第十一章　随机森林

11.1 简介

11.2 随机森林基本概况

11.3 随机森林基本理论

11.3.1 回归树基本理论

10.3.2 分类树基本理论

11.4 随机森林实践

11.1　简介

11.1 简 介

- 回归树和分类树的优点是解释性强且易于计算, 然而就像一个硬币具有两面, 基于单棵树的回归或者分类模型也具有明显的缺陷:
  - 1. 由于模型的简单性及仅用单一的常数作为最终区域的预测值, 从而使得单棵树的回归或分类模型很难具有最优的预测能力[71];
  - 2. 单棵树的预测是不稳定的[125] [70], 数据有较小的变动就可能会导致完全不一样的分裂变量和分裂点;
  - 3. 单棵树具有容易遭受选择偏差影响的问题, 即取值多的分类自变量比取值少的分类自变量更容易被选择为分裂变量[126] [127] [128].
- 为了克服单一统计模型 (单棵树) 稳定性差 (方差大) 的缺陷, 集成学习方法得到了广泛的研究和发展. 所谓集成学习, 就是指分类 (回归) 器的集成.集成学习通过构建并结合多个弱学习器来完成学习任务, 一般的方法是先产生一组个体学习器, 再用某种策略将它们结合起来, 常见的结合策略有平均法、投票法和学习法等. 例如 Bagging 方法利用 Bootstrap 方法 (有放回抽样) 对训练集进行抽样, 得到一系列新的训练集, 对每个训练集都构建一棵树,最后通过平均法、投票法组合所有预测器得到最终的预测模型.

11.1 简 介

- 后来, 为了克服单棵树的缺陷及降低每次 Bootstrap 抽样之间的相关性,Breiman [129] 综合以往集成学习的优缺点提出了一种新的集成学习方法——随机森林. 接下来我们将详细介绍随机森林的基本思想、算法步骤及变量重要性评价等内容.

11.2　随机森林基本概况

11.2 随机森林基本概况

- 在引入随机森林这一算法之前, 先考虑与之相关的概念——Bagging. Bagging 是 Bootstrap Aggregating 的缩写.
- 特别地, 我们考虑每个预测器都是决策树模型的 Bagging 算法. 在 Bagging 算法中, 我们注意到在对训练集 $N_0 = \{(\boldsymbol{X}_1, Y_1), \cdots, (\boldsymbol{X}_n, Y_n)\}$ 进行Bootstrap 抽样 (样本量为 n) 以获得 M 个新的训练集, 记为 $\{N_m, m = 1,2, \cdots, M\}$, 鉴于 Bootstrap 抽样的性质, 可以证明新训练集 $N_b, b \in \{1, 2, \cdots, M\}$大约只包含原训练集 N_0 的三分之二 (因为一个样本, 经 n 次有放回抽样, 仍未被抽中的概率是 $\lim\limits_{n\to\infty}\left(1-\dfrac{1}{n}\right)^n = \mathrm{e}^{-1} \approx \dfrac{1}{3}$). 这些未被使用的观测值称为此树的袋外观测值 (out-of-bag, OOB). 袋外观测值构成的集合称为袋外示例.
- 可以将袋外示例作为对应训练集生成的树的测试集来评估训练的结果,即可以用所有将第 i 个观测值作为 OOB 的树来预测第 i 个观测值的响应值.我们可以对这些预测响应值求平均 (回归情况下) 或执行多数投票 (分类情况下), 已得到第 i 个观测值的一个 OOB 预测. 用这种方法可以求出每个观测值的 OOB 预测, 根据这些就可以计算总体的 OOB 均方误差 (对回归问题)或分类误差 (对分类问题). 由此得到的 OOB 误差是对 Bagging 模型测试误差的有效估计. 我们将应用 OOB 误差来评估随机森林变量的重要性.

11.2 随机森林基本概况

- 随机森林的**基本思想**是：为了降低单棵树的缺陷及各次抽样间的相关性,随机森林采用有放回抽样, 每次抽取 n 个样本, 再无放回随机抽取 p 个属性(自变量) 中的 k(一般取 k 为$\sqrt{p}$) 个属性, 把这 k 个属性当成新的特征, 并结合因变量和抽取的 n 个样本, 生成一棵回归树或者分类树, 重复这一过程 M次得到 M 棵回归树或分类树, 随机森林是通过集成上述 M 棵回归树或分类树而成. 通过集成 M 棵树可以有效避免单棵树的不稳定性, 而每一棵树只用k 个属性代替 p 个属性来建模, 不仅可以有效降低树之间的相关性, 还能提高计算速度和节省计算机内存.

