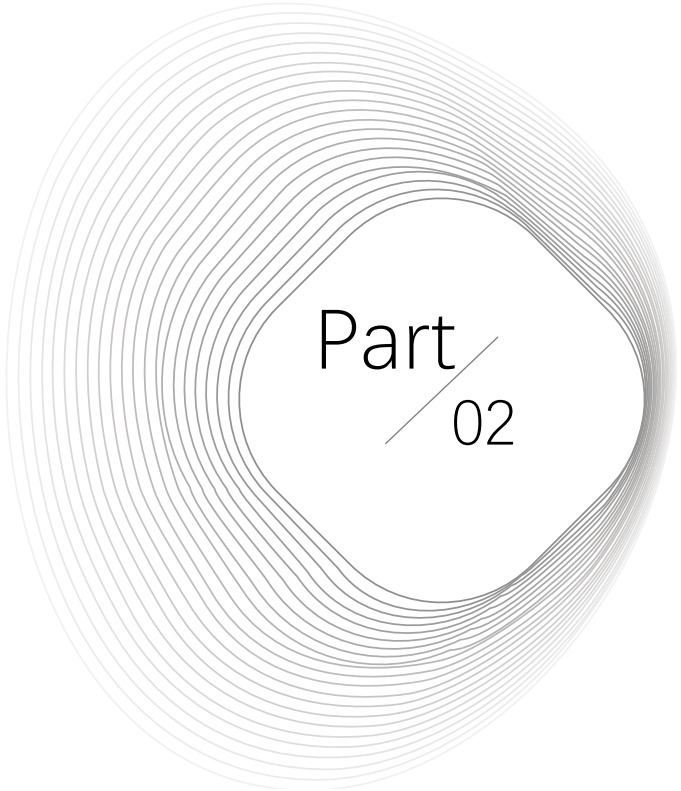


随机森林

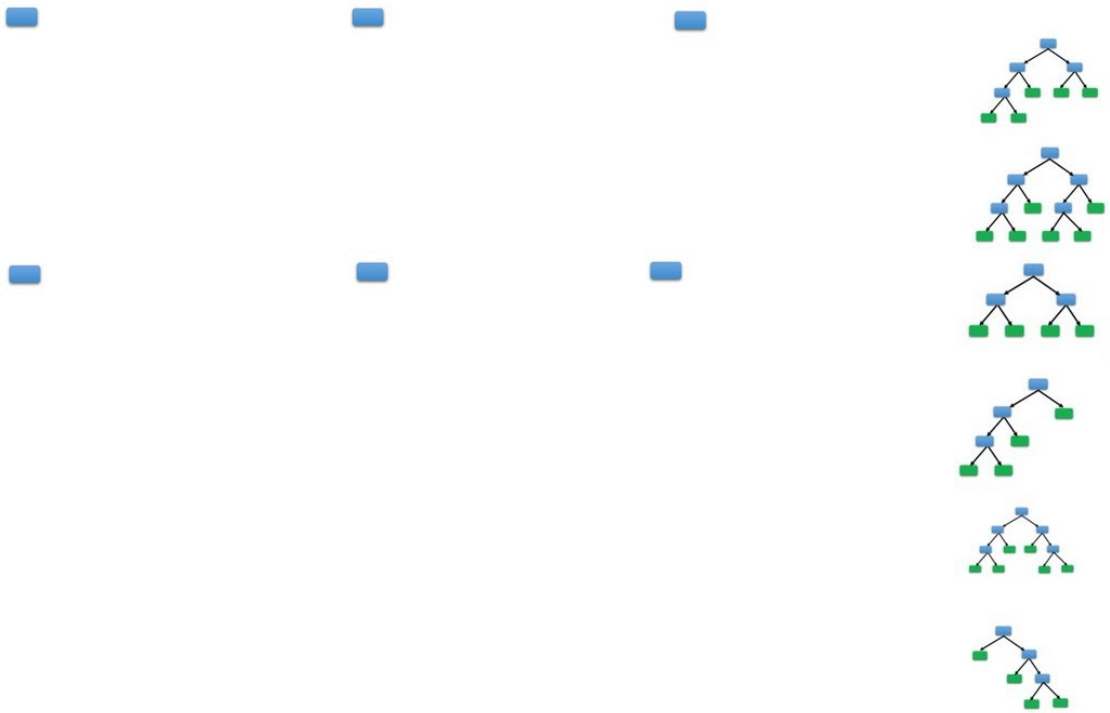
- 从树到森林
- 随机在何处
- 模型能力的度量
- 变量重要性度量

2.1

从搞笑视频到量化基金

- 找两个算命先生：
  - 一个大爷说25岁结婚，另一个说27岁
  - 为什么翻车了？
  - 搞笑弹幕：平均一下26岁嘛？
  - 算命先生“聪明”了：闪烁其词：25+，28-
  - 搞笑弹幕2：　(25,28)
- 世界是一个草台班子：
  - 量化基金怎么赚钱
  - 找一堆牛人写一堆牛的策略
  - 如何避免策略过强？
  - 策略过弱？
  - 避免尾大不掉？
- 把一堆策略放在一起做决策会怎么样？

