PART3 决策树 + 集成学习

2020-06-1

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？
2. 什么是单结点树？什么是决策树的树桩？

   分类、回归；

   分类树、回归树的每个非叶子结点的生成过程伴随着特征选择

   特征提取

3. **决策树与特征空间、训练样本集是什么关系？**
4. 本学期你学过哪些决策树模型？ID3, C4.5, CART
5. 据你所知，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)三种决策树模型在构建过程中，结点是如何进行特征选择的。

7. 分别面向分类/回归问题，掌握 CART 树的实现步骤。分类树的叶子结点预测值是什么？回归树的叶子结点预测值是什么？它们与特征空间是什么对应关系？

   分类：叶子结点两种预测结果的输出方式(类别；各类别的后验概率)

   回归：标量值

8. 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。

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

10. 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。

11. ID3、C4.5、CART、AdaBoost、RandomForest(RF)、Bagging 都是什么意思？

练习：

1. **在基于决策树的分类模型学习过程中，需要利用训练样本对决策树的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)信息熵；

(2)基尼指数.

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P_i，$i=1,2,\dots,C$

则有：$P_i=\frac{N_i}{\mathrm{N}}$，$i=1,2,\dots,C$

(1)根结点的**熵不纯度**：

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

(2)根结点的基尼不纯度：

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

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$N_j$，$j=1,2,\dots,C$.**

并且根节点使用某个特征$x^{(k)}$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$N_j^{(1)}$，$j=1,2,\dots,C$，　$N_j^{(2)}$，$j=1,2,\dots,C$

分别针对ID3,C4.5,CART三情况下分类树的构建，回答：

(1)若要构建ID3分类树，那么根结点划分导致的绝对增益=？

(2)若要构建C4.5分类树，那么根结点划分导致的信息增益比=？

(3)若是CART分类树，那么根结点划分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：

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

*ID3*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)若要构建C4.5分类树，那么根结点划分导致的信息增益比=?

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

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

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：

*C*4.5决策树根节点处的信息增益比：

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

(3)若是CART分类树，那么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. (1)什么是决策树？什么是决策树的树桩？CART回归树的树桩有什么特点？

4. 对于单变量实值函数$\boldsymbol{y=f(x)}$的回归，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：$x^{(1)}<x^{(2)}<\ldots<x^{(N)}$

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

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

对应的s^*,c_1^*,c_2^*

STEP3. 产生**CART**回归树树桩对应的预测函数：

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

$$R_1\left(s^*\right)=\left\{x_i|\,x_i\le s^*,\left(x_i,\mathrm{y}_i\right)\in D\right\},R_2\left(s^*\right)=\left\{x_i|\,x_i>s^*,\left(x_i,\mathrm{y}_i\right)\in D\right\}$$

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

5. 对于多元实值函数$\boldsymbol{y}=\boldsymbol{f}(\boldsymbol{x})$的回归，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 $\boldsymbol{x_i}\in\boldsymbol{R^d}$，$y_i\in R$.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：

(1) 基于最小二乘准则，进行决策树的根节点以及中间结点的特征选择，每次，结点一分为二. 首先，针对每个可能的切分特征，将所有训练样本关于该特征取值按照从小到大顺序排序，分别以相邻取值中间值为备选切分点，选择具有最小预测误差平方和的位置为对应于该特征的最优切分点，最后选择具有最小预测误差平方和的特征及切分点，将根结点一分为二，得到左右两个子结点.

（2）

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

使之满足最小二乘准则：

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

STEP2. 用上述$\left(\boldsymbol{j^*},\boldsymbol{s^*}\right)$对，确定划分区域$\boldsymbol{R}_1\left(\boldsymbol{j^*},\boldsymbol{s^*}\right)$，$\boldsymbol{R}_2\left(\boldsymbol{j^*},\boldsymbol{s^*}\right)$，并确定相应输出值。

$$\boldsymbol{R}_1\left(\boldsymbol{j^*},\boldsymbol{s^*}\right)=\left\{\boldsymbol{x}|\,\boldsymbol{x}^{(j^*)}\le\boldsymbol{s^*}\right\},\boldsymbol{R}_2\left(\boldsymbol{j^*},\boldsymbol{s^*}\right)=\left\{\boldsymbol{x}|\,\boldsymbol{x}^{(j^*)}>\boldsymbol{s^*}\right\}$$

$$c_{\mathrm{m}}^*=\frac{1}{\boldsymbol{N}_m}\sum_{x_i\in\boldsymbol{R}_m\left(j^*,s^*\right)}y_i,\quad\boldsymbol{x}\in\boldsymbol{R}_m,\qquad m=1,2$$

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为c_1^* c_2^*.

对于任意观测样本$\boldsymbol{x}$，若$\boldsymbol{x}^{(j)}\le\boldsymbol{s^*}$，则将该样本的输出预测为$c_1^*$，否则为$c_2^*$.

6. 对于连续特征空间的分类问题，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 $\boldsymbol{x_i}\in\boldsymbol{R^d}$，$y_i\in\{1,2,\ldots,\mathrm{C}\}$.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$\boldsymbol{x} \in \boldsymbol{R}^{\boldsymbol{d}}$，**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 $\boldsymbol{D}=\{(\boldsymbol{x}_i, y_i), \mathrm{i}=1,\cdots,N\}$，其中 $\boldsymbol{x}_i \in R^d$，$y_i \in \{1,2,...,\mathrm{C}\}$. 设样本数目 N 足够大，特征维数 d 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 M ，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$\boldsymbol{x} \in R^d$的输出进行预测。

答：(1)

输入：训练样本集$\boldsymbol{D}=\{(\boldsymbol{x}_i, \boldsymbol{y}_i), \boldsymbol{i}=1,...,\boldsymbol{N}\}$,其中$\boldsymbol{x}_i=[x_{i1},...,x_{id}]^T$；
CART决策树结点生成函数f；决策树的数目$\boldsymbol{M}$
特征子集的特征容量$\boldsymbol{p}$

模型的学习阶段：

初始化**基学习模型的集合**$\boldsymbol{E}$为空集.

