監督ML

分类

介绍

数据分析师经常需要从一组变量中预测一个分类结果，例如，

根据个人的人口统计数据和财务历史（良好的信用风险/不良的信用风险），预测个人是否会偿还贷款

根据急诊室患者的症状和生命体征（心脏病发作/无心脏病发作）来确定急诊室患者是否有心脏病发作

根据关键词、图像、超文本、标题信息和来源（垃圾邮件/非pam），决定电子邮件是否为垃圾邮件

目标是找到一种准确的方法，将新的方法从一组预测器（也称为特征）分类为两组之一。

监督学习

监督学习从一组包含预测变量和结果的值的观察结果开始。

然后将数据集分为一个训练样本和测试样本。

利用训练样本中的数据建立了一个预测模型，并利用测试样本中的数据进行了准确性测试。

这两个样本都是需要的，因为c分类技术最大限度地提高了对给定数据集的预测能力。

如果他们使用生成模型的相同数据进行重新评估，对它们有效性的估计将会过于乐观。

通过将训练样本开发的分类规则应用于单独的测试样本，您可以获得更真实的精度估计。

一个有效的预测模型可以在只有预测变量已知的情况下预测结果。

数据集信息

例如：威斯康辛州乳腺癌

来自UCI机器学习存储库

其目标将是开发一个模型，基于特征的沙发细针组织抽吸(一种组织
用一根很细的空心针从皮肤下的一个肿块中提取样本)

该数据集包含699个细针抽吸样本

458例（65.5%）为良性

241例（34.5%）为恶性肿瘤

该数据集包含11个变量，并且不包括变量名。

16个样本缺少数据，并用一个问号（？）.

数据集信息

例如：威斯康辛州乳腺癌协会

来自UCI机器学习存储库

数据文件： breast_cancer.csv

如何在Python中导入？

沙痂变量

1.身份证	*2.块状岩层*
3.细胞大小的均匀性	*4.细胞形状的均匀性*
5.边缘粘附	*6.单个e上皮细胞的大小*
7.裸核	*8.粗染色质*
9.正常圆锥体	*10.有丝分裂*
11.课	

Install the ucimlrepo package

```
pip install ucimlrepo
```

Import the dataset into your code

```
from ucimlrepo import fetch_ucirepo

# fetch dataset
breast_cancer_wisconsin_original = fetch_ucirepo(id=15)

# data (as pandas dataframes)
X = breast_cancer_wisconsin_original.data.features
y = breast_cancer_wisconsin_original.data.targets

# metadata
print(breast_cancer_wisconsin_original.metadata)

# variable information
print(breast_cancer_wisconsin_original.variables)
```

