PART3 决策树 + 集成学习

2023.06

河北师范大学 软件学院

基本内容：

1. 什么是决策树？决策树可以完成哪些可能的机器学习任务？
   - 分类、回归；
   - 决策树模型可以产生关于特征的重要性评分，有助于我们进行特征选择
   - 特征提取

2. 什么是单结点树？什么是决策树的树桩？
   - 单结点树：只有一个结点的决策树。
   - 决策树的树桩：是一种特殊的决策树模型，只含根结点和叶子结点。
     例如：CART 树的树桩，只有一个根节点和两个叶子结点。

3. 决策树与特征空间、训练样本集是什么关系？
   - 决策树的产生过程就是对特征空间的递归划分过程，决策树上含多少个叶子结点，就意味着特征空间最终划分为多少个互不相交的小区域；决策树的产生过程也是对训练集的递归划分过程，而每个叶子结点的预测结果由落入相应小区域的训练样本的标签信息来决定。

4. 本学期你学过哪些决策树模型？这些决策树可以做什么？
   以 CART 树为例，掌握它们的构建过程？
   ID3, C4.5, CART

5. 有哪些方式可以度量决策树中某个结点的不纯度？(三种方式)
   如何利用给定的训练集度量决策树根结点的不纯度？

6. 理解 ID3(分类)、C4.5(分类)、CART(分类或回归)4 种决策树模型在构建过程中，非叶子结点是如何进行特征选择的，对应的特征选择规则是什么？

7. 由于决策树的构建都是递归生成的，以根结点的特征选择为例，以决策树的根节点的特征选择为例，描述特征选择过程。

8. 分别面向分类/回归问题，掌握 CART 树的实现步骤。

   分类树的叶子结点处可以生成哪些预测值？它们是如何得到的？

   回归树的叶子结点预测值是什么？

   CART 树的叶子结点与特征空间、训练样本集是什么对应关系？

   - 分类树的叶子结点，输入样本预测其关于不同类别的后验概率，也可以产生预测类别
   - 回归树的叶子结点，可以产生该结点关于输入样本的输出值的预测。

9. 给定已知标签的训练样本集，简述基于该样本集构造 CART 回归树(或 CART 分类树)树桩的实现步骤；并指出该树桩的叶子结点输出值是如何估计出来的。

   注意：CART 树的树桩，只有一个根节点和两个叶子结点。

10. 结合课件内容：理解训练样本特征缺失时，决策树的构建过程；理解训练集的划分方式；理解部分特征缺失的样本的决策过程。

11. 面向分类或回归，掌握随机森林、Bagging 两种模型的学习步骤、以及使用方式。

12. 面向两类别分类，理解 AdaBoost 集成模型的基本思想、算法的实现步骤。

13. ID3、C4.5、CART、AdaBoost、RandomForest、Bagging 都是什么意思？

练习:

1. 在分类树的学习过程中，需要利用决策树当前结点接收的训练集对该结点进行不纯度的度量。对于C个类别的分类问题，若采用训练样本集$D=\{(x_i, \mathrm{y_i}), i=1,\dots,N\}$构建决策树模型，其中来自第$j$类的训练样本数为$N_j$，$j=1,2,\dots,C$.

   请按照如下指定的方式，估计决策树的根结点不纯度：

(1)熵不纯度；(2)基尼不纯度；(3)误差不纯度

解：

根据已知信息，分别估计各类别的概率，设第i类的概率为$\boldsymbol{P}_i$，$i = 1,2,\dots,C$

则有：$\boldsymbol{P}_i = \frac{\boldsymbol{N}_i}{\mathrm{N}}$，$i = 1,2,\dots,C$

(1)根结点的**熵不纯度：**

$$I(\boldsymbol{D}) = -\sum_{i=1}^{C} P_i log_2 P_i$$

(2)根结点的基尼不纯度：

$$I(\boldsymbol{D}) = 1 - \sum_{i=1}^{C} {P_i}^2$$

(3)根结点的误差不纯度：

$$I(\boldsymbol{D}) = 1 - max\{P_1, \dots, P_C\}$$

2. 在基于决策树的分类模型学习过程中，需要利用整个训练样本集生成决策树的根节点。对于C个类别的分类问题，若训练样本集$D = \{(x_i, \mathrm{y_i}), i = 1, \dots, N\}$其中来自第$j$类的训练样本数为$N_j, j = 1,2,\dots,C$.

