实践中常用于特征选择，但人们对这类方法的理论保证仍缺乏充分理解。本文针对使用单层决策树进行线性回归中的特征选择问题，获得了一个紧致的有限样本界。我们考察了这些"决策树桩"从 p 个特征中恢复s 个活跃特征的统计性质，其中 $s \ll p$。我们的分析为高维稀疏系统提供了紧致的样本性能保证，与 Lasso 得到的有限样本界 $O(s\log\ p)$ 一致，并改进了先前针对中位数划分和最优划分准则的界。我们的结果还可扩展至非线性情形以及任意次高斯分布，表明基于树的方法在多种设置下都具有出色的特征选择性质，并进一步阐明了这类方法在实践中取得成功的原因。作为分析的副产品，我们证明，即使活跃特征的数量 s 未知，也能保证恢复。我们还通过计算实验验证了理论结果和证明方法。

1 引言

决策树是机器学习中最受欢迎的工具之一。由于其简单且易于解释，数据科学家广泛使用决策树，既单独使用，也将其与随机森林和梯度提升等集成方法结合使用（Friedman 2001）。

除了预测准确性之外，基于树的方法也是解决变量选择问题的重要工具：从输入变量的高维特征空间中识别一个相关的小子集，以准确预测输出。当变量之间呈线性关系时，人们早已知道，LASSO 能为这一问题达到最优样本复杂度速率（Wainwright 2009a）。然而在实践中，基于树的方法被证明更具优势，因为它们能够随数据规模线性扩展，并捕捉变量之间的非线性关系（Xu 等，2014）。

值得注意的是，在众多领域中，人们采用了各种树结构系统来完成这项变量选择任务。例如，基于树的重要性度量已用于预测财务困境（Qian 等，2022）、无线信号识别（Li 和 Wang，2018）、信用评分（Xia 等，2017），以及生物信息学领域的基因发现（Breiman 2001；Bureau 等，2005；Huynh-Thu 等，2012；Lunetta 等，2004）等，仅举少数例子。

然而，尽管在实证中取得了成功，这些用于特征选择的基于树的方法的理论性质仍未得到充分理解。虽然已有几篇论文研究了这一问题的变体（概述见第3节），即使在简单的线性情形下，决策树的样本复杂度也尚未得到清晰刻画。

本文尝试弥合这一差距，并分析单层决策树（通常称为"决策树桩"（DSTUMP））的变量选择性质。通过考察线性和非线性设定，我们证明，对于多种实用的样本分布，DSTUMP 均能实现近乎紧的样本复杂度速率。与以往工作相比，我们的分析更为简洁，适用于决策树的不同变体以及更一般的函数类。

本文其余部分安排如下：第2节介绍关于 DSTUMP 有限样本保证的主要结果；第3节讨论稀疏恢复领域的重要先前结果及其不足；第4节介绍恢复算法；第 5 节在逐步放宽的问题假设下，为该过程提供理论保证。第6节给出这些结果的证明。第7节通过计算实验补充理论结果，最后在第8节总结全文。

2 我们的结果

我们假设给定一个数据集 $\mathcal{D} = \{(\mathbf{X}_{i,:}, Y_i)\}_{i=1}^n$，其中包含来自非参数回归模型 $Y_i = \sum_k f_k(\mathbf{X}_{i,k}) + W_i$ 的 n 个样本，其中 i 表示样本编号，$\mathbf{X}_{i,:} \in \mathbb{R}^p$ 是输入向量，其对应的输出为 $Y_i \in \mathbb{R}$，$W_i \in \mathbb{R}$ 是独立同分布噪声值，而 $f_k\colon \mathbb{R} \to \mathbb{R}$是单变量函数，且它们是 s-稀疏的：

定义2.1. 若存在一个大小为$s = |S| \ll p$ 的集合 $S \subseteq [p]$，使得一组单变量函数 $\{f_k\}_{k\in[p]}$ 在特征集 $[p]$ 上是 s-稀疏的，则称其为 s-稀疏：

$$f_k = 0 \iff k \notin S$$

给定对（$\mathbf{X}_{i,\ :},\ Y_i$）$_{i=1}^n$的访问权限，我们考虑稀疏恢复问题，目标是仅用尽可能少的样本揭示集合 S。请注意，这不同于

以学习函数 f_k为目标的预测问题。遵循先前的研究（Han 等，2020；Kazemitabar 等人（2017）；Klusowski，2020），我们假设$\mathbf{X}_{i,\ j}$是从 $[0,1]$ 上的均匀分布中独立同分布抽取的，并带有高斯噪声 $W_i \sim \mathcal{N}(0,\ 1)$。在第 5 节中，我们将讨论如何利用我们的非参数结果放宽这一假设，以考虑更一般的分布。

对于恢复问题，我们考虑由 Kazemitabar 等人（2017）首次提出的 DSTUMP 算法。该算法如算法 1 所示：针对每个特征，使用“中位数划分”或“最优划分”策略拟合单层决策树（树桩），并根据相应决策树的误差对特征进行排序。中位数划分显著简化了实现，因为我们无需优化树桩的构造，同时相比可比的最优切分过程，还能进一步缩短运行时间。事实上，中位数划分和最优划分的时间复杂度分别至多为 $\mathcal{O}(np)$ 和 $\mathcal{O}(np\log(n))$。

尽管进行了这一简化，我们证明，在广泛研究的线性设计情形下，其中 f_k 是线性的，DSTUMP 能够正确恢复 S，其样本复杂度上界为 $\mathcal{O}(s\log\ p)$，与 LASSO（Wainwright（2009a,b））为该问题达到的极小极大最优下界相匹配。这一结果既值得注意又出人意料，因为 DSTUMP 算法（以及一般的决策树）并非基于线性假设设计，而 LASSO 则是如此。因此，树通常因其在非线性模型中的预测能力而被采用，而我们的工作则证明了它们在相反情形下的价值。我们进一步将这些结果推广到非线性模型和一般的次高斯分布，并通过更简单的分析改进了此前已知的结果。此外，我们的结果不要求事先知道稀疏度 s。我们将主要技术结果总结如下：

- 在线性设计情况下，我们为 DSTUMP 算法得到了样本复杂度上界 $\mathcal{O}(s\log\ p)$，达到了 LASSO 的最优速率，并改进了文献中关于中位数划分和最优划分的已有上界。这是关于用于稀疏恢复的单层决策树样本复杂度的首个紧上界，并显著优于现有结果。
- 我们将结果扩展到非线性 f_k 的情况。与现有文献相比，我们对更广泛的函数类别得到了更紧的结果。我们进一步利用这些结果，通过扩展到非线性 f_k，将分析推广到次高斯分布。
- 作为改进分析的副产品，我们证明，即使活跃特征的数量 s 未知，我们的结果仍然成立。这是关于决策树桩的首个不要求已知 s 的理论保证。
- 我们通过数值模拟验证了理论结果，模拟结果表明了分析技术的必要性。

3 相关工作

尽管我们的模型框架和整体稀疏性问题已得到广泛研究（Fan 和 Lv （2006）；Lafferty 和 Wasserman （2008）；Wainwright（2009a,b)），但没有已有研究在给定的一组假设下，为分类问题复现众所周知的最优下界。我们的工作通过弱学习器，为回归问题中的 DSTUMP 提供了改进的有限样本保证，其性能几乎可与 LASSO 相媲美。

与我们工作最相关的是（Kazemitabar 等人（2017）），该研究提出了 DSTUMP 算法，并为这些弱学习器的有限样本保证建立了理论方法。不同于我们关于恢复所需样本数量的近乎紧确结果，（Kazemitabar 等人（2017））在采用中位数划分准则时给出了较弱的 $\mathcal{O}(s^2\log p)$ 界。我们在此通过强调树桩划分分析中的一个微妙细节（读者可能会对此独立感兴趣），证明该过程能够获得近乎最优的有限样本保证：我们不使用一个子树的方差作为不纯度度量，而是使用两个子树的方差。正如我们将在理论和实验中所展示的，这一考量对于获得紧确界至关重要。我们的分析也比（Kazemitabar 等人（2017））的分析更为一般，适用于中位数划分和最优划分、更广泛的一类函数 f_k以及更一般的分布。

在一项近期工作中，Klusowski 和 Tian（2021年）通过研究 DSTUMP 与 SIS 算法（Fan 和 Lv（2006年））之间的关系，对采用最优划分准则的 DSTUMP 表述进行了间接分析。SIS 专为线性模型设计，它根据特征与输出之间的皮尔逊相关系数对特征排序，并且在线性设定下具有最优样本复杂度。Klusowski 和 Tian（2021年）表明，采用最优划分准则时，DSTUMP 与 SIS等价（相差对数因子），因此样本复杂度为

$$n \gtrsim s\log(p)\cdot(\log(s)+\log(\log(p)))\,.$$

这改进了 Kazemitabar 等人（2017年）的结果，但分析过程更为复杂。类似的技术也用于研究非线性情形，但对 f_k 所作的条件难以解释，而且界也有所弱化。事实上，将非参数结果具体应用于线性情形，会得到次优的样本复杂度速率 $\mathcal{O}(s^2\log p)$。此外，Klusowski 和 Tian（2021年）还针对未知 $|S|$的情形描述了一种启发式算法，但未能证明其性能具有任何保证。

