**Decision Tree**

**1. 对于树形结构为什么不需要归一化？**

数值缩放，不影响分裂点位置。因为第一步都是按照特征值进行排序的，排序的顺序不变，那么所属的分支以及分裂点就不会有不同。对于线性模型，比如说LR，我有两个特征，一个是(0,1)的，一个是(0,10000)的，这样运用梯度下降时候，损失等高线是一个椭圆的形状，这样我想迭代到最优点，就需要很多次迭代，但是如果进行了归一化，那么等高线就是圆形的，那么SGD就会往原点迭代，需要的迭代次数较少。

另外，注意树模型是不能进行梯度下降的，因为树模型是阶跃的，阶跃点是不可导的，并且求导没意义，所以树模型（回归树）寻找最优点事通过寻找最优分裂点完成的。

**2. 熵、联合熵、条件熵、相对熵、互信息的定义和公式。**

①熵：表示一个随机变量的复杂性或者不确定性。

符号所具备的信息为

所有类别所具有的信息熵为

②联合熵：两个随机变量X，Y的两个分布，可以形成联合熵。



③条件熵：表示在知道某一条件后，某一随机变量的复杂性或不确定性。



④相对熵：KL散度，是描述两个概率分布P和Q差异的一种方法。它是非对称的，这意味着。



⑤互信息：X和Y的联合分布和独立分布成绩的相对熵。



**3. 什么是最大熵?**

对一个随机事件的概率分布进行预测时，我们的预测应当满足全部已知条件，而对未知的情况不要做任何主观假设。在这种情况下，概率分布最均匀，预测的风险最小。因为这时概率分布的信息熵最大，所以人们把这种模型叫做“最大熵模型”。





**4. 树模型的特征选择中除了信息增益、信息增益比、基尼指数这三个外，还有哪些？**

①CHAID算法通过计算类别变量与特征变量之间的相关性检验统计量的p值，即卡方统计量对应的p值，p值越小，说明特征变量与类别变量之间的关系越密切，应当被选为最佳分组特征变量。然后继续按此准则选择后续特征变量，直至所有样本被分类完毕。CHAID算法在构建决策树时具有一定的优势，它是从统计显著性的角度来确定特征变量和分割数值，对决策树的分枝过程优化明显。

②QUEST算法在选择特征变量时，对于定类属性采用卡方检验方法，这点和CHAID类似，对于定矩属性则采用F检验方法，选择对应p值最小且小于显著性水平的特征变量作为最佳分类变量。

**5. 决策树有哪些处理连续值的方法？**

以C4.5算法为代表的一系列决策树算法采用取值区间二分离散化的方法来处理连续属性。具体处理方法是：

①首先找到训练样本在该连续属性上的最大值和最小值，在最大和最小值限定的取值区间上设置多个等分断点；

②分别计算以这些断点为分裂点的信息增益值，具有最大信息增益的断点即最佳分裂点，字该分裂点把整个取值区间划分为两部分。

缺点：过于笼统武断，强行将描述性属性取值的大小与分类属性的类别联系起来，大大降低分类精度。

**6. 决策树如何防止过拟合？**

解决决策树过拟合问题的主要方法是剪枝（提前停止树的增长或者对以经生成的树按照一定的规则进行后剪枝）。

①先剪枝：通过提前停止树的构建而对树“剪枝”，一旦停止，节点就成为树叶。

定义一个高度，当决策树达到该高度时就可以停止决策树的生长，这是一种最为简单的方法；

达到某个结点的实例具有相同的特征向量，即使这些实例不属于同一类，也可以停止决策树的生长；

定义一个阈值，当达到某个节点的实例个数小于该阈值就可以停止巨册书的生长；

定义一个阈值，通过计算每次扩张对系统性能的增益，并比较增益值与该阈值的大小来决定是否停止决策树的生长。

②后剪枝：首先构造完整的决策树，允许树过拟合训练数据，然后对置信度不够的结点字数用叶子节点来代替，该叶子的类标号用该结点字数中最频繁的类标记。

**7. 信息熵、信息增益的公式。**

信息熵：

信息增益：

**8. C4.5、ID3和CART的区别和优缺点**

|  |  |  |  |
| --- | --- | --- | --- |
| 算法 | 公式 | 特点 | 缺点 |
| ID3 |  | ①只能对分类变量进行处理；  ②对缺失值敏感；  ③只能做分类；  ④节点上可以产出多叉 | ①只能处理离散型属性；  ②并且对倾向于选择取值较多的属性 |
| C4.5 |  | ①只能做分类；  ②子节点可以多分；  ③可以处理连续和分类两种自变量 | c4.5处理过程中需对数据集进行多次顺序扫描和排序，处理成本耗时较高 |
| CART |  | ①CART是一棵二叉树，采用二元切分法，每次把数据切成两份，分别进入左子树和右子树；  ②可以处理连续和分类两种自变量；  ③对缺失值可以进行多种方式的处理；  ④小样本处理下泛化误差较大  ⑤一般用于回归树。 | CART本身时一种大样本的统计分析方法，样本量较小时模型不稳定 |

ID3倾向于选择取值较多的属性的原因：信息增益反映的给定一个条件以后不确定性减少的程度,必然是分得越细的数据集确定性更高,也就是条件熵越小,信息增益越大。

**9. ID3和C4.5分裂后，节点的信息熵是变大还是变小？**

变小。决策树建立的基本思想是以信息熵为度量构造一棵熵值下降最快的树，到叶子节点处的熵值为零, 需要遍历所有特征，选择信息增益最大的特征作为当前的分裂特征，一个特征的信息增益越大，表明属性对样本的熵减少的能力更强，这个属性使得数据由不确定性变成确定性的能力越强。

**10. 决策树有哪几种？**

①ID3、C4.5、CART；

②分类树和回归数。

**11. 决策树的优缺点。**

①优点：

决策树算法中学习简单的决策规则建立决策树模型的过程非常容易理解；

决策树模型可以可视化，非常直观；

应用范围广，可用于分类和回归，而且非常容易做多类别的分类；

能够处理数值型和连续的样本特征。

②缺点：

很容易在训练数据中生成复杂的树结构，造成过拟合（overfitting）。剪枝可以缓解过拟合的负作用，常用方法是限制树的高度、叶子节点中的最少样本数量；

学习一棵最优的决策树被认为是NP-Complete问题。实际中的决策树是基于启发式的贪心算法建立的，这种算法不能保证建立全局最优的决策树。Random Forest 引入随机能缓解这个问题。

**12. 决策树处理缺失值的方法有哪些？**

①删除确实项：

删除含有特征参数缺失的数据（但是如果很多数据特征参数缺失，我们会删除大量数据，一般超过三分之一的数据被删除的话不宜采用此法）；

删除某个特征（如果仅仅是缺失项集中在了某几个特征，我们可以把所有数据的这些特征参数都删除，但是删除特征过多的时候不宜采用此法）。

②猜测参数：平均值，多数法。

③自适应法：分类误差最小

利用所有特征的非缺失值将数据划分为n个部分；

将带有缺失值的样本分别分入这n各部分计算分类误差；

选择分类误差最小的特征分支作替代缺失值。