   并且根节点使用特征$x^{(k)}$将数据集D分成了两个子集$D^{(1)}, D^{(2)}$,两个子集内包含的训练样本数目分别为$N^{(1)}, N^{(2)}$.各自包含的第j类的训练样本数为$N_j^{(1)}, j = 1,2,\dots,C$,以及$N_j^{(2)}, j = 1,2,\dots,C$

   回答如下问题：

(1)基于特征$x^{(k)}$划分根结点，导致的绝对增益=?

(2)基于特征$x^{(k)}$划分根结点，导致的信息增益比=?

(3)基于特征$x^{(k)}$划分根结点，导致的划分后基尼指数=?

(1)**划分前,**样本集D所在结点不纯度：

$$I_{Entropy}(D)=-\sum_{j=1}^{C}P_j\log_2 P_j=-\sum_{j=1}^{C}\frac{|D_j|}{|D|}\log_2\frac{|D_j|}{|D|}=-\sum_{j=1}^{C}\frac{N_j}{N}\log_2\frac{N_j}{N}$$

划分后,第i个子结点的不纯度：

$$I_{Entropy}\left(D^{(i)}\right)=-\sum_{j=1}^{C}\frac{|D_j^{(i)}|}{|D^{(i)}|}\log_2\frac{|D_j^{(i)}|}{|D^{(i)}|}=-\sum_{j=1}^{C}\frac{N_j^{(i)}}{N^{(i)}}\log_2\frac{N_j^{(i)}}{N^{(i)}}\qquad i=1,2$$

绝对增益：

$$Gain\left(D,x^{(k)}\right)=I_{Entropy}(D)-\sum_{i=1}^{2}\frac{|D^{(i)}|}{|D|}I_{Entropy}\left(D^{(i)}\right)=I_{Entropy}(D)-\sum_{i=1}^{2}\frac{N^{(i)}}{N}I_{Entropy}\left(D^{(i)}\right)$$

$$=\left[-\sum_{j=1}^{C}\frac{N_j}{N}\log_2\frac{N_j}{N}\right]-\sum_{i=1}^{2}\frac{|D^{(i)}|}{|D|}\left[-\sum_{j=1}^{C}\frac{N_j^{(i)}}{N^{(i)}}\log_2\frac{N_j^{(i)}}{N^{(i)}}\right]$$

(2)

$$Gain\left(D,x^{(k)}\right)=I_{Entropy}(D)-\sum_{i=1}^{2}\frac{|D^{(i)}|}{|D|}I_{Entropy}\left(D^{(i)}\right)=I_{Entropy}(D)-\sum_{i=1}^{2}\frac{N^{(i)}}{N}I_{Entropy}\left(D^{(i)}\right)$$

$$=\left[-\sum_{j=1}^{C}\frac{N_j}{N}\log_2\frac{N_j}{N}\right]-\sum_{i=1}^{2}\frac{|D^{(i)}|}{|D|}\left[-\sum_{j=1}^{C}\frac{N_j^{(i)}}{N^{(i)}}\log_2\frac{N_j^{(i)}}{N^{(i)}}\right]$$

特征$x^{(k)}$在训练集D的属性"**固有值**"$(Intrinsic\,Value,IV)$

$$IV\left(x^{(k)}\right)=-\sum_{i=1}^{2}\frac{|D^{(i)}|}{|D|}\log_2\frac{|D^{(i)}|}{|D|}=-\sum_{i=1}^{2}\frac{N^{(i)}}{N}\log_2\frac{N^{(i)}}{N}$$

所以：

根节点划分导致的信息增益比：

$$Gain\_ratio\left(D,x^{(k)}\right)=\frac{Gain\left(D,x^{(k)}\right)}{IV\left(x^{(k)}\right)}=\frac{\left[-\sum_{j=1}^{C}\frac{N_j}{N}\log_2\frac{N_j}{N}\right]-\sum_{i=1}^{2}\frac{|D^{(i)}|}{|D|}\left[-\sum_{j=1}^{C}\frac{N_j^{(i)}}{N^{(i)}}\log_2\frac{N_j^{(i)}}{N^{(i)}}\right]}{-\sum_{i=1}^{2}\frac{N^{(i)}}{N}\log_2\frac{N^{(i)}}{N}}$$

