随机森林与模型的可解释性与预测能力权衡

Junchen Feng

2024-12-20

课程背景与目标

- 通过随机森林为例，介绍集成学习 (ensemble learning)
- 强调模型可解释性与预测能力之间的 trade-off

回顾与引入

- 回顾机器学习的核心目标：不仅在训练集上表现好，更需在新数据上有良好泛化能力

回顾与引入

- 回顾机器学习的核心目标：不仅在训练集上表现好，更需在新数据上有良好泛化能力
- 单一模型（如决策树）在复杂数据场景中局限性：易过拟合，泛化不足

集成学习的基本思想

- Ensemble Models：**通过集成多个基学习器**（Base Learners）**提升整体性能**

集成学习的基本思想

- Ensemble Models：通过集成多个基学习器（Base Learners）提升整体性能
- 类比：专家委员会与单独专家的差异——多个决策对结果表决或平均

集成学习的基本思想

- Ensemble Models：通过集成多个基学习器（Base Learners）提升整体性能
- 类比：专家委员会与单独专家的差异——多个决策对结果表决或平均
- 常见策略：Bagging 与 Boosting（本次重点在 Bagging）

Bagging (Bootstrap Aggregating)

- 基本思想：通过对训练集进行有放回抽样（Bootstrap）生成多个子数据集。
- 对每个子数据集训练一个基学习器（如决策树），最后对预测结果取平均（回归）或投票（分类）。

数学表示：
假设有一个训练集 $D = \{(x_i, y_i)\}_{i=1}^{N}$。Bagging 通过抽样生成 M 个数据子集 $D^{(m)}$，其中 $m = 1, 2, \ldots, M$。对每个子集训练一个学习器 $h_m(x)$。最终的 Bagging 预测为：

$$H_{\text{bag}}(x) = \frac{1}{M} \sum_{m=1}^{M} h_m(x) \quad \text{(回归)}$$

或对于分类问题：

$$H_{\text{bag}}(x) = \text{majority\_vote}\big(h_1(x), h_2(x), \ldots, h_M(x)\big)$$

Boosting

- 基本思想：逐步训练一系列基学习器，每个新学习器都针对之前学习器的不足（错误样本）进行有偏重的再训练。
- 不同于 Bagging 的并行训练，Boosting 是序列式构建学习器，并不断提升整体预测精度。

数学表示（以 AdaBoost 为例）：

给定训练集 $D = \{(x_i, y_i)\}_{i=1}^{N}$，初始样本权重为 $w_i^{(1)} = \frac{1}{N}$。对于 $m = 1$ 到 M：

1. 基于当前权重分布训练基学习器 $h_m(x)$
2. 计算加权错误率 $\epsilon_m = \sum_{i=1}^{N} w_i^{(m)} \mathbf{1}(h_m(x_i) \neq y_i)$
3. 计算学习器权重 $\alpha_m = \frac{1}{2} \ln \frac{1-\epsilon_m}{\epsilon_m}$
4. 更新权重分布：

$$w_i^{(m+1)} = \frac{w_i^{(m)} \exp(-\alpha_m y_i h_m(x_i))}{Z_m}$$

其中 Z_m 是归一化因子

最终 Boosting 预测为：

$$H_{\text{boost}}(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m h_m(x)\right)$$

决策树的复习与不足

- 决策树优点：简单直观，可解释性强

决策树的复习与不足

- 决策树优点：简单直观，可解释性强
- 单数要增加预测能力，就会太复杂，容易过拟合

决策树的复习与不足

- 决策树优点：简单直观，可解释性强
- 单数要增加预测能力，就会太复杂，容易过拟合
- 能不能”聚”木成林

随机森林的构造原理（Random Forest）

- Bagging 步骤：对训练数据集有放回抽样，得到多个随机样本子集
- 每个子集训练一棵决策树，分裂时随机挑选特征子集，增加模型多样性
- 多棵树的预测结果采用投票（分类）或平均（回归）得到最终预测

随机森林的构造原理（Random Forest）

随机性体现在哪里？

随机森林的构造原理（Random Forest）

Theorem (随机森林的两重随机性)

1. 数据采样的随机性：*Bootstrap* 抽样使每棵树看到不同的训练数据
2. 特征选择的随机性：每次分裂时随机选择特征子集进行最优分裂

这两重随机性保证了森林中每棵树都具有独特性，提高了整体模型的多样性和鲁棒性。

随机森林为何提高泛化性能？

- 减少模型方差：个体决策树虽不稳定，但通过多数表决可”互相抵消”错误
- 不需对数据分布做严格假设，凭借随机性增加鲁棒性

可解释性与预测能力的平衡点

- 决策树：可解释性强（清晰的规则路径），但单树预测能力有限
- 随机森林：性能提升（更高泛化能力），但单个预测路径较复杂，可解释性下降
- 实际问题中往往在解释性与预测性能之间寻找平衡

哪个特性更重要？

- 金融信贷决策：需向客户和监管部门解释为什么拒绝贷款
- 高频金融交易：只要赚钱不需要可解释

Breiman 的"Two Cultures"

- 统计学文化：偏好可解释的参数模型，强调数据生成分布和明确的假设

Breiman 的"Two Cultures"

- 统计学文化：偏好可解释的参数模型，强调数据生成分布和明确的假设
- 机器学习文化：更注重预测性能，模型不一定有明确的分布假设

深层思考

- 面对复杂问题：究竟何为”好”模型？

深层思考

- 面对复杂问题：究竟何为”好”模型？
- 不同学科背景下，对解释与预测的侧重点不一样

深层思考

- 面对复杂问题：究竟何为”好”模型？
- 不同学科背景下，对解释与预测的侧重点不一样
- Breiman 观点的启示：当我们谈论模型时，不仅谈精度，还要谈理念、假设与实用性