蒋豆芽

机器学习——决策树 已购

来自 【机器学习面试题汇总与解析（蒋豆芽面试题总结）】 | 59 浏览 | 0 回复 | 2021-05-28

蒋豆芽

+关注

机器学习面试题汇总与解析——决策树

1. **决策树介绍一下**☆☆☆☆☆
2. **决策树优缺点**☆☆☆☆☆
3. **决策树的划分标准是什么**☆☆☆☆☆
4. **ID3和C4.5的区别**☆☆☆☆☆
5. **树模型对离散特征怎么处理的**☆☆☆☆☆
6. **树模型怎么决定一个叶子结点是否要分裂**☆☆☆☆☆
7. **决策树出现过拟合的原因及解决办法**☆☆☆☆☆
8. **如何对决策树进行剪枝?** ☆☆☆☆☆
9. **决策树需要进行归一化处理吗**☆☆☆☆☆
10. **决策树与逻辑回归的区别**☆☆☆☆☆
11. **说下决策树的损失函数**☆☆☆☆☆

==========

- 本专栏适合于**Python已经入门**的学生或人士，有一定的编程基础。
- 本专栏适合于**算法工程师、机器学习、图像处理求职**的学生或人士。
- 本专栏针对面试题答案进行了**优化，尽量做到好记、言简意赅。这才是一份面试题总结的正确打开方式。这样才方便背诵**
- 如专栏内容有错漏，欢迎在评论区指出或私聊我更改，一起学习，共同进步。
- **相信大家都有着高尚的灵魂，请尊重我的知识产权，未经允许严禁各类机构和个人转载、传阅本专栏的内容。**

==========

关于**机器学习算法**书籍，我强烈推荐一本**《百面机器学习算法工程师带你面试》**，这个就很类似面经，还有讲解，写得比较好。私聊我进群。

关于**深度学习算法**书籍，我强烈推荐一本**《解析神经网络——深度学习实践手册》**，简称CNN book，通俗易懂。私聊我进群。

蒋豆芽

决策树的内容，写得最好

参考资料

决策树：https://blog.csdn.net/u012328159/article/details/70184415

读者可以先看看参考文章

个人理解

决策树其实不难，是一种分类方法，像一棵树一样进入不同的分支，然后直到**叶子节点**得到**分类结果**。那么如何进入不同的分支？这就是最重要的内容，涉及到**节点（特征）的划分标准**，有三种：**最大信息增益**、**最大信息增益率**、**基尼系数**。而这三种不同的划分标准就对应了三种典型决策树：**ID3**（最大信息增益）、**C4.5**（最大信息增益率）、**CART**（基尼系数）。

决策树的优缺点

优点：

1. 决策树易于理解和实现. 人们在通过解释后都有能力去理解决策树所表达的意义。
2. 对于决策树，数据的准备往往是简单或者是不必要的 . 其他的技术往往要求先把数据一般化，比如去掉多余的或者空白的属性。
3. 能够同时处理数据型和常规型属性。其他的技术往往要求数据属性的单一。
4. 在相对短的时间内能够对大型数据源做出可行且效果良好的结果。
5. 对缺失值不敏感
6. 可以处理不相关特征数据
7. 效率高，决策树只需要一次构建，反复使用，每一次预测的最大计算次数不超过决策树的深度。

缺点：

1. 对连续性的字段比较难预测。
2. 对有时间顺序的数据，需要很多预处理的工作。
3. 当类别太多时，错误可能就会增加的比较快。
4. 在处理特征关联性比较强的数据时表现得不是太好

1. **决策树介绍一下**☆☆☆☆☆

   参考回答

   决策树（decision tree）是一个树结构（可以是二叉树或非二叉树）。其每个非叶节点表示一个**特征属性**上的测试，每个分支代表这个特征属性在某个值域上的输出，而**每个叶节点存放一个类**

蒋豆芽

决策数有**两大优点**：

1. 决策树模型可 读性好，具有描述性，有助于人工分析；
2. 效率高，决策树只需要一次构建，反复使用，每一次预测的最大计算次数不超过决策树的深度。

决策树涉及到**节点（特征）的划分标准**，有三种：**最大信息增益**、**最大信息增益率**、**基尼系数**。而这三种不同的划分标准就对应了三种典型决策树：**ID3**（最大信息增益）、**C4.5**（最大信息增益率）、**CART**（基尼系数）。

