Part A: 决策树

1. 信息论基础

树具有天然的分支结构。对于分类问题而言，决策树的思想是用节点代表样本集合，通过某些判定条件来对节点内的样本进行分配，将它们划分到该节点下的子节点，并且要求各个子节点中类别的纯度之和应高于该节点中的类别纯度，从而起到分类效果。

节点纯度反映的是节点样本标签的不确定性。当一个节点的纯度较低时，说明每种类别都倾向于以比较均匀的频率出现，从而我们较难在这个节点上得到关于样本标签的具体信息，其不确定性较高。当一个节点的纯度很高时，说明有些类别倾向于以比较高的频率出现，从而我们能够更有信心地把握这个节点样本标签的具体信息，即确定性较高。

那对于给定在n个状态上定义的离散分布$\mathbf{p}=[p_1,\ldots,p_n]^T$，如何定义度量不确定性的函数$H(\mathbf{p})$，即$H(p_1,\ldots,p_n)$呢？香农（1916-2001）于1948年，在创造信息论的著名论文《A Mathematical Theory of Communication》中指出如下定理：

> **定理（证明见论文的Appendix 2）**
>
> 若度量不确定性的函数H满足三个信息熵条件，则H的形式只能是
>
> $$H(p_1,\ldots,p_n)=-C\sum_{i=1}^{n}p_i\log p_i$$
>
> 其中，信息熵条件如下：
>
> - H关于p_i是连续函数。
> - 若$p_1=\ldots=p_n$，则H关于n单调递增。
> - 若将某一个p_i拆分为p_{i1}和p_{i2}，即$p_{i1}+p_{i2}=p_i$，则
>
> $$H(p_1,\ldots,p_{i-1},p_{i+1},\ldots,p_n,p_{i1},p_{i2})=H(p_1,\ldots,p_n)+p_iH(\frac{p_{i1}}{p_i},\frac{p_{i2}}{p_i})$$

从构造和计算的角度而言，条件一是容易满足的。对于条件二而言，假设原来箱子里分别有10个球和100个球，加入每次摸到的球都是等概率抽出的，那么100个球的箱子产生的不确定性必然是要大于10个球的箱子产生的不确定性，即H在等概率条件下关于n递增。

条件三看上去比较复杂，但其意义是容易理解的，即n个事件拆分为$n+1$个事件时的不确定性增加了，并且增加的不确定性与拆分时的比例和拆分事件的概率有关。举例来说：将$\mathbf{p}=[0.9,0.1]$分别拆分为$\mathbf{p}_1=[0.45,0.45,0.1]$和$\mathbf{p}_2=[0.9,0.05,0.05]$，那么显然$\mathbf{p}_1$增加的不确定性远超过$\mathbf{p}_2$；同时，将$\mathbf{p}=[0.9,0.1]$分别拆分为$\mathbf{p}_3=[0.8,0.1,0.1]$，那么显然$\mathbf{p}_1$增加的不确定性也远超过$\mathbf{p}_3$。

由于指标$H(\mathbf{p})$中的自变量$\mathbf{p}$是对于某个随机变量Y分布的描述，因此不妨将其记为信息熵$H(Y)$来反应Y的不确定性。对于定义在有限状态集合$\{y_1,\ldots,y_K\}$上的离散变量而言，对应信息熵的最大值在离散均匀分布时取到，最小值在单点分布时取到。此时，离散信息熵为

$$H(Y)=-\sum_{k=1}^{K}p(y_k)\log_2 p(y_k)$$

首先，我们需要定义当$p=0$时$p\log_2 p\triangleq 0$，原因在于

$$\lim_{p\to 0^+}p\log p=\lim_{p\to 0^+}\frac{\log p}{1/p}=\lim_{p\to 0^+}\frac{1/p}{-1/p^2}=\lim_{p\to 0^+}-p=0$$

离散熵的极值问题是带有约束的极值问题，记$p_k=P(Y=y_k)$和$\mathbf{p}=[p_1,\ldots,p_K]^T$，则约束条件为$1^T\mathbf{p}=1$，拉格朗日函数为

$$L(\mathbf{p})=-\mathbf{p}^T\log\mathbf{p}+\lambda(1^T\mathbf{p}-1)$$

求偏导数后可解得$\mathbf{p}^*=[\frac{1}{K},\ldots,\frac{1}{K}]^T$，此时$H(Y)=-\mathbb{E}_Y\log_2 p(Y)=\log K$。