- 随机森林算法步骤如下：

  - (1) 从数据集 $(\boldsymbol{X}_1, Y_1), \cdots, (\boldsymbol{X}_n, Y_n)$ 中进行 Bootstrap 抽样 (有放回抽样), 抽取 n 个样本, 得到样本集 N_m;
  - (2) 利用 N_m 建立一棵决策树, 对于树上的每个节点, 重复以下步骤, 直到节点的样本数达到指定的最小限定值 nmin :
  - a) 从全部 p 个随机变量中随机取 $k(k < p)$ 个;
  - b) 从这 k 个变量中选取最优分裂变量, 将此节点分裂成两个子节点.
  - 注：对于分类问题, 构造每棵树时默认使用 $k = \sqrt{p}$个随机变量, 节点最小样本数为 1 ; 对于回归问题, 构造每棵树时默认使用 $k = \frac{p}{3}$ 个随机变量, 节点最小样本数为 5 .

11.2 随机森林基本概况

- (3) 重复以上过程 M 次, 得到 M 棵树构成一个随机森林.
- (4) 当对新样本进行预测时, 由每个决策树得到一个预测结果, 再进行平均或“投票”得出最后的结果. a) 对于回归问题, 最后的预测结果为所有决策树预测值的平均数; b) 对于分类问题, 最终的预测结果为所有决策树预测结果中最多的那类, 即采用“投票”得出最后的分类结果.

- 从随机森林的基本思想和算法可以看出, 随机森林具有以下的**特点**:
  - (1) 与其他的集成学习如 Bagging 相比, 由于每次只选取 $k(k < p)$ 个预测, 能够有效降低树间的相关性, 从而最大限度地减少预测方差, 提高预测的精度;
  - (2) 由于每次只用到 k 个自变量, 因此能有效节省计算时间和计算机内存;
  - (3) 与 Bagging 相比, 随机森林最大的不同就在于自变量子集的规模 k .若取$k=p$建立随机森林, 则等同于建立Bagging 树. 因此, Bagging 是随机森林的特例.
- **变量重要性评估**: 与所有集成学习方法一样, 随机森林很难得到自变量(特征) 与因变量间的一个直接的显式表达关系. 因而, 很难评估自变量的重要性. 考虑到随机森林只用到了部分自变量, Breiman[129] 建议通过如下方式来度量某个特征 X_j 的重要性:

11.2 随机森林基本概况

- (1) 根据未被抽取样本 OOB 计算随机森林中第 i 棵回归树的袋外误差e_i;
- (2) 随机打乱训练集在变量 X_j 所在列的取值顺序, 并计算新的袋外误差 e_i^j ;
- (3) 重复步骤直至计算出所有决策树的误差变化, 最后变量 X_j 预测误差的平均变化, 即重要性指标: $V(X_j)=\sum_{i=1}^{M}(e_i^j-e_i)^2/M.$

- 这里袋外误差是指我们使用针对某一棵树的袋外数据得到的预测误差的均值. 因为共有 M 棵树, 故而有 M 个袋外示例. 由上述袋外示例会生成M 个袋外误差 $e_i(\mathrm{i}=1,2,\cdots,M)$. 由此可知, 若特征变量 X_j 的变化引起重要性指标增加越大, 精度减少得越多, 则说明该变量越重要.

11.3　随机森林基本理论

11.3 随机森林基本理论

- 在上一节已经详细介绍了随机森林算法的基本思想、算法及与其他集成学习相比的优缺点等, 这一节我们将详细介绍随机森林相关的基本理论. 接下来我们将分别介绍回归树和分类树的基本理论.