数据集信息

例如：威斯康辛州乳腺癌

第一个变量是ID变量（我们将删除）

最后一个变量（类）包含结果(编码2个=良性，4个=恶性）

我们还将排除包含缺失值的观测值。

对于每个样本，还记录了9个先前发现的与恶性肿瘤相关的细胞学特征。

这些变量中的每一个都从1（最接近良性）到10（大多数间变性）。

没有一个预测因子可以单独区分良性和恶性样本。

这些分类规则可以用于最近从这9种细胞特征的某些组合来快速预测恶性肿瘤。

决策树

决策树在数据挖掘上下文中很流行。

DT学习一系列的问题，将案例分成不同的类。

每个问题都有一个二进制的答案，案例将根据它们满足的标准被发送到左或右分支。

分支内可以有分支；一旦我学习了模型，它就可以被图形化地表示为树。

你是否玩过这个游戏20个问题，你必须猜出是什么对象，有人会问是或否的问题？

另一个例子是《猜猜谁》游戏，你必须通过询问其他玩家的外表来猜测他们的分歧？

决策树

沙桂园示例：

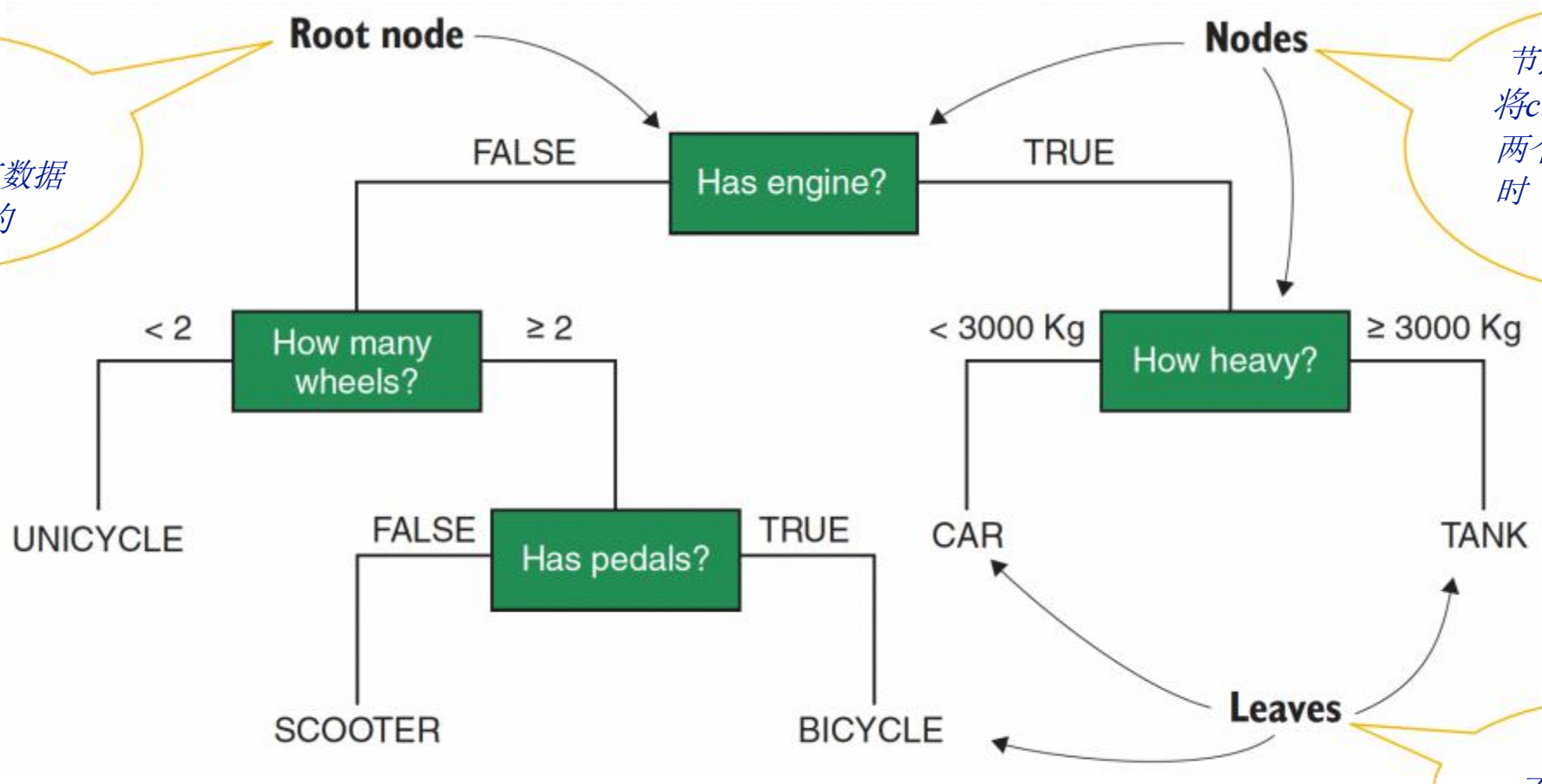
想象一下，您希望创建一个模型，以表示人们通勤到上班的方式。