> **补充材料**
>
> 有关向量求导的内容可参考这个wiki页面中的描述。

对于离散随机变量X，由于$p(Y)\in[0,1]$，故$-\log_2 p(Y)\geq 0$，从而$\mathbb{E}_Y[-\log_2 p(Y)]\geq 0$。注意到对于$\forall k\in\{1,\ldots,K\}$，当$p_k=1$，即$p_{k'}=0(k'\in\{1,\ldots,K\}/k)$时，$H(X)=0$。因此，离散信息熵的最小值为0且在单点分布时取到。由于$\mathbf{p}^*$是极值问题的唯一解，因此离散熵的最大值为$\log K$且在离散均匀分布时取到。

在决策树的分裂过程中，我们不但需要考察本节点的不确定性或纯度，而且还要考察子节点的平均不确定性或平均纯度来决定是否进行分裂。子节点的产生来源于决策树分支的条件，因此我们不但要研究随机变量的信息熵，还要研究在给定条件下随机变量的平均信息熵或条件熵（Conditional Entropy）。从名字上看，条件熵就是条件分布的不确定性，那么自然可以如下定义条件熵

Contents

1. 信息论基础
2. 分类树的节点分裂
3. CART树
4. 决策树的剪枝

知识回顾

$H(Y|X)$为

$$\mathbb{E}_X[\mathbb{E}_{Y|X}[-\log_2 p(Y|X)]]$$

对于离散条件熵，设随机变量X所有可能的取值为$\{x_1,\dots,x_M\}$，上式可展开为

$$-\sum_{m=1}^{M} p(x_m)\sum_{k=1}^{K} p(y_k|X=x_m)\log_2 p(y_k|X=x_m)$$

有了信息熵和条件熵的基础，我们就能很自然地定义信息增益（Information Gain），即节点分裂之后带来了多少不确定性的降低或纯度的提高。在得到了随机变量X的取值信息时，随机变量Y不确定性的平均减少量为

$$G(Y,X)=H(Y)-H(Y|X)$$

从直觉上说，随机变量Y关于X的信息增益一定是非负的，因为我们额外地知道了随机变量X的取值，这个条件降低了Y的不确定性。下面我们就从数学角度来证明其正确性。

定理

设Y和X为离散随机变量，Y关于X的信息增益$G(Y,X)$非负。

证明

由信息增益的定义得：

$$\begin{aligned}
G(Y,X)&=\mathbb{E}_Y[-\log_2 p(Y)]-\mathbb{E}_X[\mathbb{E}_{Y|X}[-\log_2 p(Y|X)]]\\
&=-\sum_{k=1}^{K} p(y_k)\log_2 p(y_k)+\sum_{m=1}^{M} p(x_m)\sum_{k=1}^{K} p(y_k|X=x_m)\log_2 p(y_k|X=x_m)\\
&=-\sum_{k=1}^{K}[\sum_{m=1}^{M} p(y_k,x_m)]\log_2 p(y_k)+\sum_{k=1}^{K}\sum_{m=1}^{M} p(x_m)\frac{p(y_k,x_m)}{p(x_m)}\log_2\frac{p(y_k,x_m)}{p(x_m)}\\
&=\sum_{k=1}^{K}\sum_{m=1}^{M} p(y_k,x_m)[\log_2\frac{p(y_k,x_m)}{p(x_m)}-\log_2 p(y_k)]\\
&=-\sum_{k=1}^{K}\sum_{m=1}^{M} p(y_k,x_m)\log_2\frac{p(y_k)p(x_m)}{p(y_k,x_m)}
\end{aligned}$$

从上式可以发现，信息增益$G(Y,X)$在本质上就是$p(y,x)$关于$p(y)p(x)$的KL散度，而KL散度的非负性由Jensen不等式可得：

$$\begin{aligned}
G(X,Y)&\geq-\log_2[\sum_{k=1}^{K}\sum_{m=1}^{M} p(y_k,x_m)\frac{p(y_k)p(x_m)}{p(y_k,x_m)}]\\
&=-\log_2[\sum_{k=1}^{K}\sum_{m=1}^{M} p(y_k,x_m)]=0
\end{aligned}$$

