金融机器学习算法

第六讲

金融机器学习的模型集成

本讲主要内容

- 学习本章的动机
- 模型集成视角下的偏差-方差分解
- 装袋式集成
- 助推式集成

学习本章的动机

- 模型集成是把一群算法相似的弱学习模型组织起来，生成一个强学习模型
- 为什么模型集成能够生成强学习器?
- 在金融中运用模型集成要注意哪些问题?

模型集成视角下的偏差-方差分解

复习偏差-方差分解

$$
\begin{aligned}
&\mathbb{E}(\widehat{f_i} - y_i)^2 \\
=&\mathbb{E}(\widehat{f_i} - \bar{f_i} + \bar{f_i} - y_i)^2 \\
=&\mathbb{E}(\widehat{f_i} - \bar{f_i})^2 + \mathbb{E}(\bar{f_i} - y_i)^2 \\
=&\mathbb{E}(\widehat{f_i} - \bar{f_i})^2 + \mathbb{E}(\bar{f_i} - f_i + f_i - y_i)^2 \\
=&\mathbb{E}(\widehat{f_i} - \bar{f_i})^2 + \mathbb{E}(\bar{f_i} - f_i)^2 + \mathbb{E}(f_i - y_i)^2 \\
=&Var(\widehat{f_i}) + \mathbb{E}(\mathbb{E}(\widehat{f_i}) - f_i)^2 + \mathbb{E}(y_i - f_i)^2 \\
=&[\mathbb{E}(\widehat{f_i} - f_i)]^2 + Var(\widehat{f_i}) + \sigma_\epsilon^2
\end{aligned}
$$

模型集成视角下的偏差-方差分解

- 偏差: 度量拟合能力，不现实的假设导致此值变大。机器学习算法未能识别特征和结果之间的重要关系时贡献偏差。
- 方差: 表示模型对数据的敏感性。算法错误地将噪声误认为信号，而非建模训练集中的一般模式时贡献方差
- 噪声: 表示学习的难度。这是无法由任何模型解释的不可减少的。
- 模型集成视角： 进行模型集成可以有效的减小偏差或者方差。

装袋式集成

装袋式集成 (Bagging) 的原理如下：

- 通过有放回的随机抽样生成 N 个训练数据集。
- 使用 N 个估计器，分别从一个训练集中拟合一个估计器。这些估计器是相互独立的，因此可以并行拟合模型。
- 模型集成：
  - 连续因变量：从 N 个模型中得到的单个预测进行简单平均。
  - 分类因变量：由对这个观察值进行分类为该类别的估计器数量所占比例（投票数）确定，也可以计算平均概率值

装袋式集成：降低方差

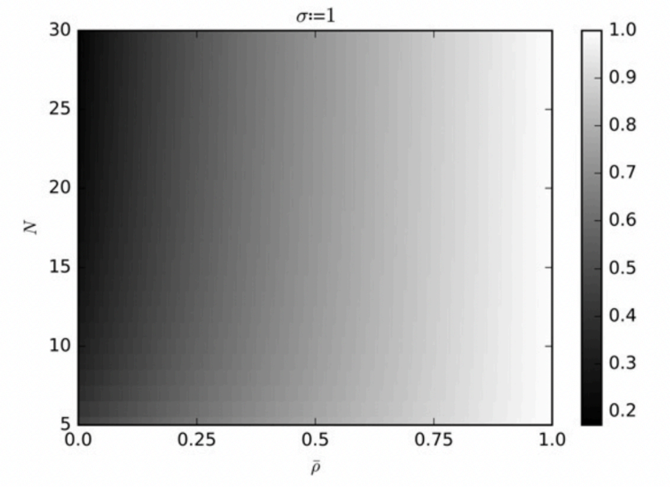
装袋式集成的最大优势就是降低模型的方差，从而缓解过拟合 $\varphi_i[c]$：第 i 个集成模型的预测，各预测间的平均关联为 $\bar{\rho}$

$$
\begin{aligned}
\mathrm{V}\left[\frac{1}{N}\sum_{i=1}^{N}\varphi_i[c]\right] &= \frac{1}{N^2}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\sigma_{i,j}\right) = \frac{1}{N^2}\sum_{i=1}^{N}\left(\sigma_i^2 + \sum_{j\neq i}^{N}\sigma_i\sigma_j\rho_{i,j}\right) \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\left(\bar{\sigma}^2 + \underbrace{\sum_{j\neq i}^{N}\bar{\sigma}^2\bar{\rho}}_{\substack{=(N-1)\bar{\sigma}^2\bar{\rho} \\ \text{for a fixed } i}}\right) = \frac{\bar{\sigma}^2 + (N-1)\bar{\sigma}^2\bar{\rho}}{N} \\
&= \bar{\sigma}^2\left(\bar{\rho} + \frac{1-\bar{\rho}}{N}\right)
\end{aligned}
$$

装袋式集成：降低方差

- $\bar{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2$
- $\bar{\rho} = \frac{\sum_{i\neq j}^{N}\sigma_i\sigma_j\rho_{i,j}}{\bar{\sigma}^2 N(N-1)}$
- $\mathbb{V}\mathrm{ar}\left[\frac{1}{N}\sum_{i=1}^{N}\varphi_i[c]\right] = \bar{\sigma}^2\left(\bar{\rho} + \frac{1-\bar{\rho}}{N}\right)$
- 当 $\bar{\rho} < 1$ 可以降低方差