你可以收集关于车辆的信息，比如它们有多少个轮子，它们是否有引擎，以及它们的重量。

我们可以把这个分类过程作为一系列的顺序问题来表述。

每个车辆在每个问题上进行评估，并根据其特征如何满足问题，在模型中向左或向右移动。

决策树



根no de是
包含的节点
之前的所有数据
爆裂似的

节点被a分割
将c迭代分解为
两个小时，每一个小时
这导致了
另一个没有de。

不进一步分裂的
节点被称为叶

决策树

基于树的模型可以同时用于分类和回归

沙痂分类树

沙桂堡回归

为了决定每个拆分时的最佳特性，使用了两种方法：

- 沙痂信息收益
- 沙痂基尼增益

基尼增益计算起来稍快，所以我们将关注它。

基尼系数增加/基尼系数指数

吉尼根用于在生长决策树时寻找特定节点的最佳分割

基尼指数用于测量杂质。

杂质是衡量阳极内各类的异质性的方法。

如果阳极只包含一个类（ora叶），它可以说是纯的。

通过估计使用每个预测变量所产生的杂质，算法可以选择将导致

最小的杂质。

换句话说，该算法选择的特征将导致后续节点的区域尽可能是均匀的。

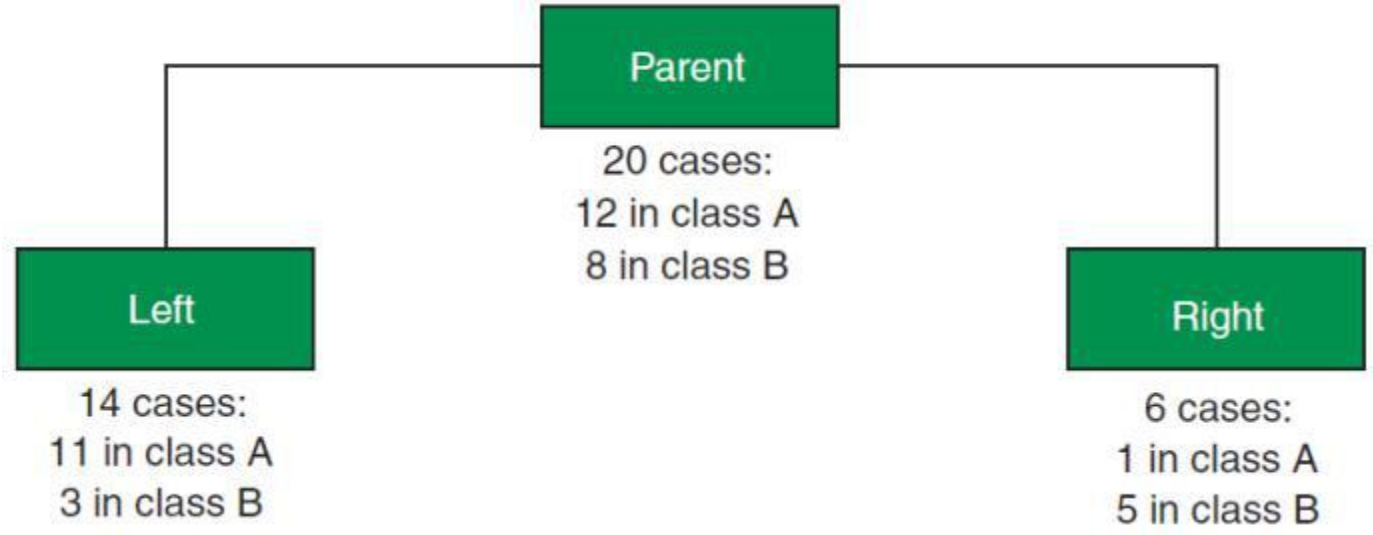
基尼系数增加/基尼系数指数

