1 [20pts] Short Answers

(1) (9pts) Under what conditions basic decision tree algorithm generates leaf nodes?

(2) (6pts) What is the principle of selecting splitting attribute? Please write down the most commonly used measure Information Gain.

(3) (5pts) What is the shortcoming of pruning according to training error?

(1) Answer:

1. All samples from current node belong to the same category.
2. Current attribute set is empty or all samples in the data set of the node take the same value
3. The data set of a child node is empty, then the child node is a leaf node.

(2) We hope that the samples contained in the branch nodes of the decision tree belong to the same category as much as possible, that is, the purity of the nodes is getting higher and higher.

Suppose that the discrete feature a has V possible values $\{a^1, a^2, \cdots, a^V\}$. Then, splitting the data set D by feature will produce V child nodes, where the v-th child node D^v includes all samples in D taking the value a^v for feature a. Then Information gain is:

$$Gain(D, a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$$

Where $Ent(D) = -\sum_{k=1}^{K} p_k \log p_k$, $\mathrm{y} \in \{1,2,\cdots,K\}$, p_k is the proportion of the k-th class in the data set D.

(3) It might cause overfitting

2 [50pts] Decision Tree

(1) (20pts) Consider the synthetic data shown in Table 1, where " 性别" and " 喜欢 ML 作业" are attributes and "ML 成绩高" is the label. Please draw all possible results of the decision tree algorithm that uses information gain as the splitting criterion. (Detailed calculation process should be explained)

表 1: 人造训练集

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

(2) (20pts) Consider the validation set shown in Table 2, what is the result of pre-pruning and post-pruning based on the previous sub-question's results? (Detailed calculation process is required.)

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

(3) (10pts) Compare the results of pre-pruning and post-pruning. What are the accuracies of the two pruning methods on the training set and validation set, respectively? Which method has stronger classification ability?

(1) Let's calculate the information entropy of root node first. At the beginning of decision tree learning, the root node includes all examples in D, where positive examples account for $p_1 = 3/5$ and negative examples account for $p_2 = 2/5$. Thus, the information entropy is

$$Ent(D) = -\sum_{k=1}^{2} p_k \log_2 p_k = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.971$$

Let D_1^1, D_1^2, D_2^1 and D_2^2 denote (性别 = 男)，(性别 = 女)，(喜欢 ML 作业 = 是) and (喜欢 ML 作业 = 否), respectively. As we did when calculating $Ent(D)$, we can get

$$Ent(D_1^1) = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) = 0.918$$

$$Ent(D_1^2) = 0$$

$$Ent(D_2^1) = 0$$

$$Ent(D_2^2) = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) = 0.918$$

Therefore, the information gain of “性别” and “喜欢 ML 作业” is

$$\begin{aligned} Gain(D, \text{性别}) &= Ent(D) - \sum_{v=1}^{2}\frac{|D^v|}{|D|}Ent(D^v) \\ &= 0.971 - \left(\frac{3}{5}\times 0.918 + \frac{2}{5}\times 0\right) \\ &= 0.420 \end{aligned}$$

$$Gain(D, \text{喜欢 ML 作业}) = 0.420$$

Finally, we get the two possible decision trees

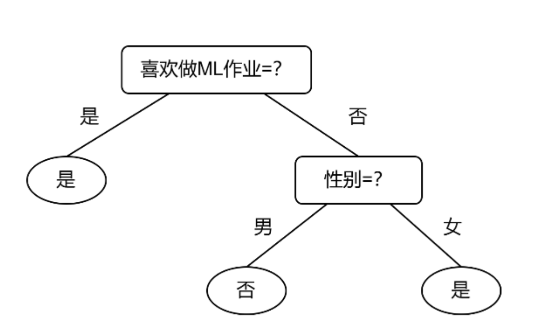


Fig. 1

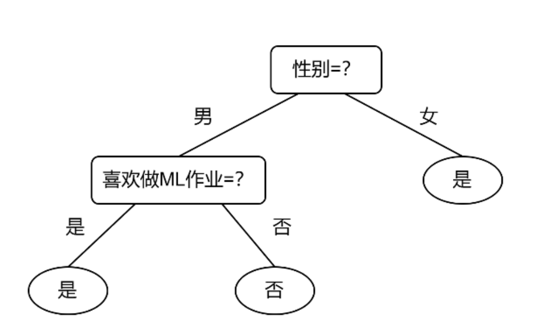


Fig. 2

(2) Pre-pruning:

Before splitting, all samples congregate at the root node. If we don’t split, the root node will be marked as a leaf node. Suppose we mark this node as “是”, then only one sample is classified correctly, so the accuracy is $\frac{1}{4}\times 100\% = 25\%$.

After splitting by attribute “喜欢 ML 作业”(shown in Fig. 3), there are 3 samples can be classified correctly, so the accuracy is 75%. Therefore, we need such a splitting scheme.

