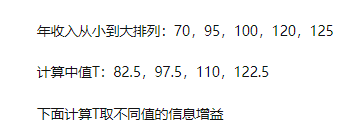
1.处理连续的属性值：



2. ID3算法由于采用的是信息增益，容易倾向于选择取值较多的属性作为节点。改良后的C4.5算法采用的是信息增益率，信息增益率=信息增益/属性熵

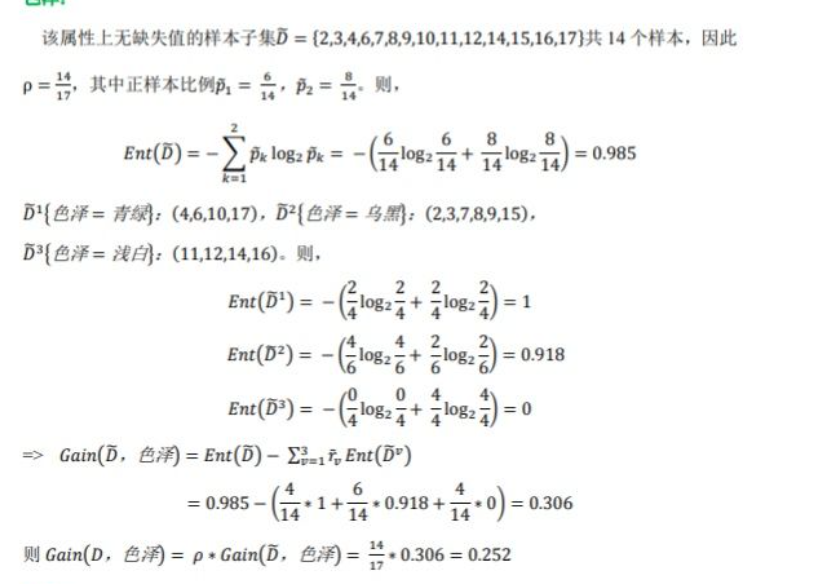
3.启发式的算法：先从候选属性中找出高于平均水平的属性，再从高于平均水平的属性中选择增益率最高的属性。

正常计算信息增益，然后求平均值。在信息增益的值在平均值之上的属性里取增益率高的属性。

一个极端的例子，如果属性取值数目较少，只取一个值。属性熵就是0。所以增益率可能会偏好取值比较少的属性。

4. 缺失值的处理

1.将无缺失值的样本取出构成子集，计算信息增益。乘P(无缺失值的样本数 / 总样本数)



2.权重处理

可以将缺失特征的样本同时划分入所有的子节点，不过将该样本的权重按各个子节点样本的数量比例来分配。比如缺失特征A的样本a之前权重为1，特征A有3个特征值A1,A2,A3。 3个特征值对应的无缺失A特征的样本个数为2,3,4.则a同时划分入A1，A2，A3。对应权重调节为2/9,3/9, 4/9。

剪枝：