相比之下，我们直接分析 DSTUMP。这使我们能够在线性情形下为两种中位数划分和最优划分获得最优界，并在非线性情形下得到改进且更一般的界。我们提出的新方法还使我们能够分析未知稀疏度的情形，因为我们首次为 (Klusowski 和 Tian（2021）) 和 (Fan、Feng 和 Song（2011）) 提出的启发式算法给出了正式证明。与以往工作相比，我们的分析简单得多，并且适用于更一般的设置。

此外，许多研究利用决策树中位数划分的简便性，为随机森林回归问题的均方误差给出了精确界（Duroux, Roxane and Scornet, Erwan 2018年；Klusowski 2021年）。在这些研究中，基于中位数划分假设进行分析，能够改进这类

集成弱学习器预测误差的渐近界和有限样本界。在本研究中，我们将这一思路拓展到高维稀疏系统的特征选择问题，并进一步强调，即使在奇异决策树桩的情形下，中位数划分仍然有用。

4 算法描述

4.1 符号与问题设置

为方便数学表达，我们采用矩阵表示法，并用 $\mathbf{X} = (\mathbf{X}_{ij}) \in \mathbb{R}^{n\times p}$ 和 $Y = (Y_i) \in \mathbb{R}^n$ 分别表示输入矩阵和输出向量。我们用 $\mathbf{X}_{i,:}$和 X^k 表示矩阵 $\mathbf{X}$ 的第 i 行和第 k 列。我们还将单变量函数 f_k 的定义扩展到向量，假设该函数分别作用于每个坐标：对于 $v \in \mathbb{R}^d$，定义 $f_k(v) \in \mathbb{R}^d$ 为 $(f_k(v))_i = f_k(v_i)$。

我们用 $\mathbb{E}[\cdot]$ 和 $\mathrm{Var}(\cdot)$ 表示随机变量的期望和方差，$\widehat{\mathbb{E}}$ 和 $\widehat{\mathrm{Var}}$，用 $\mathbb{E}[\cdot]$ 和 $\mathrm{Var}(\cdot)$ 表示它们的经验对应量，即对于任意向量 $v \in \mathbb{R}^d$,

$$\widehat{\mathbb{E}}[v] = \frac{\sum_{i=1}^d v_i}{d} \quad \text{and} \quad \widehat{\mathrm{Var}}(v) = \frac{\sum_{i=1}^d (v_i - \widehat{\mathbb{E}}[v])^2}{d}.$$

我们还用 $[d]$ 表示集合 $\{1,\dots,d\}$。最后，我们用 $\mathrm{Unif}(a,b)$ 表示 $[a,b]$ 上的均匀分布，并用 $C, c > 0$ 表示一般的普适常数。

全文中，我们将使用次高斯随机变量的概念来陈述和证明我们的结果。若存在一个 t 使得 $\mathbb{E}\left[e^{(Z/t)^2}\right] \le 2$，则称随机变量 $Z \in \mathbb{R}$ 为次高斯随机变量，其次高斯范数定义为

$$\|Z\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}\left[e^{(Z/t)^2}\right] \le 2\right\}.$$

次高斯随机变量已得到广泛研究，并具有许多理想性质（全面综述见（Vershynin，2018年））；其中一些性质将在下文概述，因为它们贯穿于我们的分析之中。

(P1)（Hoeffding不等式）若 Z 是次高斯随机变量，则对任意 $t > 0$,

$$\Pr\left(|Z - \mathbb{E}[Z]| \ge t\right) \le 2e^{-c\left(t/\|Z\|_{\psi_2}\right)^2}.$$

(P2) 若 $Z_1, \dots Z_n$ 是一列相互独立的次高斯随机变量，则 $\sum Z_i$ 也是次高斯随机变量，且其范数满足

$$\left\|\sum Z_i\right\|_{\psi_2}^2 \le \sum \|Z_i\|_{\psi_2}^2.$$

(P3) 若 Z 是次高斯随机变量，则 $Z - \mathbb{E}[Z]$ 和 $\|Z - \mathbb{E}[Z]\|_{\psi_2} \le c\|Z\|_{\psi_2}$也成立。

4.2 DSTUMP 算法

我们在此介绍恢复算法 DSTUMP。对于每个特征 $k \in [p]$，该算法会在给定特征上拟合一个单层决策树，即“树桩”，并将其不纯度定义为该特征对应的决策树的误差。直观而言，活跃特征预期能更好地预测 Y，因此具有更低的 不纯度值。于是，该算法根据这些值对特征进行排序，并输出不纯度最小的 $|S|$ 个特征。我们的方法的正式伪代码见算法 1。

算法 1：使用 DSTUMP 进行评分

输入： $\mathbf{X} \in \mathbb{R}^{n\times p}$，$Y \in \mathbb{R}^n$，$s \in \mathbb{N}$

输出： S 的估计

对 $k \in \{1,\dots,p\}$ **执行**
  $\tau^k = \mathrm{argsort}(X^k)$
  对于 $n_L \in \{1,\dots,n\}$ **执行**
    $n_R = n - n_L$
    $Y_L^k = \left(Y_{\tau^k(1)} \dots Y_{\tau^k(n_L)}\right)^T$
    $Y_R^k = \left(Y_{\tau^k(n_L+1)} \dots Y_{\tau^k(n)}\right)^T$
    $\mathrm{imp}_{k,n_L} = \frac{n_L}{n}\widehat{\text{方差}}(Y_L^k) + \frac{n_R}{n}\widehat{\text{方差}}(Y_R^k)$
  如果中位数划分 **则**
    $\mathrm{imp}_k = \mathrm{imp}_{k,\lfloor \frac{n}{2} \rfloor}$。
  else
    $\text{不纯度}_k = \text{最小}_\ell \text{不纯度}_{k,\ell}$。
返回 $\tau = \mathrm{argsort}(\text{不纯度})$

形式上，对于每个特征 k，第 k 列按升序排序，使得 $X^k_{\tau^k(1)} \le X^k_{\tau^k(2)} \cdots \le X^k_{\tau^k(n)}$，并随机打破并列值。随后将样本分成两组：左半部分，由 $X_i^k \le X^k_{\tau^k(n_L)}$ 组成；以及右半部分，由 $X_i^k > X^k_{\tau^k(n_L)}$组成。从概念上讲，这相当于在第 k 列上分割一棵单层树，比例为 n_L 比 n_R，并收集每组中的样本。然后，算法计算每组输出的方差，这表示该组用单个值$\widehat{\mathrm{Var}}\left(Y_L^k\right)$ =最小$_t \frac{1}{n_L}\cdot\sum(Y^k_{L,\,i} - t)^2$时的最优预测误差。这两个方差的平均值即为不纯度。形式上，

$$\mathrm{imp}_{k,n_L} = \frac{n_L}{n}\widehat{\mathrm{Var}}(Y_L^k) + \frac{n_R}{n}\widehat{\mathrm{Var}}(Y_R^k). \tag{1}$$

对于中位数划分 算法，n_L 的值只需选择为 $\lfloor \frac{n}{2} \rfloor$ 或 $\lceil \frac{n}{2} \rceil$；而对于最优划分，则选择该值以使划分的不纯度最小。最后，按照不纯度值对特征进行排序，并将不纯度最低的 $|S|$ 个特征预测为 S。在第5.3节中，我们讨论未知 $|S|$ 的情况，并得到一个具有几乎相匹配保证的算法。

5 主要结果

5.1 线性设计

对于我们的第一个结果，我们考虑输入服从均匀分布的最简单线性模型设定。形式化地说，我们假设存在一个向量 $\beta_k \in \mathbb{R}^p$，使得 $f_k(x) = \beta_k \cdot x$ 对所有 k均成立。这等价于考虑线性回归模型 $Y = \sum_k \beta_k X^k + W$。我们进一步假设矩阵$\mathbf{X}$ 的每个元素都是从 $[0,1]$上的均匀分布中独立同分布抽取的。这一基本设定非常重要

从理论角度来看，它使我们能够在推广到更一般的情境之前，与文献中的现有结果进行比较。定理 5.1 的这一初步结果给出了在该设定下 DSTUMP 失败概率的上界。

定理 5.1。 假设输入矩阵 X的每个元素均独立同分布地采样自 $[0,1]$上的均匀分布。进一步假设输出向量满足线性回归模型 $Y=\sum_k \beta_k X^k + W$，其中 W_i 独立同分布地采样自$N(0,\ \sigma_w^2)$。算法 *1* 以至少 S 的概率正确恢复活跃特征集

$$1-4se^{-c\cdot n}-4pe^{-c\cdot n\frac{\min_k \beta_k^2}{\|\beta\|_2^2+\sigma_w^2}}.$$

对于中位数划分，且以至少

$$1-4se^{-c\cdot n}-4npe^{-c\cdot n\frac{\min_k \beta_k^2}{\|\beta\|_2^2+\sigma_w^2}}.$$