Let’s consider on the next attribute “性别”. Before and after pruning, the accuracy is 100% and 50% respectively. Therefore, we don’t need to split.

We can also get the results corresponding to the decision tree in Fig. 2, as shown in Fig. 4

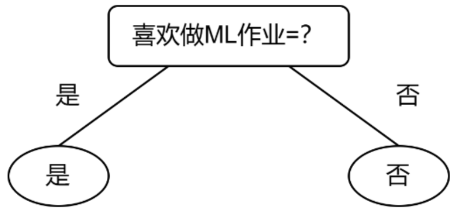


Fig. 3

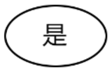


Fig. 4

Post-pruning:

Let’s start from the decision tree shown in Fig. 1. Inspect the tree from the bottom up, if we keep “性别”, the accuracy will be 25%, otherwise, the accuracy will be 75%, so we have reason to cut this attribute. How about attribute “喜欢 ML 作业“ ? With it and without it, the accuracy is 75% and 25% respectively, thus, keep it.

We can also get the results corresponding to the decision tree in Fig. 2, as shown in Fig. 6

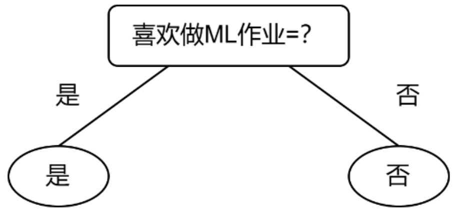


Fig. 5

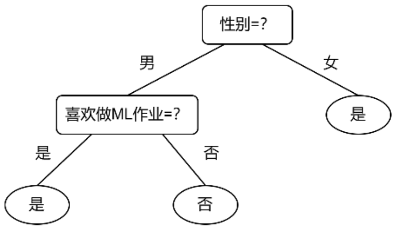


Fig. 6

(3) Post-pruning is better.

Accuracy decision tree	Pre-pruning	Post-pruning
Fig. 1	7/9	7/9
Fig. 2	4/9	7/9

3 [30pts] Regression Tree

(1) (10pts) 树也是一种线性模型，考虑图 (4) 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 上模型的输出值为 c_i，试用线性模型表示该决策树。

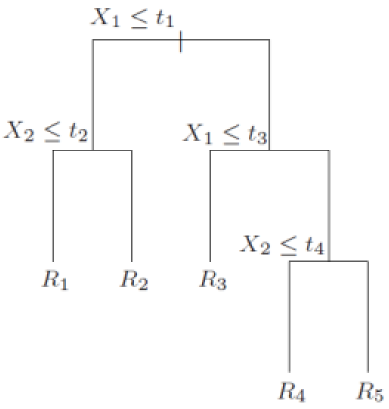


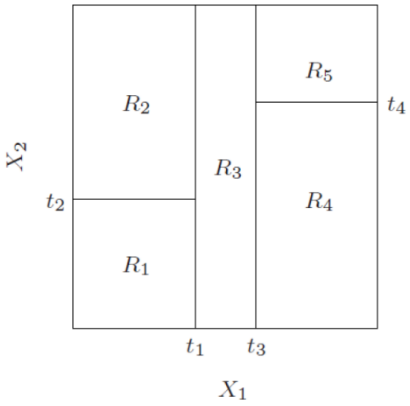
图 4: 回归决策树

(2) (20pts) 对于回归树，我们常采用平方误差来表示回归树对于训练数据的预测误差。但是找出平方误差最小化准则下的最优回归树在计算上一般是不可行的，通常我们采用贪心的算法计算切分变量 j 和分离点 s。CART 回归树在每一步求解如下优化问题

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right]$$

其中 $R_1(j,s)=\{\boldsymbol{x}|x_j\leq s\}, R_2(j,s)=\{\boldsymbol{x}|x_j>s\}$。试分析该优化问题表达的含义并给出变量 j, s 的求解思路。

(1) According to the description, we can draw the feature space.



Decision tree:

$$\sum_{i=1}^{5} c_i \mathbb{I}(x \in R_i)$$

(2) This optimization problem solves the parameters (j, s) so that the sum of the squared errors of the decision tree on the two sub-regions after splitting is the smallest.

Assuming that n sets of data are input, each set of data has m-dimensional features. When the value of the i-th dimension feature is $x_{ij_1} \leq x_{ij_2} \leq \cdots \leq x_{ij_n}$, traverse $1,2,\cdots,m$. When $j = i$, s traverses $x_{ij_1} \leq x_{ij_2} \leq \cdots \leq x_{ij_{n-1}}$. Choose the parameter pair that minimizes $f(j, s)$. For the function $f(j, s)$, the minimum points of c_1 and c_2 are respectively $c_i = avg(y_i|x_i \in R_1(j, s))$, $c_2 = avg(y_i|x_i \in R_2(j, s))$/