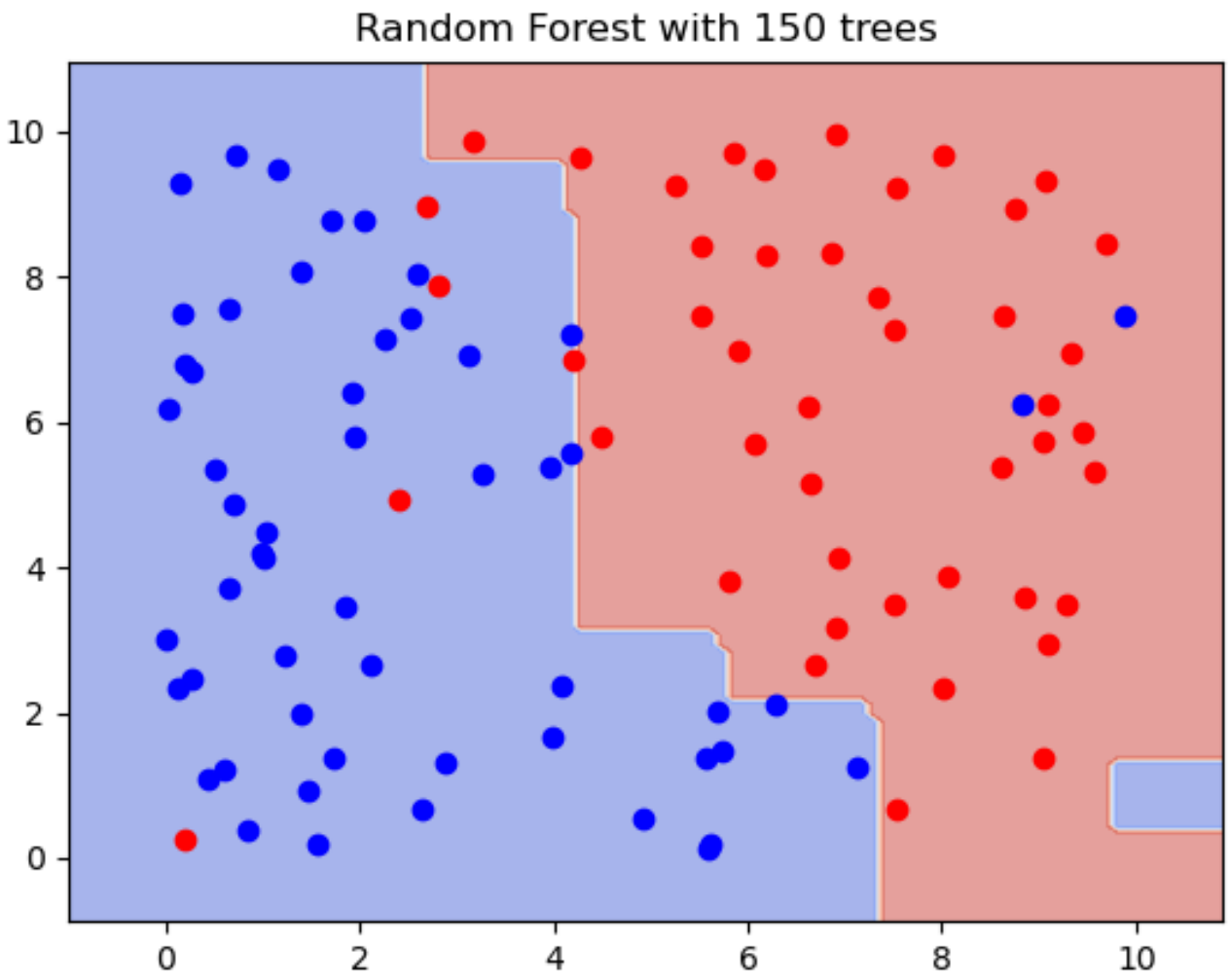
2.2 随机森林过程演示

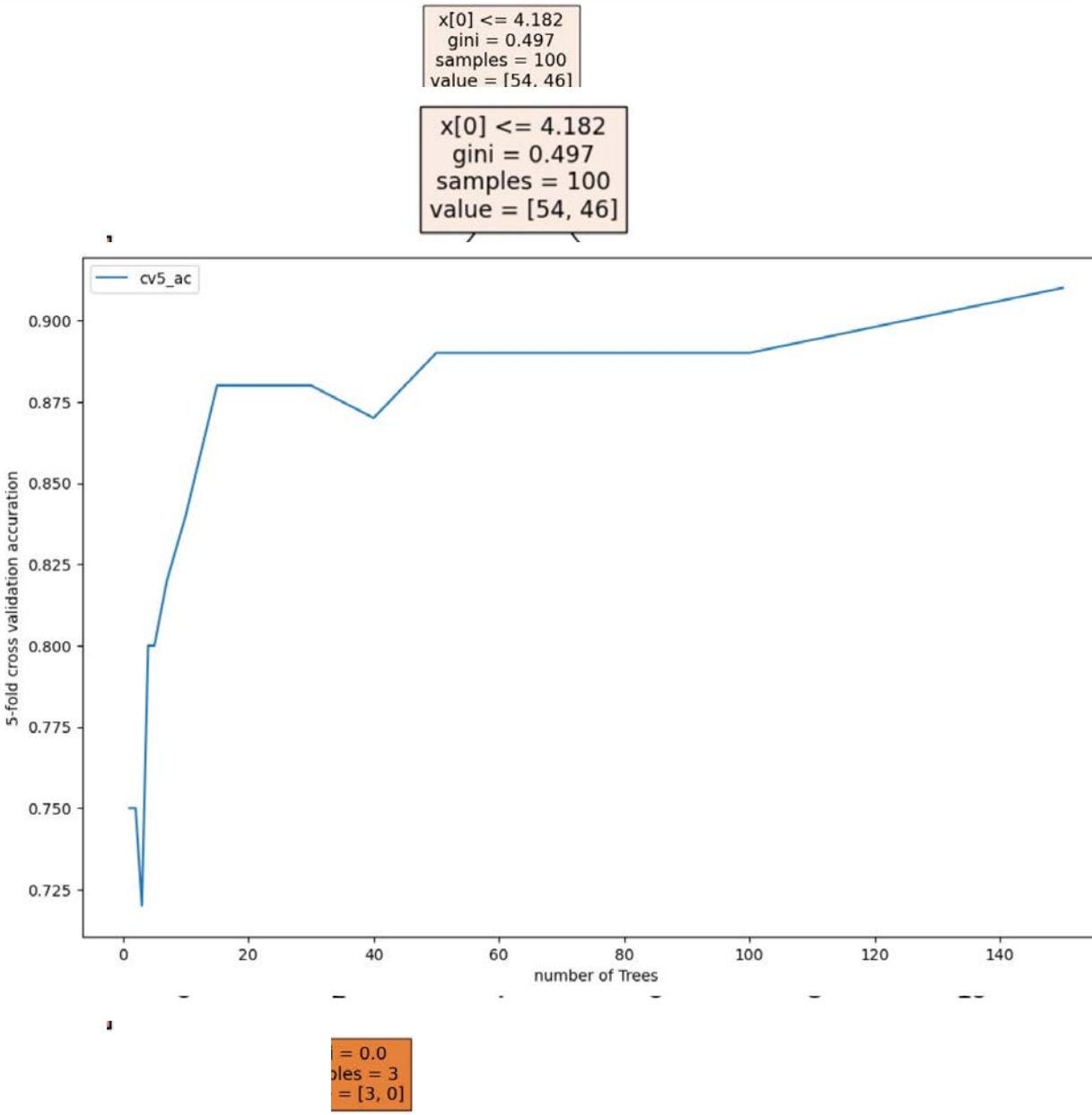


Random Forest in Action!!!