的概率适用于最优划分。

此外，上述定理为 DSTUMP 算法提供了有限样本保证，且不对参数或其渐近关系作任何假设。为了与现有文献（Kazemitabar 等人（2017）；Klusowski and Tian（2021）；Wainwright 2009a,b）进行比较，我们在渐近情形 $\min_{k\in S}\beta_k^2 \in \Omega(\frac{1}{s})$ 和 $\|\beta\|_2 \in \mathcal{O}(1)$ 下考察这些界。

推论 5.2。 在与定理 *5.1* 相同的设定下，假设 $\|\beta\|_2^2 \in \mathcal{O}$（1）且 最小值$_{k\in S}\,\beta_k^2 \in \Omega(\frac{1}{s})$。那么，$^-$只要 $n \gtrsim s$对数 p，算法 *1* 以高概率正确恢复活跃特征集 S。

推论的证明见附录。上述结果表明，当数据服从线性模型时，DSTUMP 在恢复方面是最优的。考虑到基于树的方法以擅长捕捉非线性关系而闻名，并且不像 LASSO 那样基于线性假设进行设计，这一结果很有意思。在下一节中，我们将有限样本界分析进一步扩展到非线性模型。

5.2 加性设计

我们在此考虑非线性 f_k 的情形，并为原始 DSTUMP 算法获得理论保证。我们的主要结果是下面给出的定理 5.3。

定理 5.3。 假设输入矩阵X 的每个元素都独立同分布地采样自定义在 $[0,1]$ 上的均匀分布，且 $Y=\sum_k f_k(X^k)+W$，其中 W_i 独立同分布地采样自$(0,\sigma_w^2)$。进一步假设每个 f_k 都是单调的，并且 f_k(均匀分布(0,1)) 是次高斯的，其次高斯范数为$\|f_k$(均匀分布(0,1))$\|_{\psi_2}^2 \le \sigma_k^2$。对于 $k\in S$，定义 g_k 为

$$g_k := \mathbb{E}\left[f_k(\text{Unif}(\frac{1}{2},1))\right]-\mathbb{E}\left[f_k(\text{Unif}(0,\frac{1}{2}))\right] \quad (2)$$

并定义 σ^2 为 $\sigma^2=\sigma_w^2+\sum_k \sigma_k^2$。算法 *1* 以至少.....的概率正确恢复集合 S

$$1-4se^{-cn}-4pe^{-cn\frac{\min_k g_k^2}{\sigma^2}}$$

对于中位数划分，以及

$$1-4se^{-cn}-4npe^{-cn\frac{\min_k g_k^2}{\sigma^2}}$$

对于最优划分。

注意，将定理 5.3 应用于线性模型，可得到与定理 5.1 相同的上界，这意味着上述上界在线性设定下是紧的。

定理 5.1 将结论推广到所有单调函数，具有重要的理论意义：由于 DS- TUMP 算法在输入的单调变换下保持不变，因此对于任意分布的$\mathbf{X}_{i,:}$，我们都能得到相同的结果。作为一个简单示例，考虑$\mathbf{X}_{ij} \sim \mathcal{N}(0,1)$，并假设我们关注线性模型的上界。将矩阵$\mathbf{Z}$ 定义为$\mathbf{Z}_{ij} = F_{\mathcal{N}}(\mathbf{X}^{ij})$，其中 $F_{\mathcal{N}}(.)$ 表示高斯分布的累积分布函数。由于累积分布函数是递增函数，使用 ($\mathbf{Z}$ Y) 运行 DSTUMP 算法，得到的输出与使用 ($\mathbf{X}$ Y) 运行时相同。此外，将随机变量的累积分布函数作用于该随机变量本身，会得到一个均匀随机变量。因此，$\mathbf{Z}^{ij}$ 是从 Unif(0,1) 分布中独立同分布抽取的样本。令$f_k(t) = \beta_k \cdot F_{\mathcal{N}}^{-1}(t)$，定理 5.3 对 ($\mathbf{Z}$ Y) 得出的结果意味着与定理 5.1 相同的上界。值得注意的是，对于高斯分布，我们也能得到相同的样本复杂度上界 $\mathcal{O}(s\log\ p)$。在附录中，我们将讨论这一思想的一般化形式，它实际上使我们能够去除定理 5.3 中的均匀分布条件。

5.3 未知稀疏度

前述结果的一个缺点是，它们假设已知 $|S|$，但通常情况下并非如此。然而，即使未知 $|S|$，定理 5.3 也保证在 τ中，活跃特征的排序高于非活跃特征，即对于所有 $k\in S$，$k'\notin S$，$\tau(k) < \tau(k')$。为了恢复 S，只需找到一个阈值 γ，使得 $\max_{k\in S}\text{imp}_k \le \gamma \le \min^{k\notin S}\text{imp}_k$。

为了解决这个问题，我们使用所谓的“置换算法”。这是统计学文献中一种广为人知的启发式方法（Barber and Cand`es 2015年；Chung and Romano 2013年、2016年；Fan、Feng 和 Song（2011）），Klusowski and Tian（2021）也曾讨论过这一方法（但未给出证明）。具体而言，我们对 X的各行进行随机排列 σ，得到矩阵 $\sigma(X)$，其中 $\sigma(X)_{ij} = X_{\sigma(i),\ j}$。随后，我们以 $\sigma(X)$和 Y为输入，重新运行算法 1。随机排列意味着 X 和 Y 彼此“解耦”，实际上，所有特征现在都处于非活跃状态。因此，我们预期 $\min_{i,\ t}\text{imp}_i(\sigma(X)\ y)$ 会接近 $\min^{k\notin S}\text{imp}_k(X,\ y)$。由于这个估计可能略显保守，我们会重复抽样，并取这些重复试验中的最小值。算法 2 给出了形式化的伪代码。STUMPSCORE 方法与算法 1 相同，区别在于它会在第 12 行返回 imp

算法 2：未知 $|S|$

输入： $\mathbf{X} \in \mathbb{R}^{n\times p}$，$Y \in \mathbb{R}^n$**输出：** $\gamma \in [\max_{k\in S}\text{imp}_k,\ \min^{k\notin S}\text{imp}_k]$1: **对于** $t \leftarrow 1,$ ....., $T-1$ **执行**2: $\sigma^{(t)} \leftarrow$ 对 $[n]$进行随机排列。3: $\text{imp}^{(t)} \leftarrow$ STUMPSCORE($\sigma^{(t)}(X)$ y) **返回**$\min_{i,\ t}\text{imp}_i^{(t)}$

假设我们在算法中使用了 T 次重复，则 $\min^{k\notin S}\text{imp}_k(X,\ y) \leq \gamma$ 不超过 $\frac{1}{T}$ 的概率至多为 $\frac{1}{T}$。虽然我们在附录中给出了形式化证明，但这一结果背后的主要直觉是，$\text{imp}_i^{(t)}$ 和 $\text{imp}_{k'}$ 对于 $k' \notin S$ 都是对应于非活跃特征的不纯度。因此，所有这些不纯度中的最大值落在 $[p]\backslash S$ 内的概率至多为 $\frac{p-s}{Tp} \leq \frac{1}{T}$。要确保 $\max_{k\in S}\text{imp}_k(X,\ y) \leq \gamma$，需要将额外的 T 个不纯度分数视为 $(T-1)p$ 个额外的非活跃特征。这意味着我们可以使用定理 5.3 中相同的结果，并令 p 等于 Tp，因为我们的样本复杂度上界关于 p 呈对数级 增长。形式化结果及其证明见附录。

定理 5.4. 在与定理 *5.3* 相同的设定下，令 γ 为算法 *2* 的输出，并令 S 为 $\{k:\ \text{imp}_k \leq \gamma\}$。至少有以下概率：$S = S$

$$1 - T^{-1} - 4se^{-cn} - 4Tpe^{-cn\frac{\min_k g_k^2}{\sigma^2}}$$

对于中位数划分，至少为

$$1 - T^{-1} - 4se^{-cn} - 4nTpe^{-cn\frac{\min_k g_k^2}{\sigma^2}}$$

对于最优划分。

我们注意到，若对某个常数 $c > 0$ 取 $T = n^c$，则可得到与之前相同的 $\mathcal{O}(s\log\ p)$。

6 证明

本节证明定理 5.3，因为它是定理 5.1 更一般的版本；其余证明留待附录。

为了证明该算法能够成功，我们需要证明对于所有 $k \in S$，$k' \notin S$，$\text{imp}_k < \text{imp}_{k'}$。我们首先证明一个上界 imp_k，对所有 $k \in S$