上述证明中的Jensen不等式的取等条件为$p(y,x)=p(y)p(x)$，其实际意义为随机变量Y和X独立。这个条件同样与直觉相符合，因为如果Y和X独立，那么意味着我们无论是否知道X的信息，都不会对Y的不确定性产生影响，此时信息增益为0。

用信息增益的大小来进行决策树的节点分裂时，由于真实的分布函数未知，故用$p(y)$和$p(y|x)$的经验分布（即频率）来进行概率的估计。若节点N每个分支下的样本数量为$D_1,\dots,D_M$，记$\tilde{p}(x_m)=\frac{D_m}{\sum_{m'=1}^{M}D_{m'}}(m\in\{1,\dots,M\})$，$\tilde{p}(y_k)$和$\tilde{p}(y_k|x_m)$分别为节点中第k个类别的样本占节点总样本的比例和第m个子节点中第k个类别的样本数量占该子节点总样本的比例，则节点N分裂的信息增益定义为

$$G_N(Y,X)=-\sum_{i=1}^{K}\tilde{p}(y_k)\log_2\tilde{p}(y_k)+\sum_{m=1}^{M}\tilde{p}(x_m)\sum_{k=1}^{K}\tilde{p}(y_k|x_m)\log_2\tilde{p}(y_k|x_m)$$

2. 分类树的节点分裂

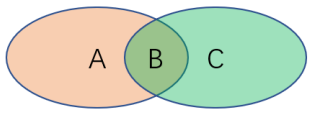
对于每个节点进行分裂决策时，我们会抽出max_features个特征进行遍历以比较信息增益的大小。特征的类别可以分为三种情况讨论：类别特征、数值特征和含缺失值的特征，它们各自的处理方法略有不同。

对于类别特征而言，给定一个阈值ϵ，树的每一个节点会选择最大信息增益$G_N^{max}(Y,X)$对应的特征进行分裂，直到所有节点的相对最大信息增益$\frac{D_N}{D_{all}}G_N^{max}(Y,X)$小于$\epsilon$，$D_N$和$D_{all}$分别指节点$N$的样本个数和整个数据集的样本个数，这种生成算法称为ID3算法。在sklearn中，ϵ即为min_impurity_decrease。

C4.5算法在ID3算法的基础上做出了诸多改进，包括但不限于：处理数值特征、处理含缺失值的特征、使用信息增益比代替信息增益以及给出树的剪枝策略。其中，剪枝策略将在第4节进行讲解，下面先对前3个改进的细节来进行介绍。

【练习】定义X,Y的联合熵为$H(Y,X)$为$\mathbb{E}_{(Y,X)\sim p(y,x)}[-\log_2 p(Y,X)]$

- 请证明如下关系：
  $G(Y,X)=H(X)-H(X|Y)$
  $G(Y,X)=H(X)+H(Y)-H(Y,X)$
  $G(Y,X)=H(Y,X)-H(X|Y)-H(Y|X)$
- 下图被分为了A、B和C三个区域。若AB区域代表X的不确定性，BC区域代表Y的不确定性，那么$H(X)$、$H(Y)$、$H(X|Y)$、$H(Y|X)$、$H(Y,X)$和$G(Y,X)$分别指代的是哪片区域?



在处理节点数值特征时，可以用两种方法来将数值特征通过分割转化为类别，它们分别是最佳分割法和随机分割法，分别对应了sklearn中splitter参数的best选项和random选项。

随机分割法下，取$s \sim U[x_{min}, x_{max}]$，其中$U[x_{min}, x_{max}]$代表特征最小值和最大值范围上的均匀分布，将节点样本按照特征$\mathbf{x}$中的元素是否超过s把样本划分为两个集合，这等价于把数值变量转换为了类别变量。此时，根据这两个类别来计算树节点分裂的信息增益，并将它作为这个数值特征分裂的信息增益。

