1.随机森林简介

随机森林(Random Forest)是一个非常灵活的机器学习方法，从市场营销到医疗保险有着众多的应用。例如用于市场营销对客户获取和存留建模或预测病人的疾病风险和易感性。随机森林能够用于**分类**和**回归**问题，可以处理大量特征，并能够帮助估计用于建模数据变量的重要性。我们先了解随机森林中**森林**和**随机**的概念。

1.1集成学习

集成学习是将多个模型进行组合来解决单一的预测问题。其原理是生成多个分类器模型，各自独立的学习并做出预测，这些预测最后结合起来得到预测结果，因此和单独分类器相比结果会更好。

单个决策树在机器学习中比作普通学习，那么成百上千棵决策树便叫做集成学习，成百上千棵树也便组成了**森林**。

1.2随机决策树

我们知道随机森林是将其他的模型进行聚合， 但具体是哪种模型呢？从其名称也可以看出，随机森林聚合的是分类（或回归） 树。

那么我们如何生成成百上千棵决策树呢？如果选择样本集N中全部数据生成众多决策树，那么生成的决策树都相同，得到预测结果便没有实际意义。因此我们采用的方法是从样本集N中有放回的随机采样选出n个样本(n<N)，然后从所有特征中选出k个特征生成单个随机决策树，这便是随机森林中**随机**的概念。

1.3随机森林算法

由于这些树是随机生成的，大部分的树对解决分类或回归问题是没有意义的，那么生成上万的树有什么好处呢?

好处便是生成的决策树中有少数非常好的决策树。当你要做预测的时候，新的观察值随着决策树自上而下的预测并被赋予一个预测值或标签。一旦森林中的每棵树都有了预测值或标签，所有的预测结果将被归总到一起，所有树的投票做为最终的预测结果。简单来说，99.9%不相关的树做出的预测结果涵盖所有的情况，这些预测结果将会彼此抵消。少数优秀的树将会脱颖而出，从而得到一个好的预测结果。随机森林算法如下所示

- 从样本集N中有放回随机采样选出n个样本。
- 从所有特征中随机选择k个特征，对选出的样本利用这些特征建立决策树(一般是CART方法)。
- 重复以上两步m次，生成m棵决策树，形成随机森林，其中生成的决策树不剪枝。
- 对于新数据，经过每棵决策树投票分类。

2.CART算法

随机森林包含众多决策树，能够用于**分类**和**回归**问题。决策树算法一般包括ID3、C4.5、CART算法，这里我们给出CART(分类与回归树)算法的详细推导过程。

2.1CART分类树算法推导

CART分类树预测离散型数据，采用基尼指数选择最优特征，同时决定该特征的最优二值切分点。分类过程中，假设有K个类，样本点属于第k个类的概率为Pk，则概率分布的基尼指数定义为

$$Gini(p)=\sum_{k=1}^{m}p_k(1-p_k)=1-\sum_{k=1}^{K}p_k^2$$

根据基尼指数定义，可以得到样本集合D的基尼指数，其中Ck表示数据集D中属于第k类的样本子集。

$$Gini(D)=1-\sum_{k=1}^{K}\left(\frac{|C_k|}{|D|}\right)^2$$

如果数据集D根据特征A在某一取值a上进行分割，得到D1,D2两部分后，那么在特征A下集合D的基尼系数如下所示。其中基尼系数Gini(D)表示集合D的不确定性，基尼系数Gini(D,A)表示A=a分割后集合D的不确定性。基尼指数越大，样本集合的不确定性越大。

$$Gain\_Gini(D,A)=\frac{|D1|}{|D|}Gini(D_1)+\frac{|D1|}{|D|}Gini(D_2)$$

对于属性A，分别计算任意属性值将数据集划分为两部分之后的Gain_Gini，选取其中的最小值，作为属性A得到的最优二分方案。然后对于训练集S，计算所有属性的最优二分方案，选取其中的最小值，作为样本及S的最优二分方案。

$$\min_{i \epsilon A}(Gain\_Gini(D,A))$$

$$\min_{A \epsilon Attribute}(\min_{i \epsilon A}(Gain\_Gini(D,A)))$$

2.2CART分类树实例详解

名称	体温	胎生	水生	类标记
人	恒温	是	否	哺乳类
巨蟒	冷血	否	否	爬行类
鲑鱼	冷血	否	是	鱼类
鲸	恒温	是	是	哺乳类
蛙	冷血	否	有时	鱼类
巨蜥	冷血	否	否	爬行类
蝙蝠	恒温	是	否	哺乳类
猫	恒温	是	否	哺乳类
豹纹鲨	冷血	是	是	鱼类
海龟	冷血	否	有时	爬行类
豪猪	恒温	是	否	哺乳类
鳗	冷血	否	是	鱼类
蝾螈	冷血	否	有时	两栖类