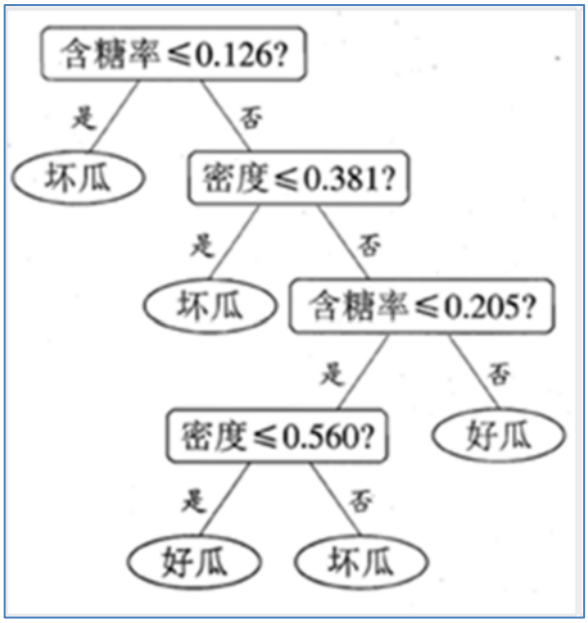
(3) 根结点划分导致的划分后基尼指数：

$$\textbf{Gini\_index}\left(D,x^{(k)}\right)=\sum_{i=1}^{2}\frac{\left|D^{(i)}\right|}{|D|}I_{Gini}\left(D^{(i)}\right)=\sum_{i=1}^{2}\frac{\left|D^{(i)}\right|}{|D|}\left[1-\sum_{j=1}^{C}\left(\frac{\left|D_j^{(i)}\right|}{\left|D^{(i)}\right|}\right)^2\right]$$

$$=\sum_{i=1}^{2}\frac{N^{(i)}}{N}\left[1-\sum_{j=1}^{C}\left(\frac{N_j^{(i)}}{N^{(i)}}\right)^2\right]$$

3. 如图所示的决策树为实用下表所示的训练集学习得到的。

编号	密度	含糖率	好瓜
1	0.697	0.460	是
2	0.774	0.376	是
3	0.634	0.264	是
4	0.608	0.318	是
5	0.556	0.215	是
6	0.403	0.237	是
7	0.481	0.149	是
8	0.437	0.211	是
9	0.666	0.091	否
10	0.243	0.267	否
11	0.245	0.057	否
12	0.343	0.099	否
13	0.639	0.161	否
14	0.657	0.198	否
15	0.360	0.370	否
16	0.593	0.042	否
17	0.719	0.103	否



完成如下工作：

(1)估计决策树根结点的熵不纯度；

(2)估计决策树根结点的基尼不纯度；

(3)上述决策树根结点的分裂，导致的绝对增益=？

(4)上述决策树根结点的分裂，导致的信息增益比=？

(5)估计该决策树最左侧叶子结点的熵不纯度；

(6若某个西瓜样本的密度值为0.7，含糖率为0.4， 该西瓜最可能为哪种类别？

3. 什么是决策树？什么是决策树的树桩？CART回归的树桩有什么特点？

4. 对于<u>一元连续实值函数$y=f(x)$的回归</u>，若采用训练样本集$\boldsymbol{D}=\{(\boldsymbol{x}_i,\mathrm{y_i}),i=1,\ldots,N\}$构建CART决策树树桩，其中 $\boldsymbol{x_i}\in R$， $y_i\in R$.按要求完成如下工作：(1)请详细描述其实现过程，并指出该树桩的所有叶子结点的预测输出如何得到；(2)对于任意

观测样本$\boldsymbol{x} \in R$，**如何基于该模型，预测可能的输出？**

思考：CART分类树的树桩如何构造？如何生成叶子结点的输出？

基于最小二乘准则的特征选择。

STEP1. 将给定的训练集中，各样本的特征取值从小到达进行排序

得：$x^{(1)} \le x^{(2)} \le \ldots \le x^{(N)}$

STEP2. 对于**切分点**$s \in \left\{ \frac{x^{(1)}+x^{(2)}}{2}, \frac{x^{(2)}+x^{(3)}}{2}, \ldots, \frac{x^{(N-1)}+x^{(N)}}{2} \right\}$

求解：$\left[s^*, c_1^*, c_2^* \right] = \arg\min_{s,c_1,c_2} \left[\sum_{x_i \le s} (y_i - c_1)^2 + \sum_{x_i > s} (y_i - c_2)^2 \right]$

