第 25 章 决策树

25.1 决策树假设

表 25.1.1: 集成策略：Blending vs. Learning

聚合类型	Blending	Learning
Uniform（均匀）	投票 / 平均	Bagging
Non-uniform（非均匀）	线性加权	AdaBoost
Conditional（条件）	Stacking	决策树

注 决策树可视为一种传统的“条件聚合”学习模型：在每个节点根据输入特征动态决定子模型的权重。

命题 25.1.1 (决策树的递归视角与路径表达)

可把决策树看作一个递归定义的函数

$$G(x) = \sum_{t=1}^{\text{叶数}} \mathbb{I}[x \text{ 落在路径} t] \cdot \mathrm{leaf}_t(x),$$

也可递归地写成

$$G(x) = \sum_{c} \mathbb{I}[b(x) = c] \cdot G_c(x),$$

其中

- $b(x)$ 为当前节点的分支准则（branching criteria）；
- $G_c(x)$ 为第 c 个子树的预测函数；
- 整棵树即“(根节点, 子树)”——正如数据结构课所言。

注[关于决策树的声明]

- 可解释性强：在商业、医疗等领域广泛使用。
- 启发式为主：理论分析有限，常用启发式选择分裂。
- 实现简单：实现轻松。
- 预测与训练高效：但无单一“标准”算法。

结论：决策树主要依赖启发式，却因其直观与高效而自成一类实用模型。

25.2 决策树算法

算法 25.2.1: 基本决策树算法

输入: 数据集 $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$

输出: 决策树模型 $G(x)$

Function `DecisionTree`($\mathcal{D}$):

  if 终止条件满足 **then**

    return 基假设 $g_t(x)$;

  else

    学习分支准则 $b(x)$;

    将 $\mathcal{D}$ 按 $b(x)$ 的值划分为 C 个子集

$$\mathcal{D}_c = \{(x_n, y_n) \mid b(x_n) = c\}, \quad c = 1, \ldots, C;$$

    递归构建子树

$$G_c \leftarrow \texttt{DecisionTree}(\mathcal{D}_c);$$

    return 组合函数

$$G(x) = \sum_{c=1}^{C} \mathbb{I}[b(x) = c] \cdot G_c(x).$$

四个关键选择

- 分支数 C（二叉或多叉）;
- 分支准则 $b(x)$（信息增益、基尼指数等）;
- 终止条件（最大深度、最小样本数、纯度阈值等）;
- 基假设 $g_t(x)$（常数、线性模型等）。

算法 25.2.2: Classification and Regression Tree (C&RT) 算法

输入: 数据集 $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$，其中 x_n 为样本特征，y_n 为对应标签/输出值，N 为样本总量

输出: 训练好的决策树模型 $G(x)$，可依据输入特征 x 输出预测结果（分类或回归值）

Function `DecisionTree`($\mathcal{D}$)

if 满足终止条件 **then**

return 基假设 $g_t(x)$，即基于当前数据集 $\mathcal{D}$ 的 E_{in}-最优常数；分类任务常用多数投票，回归任务常用均值;

else

`// 1. 选择分支准则（二分支 + 决策桩）`

采用二分支策略（$C = 2$，将数据集划分为 2 个子集），借助“决策桩（decision stump，简单的单特征判别规则）”确定划分方式：

$$b(x) = \underset{\substack{h \in \{\text{所有可能的决策桩}\} \\ \text{（单特征、简单阈值/类别划分规则）}}}{\operatorname{argmin}} \sum_{c=1}^{2} |\mathcal{D}_c| \cdot \text{impurity}(\mathcal{D}_c)$$

其中，$\mathcal{D}_c = \{(x_n, y_n) \mid b(x_n) = c\}$ 表示依据分支准则 $b(x)$ 划分后，第 c 个子集的样本集合；$|\mathcal{D}_c|$ 是该子集的样本数量；$\text{impurity}(\cdot)$ 为衡量数据集“纯度/杂质度”的函数。

`// 2. 杂质度（Impurity）计算规则`

杂质度用于量化数据集的“混乱程度”，分类与回归任务采用不同计算逻辑：

$$\text{impurity}(\mathcal{D}) = \begin{cases} \frac{1}{N} \sum\limits_{n=1}^{N} (y_n - \bar{y})^2, & \text{回归任务（预测连续值）} \\ \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}[y_n \neq y^*], & \text{分类任务（多数类占比）} \\ 1 - \sum\limits_{k=1}^{K} \left(\frac{\sum\limits_{n=1}^{N} \mathbb{I}[y_n = k]}{N} \right)^2, & \text{分类任务（Gini 指数）} \\ 1 - \max\limits_{1 \leq k \leq K} \frac{\sum\limits_{n=1}^{N} \mathbb{I}[y_n = k]}{N}, & \text{分类任务（最优类别）} \end{cases}$$