引理 6.1. 在定理 *5.3* 的设定下，对于任意活跃特征 $k \in S$，$\text{imp}_k \leq \widehat{Var}(Y) - \min_{720}^{k} g_k^2$，其成立概率至少为

$$1 - 4e^{-c\cdot n} - 4e^{-c\cdot n\cdot\frac{\min_k g_k^2}{\sigma^2}}.$$

接下来，我们需要证明对于所有 $k' \notin S$，$\text{imp}_{k'}$ 具有类似的下界。

引理 6.2. 在定理 *5.3* 的设定下，对于任意非活跃特征 $k' \notin S$，$\text{imp}_{k'} > \widehat{Var}(Y) - \min_{720}^{k} g_k^2$，其成立概率至少为

$$1 - 4e^{-c\cdot n\cdot\frac{\min_k g_k^2}{\sigma^2}}$$

对于中位数划分以及

$$1 - 4ne^{-c\cdot n\cdot\frac{\min_k g_k^2}{\sigma^2}}$$

对于最优划分。

对所有 k, k' 应用并集界，引理 6.1 和引理 6.2 证明了该定理，因为它们表明，对于所有 $k \in S, k' \notin S$，$\text{imp}_k < \text{imp}_{k'}$，且满足所需的概率。

现在我们着重证明引理 6.1。不失一般性，我们假设 f_k 随 [1] 递增。进一步假设 $\mathbb{E}\left[f_k(X_i^k)\right] = 0$，因为 DSTUMP 对输出的常数平移具有不变性。最后，我们假设 $n > 5$；对于 $n \leq 5$，通过选择较大的 c，定理的陈述可以变得平凡成立。

在整个证明中，我们将假设 $\{n_L, n_R\} = \{\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil\}$；因此，我们的结果对最优切分准则和中位数切分准则均成立。如前所述，要获得紧致界，一个关键点是在分析中同时考虑两个子树，而不是分别考察它们。形式上，尽管不纯度通常像式（1）那样通过方差定义，但它还有以下等价定义。

$$\text{imp}_k = \widehat{\text{Var}}(Y) - \frac{n_L \cdot n_R}{n^2} \cdot \left(\widehat{\mathbb{E}}\left[Y_L^k\right] - \widehat{\mathbb{E}}\left[Y_R^k\right]\right)^2. \quad (3)$$

上述恒等式常用于分析决策树及其性质（Breiman 等人，1983年；Li 等人（2019）；Klusowski，2020；Klusowski 和 Tian（2021））。从分析角度看，式（3）与式（1）的关键区别在于，经验平均是在平方之前计算的，因此我们可以更简单地分析一阶 矩，而非二阶矩。

直观地说，我们希望利用集中不等式证明 $\widehat{\mathbb{E}}\left[Y_L^k\right]$ 和 $\widehat{\mathbb{E}}\left[Y_R^k\right]$ 集中在其期望附近，并给出 $|\mathbb{E}\left[Y_L^k\right] - \mathbb{E}\left[Y_R^k\right]|$ 的下界。然而，这颇具挑战，因为集中结果通常要求独立同分布假设，但正如我们将看到的，$Y_{L,\ i}^k$ 并非独立同分布。更正式地，对每个 k，定义随机变量 $X_L^k \in \mathbb{R}^{n_L}$ 为 $(X_{\tau^k(1)}^k \cdots,\ X_{\tau^k(n_L)}^k)^T$，因此 $Y_{L,\ i}^k = f_k(X_{L,\ i}^k) + \sum_{j\neq k} f_j(X_{\tau^k(i)}^j) + W_{\tau^k(i)}$。虽然随机向量 $X^{j\neq k}$ 和 W 的各分量独立同分布，但 X_L^k 是通过对 X^k 的坐标排序得到的。因此，其坐标并非同分布，且彼此相关。为解决第一个问题，请注意，经验均值在置换下保持不变，因此我们可以随机打乱 Y_L^k 的元素，从而得到一个坐标同分布的向量。此外，根据德·菲内蒂定理，任何坐标同分布（但不一定相互独立）的随机向量，都可以看作独立同分布随机向量的混合，从而有效解决第二个问题。形式上，以下结果成立。

引理 6.3（Kazemitabar 等人（2017）中的引理 1）。令 $\tilde{\tau}: [n_L] \to [n_L]$ 为一个随机排列，且与 $[n_L]$ 独立同分布于 $(\mathbf{X}\ W)$，并将 $\widetilde{X}_L^k \in \mathbb{R}^{n_L}$ 定义为 $\widetilde{X}_{L,\ i}^k := X_{L,\ \tilde{\tau}(i)}^k$。随机向量 $\widetilde{X}_L^k$ 服从由大小为 n_L、定义在 $[0, \Theta]$ 上的均匀独立同分布向量构成的混合分布，其中 Θ 取自贝塔分布 $(n_L + 1,\ n_R)$。

将 $\widetilde{Y}_L^k \in \mathbb{R}^{n_L}$ 定义为 $\widetilde{Y}_{L,\ i}^k := Y_{L,\ \tilde{\tau}(i)}^k$，显然 $\widehat{\mathbb{E}}\left[\widetilde{Y}_L^k\right] = \widehat{\mathbb{E}}\left[Y_L^k\right]$，因此我们可以转而分析 $\widetilde{Y}_L^k$

[1] 对于递减的 f_k，可以通过对称论证，或通过映射 $X^k \to -X^k$

Y_L^k 的

$$\widetilde{Y}_{L,i}^k := f_k(\widetilde{X}_{L,i}^k) + \sum_{j\neq k} f_j(X^j_{\tau^k\circ\tilde{\tau}(i)}) + W_{\tau^k\circ\tilde{\tau}(i)}$$

根据引理 6.3，这可以看作是独立同分布随机变量的混合分布。

引理 6.3 表明，$\widetilde{Y}_{L,i}^k$的分布中存在两个随机性来源：混合变量 Θ，以及 $\widetilde{X}_L^k|\Theta$ 和（$X^{j\neq k}, W$）的次高斯随机性。对于第二个来源，可以使用标准的集中不等式证明，在给定 Θ的条件下，$\widehat{\mathbb{E}}\left[\widetilde{Y}_L^k\right]$ 会集中在 $\mathbb{E}\left[\widetilde{Y}_{L,1}^k|\Theta\right]$ 附近。我们将在引理 6.5 中对此进行正式证明。不过，在此之前，我们先关注第一个来源，以及 Θ 如何影响 $\widetilde{Y}_{L,i}^k$的分布。

由于 Θ 是从贝塔分布(n_L+1，n_R)中采样的，我们可以利用标准技术证明它集中在 $\frac{1}{2}$ 附近。更正式地说，我们可以使用以下引理，其证明见附录。

引理 6.4。如果 $n\geq 5$，则以至少 $1-2e^{-cn}$ 的概率有 $\Theta\in[\frac{1}{4},\frac{3}{4}]$。

根据上述结果，我们可以在假设 $\Theta\in[1/4,3/4]$ 的情况下分析 $\widetilde{Y}_L^k$。此时，我们可以利用集中不等式证明，以高概率，$\widehat{\mathbb{E}}\left[\widetilde{Y}_L^k\right]$ 集中在 $\mathbb{E}[f_k(\text{Unif}(0,\Theta))]$ 附近。由于假设 f_k 是递增的，因此还可以将其进一步界定为 $\mathbb{E}\left[f_k(\text{Unif}(0,\frac{3}{4}))\right]$。正式地说，我们得到以下结果。

引理 6.5 设 $\theta\in[\frac{1}{4},\frac{3}{4}]$ 为一个活跃特征。对于任意 $k\in S$，

$$\Pr\left[\widehat{\mathbb{E}}\left[\widetilde{Y}_L^k\right]-\mathbb{E}\left[\widetilde{Y}_{L,1}^k|\Theta=\theta\right]\geq t|\Theta=\theta\right]\leq 2e^{-cn\frac{t^2}{\sigma^2}}.$$

此外，令 $f_{a,b}$ 表示 $\overline{\mathbb{E}}^k[f_k(\text{均匀分布}(a,b))]$，

$$\Pr\left[\widehat{\mathbb{E}}\left[Y_L^k\right]\geq f^k_{0,\frac{3}{4}}+g_k/8\right]\leq 2e^{-c\cdot n}+2e^{-c\cdot n\cdot\frac{g_k^2}{\sigma^2}}. \quad (4)$$

证明。为简化记号，我们固定 $\theta\in[\frac{1}{4},\frac{3}{4}]$，并令随机变量 $\widehat{X}_L^k$ 和 $\widehat{Y}_L^k$ 分别表示 $\widetilde{X}_L^k|\Theta=\theta$和 $\widetilde{Y}_L^k|\Theta=\theta$。回顾一下，根据假设，对于所有 j，$f_k(X_i^j)$ 都是参数为 σ_j 的次高斯变量。容易证明（见附录），这意味着 $f_k(\widehat{X}_{L,i}^k)-\mathbb{E}\left[f(\widehat{X}_{L,i}^k)\right]$ 也是次高斯变量，且其范数至多为 $C\cdot\sigma_j^2$。因此，

$$\begin{aligned}\left\|\widehat{Y}_{L,i}^k\right\|_{\psi_2} &= \left\|f_k(\widehat{X}_{L,i}^k)+\sum_{j\neq k}f_j(X^j_{\tau^k\circ\tilde{\tau}(i)})+W_{\tau^k\circ\tilde{\tau}(i)}\right\|^2_{\psi_2}\\ &\overset{(i)}{=}\left\|f_k(\widehat{X}_{L,i}^k)+\sum_{j\neq k}f_j(X_i^j)+W_i\right\|^2_{\psi_2}\\ &\overset{(ii)}{\leq}\left\|f_k(\widehat{X}_{L,i}^k)\right\|^2_{\psi_2}+\sum_{j\neq k}\left\|f_j(X_i^j)\right\|^2_{\psi_2}+\|W_i\|^2_{\psi_2}\\ &\leq C\cdot\sigma_k^2+\sum_{j\neq k}\sigma_j^2+\sigma_w^2\leq C\cdot\sigma^2.\end{aligned}$$