相当于：求取与$\min_s \left[\min_{c_1} \sum_{x_i \le s} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i > s} (y_i - c_2)^2 \right]$

对应的s^*, c_1^*, c_2^*

STEP3. 产生**CART**回归树树桩对应的预测函数：

$$f(x) = \begin{cases} c_1^* & \text{若} x \le s^* \\ c_2^* & \text{若} x > s^* \end{cases} \quad \text{（预测过程）}$$

$$R_1(s^*) = \{x | x \le s^*\}, R_2(j^*, s^*) = \{x | x > s^*\}$$

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m(s^*)} y_i, \quad m = 1, 2$$

5. 对于<u>多元连续实值函数$y = f(x)$的回归</u>，若采用训练样本集$\boldsymbol{D} = \{(\boldsymbol{x}_i, \mathrm{y_i}), i = 1, \ldots, N\}$构建CART决策树模型，其中 $\boldsymbol{x_i} \in \boldsymbol{R^d}$，$y_i \in R$.(1)请详细描述CART决策树根结点的特征选择过程，并明确特征选择所使用的规则；(2) 若该决策树只由根结点和叶子结点组成，，对于任意观测样本$\boldsymbol{x} \in \boldsymbol{R^d}$，**如何基于该决策树，对其输出**y进行预测，请给出可能的预测结果.

解：(2)

STEP1. 从d维特征向量$\boldsymbol{x}$中选择**最优切分变量**j^*及**切分点**s^*.

使之满足最小二乘准则：

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

STEP2. 用上述$\left(j^*,s^*\right)$对，确定划分区域$R_1\left(j^*,s^*\right)$，$R_2\left(j^*,s^*\right)$，并确定相应输出值。

$$R_1\left(j^*,s^*\right)=\left\{x|\, x^{(j^*)}\le s^*\right\},\ R_2\left(j^*,s^*\right)=\left\{x|\, x^{(j^*)}> s^*\right\}$$

$$\hat{c}_m=\frac{1}{N_m}\sum_{x_i\in R_m\left(j^*,s^*\right)} y_i,\quad x\in R_m,\qquad m=1,2$$

基于上述两步，最终 CART 决策树有两个叶子结点，其中，左、右叶子结点的预测输出分别为$\hat{c}_1,\hat{c}_2$.

对于任意观测样本x，若$x^{(j)}\le s^*$，则将该样本的输出预测为$\hat{c}_1$，否则预测为$\hat{c}_2$.

6. 对于连续特征空间的分类问题，若采用训练样本集$D=\{(x_i,y_i),i=1,\ldots,N\}$构建 CART 决策树模型，其中 $x_i\in R^d$，$y_i\in\{1,2,...,C\}$.

按要求完成如下工作：(1)请详细描述 CART 决策树根结点的特征选择过程，并明确特征选择所使用的规则；(2) 若该决策树只由根结点和叶子结点组成，，对于任意观测样本$x\in R^d$，**如何基于该决策树，对其输出**y进行预测，请给出可能的预测结果.

7. 随机森林。随机森林是一种基于单一机器学习算法生成的多个个体模型的并行集成方式。给定训练样本集$D=\{(x_i,y_i),i=1,\cdots,N\}$，其中$x_i\in R^p$，$y_i\in\{1,2,...,C\}$. 设样本数目 N 足够大，特征维数 p 足够大。**若个体模型是基于 CART 的分类树**，并且个体模型数目为 J ，给定 CART 分类树结点的生成函数f。请面向分类问题，完成如下工作：

(1)简述基于随机森林的集成分类模型的实现步骤；

(2)基于随机森林模型，对任意观测样本$x\in R^p$的输出进行预测。

8. 随机森林是一种以决策树为个体模型的集成学习模型。对于实值函数的回归问题，给定训练样本集$D=\{(x_i,y_i),i=1,\cdots,N\}$，其中$x_i\in R^p$，$y_i\in R$. 设样本数目 N 及特征维数 d 足够大，**若个体模型为 CART 形式的回归树**，并且个体模型数目为 J，请设计基于随机森林的回归模型，并基于该模型对任意观测样本$x\in R^p$的输出预测.