https://medium.com/x8-the-ai-community/random-forests-an-intuitive-understanding-13cece15ba88

2.1 从树到好多棵树


Random Forest with 150 trees
0
2
4
6
8
10


x[0] <= 4.182
gini = 0.497
samples = 100
value = [54, 46]
cv5_ac
5-fold cross validation accuration
number of Trees
0.900
0.875
0.850
0.825
0.800
0.775
0.750
0.725
0
20
40
60
80
100
120
140

2.1

三个臭皮匠的数学支撑

$$G(x) \equiv \frac{1}{T}\sum_{t=1}^{T} g_t(x)$$

未知目标函数
$f: X \to Y$

计算结果
$g \approx f$

$$avg((g_t(x) - f(x))^2) = avg(g_t^2 - 2g_t f + f^2)$$

$$= avg(g_t^2) - 2Gf + f^2$$

$$= avg(g_t^2) - 2Gf + f^2 + G^2 - G^2$$

$$= avg(g_t^2) + (G - f)^2 - G^2$$

- 森林G 表现强于 树g的期望

$$= avg(g_t^2) + (G - f)^2 - 2G^2 + G^2$$

$$= avg(g_t^2 - 2G^2 + G^2) + (G - f)^2$$

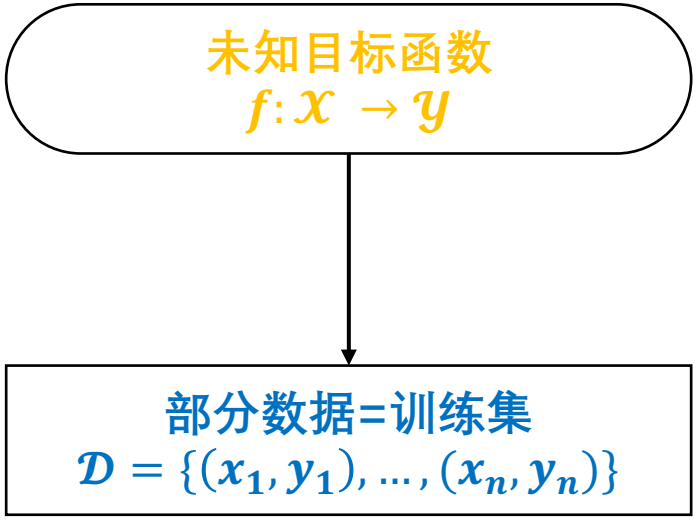
$$= avg\left((g_t - G)^2\right) + (G - f)^2$$

$$avg((g_t(x) - f(x))^2) \geq (G - f)^2 \qquad E_t(MSE(g_t - f)) \geq MSE(G - f)$$

2.1

如何找到许多棵树

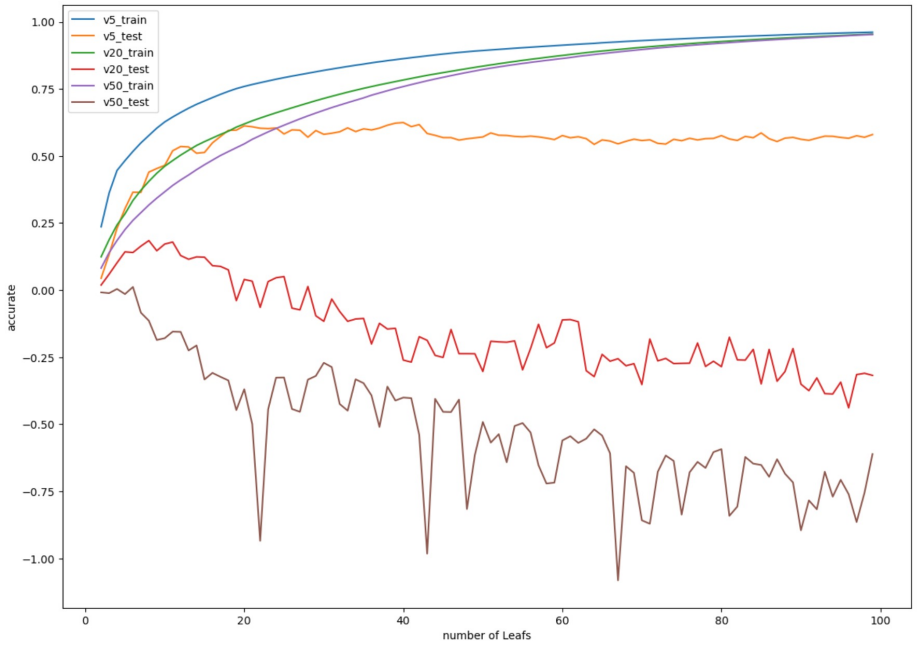
- 决策树的建立是确定的:
  - 树的规模
  - 树的损失函数
  - 数据能不能不一样?
  - 每次只选择一部分数据?
- Bootstrap方法