在上述分析中，(*i*) 根据 ($\widehat{X}_L^k$，$X^{j\neq k}$，W) 的独立性假设以及 ($X_i^{j\neq k}$，W_i) 的独立同分布假设得出。至于 (*ii*)，则由 **(P2)** 以及 ($\widehat{X}_L^k$，$X^{j\neq k}$，W) 的独立性假设得出。性质 **(P3)** 进一步表明，$\left\|\widehat{Y}_{L,i}^k-\mathbb{E}\left[\widehat{Y}_{L,i}^k\right]\right\|^2_{\psi_2}$ 的上界为 $C\cdot\sigma^2$，从而证明了引理中的第一个等式。现在，利用Hoeffding不等式，我们得到

$$\Pr\left[\widehat{\mathbb{E}}\left[\widehat{Y}_L^k\right]-\mathbb{E}\left[\widehat{Y}_{L,i}^k\right]\geq g_k/8\right]\leq 2e^{-c\cdot n\cdot\frac{g_k^2}{\sigma^2}}.$$

对任意两个事件 A,B，利用引理 6.4 以及 $\Pr(A)\leq\Pr(B)+\Pr(A|B^C)$，可得

$$\Pr\left[\widehat{\mathbb{E}}\left[Y_L^k\right]-\mathbb{E}\left[\widehat{Y}_{L,i}^k\right]\geq g_k/8\right]\leq 2e^{-cn}+2e^{-c\cdot n\cdot\frac{g_k^2}{\sigma^2}}.$$

然而请注意，$\mathbb{E}\left[\widehat{Y}_{L,i}^k\right]=f^k_{0,\theta}$正如我们在附录中所示，它还可以进一步上界为 $f^k_{0,\frac{3}{4}}$，由此完成证明。

□

利用决策树算法的对称性，我们还可以得到

$$\Pr\left[\widehat{\mathbb{E}}\left[Y_R^k\right]\geq\overline{f}^k_{\frac{1}{4},1}-g_k/8\right]\leq 2e^{-c\cdot n}+2e^{-c\cdot n\cdot\frac{g_k^2}{\sigma^2}} \quad (5)$$

由 (4) 作变量替换 $X^k\to -X^k$ 和 $f_k=-f_k$。对 (4) 和 (5) 应用并集界可知，以至少 $1-4e^{-c\cdot n}-4e^{-c\cdot n\cdot\frac{g_k^2}{\sigma^2}}$ 的概率，

$$\widehat{\mathbb{E}}\left[Y_R^k\right]-\widehat{\mathbb{E}}\left[Y_L^k\right]\geq\overline{f}^k_{\frac{1}{4},1}-\overline{f}^k_{0,\frac{3}{4}}-g_k/4.$$

然而，正如我们在附录中所示，简单应用条件期望即可推出 $f^k_{\frac{1}{4},1}-f^k_{0,\frac{3}{4}}\geq g_k/3$。因此，以至少 $1-4e^{-c\cdot n}-4e^{-c\cdot n\cdot\frac{g_k^2}{\sigma^2}}$ 的概率，我们有 $\widehat{\mathbb{E}}\left[Y_R^k\right]-\widehat{\mathbb{E}}\left[Y_L^k\right]\geq\frac{g_k}{12}$。假设 $n\geq 5$，我们还可以进一步得出 $\frac{n_L\cdot n_R}{n^2}\geq\frac{1}{5}$，结合 (3)，即可证明该引理。

7 实验结果

本节将通过模拟进一步论证我们的理论结果：考察不同情形下活跃特征恢复的有限样本数，以及单个子树相较于整棵树的预测能力。此外，我们还将 DSTUMP 与广受研究的最优 LASSO 进行对比。

我们首先验证定理 5.1 的结果，并考虑具有 $p=200$ 的线性设计，其中设计矩阵的元素独立同分布地采样自 $U(0,1)$，并叠加高斯噪声$\mathcal{N}(0,1)$。具体而言，我们考察恢复约 95% 活跃特征 s 所需的最优样本数。为此，我们对样本数进行二分搜索，找到能够恢复所需比例活跃特征集的最小样本数，并对 25 次独立重复实验的结果取平均。在图 1 最左侧的图中，我们绘制了样本数随稀疏度变化的情况

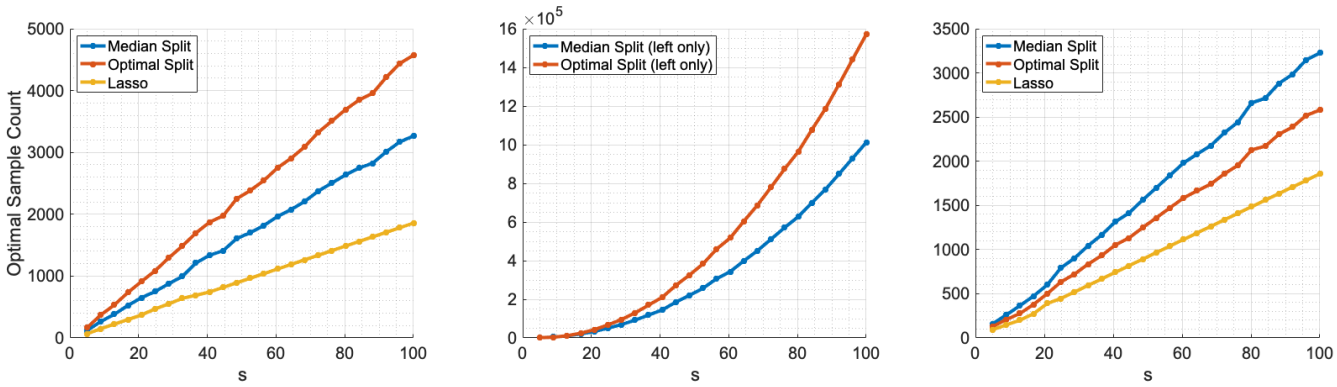

图 1：恢复 95% 活跃特征所需的最优样本数。设计矩阵的样本独立同分布地来自 $U(-1,\ 1)$ 或 $\mathcal{N}(0,\ 1)$，并带有加性高斯噪声 $\mathcal{N}(0,\ 0。1)$。比较三种方法：采用最优划分的 DSTUMP、采用中位数划分的 DSTUMP，以及 LASSO。

对于具有中位数划分的 DSTUMP、具有最优划分的 DSTUMP，以及作为基准的 LASSO（惩罚参数通过标准交叉验证选取），稀疏度分别为 $s \in [5, 100]$。固定 p 后，我们评估的仅是 n 对稀疏度 s 的依赖关系。这里的结果验证了理论上的 $\mathcal{O}(s\log\ p)$ 界，该界与 LASSO 的最优界几乎相当。还值得注意的是，中位数划分所需的样本数少于最优划分。因此，在线性设定中，我们看到采用中位数划分的 DSTUMP 不仅更简单，计算开销也更低。图 1 最右侧的图使用高斯数据样本再次展示了这一最优界结果。值得注意的是，在这种设定下，采用最优划分的决策树桩表现优于中位数划分，这表明它们在不同问题背景下具有不同的效用。

我们还要重申，Kazemitabar 等人（2017）的先前工作试图通过仅考察左子树，来简化使用 DSTUMP 解决稀疏恢复问题的分析，但由此得到的有限样本界并非最优，且为 $\mathcal{O}(s^2 \log p)$。为分析这一选择的影响，图 1 中间的图展示了在定理 5.1 的加性模型下，仅使用左子树时最优的样本恢复数量。与我们的预期及先前文献的分析一致，我们看到，在特征数 p 固定时，n 与 s 之间存在明显的二次关系。

总体而言，这些模拟进一步验证了决策树桩的实用性和预测能力。与最优 LASSO 相比，其性能略有下降，但计算量减少，分析也更为简单。

8 结论

本文针对回归情形，提出了一种基于单层决策树的简单且一致的特征选择算法，并推导了其在高维稀疏系统中的有限样本性能保证。我们的理论结果表明，与作为黄金标准的 LASSO 相比，这类非常简单的弱学习器几乎是最优的。我们为基于二叉决策树的方法在实践中的成功提供了有力的理论依据，并提出了一个框架，可据此将对这类结构的分析拓展到任意高度，这也是未来研究的一个潜在方向。

9 致谢

本研究部分由 DARPA QuICC、美国国家科学基金会 AF:Small #2218678 和美国国家科学基金会 AF:Small # 2114269 资助。Max Springer 获得美国国家科学基金会研究生研究奖学金计划资助，项目编号为 DGE 1840340。本材料中表达的任何观点、研究结果、结论或建议均属于作者个人观点，不一定反映美国国家科学基金会的立场。