随机森林分类或回归模型的学习：

Algorithm 2 Random Forests

Let $\mathscr{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ denote the training data, with $x_i = (x_{i,1}, \ldots, x_{i,p})^T$. For $j = 1$ to J:

1. Take a bootstrap sample $\mathscr{D}_j$ of size N from $\mathscr{D}$.
2. Using the bootstrap sample $\mathscr{D}_j$ as the training data, fit a tree using binary recursive partitioning
   a. Start with all observations in a single node.
   b. Repeat the following steps recursively for each unsplit node until the stopping criterion is met:
      (i) Select m predictors at random from the p available predictors.
      (ii) Find the best binary split among all binary splits on the m predictors from Step (i).
      (iii) Split the node into two descendant nodes using the split from Step (ii).

模型的学习

模型的使用：

To make a prediction at a new point x,

- $\hat{f}(x) = \frac{1}{J}\sum_{j=1}^{J} \hat{h}_j(x)$ for regression
- $\hat{f}(x) = \arg\max_y \sum_{j=1}^{J} I\left(\hat{h}_j(x) = y\right)$ for classification

where $\hat{h}_j(x)$ is the prediction of the response variable at x using the jth tree (Algorithm 1).

从p个特征中随机抽取的特征数目m经验值：

$m = \sqrt{p}$　　$m = \log_2 p$　　$m = \frac{p}{3}$

9. **Bagging 是一种基于单一机器学习算法生成的多个个体模型的并行集成方式。给定训练样本集 $\boldsymbol{D}=\{(\boldsymbol{x}_i, y_i), \mathrm{i}=1,\cdots,N\}$，其中 $\boldsymbol{x}_i \in R^d$。设样本数目 N 足够大，若个体模型是基于 CART 的决策树，并且个体模型数目为 M．请完成如下任务：若 $y_i \in R$，$\mathrm{i}=1,\cdots,N$，设计基于 Bagging 回归模型，对任意观测样本$\boldsymbol{x} \in \boldsymbol{R}^d$的输出进行预测.**

10. 可使用基于决策树的个体模型构建 Bagging 集成模型。给定训练样本集 $\boldsymbol{S}=\{(\boldsymbol{x}_i, y_i), \mathrm{i}=1,\cdots,N\}$，其中 $\boldsymbol{x}_i \in R^d$，$y_i \in \{1,2,...,C\}$。设样本数目 N 及特征维数 d 足够大。若个体模型的数目为 T，请设计基于 Bagging 的分类模型，对任意观测样本$\boldsymbol{x} \in R^d$进行类别预测。

Algorithm 1 Bagging

Inputs: Training data S; supervised learning algorithm, BaseClassifier, integer T specifying ensemble size; percent R to create bootstrapped training data.

Do $t = 1, \ldots, T$

1. Take a bootstrapped replica S_t by randomly drawing $R\%$ of S.
2. Call **BaseClassifier** with S_t and receive the hypothesis (classifier) h_t.
3. Add h_t to the ensemble, $\mathcal{E} \leftarrow \mathcal{E} \cup \quad h_t$.

End

Ensemble Combination: Simple Majority Voting—Given unlabeled instance x

1. Evaluate the ensemble $\mathcal{E} = \{h_1, \ldots, h_T\}$ on x.
2. Let $v_{t,c} = 1$ if h_t chooses class ω_c, and 0, otherwise.
3. Obtain total vote received by each class

$$V_c = \sum_{t=1}^{T} v_{t,c}, \quad c = 1, ..., C$$

Output: Class with the highest V_c.

输入： 训练样本集$D=\{(x_i,y_i),i=1,...,m\}$;

监督式基学习器算法ℓ; 基学习器的数目 T;

用于产生每个自举样本集的百分比R.

模型的学习阶段:

初始化**基学习模型的集合**E为空集.

Do $t=1,...,T$

由数据集D随机抽取$R\%$的训练样本构成数据集 D_t;

基于数据集D_t，调用基学习器算法 ℓ，学习得到个体模型 $h_t(x)$;

更新E: $E \leftarrow E \cup \{h_t(x)\}$

End

模型的使用阶段:

对于任意观测x，集成预测

$$\begin{cases} \text{若为实值函数回归，则 } \hat{y}=\frac{1}{T}\sum_{t=1}^{T}\hat{h}_t(x) \\ \text{若为分类，则 } \hat{y}=\underset{j\in\{1,2,...,C\}}{\mathbf{argmax}}\sum_{t=1}^{T}I\left(\hat{h}_t(x)=j\right) \end{cases}$$

输出: $\hat{y}$　　注意:$\hat{h}_t(x)$为个体模型$h_t(\cdot)$在x处产生的预测输出.