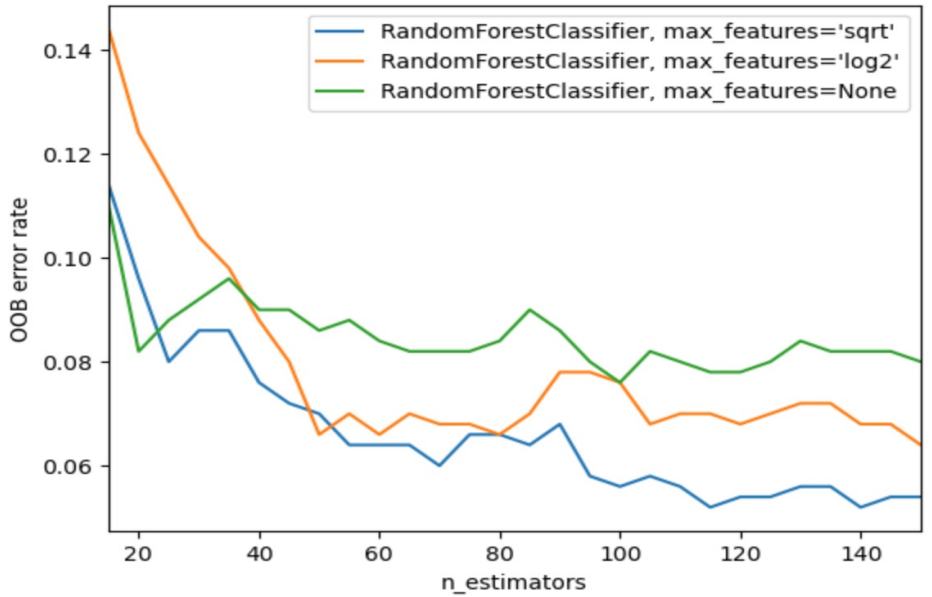


2.2

随机森林随机在哪里

- **数据**是随机的
- 想一想决策树的性质
- 我们每次只选择一部分**变量**？
  - 单棵树不能利用很多变量，但树的规模可以很大
  - Curse of dimension
  - $R^d \rightarrow R^{d'}$ $d' \ll d$空间映射
  - 原始RF，每次增加树的节点换一批d'
- **随机投影**：
  - $R^d \rightarrow R^{d'}$，本质上是一种投影，以原始值为轴
  - 如果轴换成多个原始变量的组合呢？
- 欢迎来到随机的世界
  - 虽然是白盒，位置随机，最优方向未知





Scikit learn sample

2.3

随机带来的好处

- 数据是随机的：bootstrap 抽取
- 共N条，每次抽取N'条数据
- 抽取T次，即一共生成了T棵树

	g_1	g_2	g_3	…	g_T
(**x1**,y1)	D1	*	D3		*
(**x2**,y2)	*	*	D3		DT
…					
(**xN**,yN)	D1	*	*		DT

- *意味着什么？
- g_t 没有碰过这笔数据→类似于样本外
- Out of Bag （OOB）
- 有多大概率成为OOB?

$$if\ N' = N$$

$$probability\ of\ being\ OOB$$

$$\left(1 - \frac{1}{N}\right)^{N}$$

$$\lim_{N\to\infty}\left(1 - \frac{1}{N}\right)^{\mathrm{N}} = \frac{1}{e} \approx 0.368$$

2.3

OOB 怎么做?

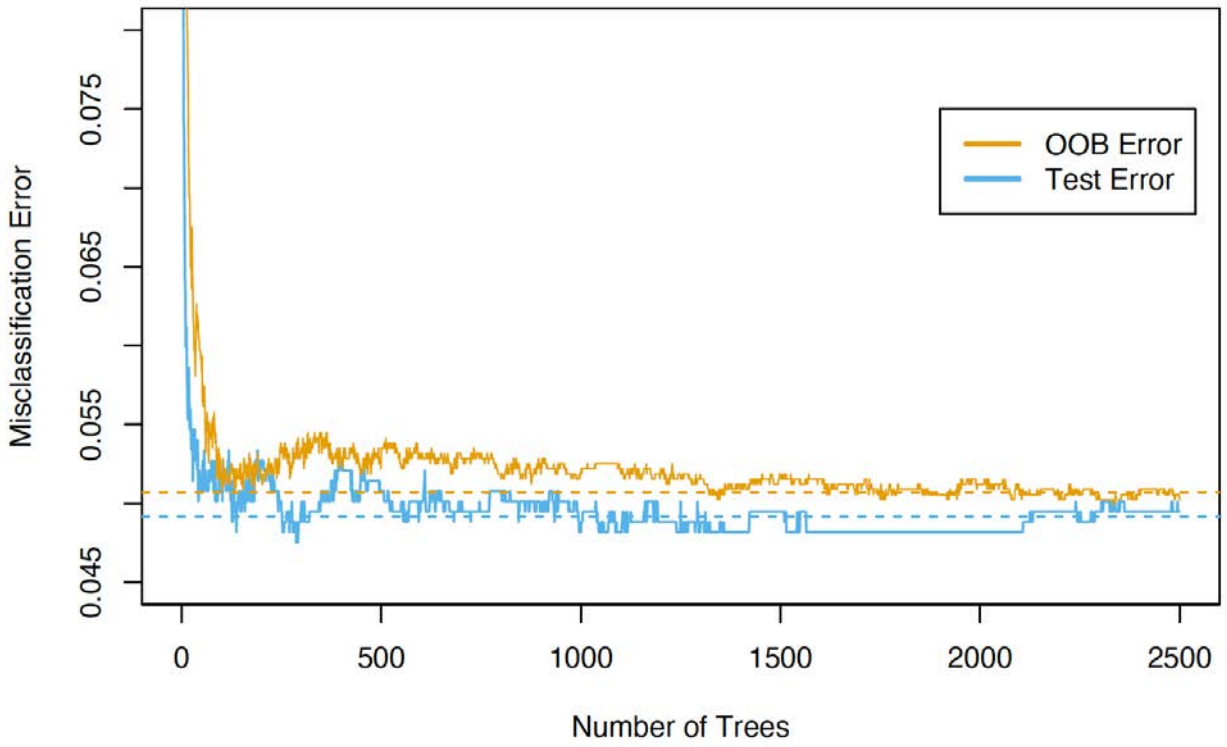
	g_1	g_2	g_3	…	g_T
(**x1**,y1)	D1	*	D3		*
(**x2**,y2)	*	*	D3		DT
…					
(**xN**,yN)	D1	*	*		DT

	g_1	g_2	g_3	…	g_T
(**x1**,y1)	Train	Train	Train		Train
(**x2**,y2)	**Test**	**Test**	**Test**		**Test**
…					
(**xN**,yN)	Train	Train	Train		Train

- 怎么用OOB，最朴素的想法，做g_t
- 直接做一部分的G^-,只要没碰过就好
- $G_1^- = average(g_2, \dots, g_T)$ $G_{\mathrm{N}}^- = average(g_2, g_3, \dots)$
- $\mathrm{Err}_{oob}(G) = \frac{1}{N}\sum_{n=1}^{N} err(y_n, G_n^-(\boldsymbol{x_n}))$
- 没事儿的，程序会算好，我们也不考卷子