参考文献

Barber, R. F.; and Cand`es, E. J. 2015年. 通过仿冒变量控制错误发现率. 《统计年刊》, 43(5): 2055–2085. Breiman, L. 2001年. 《统计建模：两种文化》（附评论及作者回应）. 《统计科学》, 16(3): 199 – 231. Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 1983年. 分类与回归树. Bureau, A.; Dupuis, J.; Falls, K.; Lunetta, K.; Hayward, B.; Keith, T.; and Eerdewegh, P. 2005年. 使用随机森林识别可预测表型的单核苷酸多态性. 遗传流行病学, 28: 171–82. Chung, E.; and Romano, J. P. 2013年. 精确且渐近稳健的置换检验. 《统计年刊》, 41(2): 484–507. Chung, E.; and Romano, J. P. 2016年. 多变量和多重置换检验. 《计量经济学杂志》, 193(1): 76–91. Duroux, Roxane; and Scornet, Erwan. 2018年. 子抽样与树深度对随机森林的影响. *ESAIM:PS*, 22: 96–128. Fan, J.; Feng, Y.; and Song, R. 2011年. 稀疏超高维加性模型中的非参数独立性筛选

模型. 美国统计学会期刊, 106(494): 544–557. Fan, J.; and Lv, J. 2006年. 超高维特征空间的确定性独立筛选. 英国皇家统计学会会刊*B*辑：统计方法学, 70: 849–911. Friedman, J. H. 2001年. 贪心函数逼近：梯度提升机. 统计年刊, 29: 1189– 1232. Han, C.; Rao, N. S.; Sorokina, D.; and Subbian, K. 2020年. （多任务）梯度提升树的可扩展特征选择. *arXiv*, abs/2109.01965. Huynh-Thu, V. A.; Irrthum, A.; Wehenkel, L.; and Geurts, P. 2012年. 使用基于树的方法从表达数据推断调控网络. *PLOS ONE*, 5(9): 1–10. Kazemitabar, S. J.; Amini, A. A.; Bloniarz, A.; and Talwalkar, A. S. 2017年. 使用决策树评估变量重要性. In神经信息处理系统大会. Klusowski, J. 2021年. 随机森林简单模型的精确分析. In Banerjee, A.; and Fukumizu, K., eds.,第*24*届人工智能与统计国际会议论文集, volume 130 of 机器学习研究会议论文集, 757–765. PMLR. Klusowski, J. M. 2020年. 基于CART的稀疏学习. *arXiv*, abs/2006.04266. Klusowski, J. M.; and Tian, P. M. 2021年. 使用最优决策树桩进行非参数变量筛选. In人工智能与统计国际会议. Lafferty, J. D.; and Wasserman, L. A. 2008年. Rodeo：稀疏贪心非参数回归. 统计年刊, 36: 28–63. Li, L.; and Wang, J. 2018年. 认知计算中基于决策树算法的无线信号识别特征重要性评估研究. 认知系统研究, 52: 882–890. Li, X.; Wang, Y.; Basu, S.; Kumbier, K.; and Yu, B. 2019年. 随机森林的去偏mDI特征重要性度量. *arXiv*, abs/1906.10845. Lunetta, K. L.; Hayward, L. B.; Segal, J.; and Eerdewegh, P. V. 2004年. 筛选大规模关联研究数据：利用随机森林挖掘交互作用. *BMC*遗传学, 5(1): 32. Qian, H.; Wang, B.; Yuan, M.; Gao, S.; and Song, Y. 2022年. 使用校正特征选择度量和梯度提升决策树预测财务困境. 《专家系统与应用》, 190: 116202. Rundel, C. 2012年. 第15讲：次序统计量. h t t p s://www2.stat.duke.edu/courses/Spring12/sta104.1/Lectures/Lec15.pdf. Vershynin, R. 2018年. 《高维概率：数据科学中的导论与应用》. Wainwright, M. J. 2009a. 高维含噪场景下稀疏性恢复的信息论极限. *IEEE* 信息论汇刊, 55: 5728– 5741.

Wainwright, M. J. 2009b. 使用 l1 约束二次规划（Lasso）进行高维、含噪稀疏性恢复的精确阈值。《IEEE 信息论汇刊》。Xia, Y.; Liu, C.; Li, Y.; and Liu, N. 2017年. 一种采用贝叶斯超参数优化的提升决策树信用评分方法。《专家系统与应用》，78: 225–241。Xu,Z.; Huang,G.;Weinberger, K.Q.;and Zheng,A.X. 2014年. 梯度提升特征选择。载于《第20届 ACM SIGKDD 国际知识发现与数据挖掘会议论文集》，522–531。

A 附录

A.1 推广到非均匀分布

在本节中，我们将主要结果（即定理 5.3）推广到非均匀分布。

定理 A.1. 假设输入矩阵 X 的元素独立采样，每列中的元素均独立同分布，且 $Y = \sum_k f_k(X^k) + W$，其中 W_i 从 $\mathcal{N}(0, \sigma_w^2)$中独立同分布地采样。进一步假设每个 f_k 均为单调函数，且 $f_k(\mathbf{X}_{i,k})$是次高斯的，其次高斯范数为 $\|f_k(\mathbf{X}_{i,k})\|_{\psi_2}^2 \le \sigma_k^2$。

令 F_k 表示 X_i^k的分布的累积分布函数。定义 $F^{-1}{}_{\cdot k}^{\ +}(z) := \inf\{t \colon F_k(t) > z\}$，并定义 $h_k := f_k \circ F^{-1}{}_{\cdot k}^{\ +}$。对于$k \in S$，定义 g_k 为

$$g_k := \mathbb{E}\left[h_k(\text{Unif}(\frac{1}{2}, 1))\right] - \mathbb{E}\left[h_k(\text{Unif}(0, \frac{1}{2}))\right], \tag{6}$$

并将 σ^2 定义为 $\sigma^2 = \sigma_w^2 + \sum_k \sigma_k^2$。算法 *1* 以至少.....的概率正确恢复集合 S

$$1 - 4se^{-cn} - 4pe^{-cn\frac{\min_k g_k^2}{\sigma^2}}$$

对于中位数划分和

$$1 - 4se^{-cn} - 4npe^{-cn\frac{\min_k g_k^2}{\sigma^2}}$$

对于最优划分。

证明。我们首先断言 i f $Z \sim$ Unif(0，1)，则 $F^{-1}{}_{\cdot\ k}^{\ +}(Z)$ 与 X_i^k 具有相同的分布。形式化地，对所有 $z \in [0, 1]$,

$$\begin{aligned} F_k^{-1,+}(z) \le t \iff \inf\{\tilde{t} : F_k(\tilde{t}) > z\} \le t &\overset{(i)}{\iff} \forall \delta > 0 : \exists \tilde{t} \le t + \delta : F_k(\tilde{t}) > z \\ &\overset{(ii)}{\iff} \lim_{\epsilon \to 0, \epsilon > 0} F_k(t + \epsilon) > z \\ &\overset{(iii)}{\iff} F_k(t) > z. \end{aligned}$$

其中 (i) 根据 inf 的定义得出，(ii) 根据 F_k 递增这一事实得出，而 (iii) 根据 F_k 右连续这一事实得出。因此,

$$\Pr\left(F_k^{-1,+}(Z) \le t\right) = \Pr\left(Z < F_k(t)\right) \overset{(i)}{=} F_k(t).$$

其中，对于 (i)，我们使用了 这样一个事实： Z 在 $[0, 1]$ 上均匀采样。由于 $F_k(t) = \Pr(X_i^k \le t)$，命题得证 .

根据这一结果，我们可以假设 X_i^k 被设为 $F^{-1}{}_{\cdot\ k}^{\ +}(Z_i^k)$，其中 $Z_i^k \sim$ 是服从 Unif$(0, 1)$ 分布的潜变量。虽然这意味着 Z_i^k 不再是 X_i^k 的 函数，但 argsort(Z^k) 的分布仍与 argsort(X^k) 相同。为了解释原因，考虑 X^k 的一个固定取值，并基于条件分布对 Z^k 进行采样。如果对于某个 i， j， $X_i^k < X_j^k$，那么 $Z_i^k < Z_j^k$ 也成立，因为 $F_{-1,\ +}^k$ 是一个函数。至于 $X_i^k = X_j^k$ 的情况， $Z_i^k < Z_j^k$ 和 $Z_i^k > Z_j^k$ 等概率，这与 argsort 会随机打破并列值的假设一致。现在我们注意到

$$f_k(X_i^k) = f_k(F_k^{-1,+}(Z_i^k)) = h_k(Z_i^k).$$

由于 argsort(Z^k) 和 argsort(X^k) 具有相同的分布，算法 1 在 (**X** f_k) 上的输出与在 (**Z** h_k) 上的输出相同。因此，应用定理 5.3 即完成证明。