实际应用中，回归常选均方误差，分类常选 Gini 指数。

`// 3. 终止条件细化说明`

所有样本标签/输出值相同：$\text{impurity}(\mathcal{D}) = 0$，无需再分裂，直接以该值作为叶节点预测结果;

所有样本特征完全相同（但标签可能不同，即存在杂质）：无法再用决策桩划分，强制终止分裂;

`// 4. 递归构建子树与组合`

对划分得到的 2 个子集 $\mathcal{D}_1$、$\mathcal{D}_2$，分别递归调用 `DecisionTree`($\mathcal{D}_c$)，生成子树 $G_1(x)$、$G_2(x)$。最终：

$$G(x) = \sum_{c=1}^{2} \mathbb{I}\big[b(x) = c\big] \cdot G_c(x)$$

其中，$\mathbb{I}[\cdot]$ 是指示函数，条件满足时为 1，否则为 0。

例题 25.1 选择题：基尼指数（二分类场景）

已知基尼指数公式为 $G = 1 - \sum_{k=1}^{K} \left(\frac{\sum_{n=1}^{N} [\![y_n = k]\!]}{N} \right)^2$。当 $K = 2$（二分类）时，设 $\mu = \frac{N_1}{N}$（N_1 是标签 $y_n = 1$ 的样本数量，N 是总样本数），则此时基尼指数用 μ 表示的公式为？

1) $2\mu(1-\mu)$
2) $2\mu^2(1-\mu)$
3) $2\mu(1-\mu)^2$
4) $2\mu^2(1-\mu)^2$

解答 正确选项为 $\boxed{1}$。基尼指数定义为：

$$G = 1 - \sum_{k=1}^{K} \left(\frac{\text{标签为}k\text{ 的样本数}}{\text{总样本数}} \right)^2$$

当 $K = 2$ 时，设标签为 1 的样本比例为 μ，则标签为 2 的样本比例为 $1-\mu$。代入公式得：

$$G = 1 - \mu^2 - (1-\mu)^2$$

展开并化简：

$$G = 1 - \mu^2 - (1 - 2\mu + \mu^2) = 1 - \mu^2 - 1 + 2\mu - \mu^2 = 2\mu - 2\mu^2 = 2\mu(1-\mu)$$

■

25.3 C&RT 决策树启发式

命题 25.3.1 (通过剪枝的正则化)

完全生长的决策树满足 $E_{\text{in}}(G) = 0$（若所有 x_n 互不相同），却因在极小子集 $\mathcal{D}_c$ 上继续分裂而严重过拟合（E_{out} 很大）。

正则化目标 引入正则项 $\Omega(G) = \text{NumberOfLeaves}(G)$，求解

$$G^* = \arg\min_{G} \left[E_{\text{in}}(G) + \lambda\, \Omega(G) \right].$$

所得 G^* 称为 剪枝决策树 **(pruned decision tree)**。

计算策略 枚举全部子树不可行。通常采用 后剪枝：

1. 令 $G^{(0)}$ 为完全生长树；
2. 对 $i = 1, 2, \ldots$，令 $G^{(i)} = \arg\min_{G} E_{\text{in}}(G)$ 其中 G 仅比 $G^{(i-1)}$ 少一片叶子；
3. 在验证集上选择最优剪枝序列 $\{G^{(i)}\}$。

♠

命题 25.3.2 (C&RT 对类别特征与缺失值的处理)