沙桂园示例：

我们有20个父节点属于两个类，A类和b类。我们根据一些Ω分类将该节点分成两个叶子。

在左叶中，我们有11个来自A类和3个来自B类。

在右边的叶子中，我们有5个来自B类，1个来自A类。



基尼系数增加/基尼系数指数

例如：我们想知道从分裂中得到的基尼系数的收益

基尼增益表示父节点的基尼指数和分裂的基尼指数之间的差异。

查看我们的示例，任何节点的Gin i索引都被计算为：

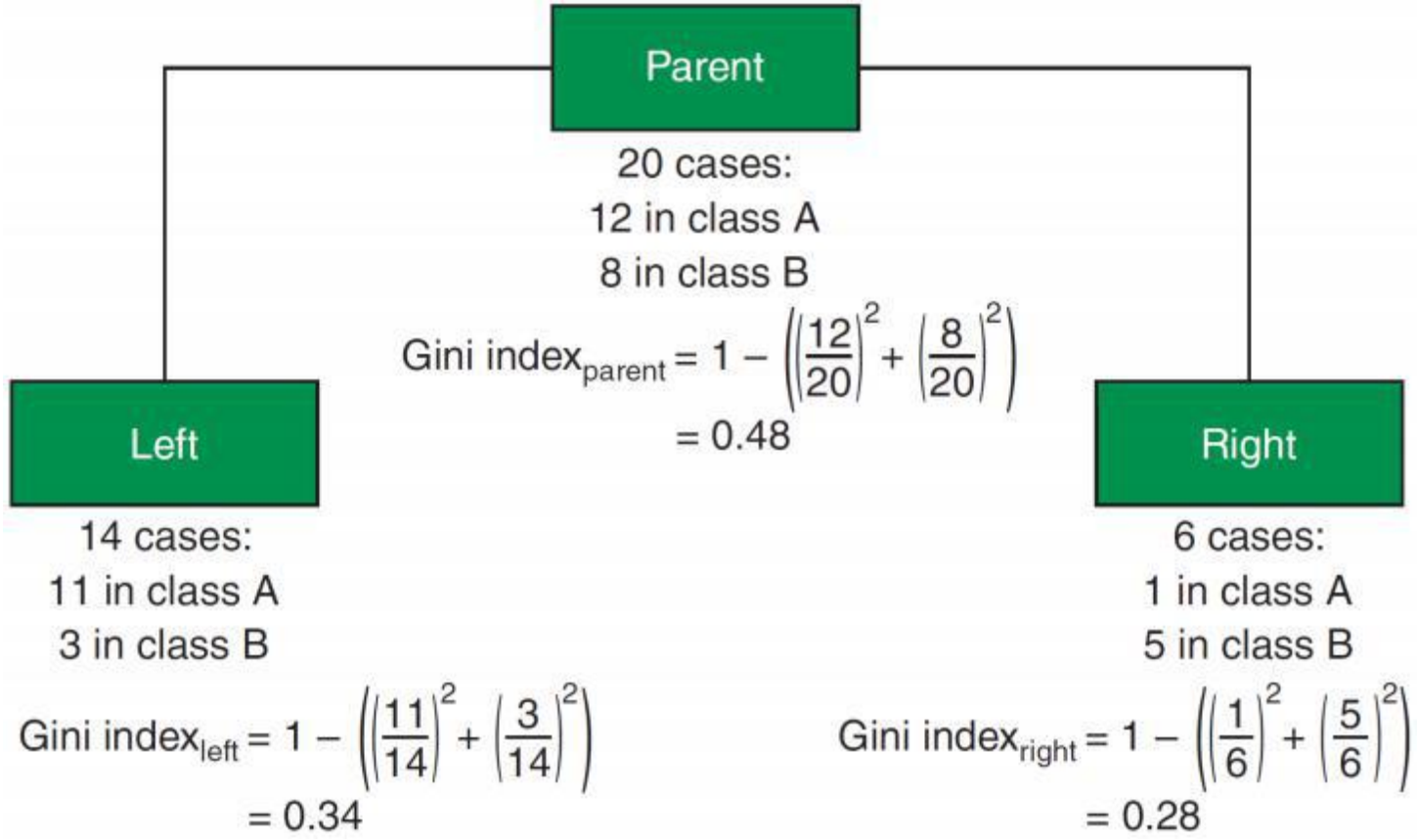
$$基尼指数=1-(pA2+pB2)$$

其中，p (A)和p (B)分别为属于A类和B类的病例的比例。

基尼系数增加/基尼系数指数

例如：我们想知道从分裂中得到的基尼系数的收益

基尼指数=1−（pA2(+)pB2( )



基尼系数增加/基尼系数指数

沙桂园示例：

一旦我们有了左右两叶的基尼指数，我们就可以计算出分裂一个sa整体的基尼指数。

分裂的基尼指数是根据父母接受的病例比例计算的左右基尼指数的总和。

基尼指数分裂=p左×基尼指数左+p right×基尼指数right

( ) 14 6 ( )