【练习】假设当前我们需要处理一个分类问题，请问对输入特征进行归一化会对树模型的类别输出产生影响吗？请解释原因。

最佳分割法下，依次令s取遍所有的$x_i(i = 1, \dots, D_N)$，将其作为分割点，按照特征$\mathbf{x}$中的元素是否超过s把样本划分为两个集合，计算所有s对应信息增益的最大值，并将其作为这个数值特征分裂的信息增益。

C4.5算法处理缺失数据的思想非常简单，样本的缺失值占比越大，那么对信息增益的惩罚就越大，这是因为缺失值本身就是一种不确定性成分。设节点N的样本缺失值比例为γ，记非缺失值对应的类别标签和特征分别为$\tilde{Y}$和$\tilde{X}$，则修正的信息增益为

【练习】如果将系数替换为$1-\gamma^2$，请问对缺失值是加强了还是削弱了惩罚？

$$\tilde{G}(Y, X) = (1 - \gamma)G(\tilde{Y}, \tilde{X})$$

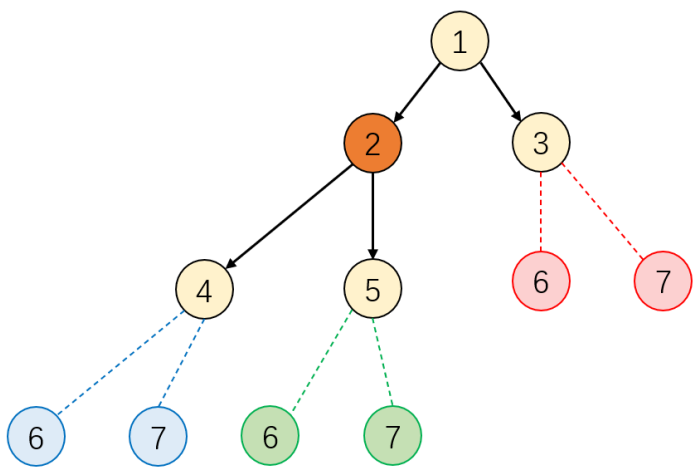
当数据完全缺失时$\gamma = 1$，信息增益为0；当数据没有缺失值时$\gamma = 0$，信息增益与原来的值保持一致。

在C4.5算法中，使用了信息增益比来代替信息增益，其原因在于信息增益来选择的决策树对类别较多的特征具有天然的倾向性，例如当某一个特征X（身份证号码、学号等）的类别数恰好就是样本数量时，此时由于$H(Y|X) = 0$，即$G(Y, X)$达到最大值，因此必然会优先选择此特征进行分裂，但这样的情况是非常不合理的。

我们在第1节已经证明了，在类别占比均匀的情况下，类别数越多则熵越高，因此我们可以使用特征对应的熵来进行惩罚，即熵越高的变量会在信息增益上赋予更大程度的抑制，由此我们可以定义信息增益比为

$$G^R(Y, X) = \frac{G(Y, X)}{H(X)}$$

在前面的部分中，我们讨论了单个节点如何选取特征进行分裂，但没有涉及到树节点的分裂顺序。例如下图所示，假设当前已经处理完了节点2的分裂，所有黄色节点（包括2号节点）都是当前已经存在的树节点，那么我们接下来究竟应该选取叶节点3号、4号和5号中的哪一个节点来继续进行决策以生成新的叶节点6号和7号？



在sklearn中提供了两种生长模式，它们分别被称为深度优先生长和最佳增益生长，当参数max_leaf_nodes使用默认值None时使用前者，当它被赋予某个数值时使用后者。

深度优先生长采用深度优先搜索的方法：若当前节点存在未搜索过的子节点，则当前节点跳转到子节点进行分裂决策；若当前节点为叶节点，则调转到上一层节点，直到根节点不存在未搜索过的子节点为止。对上图而言，当前节点为2号，它的两个子节点4号和5号都没有被搜索过，因此下一步则选择两个节点中的一个进行跳转。当决策树使用最佳增益生长时，每次总是选择会带来最大相对信息增益的节点进行分裂，直到叶节点的最大数量达到max_left_nodes。

【练习】如果将树的生长策略从深度优先生长改为广度优先生长，假设其他参数保持不变的情况下，两个模型对应的结果输出可能不同吗？

3. CART树

CART（Classification And Regression Tree）是一棵二叉树，它既能处理分类问题，又能够处理回归问题。值得注意的是，在sklearn中并没有实现处理类别特征和处理缺失值的功能，前者是因为多个类别的特征会产生多叉树，后者是因为sklearn认为用户应当自己决定缺失值的处理而不是交给模型来决定。

