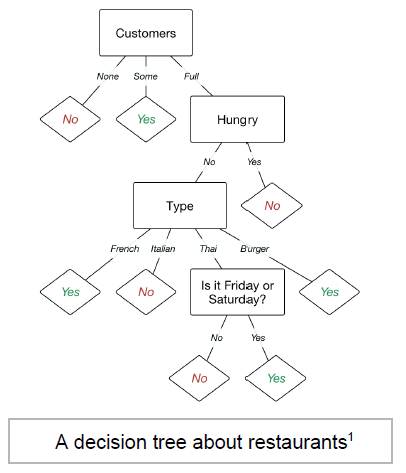
# Decision Tree

在这次的夏令营线下讨论中，我们对课程中的监督学习算法进行了小结。我负责写第一部分，相对简单。把它记录下来，希望能够为大家回归课程内容做一个简单回顾，同时也方便大家看下自己是否还遗漏了什么。

## 1. 什么是决策树

决策树是一个极其直观的算法（这也使得它的可解释性非常好），比如下面这个决策树  


## 2. 基本决策树算法

从上面可以看出来这个形式非常的简单，那么如何生成一颗决策树？相信开始了nano degree并完成了P0的同学们都已经自己实现过一个决策树了，也就是大家做的Titanic的project，在第三个问题中，大家自己通过不断的选择特征，产看结果，最终得到了一个准确率超过80%的模型，这就是一个决策树。  
可是这种方法明显是没有可扩展性可言的，大家可以体验到对于这么小的一个数据集大家都构建的很辛苦，并且准确率也不是太高。所以我们需要一个自动构建的算法。  
其实这个算法也十分简单，对于我们的数据集来说我们关键是按照下面两步来完成决策树的构建：

* 选择一个最佳属性进行划分
* 对每一个子节点迭代进行划分，知道完全划分开

### 划分准则的选择

从上面关键的两步可以看出划分准则的选择是一个很重要的问题，因为理论上来说找到一个最佳的划分是一个NP-hard的问题，所以说我们不能够通过某一种算法得到最优划分，那么现在的做法就是通过一些启发式的指标来选择合适的划分准则，常用的有（下面我就不一一列公式了，大家可以从其他地方查一下）：

* 信息增益（ID3使用这种）  
  计算划分前后信息熵的差值，增益大的划分好
* 信息增益率（C4.5使用这种）  
  信息增益存在的一个问题是，它对于可取值数目多的属性有偏好（极端情况，我把每一个值都作为一个划分将获得一个很大的信息增益）。所以定义了增益率（不过增益率会对取指少的有偏好，所以C4.5实际上对选择还有一个权衡，结合了增益和增益率考虑）。
* 基尼系数（CART使用这种）  
  暂时不了解它有什么明显缺陷，应该是现在相对最好的一个准则吧。

## 3. 常见决策树算法

常见的三个算法有：

* ID3
* CART
* C4.5

这三个算法都符合上面的简单形式，他们的不同点主要体现在1.如何选择最佳属性？2.对于各种不同属性的处理3.剪枝。大家可以学习下，我就不一一写出来了。

## 4. 决策树的剪枝

### 为什么需要剪枝？

上面没有太谈剪枝问题，那么首先看下为什么要剪枝？因为如果我们不对决策树剪枝的话，决策树会不断生长，这样就会出现我们熟悉的过拟合问题了。所以对决策树剪枝是十分必要的。

### 剪枝方式

* 前剪枝  
  所谓前剪枝就是我们可以在树还没有构建的情况下，选择一些先验的准则对树的生长做限制，比如：限制树最多生长多少层，限制树有多少节点等。这个部分大家如果用sklearn的话，实际上它后面的很多参数就是做了一个前剪枝的工作。
* 后剪枝  
  所谓后剪枝就是我们在树已经构建完成之后，我们对树进行修剪。这时候修剪当然就要考虑到生成的数和原始的数据集。所以一般的做法就是我们专门为决策树构建出一个剪枝测试集。然后对于树上的节点（可以一次考虑一个或者多个）我们都考虑保留这些节点和不保留这些节点，如果过测试出来不保留更佳的话我们就可以把这个剪掉。一般来说后剪枝的效果会更好，但是代价也会大很多所以是否采用也需要权衡。

## 5. 决策树的优缺点

我们也可以考虑一下决策树的优缺点，这里也能够帮助理解他的思想。这里简单列出几点：

### 优点

* 可解释性好
* 只需要很少的数据预处理工作。并且能够处理各种类型的特征
* 原生支持多分类

### 缺点

* 决策树容易过拟合
* 决策树不稳定，也就是说可能数据集的一个很小的变动就会导致树发生很大的改变。
* 构建一个最佳的决策树是一个NP-hard问题，现有的启发式算法不能够保证最优（也没有一个理论的上下界）。
* 简单的决策树模型比较简单，很难表示数据集中的复杂关系。

## 6. 建议学习方向

* 看一下具体的ID3（这个在Udacity提供的附加资料中就有）、C4.5、CART算法（其实这三个算法并不复杂，而且只是某个算法在某一方面做了一些优化）
* 自己实现一个决策树？（比如可以看[《机器学习实战》](https://discussions.youdaxue.com/t/topic/2194/4)，很简单）
* 看一下[sklearn介绍6](http://scikit-learn.org/stable/modules/tree.html)

## 7. 总结

决策树的关键在于如何自动构建和防止过拟合。

如果大家觉得还有一些知识是课程中提到或者是没提到但是十分重要的部分也欢迎补充。