11.3.1 回归树基本理论

- 假设随机森林是通过树的预测 $h(\boldsymbol{X}, \boldsymbol{\Theta})$ 生成的, 其中, $\boldsymbol{\Theta}$ 是 q 维随机向量, $h(\boldsymbol{X}, \boldsymbol{\Theta})$ 是 $\boldsymbol{R}^{p+q} \to \boldsymbol{R}$ 上的实值函数. 不妨假设 $(\boldsymbol{X}_1, Y_1), \cdots, (\boldsymbol{X}_n, Y_n)$独立同分布地来自 $(\boldsymbol{X}, Y)$, 并定义均方误差为 $E_{\boldsymbol{X},Y}(Y - h(\boldsymbol{X}, \boldsymbol{\Theta}))^2$.
- 通过对 M 棵单一回归树 $h(\boldsymbol{X}, \boldsymbol{\Theta}_i)$ 取平均来生成随机森林的预测, 当$M \to \infty$ 时, 我们可以得到定理 11.1的结论, 即随机森林预测是均方收敛的.
- **定理 11.1** 随着随机森林中树的数目趋向无穷, 如下结论几乎处处成立:

$$\lim_{M\to\infty} E_{\boldsymbol{X},Y}\left(Y - \frac{1}{M}h\left(\boldsymbol{X}, \boldsymbol{\Theta}_i\right)\right)^2 \to E_{\boldsymbol{X},Y}\left(Y - E_{\Theta}h\left(\boldsymbol{X}, \boldsymbol{\Theta}\right)\right)^2.$$

- 由定理 11.1可知当树的数目 $M \to \infty$ 时, 随机森林的预测误差趋向于总体预测误差. 接下来我们将推导出随机森林预测误差的上界, 其结果在定理11.2中给出.

11.3.1 回归树基本理论

- 定理 11.2 假设对所有 $\boldsymbol{\Theta}$, $E_Y = E_{\boldsymbol{X}}(h(\boldsymbol{X}, \boldsymbol{\Theta}))$, 则如下结论成立:

$$E_{\boldsymbol{X},Y}\left[E_{\boldsymbol{\Theta}}\left(Y-h\left(\boldsymbol{X},\boldsymbol{\Theta}\right)\right)\right]^2 \leq \rho E_{\boldsymbol{\Theta}}\left[E_{\boldsymbol{X},Y}\left(Y-h\left(\boldsymbol{X},\boldsymbol{\Theta}\right)\right)^2\right],$$

  - 其中, ρ 是 $Y-h(\boldsymbol{X}, \boldsymbol{\Theta})$ 和 $Y-h(\boldsymbol{X}, \boldsymbol{\Theta}')$ 间的加权相关系数, $\boldsymbol{\Theta}$ 和 $\boldsymbol{\Theta}'$ 相互独立.

- **证明**

$$E_{\boldsymbol{X},Y}\left[E_{\boldsymbol{\Theta}}\left(Y-h\left(\boldsymbol{X},\boldsymbol{\Theta}\right)\right)\right]^2 = E_{\boldsymbol{\Theta}}E_{\boldsymbol{\Theta}'}E_{\boldsymbol{X},Y}\left(Y-h\left(\boldsymbol{X},\boldsymbol{\Theta}\right)\right)\left(Y-h\left(\boldsymbol{X},\boldsymbol{\Theta}'\right)\right), \quad (11.3.1)$$

  - 式 (11.3.1) 的右边是一个协方差并且可以写成:

$$E_{\boldsymbol{\Theta}}E_{\boldsymbol{\Theta}'}\left(\rho\left(\boldsymbol{\Theta},\boldsymbol{\Theta}'\right)\mathrm{sd}\left(\boldsymbol{\Theta}\right)\mathrm{sd}\left(\boldsymbol{\Theta}'\right)\right), \quad (11.3.2)$$

  - 其中, $\mathrm{sd}(\boldsymbol{\Theta}) = \sqrt{E_{\boldsymbol{X},Y}\left(Y-h\left(\boldsymbol{X},\boldsymbol{\Theta}\right)\right)^2}$. 加权相关系数的定义为

$$\rho = E_{\boldsymbol{\Theta}}E_{\boldsymbol{\Theta}'}\left(\rho\left(\boldsymbol{\Theta},\boldsymbol{\Theta}'\right)\mathrm{sd}\left(\boldsymbol{\Theta}\right)\mathrm{sd}\left(\boldsymbol{\Theta}'\right)\right) / \left(E_{\boldsymbol{\Theta}}\mathrm{sd}\left(\boldsymbol{\Theta}\right)\right)^2. \quad (11.3.3)$$

11.3.1 回归树基本理论

▶ 因此,

$$E_{\boldsymbol{X},Y}\left[E_{\boldsymbol{\Theta}}\left(Y-h\left(\boldsymbol{X},\boldsymbol{\Theta}\right)\right)\right]^2=\rho\left(E_{\boldsymbol{\Theta}}\mathrm{sd}\left(\boldsymbol{\Theta}\right)\right)^2$$
$$\leq \rho E_{\boldsymbol{\Theta}}\left[E_{\boldsymbol{X},Y}\left(Y-h\left(\boldsymbol{X},\boldsymbol{\Theta}\right)\right)^2\right]. \tag{11.3.4}$$

▶ 证毕.