对于回归树而言，每个叶节点输出的不再是类别而是数值，其输出值为该叶节点所有样本标签值的均值。在每次分裂时，我们希望不同的子节点之间的差异较大，但每个子节点内部的差异较小。此时，分割策略仍然可以采用随机分割法或最佳分割法，只是现在不再以熵（条件熵）来评价节点（子节点）的纯度。

我们应当如何定义回归树的节点纯度？对于数值标签而言，我们可以认为节点间元素大小越接近则纯度越高，因此可以考虑使用均方误差（MSE）或平均绝对误差（MAE）来替换熵和条件熵的位置。

设节点N的样本标签为$y_1^{(D)}, \dots, y_N^{(D)}$，左右子节点的样本个数分别为$y_1^{(L)}, \dots, y_{N_R}^{(L)}$和$y_1^{(R)}, \dots, y_{N_R}^{(R)}$，记$\bar{y}^{(D)} = \frac{1}{N}\sum_{i=1}^{N} y_i^{(D)}$、$\bar{y}^{(L)} = \frac{1}{N_L}\sum_{i=1}^{N_L} y_i^{(L)}$和$\bar{y}^{(R)} = \frac{1}{N_R}\sum_{i=1}^{N_R} y_i^{(R)}$分别为节点$N$的样本标签均值、左子节点的样本标签均值和右子节点的样本标签均值，再记$\tilde{y}^{(D)}$、$\tilde{y}^{(L)}$和$\tilde{y}^{(R)}$分别为节点N的样本标签中位数、左子节点的样本标签中位数和右子节点的样本标签中位数。

此时，两者的信息增益可以分别定义为

$$G^{MSE}(Y, X) = \frac{1}{N}\sum_{i=1}^{N}(y_i^{(D)} - \bar{y}^{(D)})^2 - \frac{N_L}{N}\frac{1}{N_L}\sum_{i=1}^{N_L}(y_i^{(L)} - \bar{y}^{(L)})^2 - \frac{N_R}{N}\frac{1}{N_R}\sum_{i=1}^{N_R}(y_i^{(R)} - \bar{y}^{(R)})^2$$

【练习】在一般的机器学习问题中，我们总是通过一组参数来定义模型的损失函数，并且在训练集上以最小化该损失函数为目标进行优化。请问对于决策树而言，模型优化的目标是什么？

$$G^{MAE}(Y,X)=\frac{1}{N}\sum_{i=1}^{N}|y_i^{(D)}-\tilde{y}^{(D)}|-\frac{N_L}{N}\frac{1}{N_L}\sum_{i=1}^{N_L}|y_i^{(L)}-\tilde{y}^{(L)}|-\frac{N_R}{N}\sum_{i=1}^{N_R}\frac{1}{N_R}|y_i^{(R)}-\tilde{y}^{(R)}|$$

当处理分类问题时，虽然ID3或C4.5定义的熵仍然可以使用，但是由于对数函数log的计算代价较大，CART将熵中的log在$p=1$处利用一阶泰勒展开，基尼系数定义为熵的线性近似，即由于

$$H(Y)=\mathbb{E}_YI(p)=\mathbb{E}_Y[-\log_2 p(Y)]\approx\mathbb{E}_Y[1-p(Y)]$$

从而定义基尼系数为

$$\begin{aligned}\text{Gini}(Y)&=\mathbb{E}_Y[1-p(Y)]\\&=\sum_{k=1}^{K}\tilde{p}(y_k)(1-\tilde{p}(y_k))\\&=1-\sum_{k=1}^{K}\tilde{p}^2(y_k)\end{aligned}$$

类似地定义条件基尼系数为

$$\begin{aligned}\text{Gini}(Y|X)&=\mathbb{E}_X[\mathbb{E}_{Y|X}1-p(Y|X)]\\&=\sum_{m=1}^{M}\tilde{p}(x_m)\sum_{k=1}^{K}[\tilde{p}(y_k|x_m)(1-\tilde{p}(y_k|x_m))]\\&=\sum_{m=1}^{M}\tilde{p}(x_m)[1-\sum_{k=1}^{K}\tilde{p}^2(y_k|x_m)]\end{aligned}$$