答案解析

无。

类似的问题还有：

2. **决策树优缺点** ☆☆☆☆☆

参考回答

决策树的优缺点

优点：

1. 决策树易于理解和实现. 人们在通过解释后都有能力去理解决策树所表达的意义。
2. 对于决策树，数据的准备往往是简单或者是不必要的 . 其他的技术往往要求先把数据一般化，比如去掉多余的或者空白的属性。
3. 能够同时处理数据型和常规型属性。其他的技术往往要求数据属性的单一。
4. 在相对短的时间内能够对大型数据源做出可行且效果良好的结果。
5. 对缺失值不敏感
6. 可以处理不相关特征数据
7. 效率高，决策树只需要一次构建，反复使用，每一次预测的最大计算次数不超过决策树的深度。

缺点：

1. 对连续性的字段比较难预测。
2. 对有时间顺序的数据，需要很多预处理的工作。
3. 当类别太多时，错误可能就会增加的比较快。
4. 在处理特征关联性比较强的数据时表现得不是太好

蒋豆芽

3. **决策树的划分标准是什么**☆☆☆☆☆

   参考回答

   有三种：**最大信息增益**、**最大信息增益率**、**基尼系数。**而这三种不同的划分标准就对应了三种典型决策树：**ID3**（最大信息增益）、**C4.5**（最大信息增益率）、**CART**（基尼系数）。

   信息增益：指的是使用某一个属性a进行划分后，所带来的**纯度**（**信息熵**用来度量样本集合的**纯度**）提高的大小。一般而言，**信息增益越大**，意味着使用属性a来进行划分所获得的“**纯度提升**”越大。但信息增益对可取值较多的属性**有所偏好。**

   而信息增益率则解决了特征偏好的问题。

   但是不论是信息增益还是信息增益率，存在的问题是涉及对数运算，**计算量大**，为了解决这个问题。可以采用**基尼系数**作为节点划分的标准。

   答案解析

   无。

4. **ID3和C4.5的区别**☆☆☆☆☆

   参考回答

   最大的区别是划分标准的不同：**ID3**采用**信息增益**，而**C4.5**采用的是**信息增益率。**

   C4.5继承了**ID3**的优点，并在以下几个方面对**ID3**算法进行了改进：

   1. 用**信息增益率**来选择属性，克服了用信息增益选择属性是偏向选择去之多的属性的不足
   2. 在树的构造过程中进行剪枝
   3. 能够对连续的属性进行离散化处理
   4. 能够对不完整的数据进行处理

   答案解析

   无。

5. **树模型对离散特征怎么处理的**☆☆☆☆☆

   参考回答

   树模型是要寻找**最佳分裂点**，对于离散特征，树模型会评估每个离散值的**信息增益**，将信息增益最大的数值作为分裂点，因此，树模型不需要对离散特征进行事先one-hot处理，否则会使特征维度

答案解析

无。

6. **树模型怎么决定一个叶子结点是否要分裂** ☆ ☆ ☆ ☆ ☆

参考回答

答案参考上面。

答案解析

无。

7. **决策树出现过拟合的原因及解决办法** ☆ ☆ ☆ ☆ ☆

参考回答

原因

1. 在决策树构建的过程中，对决策树的生长没有进行合理的限制（剪枝）；
2. 样本中有一些噪声数据，没有对噪声数据进行有效的剔除；

解决办法

1. 选择合理的参数进行剪枝，可以分为预剪枝和后剪枝，我们一般采用后剪枝的方法；
2. 利用K-folds交叉验证，将训练集分为K份，然后进行K次交叉验证，每次使用K-1份作为训练样本数据集，另外一份作为测试集；
3. 减少特征，计算每一个特征和响应变量的相关性，常见得为皮尔逊相关系数，将相关性较小的变量剔除；