装袋式集成：降低方差 N 与 ρ



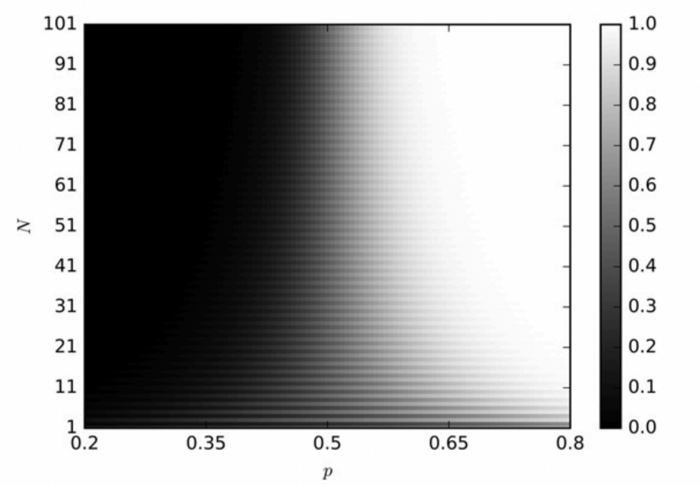
装袋式集成：降低偏差

对于分类模型，偏差相当于准确率的数学期望。设单个模型的准确率为 p, 若有 k 个类别，按照少数服从多数原则，分类正确的必要条件为

$$\mathbb{P}\left(Vote_i > \frac{N}{k}\right) = 1 - \mathbb{P}\left(Vote_i \leq \frac{N}{k}\right) = 1 - \sum_{i=0}^{\lfloor N/k \rfloor} C_N \, p^i (1-p)^{N-i}$$

- $N > \frac{p}{(p-\frac{1}{k})^2}$, 则如果 $p > \frac{1}{k}$, $\mathbb{P}\left(Vote_i > \frac{N}{k}\right) > p$
- 如果 $p << \frac{1}{k}$, 无论 N 多大，$\mathbb{P}\left(Vote_i > \frac{N}{k}\right) << p$ 弱分类器无法变强

装袋式集成：降低偏差 N 与 p



装袋式集成：样本重叠问题

- 可放回随机抽样产生相同样本，$\bar{\rho} \approx 1$，无法削减方差。通过 DSB 来缓解
- 袋内（训练集）与袋外（验证集）样本非常相似，造成准确率被高估。通过不 shuffling 的 k 折交叉验证来缓解。
- 优先选择较低的 k 值。较高的 k 值可能带来较多袋内与袋外的相似性。

装袋式集成：随机森林

- 随机森林（Random Forest, RF）本质上是袋装式集成的决策树
- 引入了第二层的随机性：在优化每一个树节点时，仅使用特征 X_i 的一个子集
- RF 能够降低模型方差，缓解过拟合
- RF 能够计算出每个特征的重要程度
- RF 不一定能降低偏差

装袋式集成：如何正确使用 RF

应对金融机器学习中的样本的重叠性问题，RF 的参数需要合理选择

- 将最大特征数 (Python: max_features， R: mtry) 设为较低的值，强制树之间产生差异
- 将叶节点的最小样本数 (Python: min_weight_fraction_leaf ,R: min.node.size) 设为一个较大的值，形成 Early stop
- 将每个模型的样本量设定为样本间的平均独特度与总样本量的乘积：$avgU \times N$。
- 使用 DSB 进行抽样

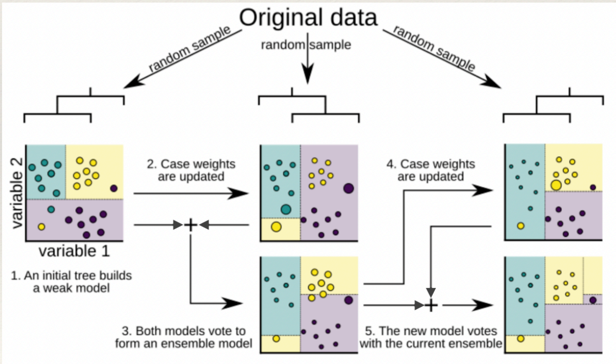
助推式集成

助推式集成的思路是逐步通过弱学习器生成偏差小的强学习器。步骤如下

- (1) 初始化为均匀权重，使用随机抽样生成一个训练集
- (2) 使用该训练集拟合一个学习器；
- (3) 如果单个学习器的准确率大于接受阈值 (例如在二元分类器中为 50%，即比随机分类好)，则保留该学习器，否则丢弃；
- (4) 给误分类的样本更多的权重，给正确分类的样本更少的权重；
- (5) 重复前面的步骤直到生成 N 个学习器；
- (6) 合成的预测值是 N 个模型的个体预测值的加权平均，其中权值由个体学习器的准确性确定

助推式集成: AdaBoost 算法

将弱学习器它们合并到一个强的学习器中。每个弱学习器在训练集中被赋予一个权重，该权重随着每轮迭代进行更新，同时给给误分类的样本增加权重，以提高它们的“重要性”。



助推式集成: 与装袋式集成的比较

- 助推式式串行计算，无法并行
- 很弱的分类器将被放弃
- 每轮迭代时样本的权重都不相同
- 每个学习器都有一个不同的权重
- 主要用于解决欠拟合问题。金融应用中主要面临过拟合问题，应以装袋式集成为主。