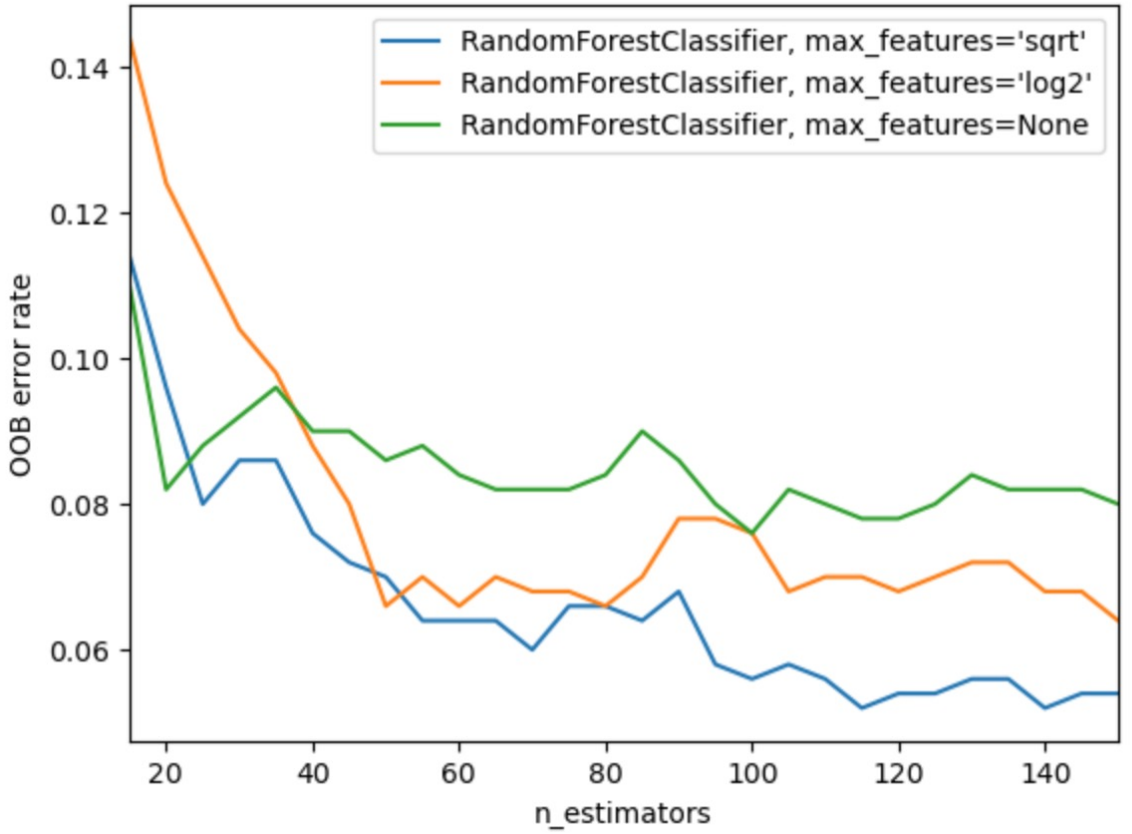
2.3 OOB 效果

OOB可以作为Test Error的平替



ESL p592

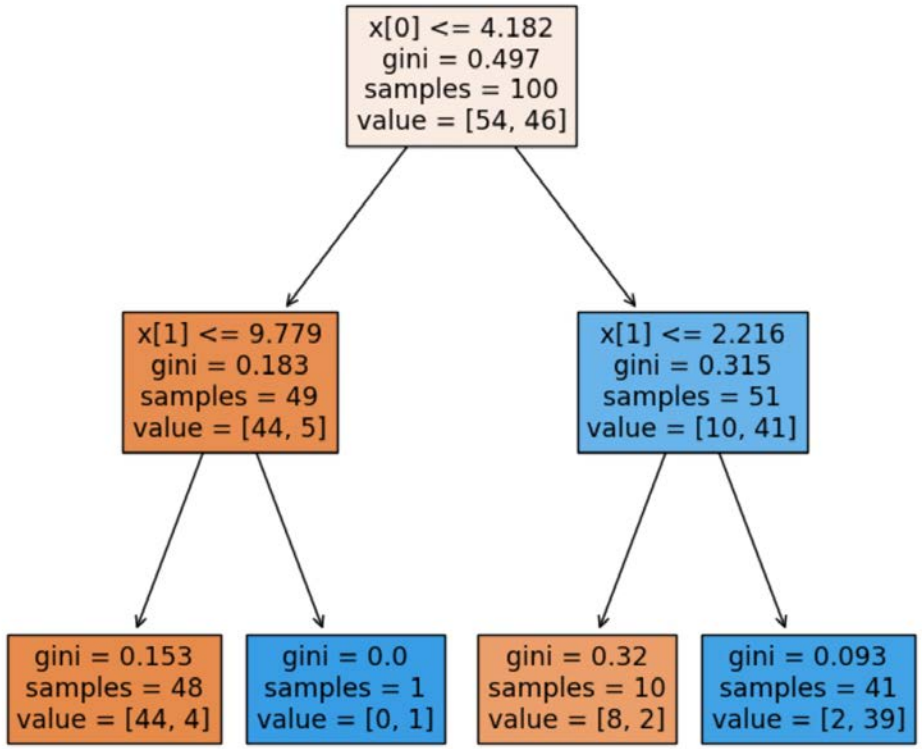
使用部分特征很重要



Scikit learn sample

2.4

变量重要性



- 线性回归：系数大小
- 这里有个问题
- Lasso

2.4

变量重要性：如果都是随机

- 都是随机怎么办？
- 实践出真知：量化基金怎么办？

- 分析师A很重要？