从而引出基于基尼系数的信息增益为

$$G(Y,X)=\text{Gini}(Y)-\text{Gini}(Y|X)$$

【练习】对信息熵中的log函数在$p=1$处进行一阶泰勒展开可以近似为基尼系数，那么如果在$p=1$处进行二阶泰勒展开我们可以获得什么近似指标？请写出对应指标的信息增益公式。

【练习】除了信息熵和基尼系数之外，我们还可以使用节点的$1-\max_k p(Y=y_k)$和第m个子节点的$1-\max_k p(Y=y_k|X=x_m)$来作为衡量纯度的指标。请解释其合理性并给出相应的信息增益公式。

4. 决策树的剪枝

决策树具有很强的拟合能力，对于任何一个没有特征重复值的数据集，决策树一定能够在训练集上做到分类错误率或均方回归损失为0，因此我们应当通过一些手段来限制树的生长，这些方法被称为决策树树的剪枝方法。其中，预剪枝是指树在判断节点是否分裂的时候就预先通过一些规则来阻止其分裂，后剪枝是指在树的节点已经全部生长完成后，通过一些规则来摘除一些子树。

【练习】为什么对没有重复特征值的数据，决策树能够做到损失为0?

在sklearn的CART实现中，一共有6个控制预剪枝策略的参数，它们分别是最大树深度max_depth、节点分裂的最小样本数min_samples_split、叶节点最小样本数min_samples_leaf、节点样本权重和与所有样本权重和之比的最小比例min_weight_fraction_leaf、最大叶节点总数max_leaf_nodes以及之前提到的分裂阈值min_impurity_decrease。

【练习】如何理解min_samples_leaf参数能够控制回归树输出值的平滑程度?

后剪枝过程又称作MCCP过程，即Minimal Cost-Complexity Pruning，它由参数ccp_alpha控制，记其值为α。一般而言，树的叶子越多就越复杂，为了抑制树的生长，我们定义以节点N为根节点的树T^N的复杂度为该树的叶节点数量$|T^N|$。设树T的剪枝度量为

$$R_\alpha(T^N)=R(T^N)+\alpha|T^N|$$

其中，$R(T^N)$代表各个叶子节点的平均不纯度，此处的不纯度即指分类中的信息熵或者回归中的均方误差或平均绝对误差，即MCCP中的Cost部分，$\alpha|T^N|$对应的就是Complexity部分。

对于树的单个节点而言，由于此时节点数为1，故其剪枝度量为$R_\alpha(Node^N)=R(Node^N)+\alpha$。树剪枝的思想在于，如果对于决策树某一个节点为根的子树，其根的剪枝度量低于该子树的剪枝度量，那么这个根节点就没有必要分裂，即砍掉这棵子树中除了根节点以外的所有节点。

此时，我们可以得到剪枝的依据为

$$R_\alpha(Node^N)\leq R_\alpha(T^N)$$

这等价于

$$R(Node^N)+\alpha\leq R(T^N)+\alpha|T^N|$$

对上式进行移项后可得

$$E(Node^N)=\frac{R(Node^N)-R(T)}{|T^N|-1}\leq\alpha$$

这个条件表明只要$E(Node^N)$的值小于给定的参数cpp_alpha，那么这个节点下的所有节点都会被删除。事实上在sklearn中，在树完全生成后就会把所有节点的$E(Node^N)$值进行记录，每次剪枝都会分别查看所有非叶子节点的树节点对应的$E(Node^N)$值，并且对具有最小$E(Node^N)$值的非叶子节点进行剪枝，直到所有节点的$E(Node^N)$值都大于给定的cpp_alpha。

知识回顾

1. ID3树算法、C4.5树算法和CART算法之间有何异同?
2. 什么是信息增益?它衡量了什么指标?它有什么缺陷?
3. sklearn决策树中的random_state参数控制了哪些步骤的随机性?
4. 决策树如何处理连续变量和缺失变量?
5. 基尼系数是什么?为什么要在CART中引入它?
6. 什么是树的预剪枝和后剪枝?具体分别是如何操作的?

By GYH