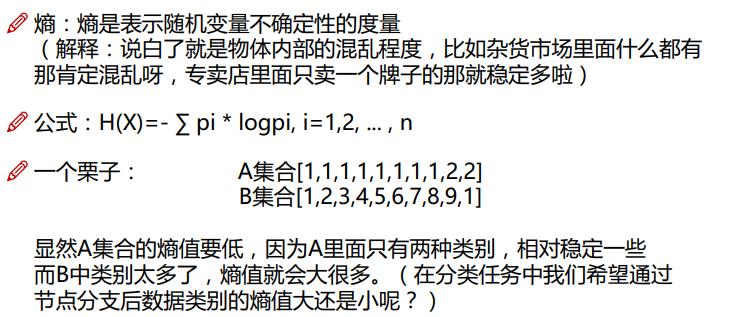
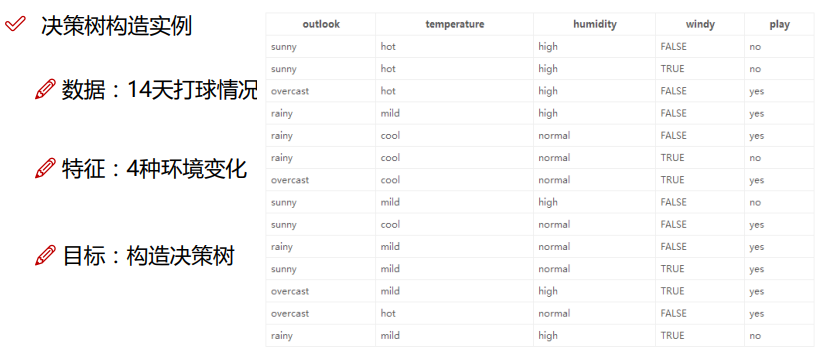


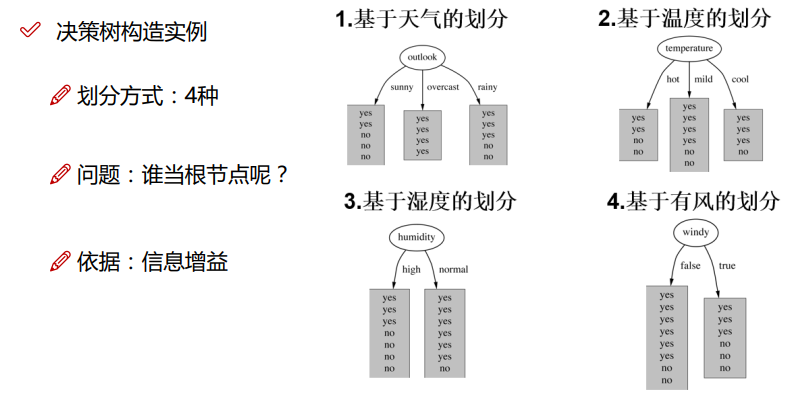
决策树的分裂过程：思考决策树生成过程，从父节点到若干子节点，我们按照正常的逻辑去解决这个问题，无非就是找到从若干个特征中找出一个特征，再从这个特征的分布中找出一个或多个切割点，不同的子树代表该特征不同的取值范围，不同的子树分得的样本有好有坏，假设一个二分类问题，一个好的决策树其叶子节点中的好坏样本比越高越好，因为样本的决策树在最终的预测中，效果会越好，那么怎么能让每次分裂的子节点的好坏样本比更高，这是我们追求的目标。决策树引入了一个衡量标准-熵，熵可以评价子树中样本的不确定性程度，对于A集合明显的我们确定1事件发生的概率很大，对于B我们不太确定哪个事件可能会发生，显然我们追求高确定性，那么熵越小越好