□

A.2 辅助引理

引理 A.2。 令 $U \sim$Unif$(0, 1)$为 $[0, 1]$上的均匀随机变量，并令 f为一个函数，使得 $f(U)$是次高斯的，其范数为 $\|f(U)\|_{\psi_2}$。令 δ为值域 $[\frac{1}{4}, \frac{3}{4}]$,中的一个值，并令 $\tilde{U} \sim$Unif$(0, \delta)$为 $[0, \delta]$上的随机变量。那么 $f(U)$也是次高斯的，其范数至多为 $f(U)$的常数倍，即，

$$\left\|f(\tilde{U})\right\|_{\psi_2} \le C \cdot \|f(U)\|_{\psi_2}$$

证明。定义 $K := \|f(U)\|_{\psi_2}$，

$$\begin{aligned}
2 &\overset{(i)}{\geq} \mathbb{E}\left[e^{\frac{f(U)^2}{K^2}}\right] \\
&= \mathbb{E}\left[e^{\frac{f(U)^2}{K^2}}\middle| U \leq \delta\right] \cdot \Pr\left[U \leq \delta\right] + \mathbb{E}\left[e^{\frac{f(U)^2}{K^2}}\middle| U > \delta\right] \cdot \Pr\left[U > \delta\right] \\
&\overset{(ii)}{\geq} \mathbb{E}\left[e^{\frac{f(U)^2}{K^2}}\middle| U \leq \delta\right] \cdot \Pr\left[U \leq \delta\right] \\
&= \mathbb{E}\left[e^{\frac{f(\tilde{U})^2}{K^2}}\right] \cdot \delta
\end{aligned} \tag{7}$$

其中，**(*i*)** 根据次高斯范数的定义可得，**(*ii*)** 根据以下事实可得：$e^x \geq 0$ 对所有 x 均成立。定义$\widetilde{K} := K \cdot {}^{ln}{}_{(2/\delta)\, ln\,(2)}$。注意到~~$\delta <$~~ 1，因此 $\frac{K}{\widetilde{K}} < 1$。由此可得。

$$\begin{aligned}
\mathbb{E}\left[e^{\frac{f(\tilde{U})^2}{\widetilde{K}^2}}\right] &= \mathbb{E}\left[e^{\frac{f(\tilde{U})^2}{K^2} \cdot \frac{K^2}{\widetilde{K}^2}}\right] \\
&= \mathbb{E}\left[\left(e^{\frac{f(\tilde{U})^2}{K^2}}\right)^{\frac{K^2}{\widetilde{K}^2}}\right] \\
&\overset{(i)}{\leq} \mathbb{E}\left[\left(e^{\frac{f(\tilde{U})^2}{K^2}}\right)\right]^{\frac{K^2}{\widetilde{K}^2}} \\
&\overset{(ii)}{\leq} (2/\delta)^{\frac{K^2}{\widetilde{K}^2}} \\
&\overset{(iii)}{=} e^{\ln(2/\delta) \cdot \frac{ln(2)}{ln(2/\delta)}} = 2
\end{aligned}$$

其中，对于 **(*i*)**，我们使用了凹函数 $t \to t^{\frac{K^2}{\widetilde{K}^2}}$ 的詹森不等式；对于 **(*ii*)**，我们使用了 **(7)**；对于 **(*iii*)**，我们使用了 $\widetilde{K}$ 的定义。因此，取 $C = {}^{ln}{}_{(2/\delta)\, ln(2)}$，即可证明该结论。 □

引理 A.3。若 f_k 是递增的，则对任意 $0 \leq a \leq c \leq b \leq 1$、$f^k_{a,c} \leq f^k_{c,b}$。

证明。

$$\overline{f}^k_{a,c} = \mathbb{E}_{Z \sim \text{Unif}(a,c)}\left[f_k(Z)\right] \leq f_k(c) \leq \mathbb{E}_{Z \sim \text{Unif}(c,b)}\left[c\right] = \overline{f}^k_{c,b}$$

□

引理 A.4。若 f_k 是递增的，则对任意 $0 \leq a \leq c \leq b \leq 1$、$f^k_{a,c} \leq f^k_{a,b}$。

证明。

$$\begin{aligned}
\overline{f}^k_{a,b} = \mathbb{E}_{Z \sim \text{Unif}(a,b)}\left[f_k(Z)\right] &= \Pr\left[Z \in [a,c]\right] \cdot \mathbb{E}_{Z \sim \text{Unif}(a,c)}\left[f_k(Z)\right] + \Pr\left[Z \in [c,b]\right] \cdot \mathbb{E}_{Z \sim \text{Unif}(c,b)}\left[f_k(Z)\right] \\
&= \Pr\left[Z \in [a,c]\right] \cdot \overline{f}^k_{a,c} + \Pr\left[Z \in [c,b]\right] \cdot \overline{f}^k_{c,b} \\
&\geq \overline{f}^k_{a,c}
\end{aligned}$$

□

引理 A.5。

$$\overline{f}^k_{\frac{1}{4},1} - \overline{f}^k_{0,\frac{3}{4}} \geq g_k/3$$

证明。我们有

$$\overline{f}^k_{0,\frac{3}{4}} = \frac{2}{3} \cdot \overline{f}^k_{\frac{1}{4},\frac{3}{4}} + \frac{1}{3} \cdot \overline{f}^k_{0,\frac{1}{4}},$$

and

$$\overline{f}^k_{\frac{1}{4},1} = \frac{2}{3}\cdot\overline{f}^k_{\frac{1}{4},\frac{3}{4}} + \frac{1}{3}\cdot\overline{f}^k_{\frac{3}{4},1}.$$

因此可得

$$\overline{f}^k_{\frac{1}{4},1} - \overline{f}^k_{0,\frac{3}{4}} = \frac{1}{3}\cdot\left(\overline{f}^k_{\frac{3}{4},1} - \overline{f}^k_{0,\frac{1}{4}}\right)$$

不过请注意,

$$\begin{aligned} g_k = \overline{f}^k_{\frac{1}{2},1} - \overline{f}^k_{0,\frac{1}{2}} &= \left(\frac{1}{2}\cdot\overline{f}^k_{\frac{1}{2},\frac{3}{4}} + \frac{1}{2}\cdot\overline{f}^k_{\frac{3}{4},1}\right) - \left(\frac{1}{2}\cdot\overline{f}^k_{0,\frac{1}{4}} + \frac{1}{2}\cdot\overline{f}^k_{\frac{1}{4},\frac{1}{2}}\right) \\ &= \left(\frac{1}{2}\cdot\overline{f}^k_{\frac{3}{4},1} - \frac{1}{2}\cdot\overline{f}^k_{0,\frac{1}{4}}\right) + \left(\frac{1}{2}\cdot\overline{f}^k_{\frac{1}{2},\frac{3}{4}} - \frac{1}{2}\cdot\overline{f}^k_{\frac{1}{4},\frac{1}{2}}\right) \\ &\leq \overline{f}^k_{\frac{3}{4},1} - \overline{f}^k_{0,\frac{1}{4}} \end{aligned}$$

由此即可证明该结论。 □

A.3 推论 5.2 的证明

证明。 我们需要证明，对于 $n \gtrsim s\log(p)$,

$$\lim_{p\to\infty} 1 - 4se^{-c\cdot n} - 4npe^{-c\cdot n\frac{\min_k \beta_k^2}{\|\beta\|_2^2+\sigma_w^2}} = 1$$

首先注意到，由于 $\min_k \beta_k^2 \in \Omega(\frac{1}{s})$ 且 $\|\beta\|_2 \in \mathcal{O}(1)$，可得 $\frac{\min_k \beta_k^2}{\|\beta\|_2^2+\sigma_w^2} \in \Omega(\frac{1}{s})$。假设对于某个 C，有 $\frac{\min_k \beta_k^2}{\|\beta\|_2^2+\sigma_w^2} \geq \frac{C}{s}$，令 $n = c^{\{4/C,2\}} s\log(p)$。这意味着

$$4se^{-cn} \leq 4\frac{s}{p^{2s}} \leq \frac{4}{p}.$$

当 p 足够大时，它趋于零。因此，接下来只需证明 $4npe^{-cn\frac{\min_k \beta_k^2}{\|\beta\|_2^2+\sigma_w^2}}$ 趋于零，这等价于证明 $cn\frac{\min_k \beta_k^2}{\|\beta\|_2^2+\sigma_w^2} - \log(np)$ 趋于 ∞。不过请注意,

$$\begin{aligned} cn\frac{\min_k \beta_k^2}{\|\beta\|_2^2+\sigma_w^2} - \log(np) &= cn\frac{\min_k \beta_k^2}{\|\beta\|_2^2+\sigma_w^2} - \log(n) - \log(p) \\ &\geq c\cdot\frac{4}{c\cdot C}s\log(p)\cdot\frac{C}{s} - \log(p) - \log(n) \\ &= 3\log(p) - \log(n) \end{aligned}$$

现在注意，根据我们对 n 的选择,

$$n \leq (\frac{4}{c\cdot C} + \frac{2}{c})\cdot p\log(p) \leq (\frac{4}{c\cdot C} + \frac{2}{c})p^2.$$