针对上述离散型数据，按照**体温为恒温和非恒温**进行划分。其中恒温时包括哺乳类5个、鸟类2个，非恒温时包括爬行类3个、鱼类3个、两栖类2个，如下所示我们计算D1,D2的基尼指数。

$$Gini(D_1) = 1 - [(\frac{5}{7})^2 + (\frac{2}{7})^2] = \frac{20}{49}$$

$$Gini(D_2) = 1 - [(\frac{3}{8})^2 + (\frac{3}{8})^2 + (\frac{2}{8})^2] = \frac{42}{64}$$

然后计算得到特征**体温**下数据集的Gini指数，最后我们选择Gain_Gini最小的特征和相应的划分。

$$Gain\_Gini(D, \text{体温}) = \frac{7}{15} * \frac{20}{49} + \frac{8}{15} * \frac{42}{64}$$

2.3CART回归树算法详解

CART回归树预测连续型数据，假设X与Y分别是输入和输出变量，并且Y是连续变量。在训练数据集所在的输入空间中，递归的将每个区域划分为两个子区域并决定每个子区域上的输出值，构建二叉决策树。

$$D=\{(x_1,y_1),(x_2,y_2),(x_3,y_3),\dots(x_n,y_n)\}$$

选择最优切分变量j与切分点s：遍历变量j，对规定的切分变量j扫描切分点s，选择使下式得到最小值时的(j,s)对。其中Rm是被划分的输入空间，cm是空间Rm对应的固定输出值。

$$\min_{j,s}[\min_{c_1}\sum_{x_i\epsilon R_i(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\epsilon R_i(j,s)}(y_i-c_2)^2]$$

用选定的(j,s)对，划分区域并决定相应的输出值

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

$$\hat{c}_m=\frac{1}{N_m}\sum_{x_i\epsilon R_m(j,s)}y_i$$

$$x\epsilon R_m,m=1,2$$

继续对两个子区域调用上述步骤，将输入空间划分为M个区域R1,R2,...,Rm，生成决策树。

$$f(x)=\sum_{m=1}^{M}\hat{c}_mI(x\epsilon R_m)$$

当输入空间划分确定时，可以用**平方误差**来表示回归树对于训练数据的预测方法，用平方误差最小的准则求解每个单元上的最优输出值。

$$\sum_{x_i\epsilon R_m}(y_i-f(x_i))^2$$

2.4CART回归树实例详解

x_i	1	2	3	4	5	6	7	8	9	10
y_i	5.56	5.70	5.91	6.40	6.80	7.05	8.90	8.70	9.00	9.05

考虑如上所示的连续性变量，根据给定的数据点，考虑**1.5,2.5,3.5,4.5,5.5,6.5,7.5,8.5,9.5**切分点。对各切分点依次求出**R1,R2,c1,c2及m(s)**，例如当切分点s=1.5时，得到R1={1},R2={2,3,4,5,6,7,8,9,10}，其中c1,c2,m(s)如下所示。

$$c_1=\frac{1}{N_m}\sum_{x_i\epsilon R_m(j,s)}y_i=\frac{1}{1}\sum_{x_i\epsilon R_1(1,1.5)}5.56=5.56$$

$$c_2=\frac{1}{N_m}\sum_{x_i\epsilon R_m(j,s)}y_i=\frac{1}{9}\sum_{x_i\epsilon R_2(1,1.5)}(5.70+5.91+\dots+9.05)=7.50$$

$$m(s)=\min_{j,s}[\min_{c_1}\sum_{x_i\epsilon R_i(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\epsilon R_i(j,s)}(y_i-c_2)^2]=0+15.72=15.72$$

依次改变(j,s)对，可以得到s及m(s)的计算结果，如下表所示。

s	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5
$m(s)$	15.72	12.07	8.36	5.78	3.91	1.93	8.01	11.73	15.74

当x=6.5时，此时R1={1,2,3,4,5,6},R2={7,8,9,10},c1=6.24,c2=8.9。**回归树T1(x)**为

$$T_1(x)=\begin{cases}6.24, x<6.5\\8.91, x\geq 6.5\end{cases}$$

$$f_1(x)=T_1(x)$$

然后我们利用f1(x)拟合训练数据的残差，如下表所示

x_i	1	2	3	4	5	6	7	8	9	10
y_i	-0.68	-0.54	-0.33	0.16	0.56	0.81	-0.01	-0.21	0.09	0.14