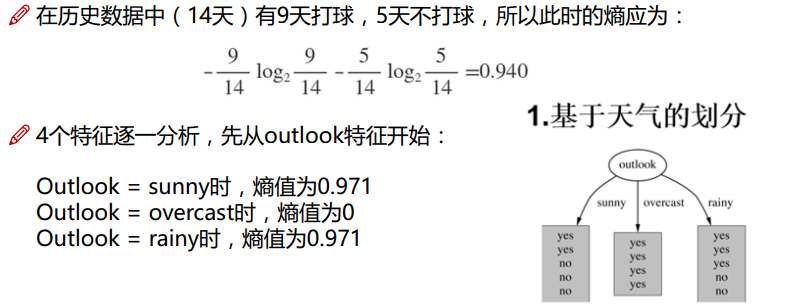


显然我们可以通过比较父节点和子树中样本的熵值来评估这次划分的好还，即信息增益，信息增益当然越大越好。p代表该事件发生的概率



这里特征全部是离散的且类别较少，对于连续特征需要进行离散化



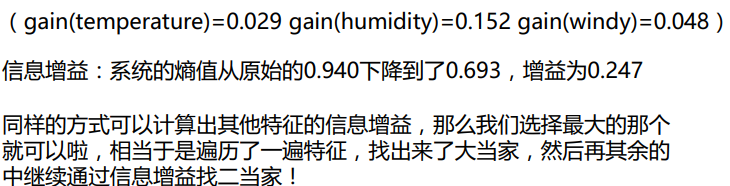


这里是根节点划分，故父节点的熵应该是所有样本的熵值

Eall = 0.94

子节点的样本的熵值分别是0.972、0、0.971，取加权平均即可求出子节点整体的信息熵





在根节点的分裂中发现，outlook划分之后的信息增益最大，故按照outlook进行划分。

评估标准：





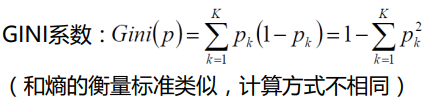
缺点：思考有这么一列特征，是连续特征，从1到n，如果不按照范围划分的话，每一个值都可以划分为一个子树，而且该子树中只有对应的一个样本，该层可以划分为n个子树，而且这次划分的信息熵是0，因为任何一个子树的样都只有一个，这样的划分其信息增益无疑相比于其他特征的划分都是最好的，那么算法就会这么划分，但是显然是不合理的，试想你的特征数据中混入了样本的ID数据，其结果就是这样的，还有一些连续特征比如年龄，如果不进行离散化，其结果也是这样的，这是信息增益所带来的问题



Hx = 信息增益 / 信息熵

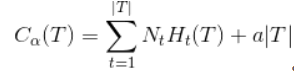
针对上述问题，虽然其信息增益很大，但是该列本身的信息熵非常大，用信息增益率来代替信息增益，这样可以解决这种问题





基尼系数用来表征序列的纯度，效果与熵类似

损失函数：



其中T代表叶子节点个数，Nt代表该叶子节点样本数，Ht代表该叶子节点的信息熵，最终再加一个正则项。

直观理解的话，使得损失函数最小，那么每一个叶子节点的信息熵应该尽可能的小，这样想是合理的，但是无法解释决策树损失函数的来源。

<https://blog.csdn.net/wzk1996/article/details/81941572>

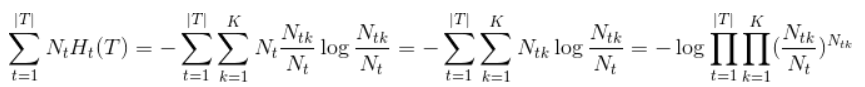
我们知道正则化的损失函数中前一项代表经验误差，而在概率模型中(决策树模型是一种概率模型)，经验误差函数的获得往往通过将极大似然函数取反，即将求极大化为求极小而获得。因此，在概率模型中，极大似然函数与经验误差函数可以认为是相同的概念，那么必然就可以通过经验误差函数来推导出极大似然函数，以此来加深对决策树损失函数的理解。

表面上来看：将每个叶节点的实例个数与其信息熵的乘积相加，这究竟代表个什么玩意呢？现在，我将利用该损失函数反向推导出极大似然函数，将信息熵Ht分解

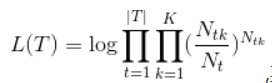
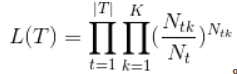


Pi = Ntk / Nt 该类别占比





至此，损失函数被化简为这样，可以看到它与我们所要求的极大似然函数仅一步之遥：将求极小转为求极大，即去掉上式的负号，我们得到对数极大似然函数

 ---🡪 

从极大似然的角度思考,