- 由定理 11.2可以看出随机森林预测的准确性取决于单棵树的预测能力及树之间相关性的强弱. 补偿一点, 树 $\boldsymbol{\Theta}$ 和 $\boldsymbol{\Theta}'$ 的建立是相互独立的, 但两颗树在拟合数据, 即误差表现 $Y-h(\boldsymbol{X},\boldsymbol{\Theta})$ 和 $Y-h(\boldsymbol{X},\boldsymbol{\Theta}')$ 有相关性.

11.3.2 分类树基本理论

- 假设随机森林是树型分类器 $\{h(\boldsymbol{X}, \boldsymbol{\Theta}_i), i = 1, 2, \cdots, M\}$ 的集合, 其中,$\boldsymbol{X}$ 是预测向量; $\boldsymbol{\Theta}_i$ 是独立同分布的随机向量, 决定了单棵分类树的生成过程; 元分类器 $h(\boldsymbol{X}, \boldsymbol{\Theta}_i)$ 是用 CART[67] 算法构建的无剪枝的分类决策树. 则当 M 趋向无穷时, 对分类树也是收敛的, 即有如下结论:

- **定理 11.3** 随着随机森林中树的数目趋向无穷, 如下结论成立:

$$\lim_{M\to\infty} P_{\boldsymbol{X},Y}\left(\left[\frac{1}{M}\sum_{i=1}^{M} I\left(h\left(\boldsymbol{X},\boldsymbol{\Theta}_i\right)=Y\right)-\max_{j\neq Y}\frac{1}{M}I\left(h\left(\boldsymbol{X},\boldsymbol{\Theta}_i\right)=j\right)\right]<0\right)$$

$$\to P_{\boldsymbol{X},Y}\left(\left[P_{\boldsymbol{\Theta}}\left(h\left(\boldsymbol{X},\boldsymbol{\Theta}\right)=Y\right)-\max_{j\neq Y}P_{\boldsymbol{\Theta}}\left(h\left(\boldsymbol{X},\boldsymbol{\Theta}\right)=j\right)\right]<0\right),$$

  - 其中, $I(\cdot)$ 是示性函数.

- **证明** 容易证明, 参数空间$\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\cdots,\boldsymbol{\Theta}_M$上存在一个零概率集合 C,在 C 之外, 对于所有的 $\boldsymbol{X}$, 有下式成立

$$\frac{1}{M}\sum_{i=1}^{M} I\left(h\left(\boldsymbol{X},\boldsymbol{\Theta}_i\right)=g\right)\to P_{\boldsymbol{\Theta}}\left(h\left(\boldsymbol{X},\boldsymbol{\Theta}\right)=g\right).$$

11.3.2 分类树基本理论

- 在一个固定的训练集和参数空间 $\boldsymbol{\Theta}$ 上, 所有满足 $h(\boldsymbol{\Theta}, \boldsymbol{X}) = g$ 的 $\boldsymbol{X}$ 构成的集合是一个超矩形单元. 对于所有 $h(\boldsymbol{\Theta}, \boldsymbol{X})$ 只有有限的 K 个这种超矩阵单元, 记作 $S_1, \cdots, S_K$. 若 $\{\boldsymbol{X} : h(\boldsymbol{\Theta}, \boldsymbol{X}) = g\} = S_g$, 此时定义 $\phi(\boldsymbol{\Theta}) = g$, 并令$N_g$ 为前 N 次实验中 $\phi(\boldsymbol{\Theta}_i) = g$ 的次数. 那么有

$$\frac{1}{M}\sum_{i=1}^{M} I\left(h\left(\boldsymbol{X}, \boldsymbol{\Theta}_i\right) = g\right) = \frac{1}{M}\sum_{g=1}^{M} N_g I\left(\boldsymbol{X} \in S_g\right).$$

- 再由大数定理可得

$$N_g = \frac{1}{M}\sum_{i=1}^{M} I\left(\phi\left(\boldsymbol{\Theta}_i\right) = g\right),$$