答案解析

无。

8. **如何对决策树进行剪枝?** ☆ ☆ ☆ ☆ ☆

参考回答

剪枝是防止决策树过拟合的方法。一棵完全生长的决策树很可能失去泛化能力，因此需要剪枝。

剪枝的策略

蒋豆芽

预剪枝

1. 设置一个树的最大高度/深度或者为树设置一个最大节点数，达到这个值即停止生长
2. 对每个叶子节点的样本数设置最小值，生长时叶子节点样本数不能小于这个值
3. 判断每次生长对系统性能是否有增益

后剪枝

1. 错误率降低剪枝（Reduced-Error Pruning）
2. 悲观剪枝（Pessimistic Error Pruning）
3. 代价复杂度剪枝（Cost-Complexity Pruning）

预剪枝和后剪枝的优缺点比较

1. 时间成本方面，预剪枝在训练过程中即进行剪枝，后剪枝要在决策树完全生长后自底向上逐一考察。显然，后剪枝训练时间更长。预剪枝更适合解决大规模问题。
2. 剪枝的效果上，预剪枝的常用方法本质上是基于贪心的思想，但贪心法却可能导致欠拟合，后剪枝的欠拟合风险很小，泛化性能更高。
3. 另外，预剪枝的有些方法使用了阈值，如何设置一个合理的阈值也是一项挑战。

答案解析

后剪枝错误率降低剪枝的方法比较直观，从下至上遍历所有非叶子节点的子树，每次把子树剪枝（所有数据归到该节点，将数据中最多的类设为结果），与之前的树在验证集上的准确率进行比较，如果有提高，则剪枝，否则不剪，直到所有非叶子节点被遍历完。

9. **决策树需要进行归一化处理吗** ☆ ☆ ☆ ☆ ☆

参考回答

概率模型不需要归一化，因为他们不关心变量的值，而是关心变量的分布和变量之间的**条件概率**。**决策树**是一种概率模型，数值缩放，不影响分裂点位置。所以一般不对其进行**归一化**处理。

答案解析

无。

10. **决策树与逻辑回归的区别** ☆ ☆ ☆ ☆ ☆

参考回答

2. 逻辑回归对数据整体结构的分析优于决策树，而决策树对局部结构的分析优于逻辑回归；
3. 逻辑回归擅长分析线性关系，而决策树对线性关系的把握较差。线性关系在实践中有很多优点：简洁，易理解，可以在一定程度上防止对数据的过度拟合。
4. 逻辑回归对极值比较敏感，容易受极端值的影响，而决策树在这方面表现较好。
5. 执行速度上：当数据量很大的时候，逻辑回归的执行速度非常慢，而决策树的运行速度明显快于逻辑回归。

答案解析

无。

11. **说下决策树的损失函数**☆☆☆☆☆

参考回答

决策树的剪枝往往通过极小化决策树整体的损失函数（loss function）或代价函数（cost function）来实现．设树T的叶结点个数为$|T|$，t是树T的叶结点，该

叶结点有N_t个样本点，其中k类的样本点有N_{tk}个，$k=1,2,\cdots,K$，$H_t(T)$为叶结点t上的经验熵，$\alpha \geqslant 0$为参数，则决策树学习的损失函数可以定义为

$$C_\alpha(T)=\sum_{t=1}^{|T|}N_tH_t(T)+\alpha|T| \tag{5.11}$$

其中经验熵为

$$H_t(T)=-\sum_k\frac{N_{tk}}{N_t}\log\frac{N_{tk}}{N_t} \tag{5.12}$$

答案解析

公式中H(X)可以理解为这个叶子节点的熵。如果把决策树一直划分下去，叶子节点的熵应该为0，只有一个类。但是如果使用一些剪枝规则，每个节点中仍然可以有熵值，也就是可以继续划分。

Nt是这个节点中的样本的个数，可以看做这个节点的权重。节点中样本数越多，权重越大。

蒋豆芽

后面一项是对整棵决策树的复杂度的惩罚项，结点数越多，越复杂。相当于一个正则项，也可以理解为先验概率：较小的树有较大的先验概率。

资源分享 python 机器学习 算法工程师 春秋招 面试题 软件开发 面经 举报

收藏 赞

相关专栏

机器学习面试题汇总与解析（蒋豆芽面试题总结）

27篇文章 | 90订阅 已订阅

0条评论 默认排序

没有回复

请留下你的观点吧~

发布