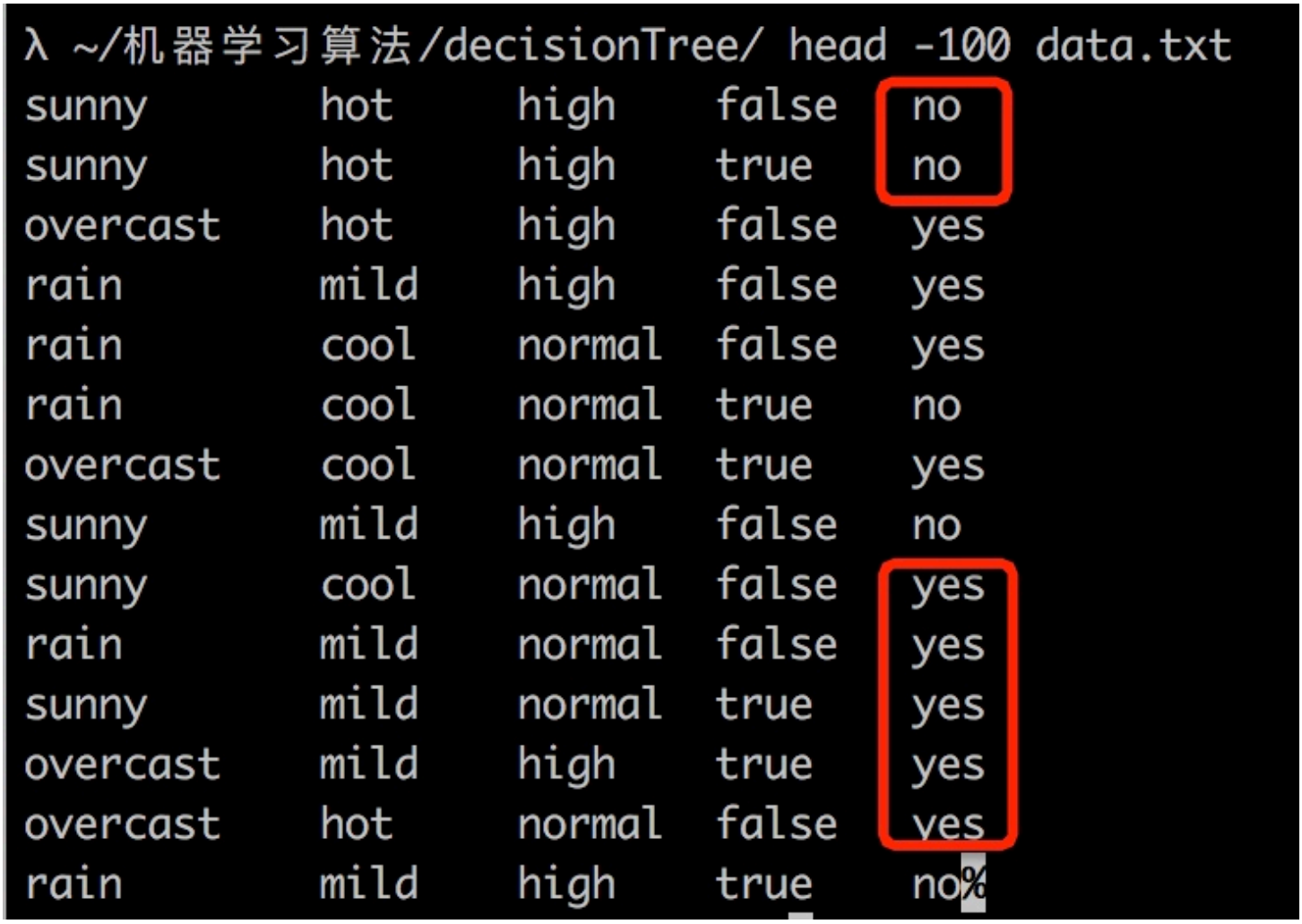
决策树

1、ID3【信息增益】
2、C4.5【信息增益率】
3、CART【基尼指数】



其中第一列表示天气情况：晴天、阴天和雨天；第二列比表示温度：炎热和凉爽；第三列表示湿度：高和正常；第四列表示风速：弱和强；最后一列表示是否出行：no占5/14，yes占9/14
属性集合{天气，温度，湿度，风速}，类别标签{出行，取消}
1、计算类别信息熵
类别信息熵表示的是在所有样本中各类别出现的不确定性之和。熵越大，不确定性越大，那么其所附带的信息量越大。