- 会收敛到 $P_{\boldsymbol{\Theta}}(\phi(\boldsymbol{\Theta}) = g)$. 对于 g 的某个值, 所有集合的并集都不会发生收敛,得到一个概率为零的集合 C, 因此在 C 之外有

$$\frac{1}{M}\sum_{i=1}^{M} I\left(h\left(\boldsymbol{X}, \boldsymbol{\Theta}_i\right) = g\right) \to \sum_{g=1}^{M} P_{\boldsymbol{\Theta}}\left(\phi\left(\boldsymbol{\Theta}\right) = g\right) N_g I\left(\boldsymbol{X} \in S_g\right).$$

- 上式右边即是 $P_{\boldsymbol{\Theta}}(h(\boldsymbol{X}, \boldsymbol{\Theta}) = g)$. 这样我们就证明了定理.

11.3.2 分类树基本理论

- 由定理 11.3可知, 随着随机森林中树数量增加, 模型的分类误差上限趋于一个固定值. 即随机森林不会随着分类树数目的增加而产生过度拟合的问题, 将对未知实例预测提供较好的参考思路和应用性. 类似定理 11.2, 我们将给出随机森林分类误差的一个上界, 为叙述方便, 先给出如下定义:

- 给定一组分类器 $h(\boldsymbol{X}, \boldsymbol{\Theta}_1), h(\boldsymbol{X}, \boldsymbol{\Theta}_2), \cdots, h(\boldsymbol{X}, \boldsymbol{\Theta}_M)$, 并使用从随机向量$(\boldsymbol{X},Y)$ 的分布中随机抽取的训练集, 将边际函数定义为

$$\mathrm{mg}(\boldsymbol{X},Y)=\frac{1}{M}\sum_{i=1}^{M} I\left(h(\boldsymbol{X},\boldsymbol{\Theta}_i)=Y\right)-\max_{j\neq Y}\frac{1}{M}\sum_{i=1}^{M} I\left(h(\boldsymbol{X},\boldsymbol{\Theta}_i)=j\right),$$

  - 其中, I(·) 是示性函数.

- 边际函数衡量的是正确分类在 $(\boldsymbol{X},Y)$ 的平均投票数超过任何其他分类的平均投票数的程度. 差距越大, 对分类的信心就越大. 一般泛化误差由下式给出:

$$\mathrm{PE}^{*}=P_{X,Y}\left(\mathrm{mg}(\boldsymbol{X},Y)<0\right).$$

11.3.2 分类树基本理论

- 随机森林的边际函数为

$$\mathrm{mg}(\boldsymbol{X},Y)=P_{\boldsymbol{\Theta}}\left(h(\boldsymbol{X},\boldsymbol{\Theta})=Y\right)-\max_{j\neq Y}P_{\boldsymbol{\Theta}}\left(h(\boldsymbol{X},\boldsymbol{\Theta})=j\right).$$

- 分类器集 $h(\boldsymbol{X},\boldsymbol{\Theta})$ 的强度定义为

$$s=E_{\boldsymbol{X},Y}\mathrm{mg}(\boldsymbol{X},Y).$$

- 不妨设 $s\geqslant 0$, 由切比雪夫不等式可得下式成立:

$$\mathrm{PE}^{*}\leq \mathrm{Var}\left(\mathrm{mg}(\boldsymbol{X},Y)\right)/s^{2}. \tag{11.3.5}$$

- $\mathrm{mg}(\boldsymbol{X},Y)$ 方差的一个显式表达为:

$$\hat{\Theta}(\boldsymbol{X},Y)=\arg\max_{j\neq Y}P_{\boldsymbol{\Theta}}\left(h(\boldsymbol{X},\boldsymbol{\Theta})=j\right),$$

- 故

$$\begin{aligned}\mathrm{mg}(\boldsymbol{X},Y)&=P_{\boldsymbol{\Theta}}\left(h(\boldsymbol{X},\boldsymbol{\Theta})=Y\right)-P_{\boldsymbol{\Theta}}\left(h(\boldsymbol{X},\boldsymbol{\Theta})=\hat{\Theta}(\boldsymbol{X},Y)\right)\\&=E_{\boldsymbol{\Theta}}\left[I\left(h(\boldsymbol{X},\boldsymbol{\Theta})=Y\right)-I\left(h(\boldsymbol{X},\boldsymbol{\Theta})=\hat{\Theta}(\boldsymbol{X},Y)\right)\right]\end{aligned}$$