- 如果**没有**ta？不能直接开了

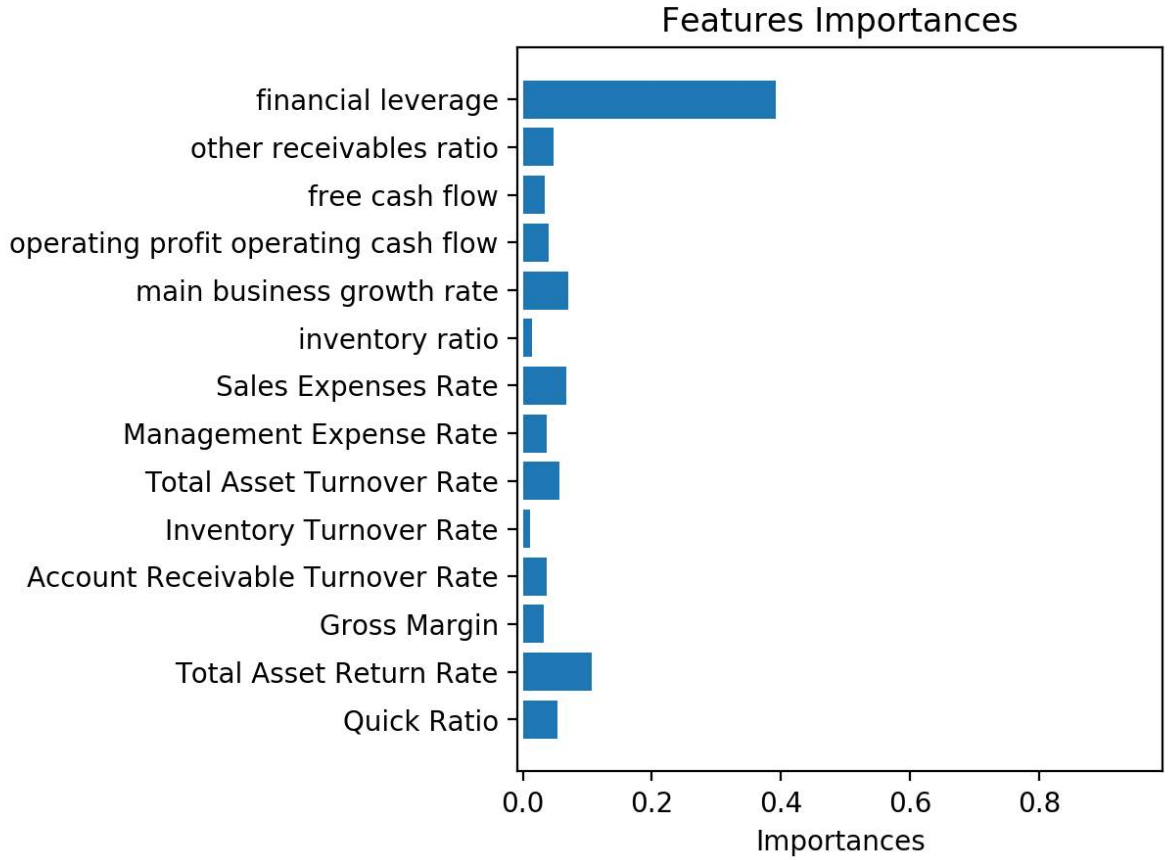
- 变量A很重要？
- 排除变量A？加噪音？
- 改变数据分布是糟糕的
- Permutation： 把这个变量扰乱

2.4

变量重要性：排列测试 permutation

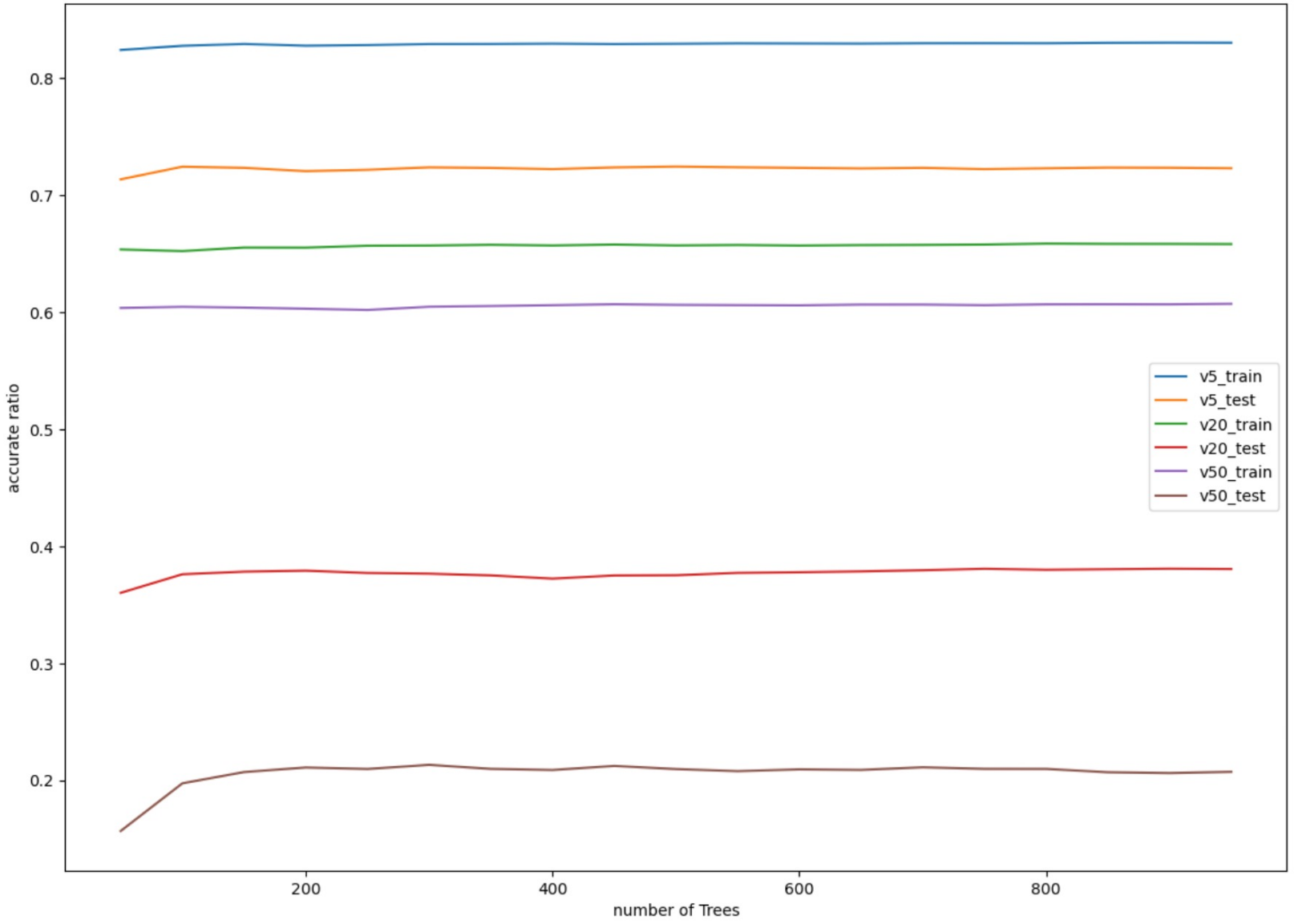
$$\mathrm{Import}ance(i) = Performance(D) - Performance(D_p^i)$$