因此，$\log(n) \leq 2\log(p) + \log(C')$，其中 $C' := \frac{4}{c\cdot C} + \frac{2}{c}$，这意味着

$$3\log(p) - \log(n) \geq \log(p) - \log(C'),$$

当 p 足够大时，其趋于 ∞。 □

A.4 引理 6.2 的证明

在本节中， 我们给出引理 6.2 的证明，该引理是 (Li 等人（2019）) 中引理 1 的推广。

证明。 由于 k' 与 Y 独立，因此置换 $\tau^{k'}$ 也与之独立。由于假设 Y_i 是独立同分布的，这意味着 $Y_{\tau^k(i)}$，进而 $Y^k_{L,i}$ 也都是独立同分布的，并且与 Y_i 具有相同的分布。Y_i^k 均为零均值次高斯变量，其范数至多为 σ，然而,

$$\begin{aligned} \|Y_i\|_{\psi_2}^2 = \left\| f_k(X_i^k) + \sum_{j\neq k} f_j(X_i^j) + W \right\|_{\psi_2}^2 &\overset{(i)}{\leq} \left\| f_k(X_i^k)\right\|_{\psi_2} + \sum_{j\neq k}\left\| f_j(X_i^j)\right\|_{\psi_2} + \|W\|_{\psi_2}^2 \\ &= \sigma^2. \end{aligned}$$

其中 (i) 根据 **(P2)** 得出。聚焦于中位数划分，Hoeffding不等式因此表明，对于任意 $t > 0$，$\Pr\left(\left|\widehat{\mathbb{E}}\left[Y_L^k\right]\right| > t\right) \le 2e^{-c\cdot n_L t^2/\sigma^2}$ 且 $\Pr\left(\left|\widehat{\mathbb{E}}\left[Y_R^k\right]\right| > t\right) \le 2e^{-c\cdot n_R t^2/\sigma^2}$。令 $t = \frac{g_k}{30}$，并对两个子树应用并集界可知，以至少 $1 - 4e^{-c\cdot n\cdot \frac{\min_k g_k^2}{\sigma^2}}$ 的概率，

$$\left|\widehat{\mathbb{E}}\left[Y_R^k\right] - \widehat{\mathbb{E}}\left[Y_L^k\right]\right| \le \frac{g_k}{15}.$$

由于 $\frac{n_L\cdot n_R}{n^2} \le \frac{1}{4}$，结论由 (3) 得出。

对于最优划分，分析需要有所改变，因为划分点不一定在中间，而且还取决于数据。不过，对于固定的 n_L，可以采用上述相同的分析来界定 $\mathrm{imp}_{k,\ n_L}$，只需稍作调整，如 Li 等人（2019）所示；虽然对 $\widehat{\mathbb{E}}\,[Y]_L^k$ 集中性的界较弱，因为 $\frac{n_L}{n}$ 可能很小，但 (3) 中的 $\frac{n_L\cdot n_R}{n^2}$ 项弥补了这一点，最终仍可得到相同的 $\mathrm{imp}_{k,\ n_L}$ 界。对所有可能的划分点应用并集界，即可证明该结果。 □

A.5 引理 6.4 的证明

引理 A.6。设随机变量 Θ 服从 *Beta* 分布(n_L+1, n_R)，其中 $\{n_L, n_R\} = \{\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil\}$。对于 $n \ge 5$ 和 $t \ge \frac{2}{3}$，

$$\Pr\left(\Theta \ge t\right) \le e^{-n(t-\frac{2}{3})^2}.$$

Proof. 设 $U_1, \ldots, U_n$ 是独立同分布的 Unif(0, 1) 随机变量，令 τ 为它们的排序置换，i.e.,

$$U_{\tau(1)} \le \cdots \le U_{\tau(n)}$$

众所周知（例如，参见（Rundel, 2012））$U_{\tau(k)}$ 服从贝塔分布$(k, n+1-k)$。因此，Θ 与 $U_{\tau(n_L+1)}$ 具有相同的分布。这意味着对于任意 $t \in [0,1]$,

$$\begin{aligned}
\Pr\left[\Theta \ge t\right] &= \Pr\left[U_{\tau(n_L+1)} \ge t\right] \\
&= \Pr\left[\left|\{k : U_k < t\}\right| \le n_L\right] \\
&= \Pr\left[\left|\{k : U_k \ge t\}\right| \ge n_R\right] \\
&= \Pr\left[\sum_k \mathbb{1}\left[U_k \ge t\right] \ge n_R\right] \\
&= \Pr\left[\frac{1}{n}\cdot\left(\sum_k \mathbb{1}\left[U_k \ge t\right]\right) \ge \frac{n_R}{n}\right]
\end{aligned}$$

定义 $Z_k := \mathbb{1}\left[U_k \ge t\right]$，显然 Z_k 是参数为 $1-t$ 的独立同分布伯努利随机变量。若 $1 - t \le \frac{n_R}{n}$，则由 Hoeffding不等式可得

$$\begin{aligned}
\Pr\left[\Theta \ge t\right] &= \Pr\left[\frac{1}{n}\cdot\left(\sum_k Z_k\right) \ge \frac{1}{3}\right] \\
&\le 2e^{-n\left(\frac{n_R}{n} - (1-t)\right)^2} \\
&= 2e^{-n\cdot(t-\frac{2}{3})^2}
\end{aligned}$$

由此完成证明。 □

由对称性，对于任意 $t \le \frac{1}{3}$，我们也有 $\Pr\ \Theta \le t \le e^{-n(t-\frac{1}{3})^2}$，由此通过并集界可得引理 6.4 nd.

A.6 定理 5.4 的证明

证明。考虑非活跃 k 的 imp_k 分布。由于 k 处于非活跃状态，排序置换 τ_k 与**X** Y 独立。因此，imp_k 的分布就是 $\mathrm{imp}(\tau(Y))$ 的分布，其中 τ 是 $[n]$ 上的随机排列，且 $\mathrm{imp}(Y)$ 的定义同第 7、9、11 行，即

$$\mathrm{imp}(Y) = \frac{\lfloor \frac{n}{2} \rfloor}{n}\widehat{\mathrm{Var}}(Y_{\le \lfloor \frac{n}{2} \rfloor}) + \frac{n - \lfloor \frac{n}{2} \rfloor}{n}\widehat{\mathrm{Var}}(Y_{> \lfloor \frac{n}{2} \rfloor})$$

对于中位数划分和

$$\mathrm{imp}(Y) = \min_{n_L} \frac{n_L}{n}\widehat{\mathrm{Var}}(Y_{\le n_L}) + \frac{n - n_L}{n}\widehat{\mathrm{Var}}(Y_{> n_L})$$

对于最优划分。在上述公式中，$Y_{\leq i}$ 和 $Y_{>i}$ 分别指 $(Y_1, \dots Y_i)^T$ 和 $(Y_{i+1}, \dots, Y_n)^T$。

同样地，对于 $t \in [T]$ 和 $k \in [p]$，$\tau_i^{(t)}$ 是一个与 $\mathbf{X}$ Y 独立的随机排列。因此，$\text{imp}_i^{(t)}$ 的分布也与随机排列 τ 作用于 $[n]$ 后得到的 $\text{imp}(\tau(Y))$ 相同。此外，所有这些随机排列（因此，对于非活跃 k 而言，所有 $\text{imp}_i^{(t)}$ 和 imp_k 的取值）彼此独立。这是因为 **(a)** 根据假设，所有非活跃特征彼此独立，并且 **(b)** 排列 $\sigma^{(t)}$ 在设计时就是相互独立的，因为它们是在第 2 行中独立抽样得到的。现在我们注意到，$\min_{k \notin S} \text{imp}_k \leq \min_{i,\ t} \text{imp}_i^{(t)}$ 发生，当且仅当所有这些 $(T-1)p + p - s$ 个数中的最小值位于 $p - s$ 个 imp_k 数值中，且这些数值对应于 $k \notin S$。根据对称性，该概率恰好为 $\frac{p-s}{(T-1)p+p-s} \leq \frac{1}{T}$。

至于如何界定 $\max_{k \in S} \text{imp}_k \geq \min_{i,\ t} \text{imp}_i^{(t)}$ 的概率，根据上述讨论，我们注意到，额外的 $\text{imp}_{i,\ t}$ 可以视为针对由 i，t 索引的 额外的 非活跃特征的 imp_k。$\max_{k \in S} \text{imp}_k \geq \min_{k \notin S} \text{imp}_k$ 的概率可按照定理 5.3 中的方法界定。结合 $\Pr\left[\min_{k \notin S} \text{imp}_k \leq \min_{i,\ t} \text{imp}_i^{(t)}\right]$ 使用并集界，即可证明该定理的结论。 □

A.7 实验配置

所有数值模拟均使用 Python 和 SkLearn 软件包完成，以实现 Lasso 和单层决策树。Lasso 回归采用正则化强度 $C = 2$，并使用 *liblinear* 求解器。单层决策树的分裂点按照算法 1 的伪代码所述，通过方差缩减确定。数据使用 PyTorch 软件包随机生成，所有模拟均在配备 NVIDIA RTX A4000 GPU 的设备上，借助 Python 的 CUDA 生态系统完成。