用f1(x)拟合训练数据得到平方误差

$$L(y,f_1(x))=\sum_{i=1}^{10}(y_i-f_1(x_i))^2=1.93$$

第二步求T2(x)与求T1(x)方法相同，只是拟合的数据是上表的残差。可以得到

$$T_2(x)=\begin{cases}-0.52, x<3.5\\0.22, x\geq 3.5\end{cases}$$

$$f_2(x)=f_1(x)+T_2(x)=\begin{cases}5.72, x<3.5\\6.46, 3.5\leq x\leq 6.5\\9.13, x\geq 6.5\end{cases}$$

用f2(x)拟合训练数据的平方误差

$$L(y,f_2(x))=\sum_{i=1}^{10}(y_i-f_2(x_i))^2=0.79$$

继续求得T3(x)、T4(x)、T5(x)、T6(x)，如下所示

$$T_3(x)=\begin{cases}0.15, x<6.5\\-0.22, x\geq 6.5\end{cases}\ L(y,f_3(x))=0.47$$

$$T_4(x)=\begin{cases}-0.16, x<4.5\\0.11, x\geq 4.5\end{cases}\ L(y,f_4(x))=0.30$$

$$T_5(x)=\begin{cases}0.07, x<6.5\\-0.11, x\geq 6.5\end{cases}\ L(y,f_5(x))=0.23$$

$$T_6(x)=\begin{cases}-0.15, x<2.5\\0.04, x\geq 2.5\end{cases}$$

$$f_6(x)=f_5(x)+T_6(x)=T_1(x)+\ldots+T_6(x)=\begin{cases}5.63, x<2.5\\5.82, 2.5\leq x\leq 3.5\\6.56, 3.5\leq x\leq 4.5\\6.83, 4.5\leq x\leq 6.5\\8.95, x\geq 6.5\end{cases}$$

用f6(x)拟合训练数据的平方损失误差如下所示，假设此时已经满足误差要求，那么f(x)=f6(x)便是所求的回归树。

$$L(y, f_6(x)) = \sum_{i=1}^{10}(y_i - f_6(x_i))^2 = 0.71$$

3.Sklearn实现随机森林

我们经常需要通过改变参数来让模型达到更好的分类或回归结果，具体参数设置可参考sklearn官方教程。

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.datasets import make_classification

X,y=make_classification(n_samples=1000,n_features=4,
                        n_informative=2,n_redundant=0,
                        random_state=0,shuffle=0)
print(X[:10],y[:10])
#  X
# [[-1.66853167 -1.29901346  0.2746472  -0.60362044]
#  [-2.9728827  -1.08878294  0.70885958  0.42281857]
#  [-0.59614125 -1.37007001 -3.11685659  0.64445203]
#  [-1.06894674 -1.17505738 -1.91374267  0.66356158]
#  [-1.30526888 -0.96592566 -0.1540724   1.19361168]
#  [-2.18261832 -0.97011387 -0.09816121 -0.88661426]
#  [-1.24797892 -1.13094525 -0.14735366  1.05980629]
#  [-1.35308792 -1.06633681  0.02624662 -0.11433516]
#  [-1.13449871 -1.27403448  0.74355352  0.21035937]
#  [-0.38457445 -1.08840346 -0.00592741  1.36606007]]

#  y
# [0 0 0 0 0 0 0 0 0 0]

clf=RandomForestClassifier(max_depth=2,random_state=0)
clf.fit(X,y)
print(clf.feature_importances_)
# [ 0.17287856  0.80608704  0.01884792  0.00218648]
print(clf.predict([[0,0,0,0]]))
# [1]
```

4.随机森林优缺点

4.1优点

- 决策树选择部分样本及部分特征，一定程度上避免过拟合。
- 决策树随机选择样本并随机选择特征，模型具有很好的抗噪能力，性能稳定。
- 能够处理高维度数据，并且不用做特征选择，能够展现出哪些变量比较重要。
- 对缺失值不敏感，如果有很大一部分的特征遗失，仍可以维持准确度。

- 训练时树与树之间是相互独立的，训练速度快，容易做成并行化方法。
- 随机森林有oob，不需要单独划分交叉验证集。

4.2缺点

- 可能有很多相似决策树，掩盖真实结果。
- 对小数据或低维数据可能不能产生很好分类。
- 产生众多决策树，算法较慢。

5.推广

更多内容请关注公众号**谓之小一**，若有疑问可在公众号后台提问，随时回答，欢迎关注，内容转载请注明出处。

长按关注微信公众号