- D_p^i是将变量i进行随机重新排列后的结果
- Performance 是算法的效果
- 拟合f的能力，一般是样本外误差
- 随机森林有没有捷径?
- 基于OOB的变量重要性测试
- 如果一个变量所有取值是相同的?



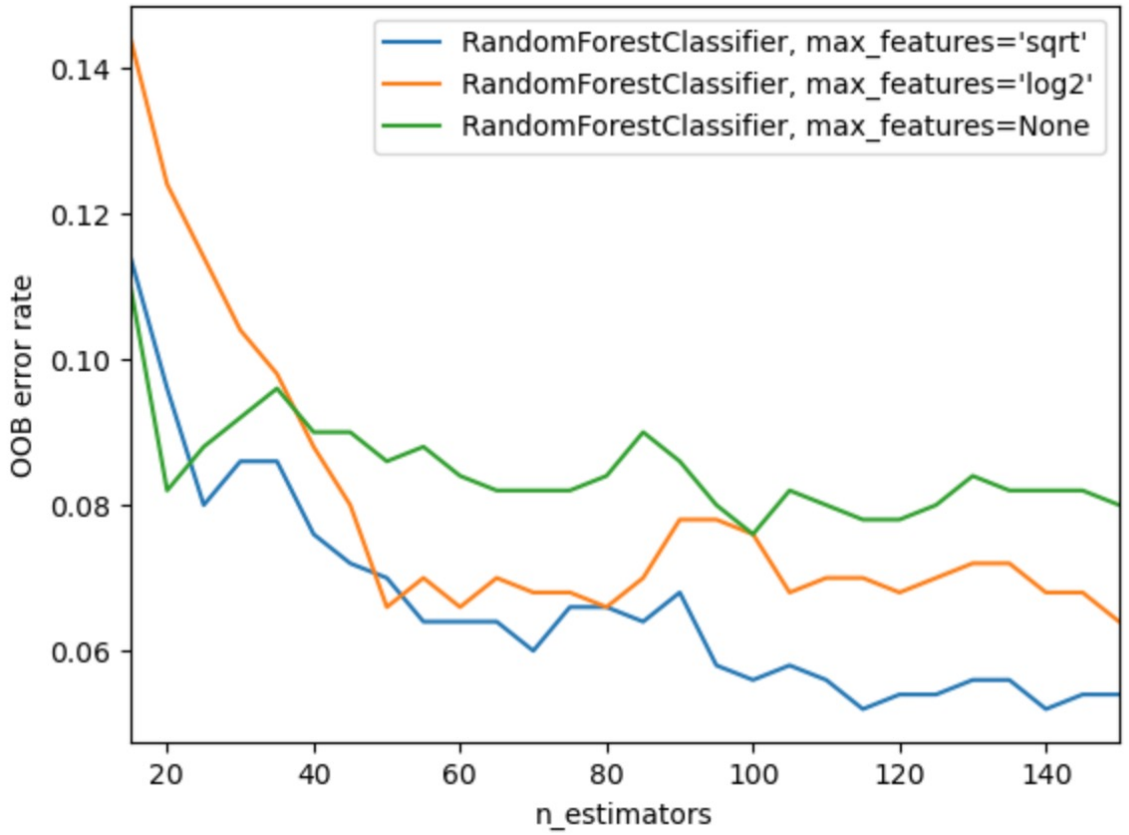
2.4

随机森林的力量


0.8
0.7
0.6
0.5
0.4
0.3
0.2
accurate ratio
200
400
600
800
number of Trees
v5_train
v5_test
v20_train
v20_test
v50_train
v50_test

