机器学习第三章决策树课后习题

162050127 颜劭铭

2022 年 5 月 12 日

1 课后习题

1.1 试分析使用”最小训练误差”作为决策树划分选择的缺陷.

首先结合第二章模型评估与选择，我们可以给出泛化误差的定义：

$$E(f;D)(\text{误差}) = bias^2(x)(\text{偏差}) + var(x)(\text{方差}) + \varepsilon^2(\text{噪声})$$

而泛化误差包括训练误差和测试误差，偏差指的是模型在样本上的输出与真实值之间的误差，即模型本身的精准度，方差指的是使用样本数相同的不同训练集所训练出的模型每一次输出结果与模型输出期望之间的误差，即模型的稳定性。

因此题目中的最小化训练误差应该是指通过训练得到不同模型的参数，并进行比较，得到最小的训练误差的模型。

虽然最小化训练误差得到的是最优解，但是我觉得这不能作为决策树的划分选择，主要有两方面的问题：

① 一方面是因为当训练误差达到最小的时候，模型有可能学习了训练集上的某些特殊、异常、偶然的特征，并不适用于大部分情况，这会使模型在测试集上或者其他的数据集上有很大的误差，导致过拟合问题。

② 另一方面我觉得得从决策树算法出发进行分析，因为决策树是一种贪心算法，它是一层层生成的，并不能保证是一个全局最优解，而最小化训练误差应该是在树结构确定以后才能进行的（除非决策树只需要进行一次 if-else），当然也可以在确定树结构以后用最小训练误差对于树的结构进行调整，但是这个代价和收益我觉得是差别非常大的，也不满足贪心算法的出发点，所以我觉得如果从这一方面考虑的话决策树是不适合（甚至可以说不应该）使用最小训练误差作为决策树的划分选择。

1.2 树也是一种线性模型，考虑图 1 所示回归决策树，X_1, X_2 均在单位区间上取值，t_1, t_2, t_3, t_4 满足 $0 < t_1 < t_3 < 1, 0 < t_2, t_4 < 1$，试绘制出该决策树对于特征空间的划分。假设区域 R_i 上模型的输出为 c_i，试用线性模型表示该决策树。

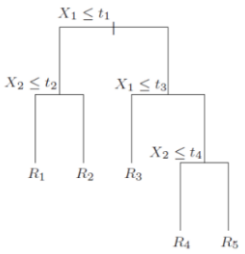


图 1: 回归决策树

由题意我们可以得到，该决策树中的每个样本都有两个属性 (X_1, X_2)，并且均在 $[0, 1]$ 上取值，因此我们可以用一个二维坐标图进行表示，其中横轴为 X_1，纵轴为 X_2。

对于图 1 回归决策树进行分析我们可以得到：

① 在根节点，当 $X_1 \leqslant t_1$ 时，进入左子树，并且对于 X_2 进行分析，当 $X_2 \leqslant t_2$ 时，进入该结点的左子树 R_1 区域并输出值为 c_1.

② 在根节点，当 $X_1 \leqslant t_1$ 时，进入左子树，并且对于 X_2 进行分析，当 $X_2 > t_2$ 时，进入该结点的右子树 R_2 区域并输出值为 c_2.

③ 在根节点，当 $X_1 > t_1$ 时，进入右子树，并继续对于 X_1 进行分析，当 $X_1 \leqslant t_3$ 时，进入该结点的左子树 R_3 区域并输出值为 c_3.

④ 在根节点，当 $X_1 > t_1$ 时，进入右子树，并继续对于 X_1 进行分析，当 $X_1 > t_3$ 时，进入该结点的右子树，并对于 X_2 进行分析，当 $X_2 \leqslant t_4$ 时，进入该结点的左子树 R_4 区域并输出值为 c_4.