$$I(D) = -9/14 * log(9/14) - 5/14 * log(5/14) = 0.940$$

2、计算每个属性的信息熵
每个属性的信息熵即条件熵：即表示在某种属性条件下，各类被出现的不确定性之和。属性的信

息熵越大，就表示该属性的信息量越大，其包含的样本类别就越不纯。
对于天气，存在3种状态：{sunny: 5, overcast: 4, rain: 5}
sunny中对应3个no，2个yes
overcast对应4个yes
rain对应3个yes，2个no

$$I(Weather) = 5/14 * [-2/5log(2/5) - 3/5log(3/5)] + 4/14 * [-4/4log(4/4)] + 5/14 * [-3/5log(3/5) - 2/5log(2/5)]] = 0.694$$

$$I(Temperature) = 4/14 * [-2/4log(2/4) - 2/4log(2/4)] + 6/14 * [-4/6log(4/6) - 2/6log(2/6)] + 4/14 * [-3/4log(3/4) - 1/4log(1/4)]] = 0.911$$

$$I(Humidity) = 4/14 * [-2/4log(2/4) - 2/4log(2/4)] + 6/14 * [-4/6log(4/6) - 2/6log(2/6)] + 4/14 * [-3/4log(3/4) - 1/4log(1/4)]] = 0.789$$

$$I(WindSpeed) = 6/14 * [-3/6log(3/6) - 3/6log(3/6)] + 8/14 * [-6/8log(6/8) - 2/8log(2/8)] = 0.892$$

3、计算信息增益
信息增益 = 熵 - 条件熵
此处即为：类别熵减去属性熵；信息增益即表示不确定性减少的程度。如果一个属性的信息增益越大，就表示采用这个属性进行划分就会减少更多样本的不确定性，也就能更快的完成分类的目标。

$$Gain(Weather) = I(D) - I(Weather) = 0.940 - 0.694 = 0.246$$

$$Gain(Temperature) = I(D) - I(Temperature) = 0.940 - 0.911 = 0.029$$

$$Gain(Humidity) = I(D) - I(Humidity) = 0.940 - 0.789 = 0.15$$

$$Gain(WindSpeed) = I(D) - I(WindSpeed) = 0.940 - 0.892 = 0.048$$

到这里如果直接采用信息增益作为子节点的划分依据，该算法就称之为

4、计算属性分裂信息度量

本质上是对信息增益的补偿机制。简单来说就是，假设一种情况：每个属性中每种类别都只包含一个样本，那么该属性的信息熵就为零，根据信息增益就无法选择出有效的分类特征。而C4.5算法采用信息增益率来进行改进：具体就是加入了对于属性分裂时分支的数量和尺寸信息考量，这些信息是属性的内在信息，而信息增益率 = 信息增益 / 分裂信息度量，如果内在信息越多，也即属性本身的不确定性就越大，那么就会越不倾向于选择它，那么这样就可以对信息增益做一定的补偿。

$$H(Weather) = -5/14 * log(5/14) - 5/14 * log(5/14) - 4/14 * log(4/14) = 1.577$$

$$H(Temperature) = -4/14 * log(4/14) - 6/14 * log(6/14) - 4/14 * log(4/14) = 1.556$$

$$H(Humidity) = -7/14 * log(7/14) - 7/14 * log(7/14) = 1.0$$

$$H(WindSpeed) = -6/14 * log(6/14) - 8/14 * log(8/14) = 0.985$$

5、计算信息增益率

$$IGR(Weather) = Gain(Weather)/H(Weather) = 0.246/1.577 = 0.155$$

$$IGR(Temperature) = Gain(Temperature)/H(Temperature) = 0.029/1.556 = 0.0186$$

$$IGR(Humidity) = Gain(Humidity)/H(Humidity) = 0.151/1.0 = 0.151$$

$$IGR(WindSpeed) = Gain(WindSpeed)/H(WindSpeed) = 0.048/0.985 = 0.048$$

这样根据信息增益率就找到了最优的子节点Weather，然后分别根据Weather对应的几个子节点分别执行上述1~5的过程。最终就构建出了决策树。