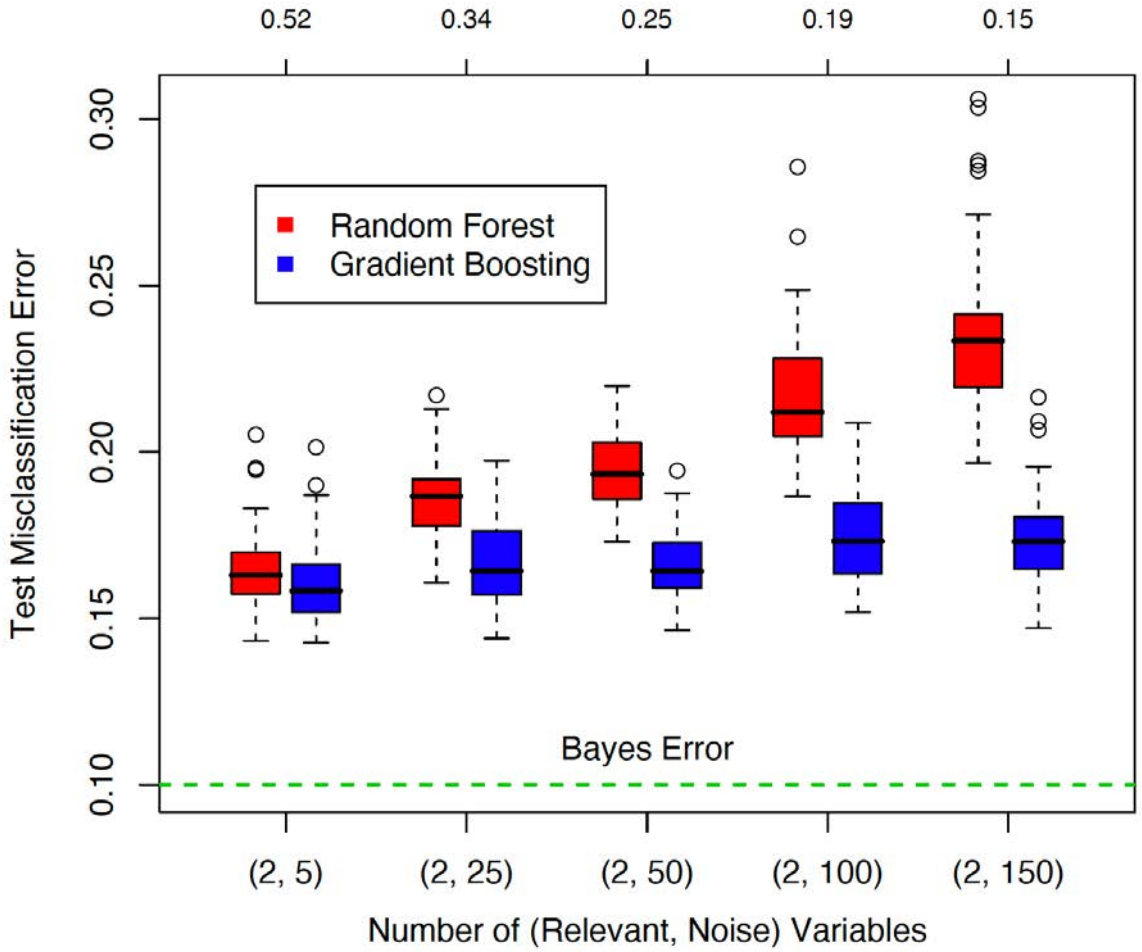
2.4 随机森林的正则化

使用部分特征很重要



Scikit learn sample

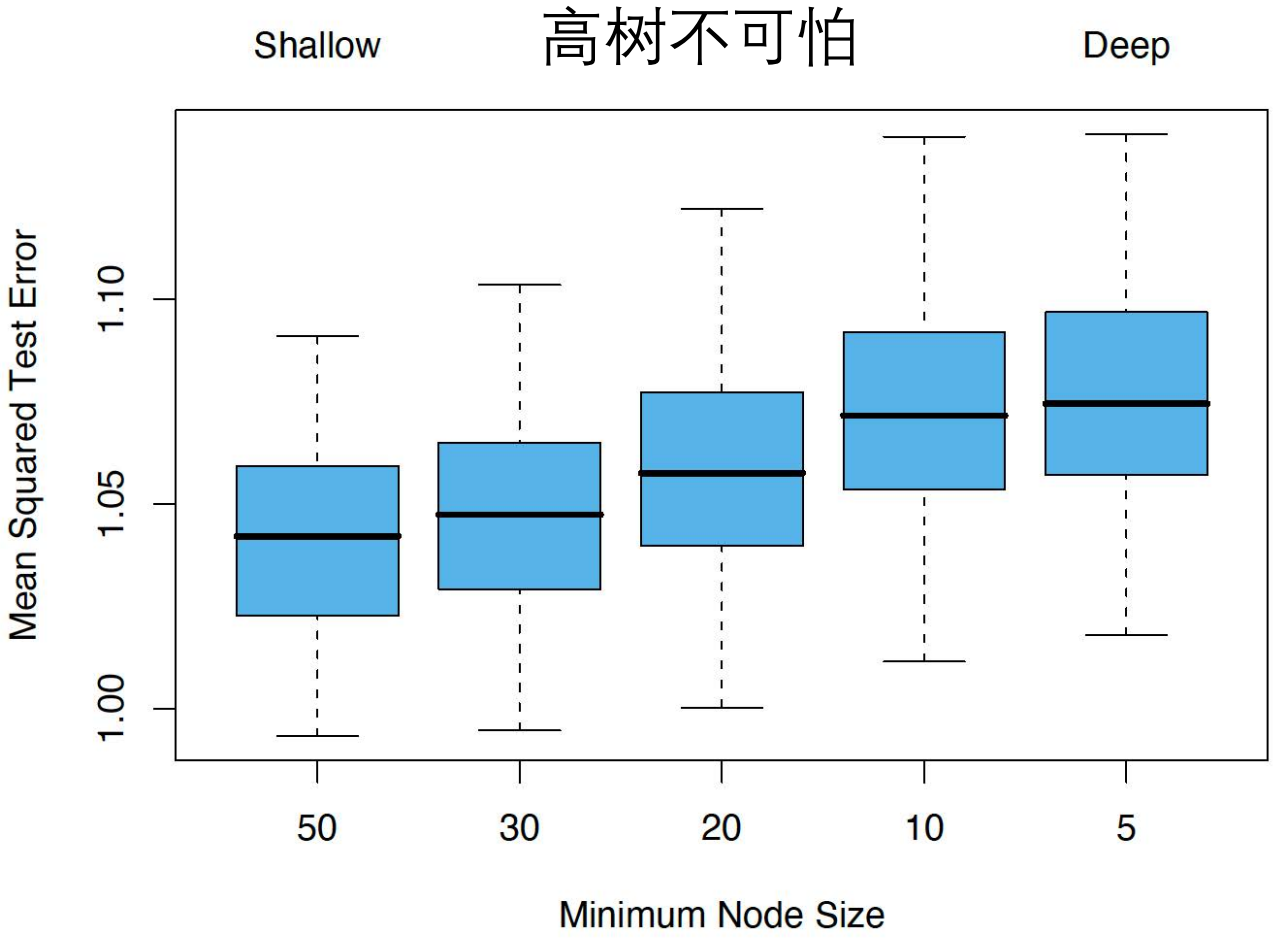
对噪音没有那么敏感



ESL p597

2.4

随机森林的正则化



ESL p598

- 最艰难的问题：变量个数
  - max_features函数
- 无关变量的问题
  - 近似于噪音
  - 没有必要去排除
  - 降维可以更好
- 树的限制？
  - max_depth 子树的最大深度
  - min_samples_split 可细分的节点下限
  - min_samples_leaf 最小叶子节点
  - max_samples 每次用的样本数
- 树的数量要不要限制？