Do $\mathrm{t}=1,...,\mathrm{M}$

(1)由数据集$\boldsymbol{D}$自举重采样得容量为$\boldsymbol{N}$的数据集$\boldsymbol{D}_t$

(2)重复如下过程，递归划分决策树的结点，直到满足终止条件

(2.1)由$\boldsymbol{d}$个特征中随机抽取$\boldsymbol{p}$个特征构成特征子集

(2.2)调用决策树结点生成函数f，基于该特征子集以及$\boldsymbol{D}_t$，寻找最好的切分特征及切分点

(2.3)基于最好的切分特征及切分点，分裂该结点，得到两个后继子结点。

(3)返回决策树模型$h_t(\boldsymbol{x})$，并更新$\boldsymbol{E}$: $\boldsymbol{E} \leftarrow \boldsymbol{E} \cup \{h_t(\boldsymbol{x})\}$

End

(2)

对于任意观测$\boldsymbol{x}$，采用投票法预测该样本的类别：

$$y^* = \underset{j \in \{1,2,...,C\}}{\mathbf{argmax}} \sum_{t=1}^{M} I\left(\hat{h}_t(\boldsymbol{x}) = \boldsymbol{j}\right)$$

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 $\boldsymbol{D}=\{(\boldsymbol{x}_i,y_i),\mathrm{i}=1,\cdots,N\}$，其中 $\boldsymbol{x}_i\in R^d$， $y_i\in R$. 设样本数目 N 及特征维数 d 足够大, 若个体模型为 CART 形式的决策树，并且个体模型数目为 M，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$\boldsymbol{x}\in R^d$的输出预测.

解：

输入： 训练样本集$\boldsymbol{D}=\{(\boldsymbol{x}_i,y_i),\boldsymbol{i}=1,...,\boldsymbol{N}\}$,其中$\boldsymbol{x}_i=[x_{i1},...,x_{id}]^T$；

CART回归树的结点生成函数f；决策树的数目$\boldsymbol{M}$

特征子集的特征容量$\boldsymbol{p}$

模型的学习阶段：

初始化**基学习模型的集合**$\boldsymbol{E}$为空集.

Do t = 1,...,M

(1)由数据集$\boldsymbol{D}$自举重采样得容量为$\boldsymbol{N}$的数据集$\boldsymbol{D}_t$

(2)重复如下过程，递归划分决策树的结点，直到满足终止条件

(2.1)由$\boldsymbol{d}$个特征中随机抽取$\boldsymbol{p}$个特征构成特征子集

(2.2)调用回归树结点生成函数f，基于该特征子集以及$\boldsymbol{D}_t$，寻找最好的切分特征及切分点

(2.3)基于最好的切分特征及切分点，分裂该结点，得到两个后继子结点。

(3)返回决策树模型$h_t(\boldsymbol{x})$，并更新$\boldsymbol{E}$: $\boldsymbol{E}\leftarrow\boldsymbol{E}\cup\{h_t(\boldsymbol{x})\}$

End

模型的使用阶段：

对于任意观测$\boldsymbol{x}$，输出预测$\hat{\boldsymbol{y}}=\frac{1}{\boldsymbol{M}}\sum_{t=1}^{\boldsymbol{M}}\hat{h}_t(\boldsymbol{x})$

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 集成回归模型，对任意观测样本$\boldsymbol{x}\in\boldsymbol{R}^{\boldsymbol{d}}$的输出进行预测.**

解：

模型的学习阶段：

初始化CART决策树**基学习模型的集合**$\boldsymbol{E}$为空集.

Do $t=1,\ldots,M$

由数据集$\boldsymbol{D}$自举重采样得容量为$\boldsymbol{N}$的数据集$\boldsymbol{D}_t$;

基于数据集$\boldsymbol{D}_t$，调用CART决策树基学习器算法ℓ，得个体回归树模型$h_t(\boldsymbol{x})$;

更新$\boldsymbol{E}$: $\boldsymbol{E} \leftarrow \boldsymbol{E} \cup \{h_t(\boldsymbol{x})\}$

End

模型的使用阶段：

对于任意观测$\boldsymbol{x}$，集成预测输出$\hat{\boldsymbol{y}}=\frac{1}{\boldsymbol{M}}\sum_{t=1}^{\boldsymbol{M}} h_t(\boldsymbol{x})$

10. Bagging 和随机森林均是以决策树为个体模型的两种并行集成学习模型。给定训练样本集 $\boldsymbol{D}=\{(\boldsymbol{x}_i, y_i), \mathrm{i}=1,\cdots,N\}$，其中 $\boldsymbol{x}_i \in R^d$， $y_i \in \{1,2,...,C\}$。设样本数目 N 及特征维数 d 足够大。若两种集成模型中，个体模型的数目为 M。请分别进行如下算法描述。

(1)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。

(2)设计基于随机森林的分类模型，对任意观测样本$\boldsymbol{x} \in R^d$进行类别预测。