1. 类别特征的分支

- 数值特征：使用决策桩

$$b(x) = \mathbb{I}[x_i \le \theta] + 1, \quad \theta \in \mathbb{R}.$$

- 类别特征：使用决策子集

$$b(x) = \mathbb{I}[x_i \in S] + 1, \quad S \subseteq \{1, 2, \ldots, K\}.$$

C&RT（及一般决策树）天然支持类别特征，无需额外编码。

2. 缺失值的代理分支

假设主分支为

$$b(x) = \mathbb{I}[\text{weight} \le 50\ \text{kg}] + 1.$$

若预测时 `weight` 缺失：

- 人类做法：用 `height` 阈值近似 `weight` 阈值；
- 算法做法：训练时为 $b(x)$ 预先学习 代理分支

$$b_1(x),\ b_2(x),\ \ldots \text{（按与主分支一致性排序）；}$$

  预测时缺失即用最佳代理分支替代。

C&RT 通过代理机制轻松处理缺失特征。

25.4 决策树实践

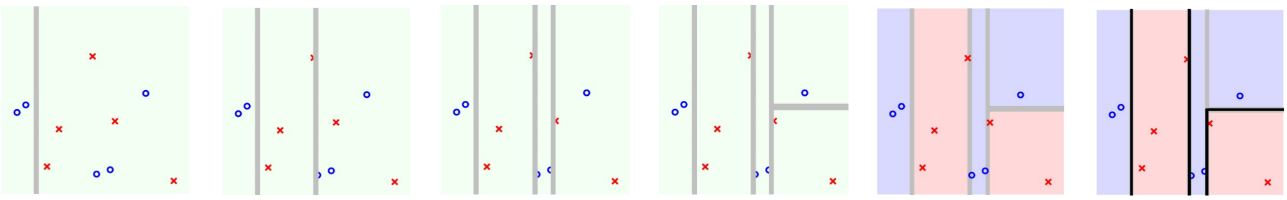

图 25.4.1: C&RT 实际应用

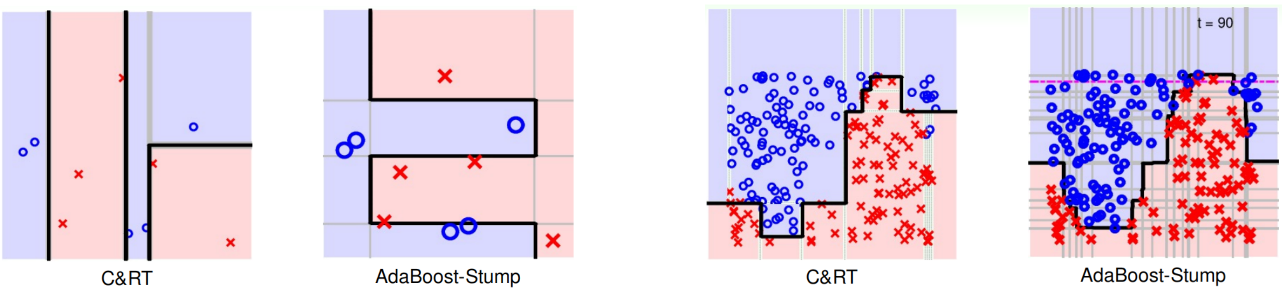

图 25.4.2: C&RT 与 AdaBoost-Stump 的对比图

命题 25.4.1 (C&RT 的实用特性)

分类与回归树（C&RT）具有以下独特优势：

- 易于解释：决策树的结构直观，易于人类理解；
- 多分类友好：天然支持多分类任务，无需额外调整；
- 类别特征友好：直接处理类别特征，无需编码；
- 缺失值友好：通过代理分支处理缺失特征，无需丢弃样本；
- 高效非线性建模：训练与测试均高效，能自动捕捉非线性关系。

总结　几乎无其他单一学习模型具备上述全部特性，唯有其他决策树算法（即除了基础决策树之外的其他决策树相关算法）与之类似。

例题 25.2 选择题：未剪枝 CART 树的特性

以下哪项不是未剪枝（without pruning）的 CART（分类与回归树）的特性？

1) 易于处理缺失特征
2) 生成可解释的假设
3) 达到低训练误差 E_{in}
4) 达到低泛化误差 E_{out}

解答 正确选项为 $\boxed{4}$。未剪枝的 CART 树具有以下特性：

- 可通过替代分裂等方法轻松处理缺失特征（对应选项 1）；
- 树结构决策逻辑直观，生成的假设可解释（对应选项 2）；
- 会完全拟合训练数据，因此训练误差 E_{in} 低（对应选项 3）。

未剪枝的 CART 树因过度拟合训练数据，泛化误差 E_{out} 通常较高，无法保证低 E_{out}（选项 4 不符合）。

■

25.5 总结

笔记 [决策树]

- 决策树假设：以路径条件方式表达的聚合模型。
- 决策树算法：递归地分裂节点，直至满足终止条件并返回叶节点（基预测）。
- C&RT 决策树启发式：剪枝、类别型特征分支、代理分裂。
- 决策树实践：模型可解释且计算高效。