⑤ 在根节点，当 $X_1 > t_1$ 时，进入右子树，并继续对于 X_1 进行分析，当 $X_1 > t_3$ 时，进入该结点的右子树，并对于 X_2 进行分析，当 $X_2 > t_4$ 时，进入该结点的右子树 R_5 区域并输出值为 c_5.

因此，以上的文字形式可由下图的线性模型进行表示（由于图中绘制的时候不能出现未知量，这里的 t_1, t_2, t_3, t_4 用固定数值进行表示，我采用均分，取了 $t_1 = 0.33, t_2 = 0.5, t_3 = 0.66, t_4 = 0.5$）：

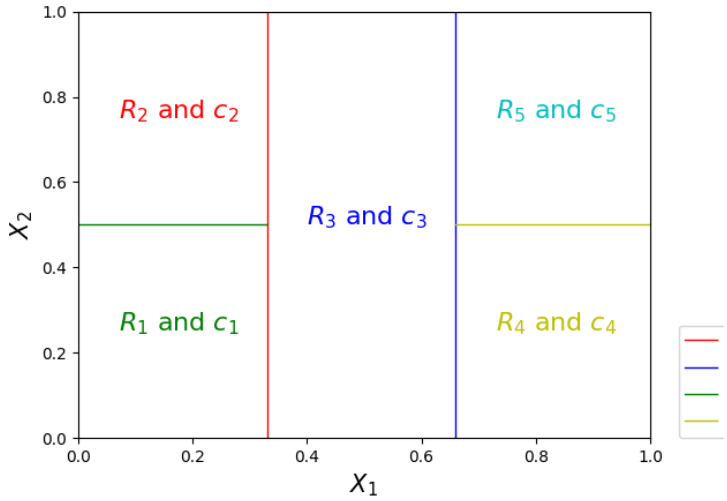


图 2: 回归决策树的线性模型

注：其中图 2 红色的线是指 t_1，蓝色的线是 t_3，绿色的线是 t_2，黄色的线是 t_4.

1.3 考虑如表 1 所示的人造数据，其中“性别”、“喜欢 ML 作业”是属性，“ML 成绩高”是标签。请画出使用信息增益为划分准则的决策树算法所有可能的结果（需说明详细计算过程）

表 1: 训练集

编号	性别	喜欢 ML 作业	ML 成绩高
1	男	是	是
2	女	是	是
3	男	否	否
4	男	否	否
5	女	否	是

把该训练集用 $\mathcal{D}$ 表示。在决策树学习开始时，根结点包含 $\mathcal{D}$ 中所有样例，显然，样本类别数目 $|\gamma| = 2$，其中正例占 $p_1 = \frac{3}{5}$，反例占 $p_2 = \frac{2}{5}$，于是可以计算出根结点的信息熵：

$$Ent(\mathcal{D}) = -\sum_{k=1}^{2} p_k \log_2 p_k = -(\frac{3}{5} \times \log_2 \frac{3}{5} + \frac{2}{5} \times \log_2 \frac{2}{5}) = 0.9710$$

当前属性集合为 {性别，喜欢 ML 作业}，其中性别有两种可能的取值，分别为 {男，女} 因此若以性别对 $\mathcal{D}$ 进行划分，可以得到 2 个子集，分别记为 $\mathcal{D}^1$(性别 = 男) 和 $\mathcal{D}^2$(性别 = 女)，因此可以计算出用“性别”进行划分之后获得的两个结点的信息熵为：

$$\begin{aligned} Ent(\mathcal{D}^1) &= -(\frac{1}{3} \times \log_2 \frac{1}{3} + \frac{2}{3} \times \log_2 \frac{2}{3}) = 0.9183 \\ Ent(\mathcal{D}^2) &= -(1 \times \log_2 1 + 0 \times \log_2 0) = 0 \end{aligned}$$