$Gini\ index_{split} = \frac{}{} \times 0.34 + \frac{}{20} \times 0.28 = 0.32$

基尼系数增益（父节点和分割节点的基尼系数指数之间的差异）

基尼系数增益=值为0。$48 - 0.32 = 0.16$

其中0.48是父系的基尼指数

基尼系数增加/基尼系数指数

对每个预测变量计算特定节点的基尼系数增益。

产生最大基尼系数增益的预测器被用于分割该节点。

这个过程对每个节点重复。

将Gi ni索引推广到任意数量的类：

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

这就是说我们从1开始计算每个类的p类k2到K（类的数量），将它们全部相加，然后从1中减去这个值。

优势与劣势

沙痂的优势

造树背后的直觉非常简单，而且易于解释。

它可以处理分类的和连续的预测变量。

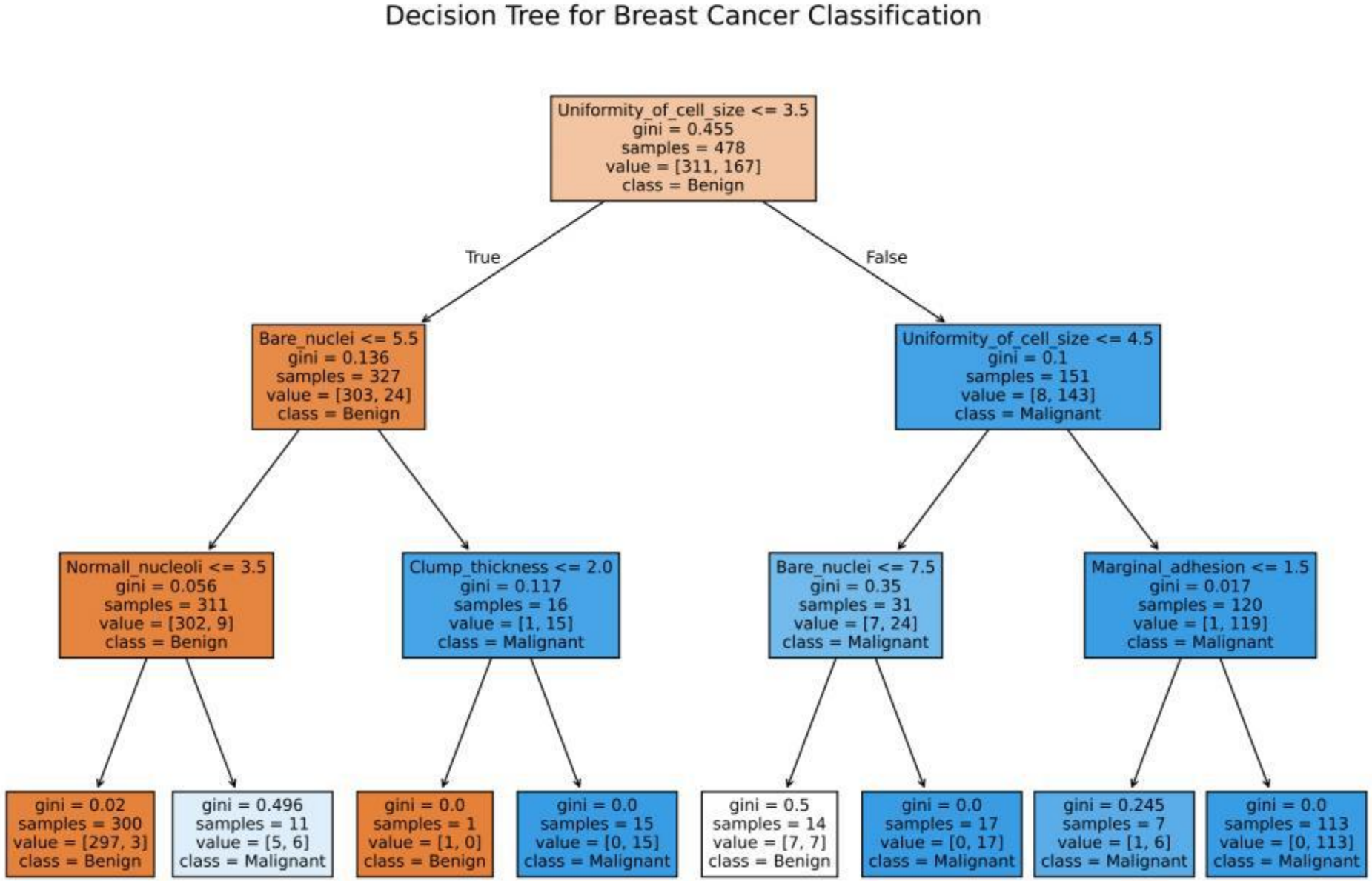
它对预测或变量的分布没有任何假设。

它可以处理缺失值的年代不合理的方法。

它可以处理不同尺度上的连续v个变量。

沙痂的弱点

个别的树非常容易被过度拟合，以至于它们很少被使用。


Decision Tree for Breast Cancer Classification
Uniformity_of_cell_size <= 3.5
gini = 0.455
samples = 478
value = [311, 167]
class = Benign
True
False
Bare_nuclei <= 5.5
gini = 0.136
samples = 327
value = [303, 24]
class = Benign
Uniformity_of_cell_size <= 4.5
gini = 0.1
samples = 151
value = [8, 143]
class = Malignant
Normall_nucleoli <= 3.5
gini = 0.056
samples = 311
value = [302, 9]
class = Benign
Clump_thickness <= 2.0
gini = 0.117
samples = 16
value = [1, 15]
class = Malignant
Bare_nuclei <= 7.5
gini = 0.35
samples = 31
value = [7, 24]
class = Malignant
Marginal_adhesion <= 1.5
gini = 0.017
samples = 120
value = [1, 119]
class = Malignant
gini = 0.02
samples = 300
value = [297, 3]
class = Benign
gini = 0.496
samples = 11
value = [5, 6]
class = Malignant
gini = 0.0
samples = 1
value = [1, 0]
class = Benign
gini = 0.0
samples = 15
value = [0, 15]
class = Malignant
gini = 0.5
samples = 14
value = [7, 7]
class = Benign
gini = 0.0
samples = 17
value = [0, 17]
class = Malignant
gini = 0.245
samples = 7
value = [1, 6]
class = Malignant
gini = 0.0
samples = 113
value = [0, 113]
class = Malignant

问题？