11.3.2 分类树基本理论

- 原始边际函数定义为

$$\mathrm{rmg}(\boldsymbol{\Theta}, \boldsymbol{X}, Y) = I\left(h(\boldsymbol{X}, \boldsymbol{\Theta}) = Y\right) - I\left(h(\boldsymbol{X}, \boldsymbol{\Theta}) = \hat{Y}(\boldsymbol{X})\right),$$

- 因此, $\mathrm{mg}(\boldsymbol{X}, Y)$ 是 $\mathrm{rmg}(\boldsymbol{\Theta}, \boldsymbol{X}, Y)$ 关于 $\boldsymbol{\Theta}$ 的期望. 对于任意函数 f, 当 $\boldsymbol{\Theta}$, $\boldsymbol{\Theta}'$ 独立同分布时, 有下式成立:

$$\left[E_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta})\right]^2 = E_{\boldsymbol{\Theta}, \boldsymbol{\Theta}'} f(\boldsymbol{\Theta}) f(\boldsymbol{\Theta}'),$$

- 即有

$$\mathrm{mg}(\boldsymbol{X}, Y) = E_{\boldsymbol{\Theta}, \boldsymbol{\Theta}'} \mathrm{rmg}(\boldsymbol{\Theta}, \boldsymbol{X}, Y) \mathrm{rmg}(\boldsymbol{\Theta}', \boldsymbol{X}, Y) \tag{11.3.6}$$

- 再由式 (11.3.2) 可得

$$\begin{aligned} \mathrm{Var}\left(\mathrm{mg}(\boldsymbol{X}, Y)\right) &= E_{\boldsymbol{\Theta}, \boldsymbol{\Theta}'}\left(\mathrm{Cov}_{X,Y}\left(\mathrm{rmg}(\boldsymbol{\Theta}, \boldsymbol{X}, Y) \mathrm{rmg}(\boldsymbol{\Theta}', \boldsymbol{X}, Y)\right)\right) \\ &= E_{\boldsymbol{\Theta}, \boldsymbol{\Theta}'}\left(\rho(\boldsymbol{\Theta}, \boldsymbol{\Theta}') \mathrm{sd}(\boldsymbol{\Theta}) \mathrm{sd}(\boldsymbol{\Theta}')\right), \end{aligned}$$

- 其中, $\rho(\boldsymbol{\Theta}, \boldsymbol{\Theta}')$ 是 $\mathrm{rmg}(\boldsymbol{\Theta}, \boldsymbol{X}, Y)$ 和 $\mathrm{rmg}(\boldsymbol{\Theta}', \boldsymbol{X}, Y)$ 间的相关系数, $\mathrm{sd}(\boldsymbol{\Theta})$ 是 $\mathrm{rmg}(\boldsymbol{\Theta}, \boldsymbol{X}, Y)$ 的标准差.

11.3.2 分类树基本理论

- 那么,

$$\mathrm{Var}\left(\mathrm{mg}\left(\boldsymbol{X},Y\right)\right)=\rho\left(E_{\boldsymbol{\Theta}}\mathrm{sd}\left(\boldsymbol{\Theta}\right)\right)^{2}\leq\rho E_{\boldsymbol{\Theta}}\mathrm{Var}\left(\boldsymbol{\Theta}\right). \tag{11.3.7}$$

- 其中, ρ 是相关系数的平均值, 即

$$E_{\boldsymbol{\Theta}}\mathrm{Var}\left(\boldsymbol{\Theta}\right)\leq E_{\boldsymbol{\Theta}}\left(E_{\boldsymbol{X},Y}\mathrm{rmg}\left(\boldsymbol{\Theta},\boldsymbol{X},Y\right)\right)^{2}-s^{2}\leq 1-s^{2}. \tag{11.3.8}$$

- 由式 (11.3.7) 及 (11.3.8) 可得到随机森林分类误差的上界.

定理 11.4 随机森林泛化误差的上界由下式给出

$$\mathrm{PE}^{*}\leq\frac{\rho\left(1-s^{2}\right)}{s^{2}}.$$

由定理 11.4 可知, 分类错误的上界受随机森林中单个分类器强弱及分类间的相关性影响.

11.4　随机森林实践

实践代码