继续计算属性“性别”的信息增益为：

$$\begin{aligned} Gain(\mathcal{D}, \text{性别}) &= Ent(\mathcal{D}) - \sum_{v=1}^{3} \frac{|D^v|}{|D|} Ent(D^v) \\ &= 0.9710 - (\frac{3}{5} \times 0.9183 + \frac{2}{5} \times 0) \\ &= 0.42 \end{aligned}$$

而对于喜欢 ML 作业有两种可能的取值，分别为 {是，否} 因此若以喜欢 ML 作业对 $\mathcal{D}$ 进行划分，可以得到 2 个子集，分别记为 $\mathcal{D}^1$(喜欢 ML 作业 = 是) 和 $\mathcal{D}^2$(喜欢 ML 作业 = 否)，因此可以计算出用“喜欢 ML 作业”进行划分之后获得的两个结点的信息熵为：

$$\begin{aligned} Ent(\mathcal{D}^1) &= -(1 \times \log_2 1 + 0 \times \log_2 0) = 0 \\ Ent(\mathcal{D}^2) &= -(\frac{1}{3} \times \log_2 \frac{1}{3} + \frac{2}{3} \times \log_2 \frac{2}{3}) = 0.9183 \end{aligned}$$

继续计算属性“喜欢 ML 作业”的信息增益为：

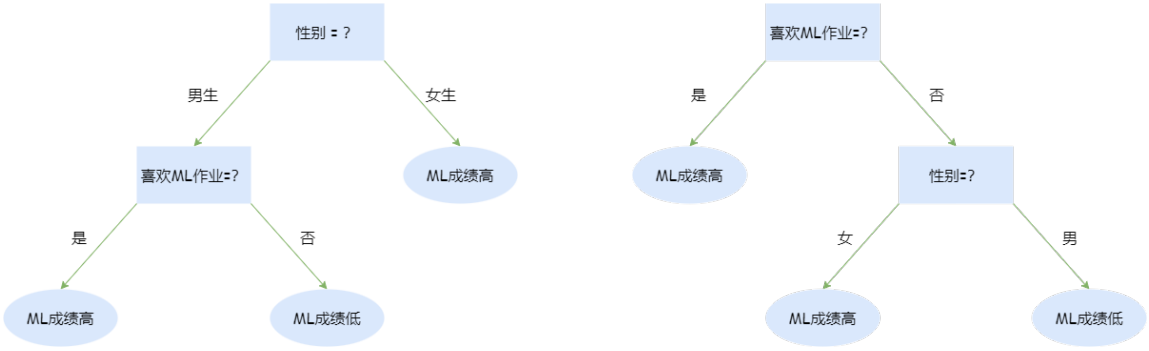
$$\begin{aligned} Gain(\mathcal{D}, \text{喜欢 ML 作业}) &= Ent(\mathcal{D}) - \sum_{v=1}^{3} \frac{|D^v|}{|D|} Ent(D^v) \\ &= 0.9710 - (\frac{2}{5} \times 0 + \frac{3}{5} \times 0.9183) \\ &= 0.42 \end{aligned}$$

因此，我们可以发现属性“性别”和“喜欢 ML 作业”的信息增益是相同的，所以我们可以选择其中任意一个属性作为划分属性，剩下那个属性是第二次的划分属性。

若使用性别作为第一次划分属性，我们可以发现女生一定是分数高的，所以只需要对男生以“喜欢 ML 作业”为划分属性进行第二次划分；若使用喜欢 ML 作业作为第一次划分属性，

我们可以发现喜欢 ML 作业为是的一定是分数高的，所以只需要对不喜欢 ML 作业为否的以“性别”为划分属性进行第二次划分。

以上的划分方法可用如下两个图进行表示：



(a) 使用性别作为第一次划分属性　　(b) 使用喜欢 ML 作业作为第一次划分属性

图 3: 划分决策树结果

1.4 考虑如表 2 所示的验证集，对上一小问的结果基于该验证集进行预剪枝、后剪枝，剪枝结果是什么？（需给出计算过程）比较预剪枝、后剪枝的结果，哪种剪枝方法在训练集、验证集上的准确率分别为多少？哪种方法拟合能力较强？

表 2: 人造验证集

编号	性别	喜欢 ML 作业	ML 成绩高
6	男	是	是
7	女	是	否
8	男	否	否
9	女	否	否

预剪枝：

① 考虑图 3.a，以性别作为第一次划分属性的决策树. 在划分之前，所有样例都集中在根结点，其类别被标记为训练样例数最多的类别，因此将这个叶结点标记为”ML 成绩高”。所以使用表 2 所示的验证集对这个单结点决策树进行评估，则编号为 {6} 的样例被分类正确，因此，此时在验证集上精度为 $\frac{1}{4} \times 100\% = 25\%$。

在用属性” 性别” 划分之后，我们可以发现男生包含编号为 {1,3,4} 的样例，女生包含编号为 {2,5} 的样例，因此这两个结点分别被标记为叶结点”ML 成绩低”，”ML 成绩高”。此时，验证集中编号为 {8} 的样例被分类正确，所以验证集精度为 $\frac{1}{4} \times 100\% = 25\% = 25\%$。因此预剪枝的决策结果为禁止划分，所以最终所有样例都被划分到叶结点”ML 成绩高” 中。

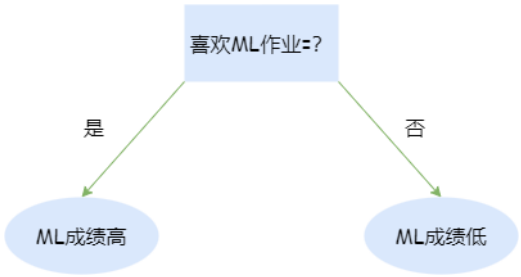
② 考虑图 3.b，以喜欢 ML 作业作为第一次划分属性的决策树. 在划分之前，所有样例都集中在根结点，其类别被标记为训练样例数最多的类别，因此将这个叶结点标记为”ML 成绩高”。所以使用表 2 所示的验证集对这个单结点决策树进行评估，则编号为 {6} 的样例被分类正确，因此，此时在验证集上精度为 $\frac{1}{4} \times 100\% = 25\%$。

在用属性” 喜欢 ML 作业” 划分之后，我们可以发现喜欢 ML 作业包含编号为 {1,2} 的样例，不喜欢 ML 作业包含编号为 {3，4,5} 的样例，因此这两个结点分别被标记为叶结点”ML 成绩高”，”ML 成绩低”。此时，验证集中编号为 {6,8,9} 的样例被分类正确，所以验证集精度为 $\frac{3}{4} \times 100\% = 75\% > 25\%$。因此预剪枝的决策结果为划分，所以运用” 喜欢 ML 作业” 进行划分得以确定。由于划分以后的左结点已经是”ML 成绩高”，所以应该对于右结点继续进行划分，且划分属性为” 性别”。

在继续用” 性别” 进行划分后，我们可以将” 不喜欢 ML 作业” 的样例继续分为两个叶结点，” 男生” 和” 女生”，男生包含编号为 {3,4} 的样例，女生包含编号为 {5} 的样例，因此这两个结点分别被标记为叶结点”ML 成绩低” 和”ML 成绩高”。此时，验证集中编号为 {6,8} 的样例被分类正确，所以验证集精度为 $\frac{1}{2} \times 100\% = 50\% < 75\%$。因此预剪枝的决策结果为禁止划分，所以最终所有样例仅仅对于属性” 喜欢 ML 作业” 进行划分。

基于预剪枝生成的决策树如图 4所示：

(a) 图 3.a 基于预剪枝生成的决策树



(b) 图 3.b 基于预剪枝生成的决策树

图 4: 基于预剪枝生成的决策树

后剪枝：

① 考虑图 3.a，以性别作为第一次划分属性的决策树. 在完成划分后，验证集中编号为 {6,8} 的样例被分类正确，所以验证集精度为 $\frac{1}{2} \times 100\% = 50\%$。

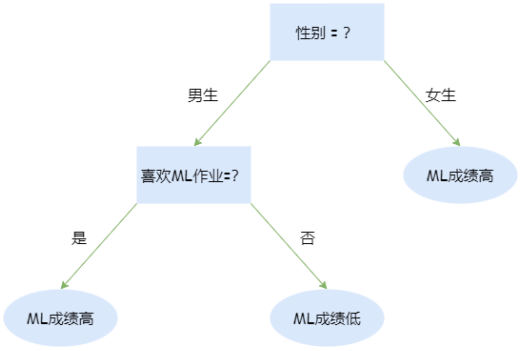
后剪枝首先考察结点” 喜欢 ML 作业 =？ ”，若将其领衔的分支剪除，则该结点变为叶结点，且该叶结点包含编号为 {1,3,4} 的训练样本，于是该叶结点的类别被标记为”ML 成绩低”，此时验证集的精度为 $\frac{1}{4} \times 100\% = 25\% < 75\%$，所以不进行剪枝。

② 考虑图 3.b，以喜欢 ML 作业作为第一次划分属性的决策树. 在完成划分后，验证集中编号为 {6,8} 的样例被分类正确，所以验证集精度为 $\frac{1}{2} \times 100\% = 50\%$。

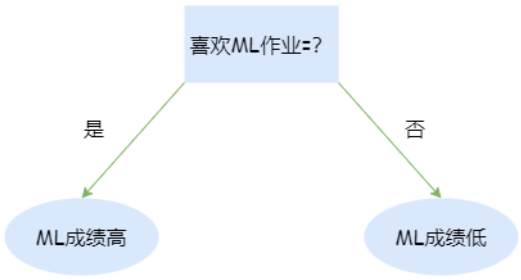
后剪枝首先考察结点” 性别 =？ ”，若将其领衔的分支剪除，则该结点变为叶结点，且该叶结点包含编号为 {3,4,5} 的训练样本，于是该叶结点的类别被标记为”ML 成绩低”，此时验证集的精度为 $\frac{3}{4} \times 100\% = 75\% > 50\%$，所以后剪枝策略决定进行剪枝。

后剪枝再次考虑结点” 喜欢 ML 作业 =？ ”，若将其领衔的分支剪除，则该结点变为叶结点，且该叶结点包含编号为 {1,2,3,4,5} 的训练样本，于是该叶结点的类别被标记为”ML 成绩高”,，此时验证集的精度为 $\frac{1}{4} \times 100\% = 25\% < 75\%$，所以不进行剪枝。

基于后剪枝生成的决策树如图 5所示：



(a) 图 3.a 基于后剪枝生成的决策树



(b) 图 3.b 基于后剪枝生成的决策树

图 5: 基于后剪枝生成的决策树

综上，我们可以得出对于图 3.a 进行预剪枝后，训练集的精度为 $\frac{3}{5} \times 100\% = 60\%$，验证集的精度为 $\frac{1}{4} \times 100\% = 25\%$。对于图 3.b 进行预剪枝后，训练集的精度为 $\frac{4}{5} \times 100\% = 80\%$，验证集的精度为 $\frac{3}{4} \times 100\% = 75\%$。

而对于图 3.a 进行后剪枝后，训练集的精度为 $1 \times 100\% = 100\%$，验证集的精度为 $\frac{1}{2} \times 100\% = 50\%$。对于图 3.b 进行后剪枝后，训练集的精度为 $\frac{4}{5} \times 100\% = 80\%$，验证集的精度为 $\frac{3}{4} \times 100\% = 75\%$。

因此我们可以给出结论：

对于图 3.a 进行后剪枝的拟合效果在训练集和测试集上都要好于预剪枝。

对于图 3.b 进行后剪枝和预剪枝的拟合效果在训练集和测试集上相同。