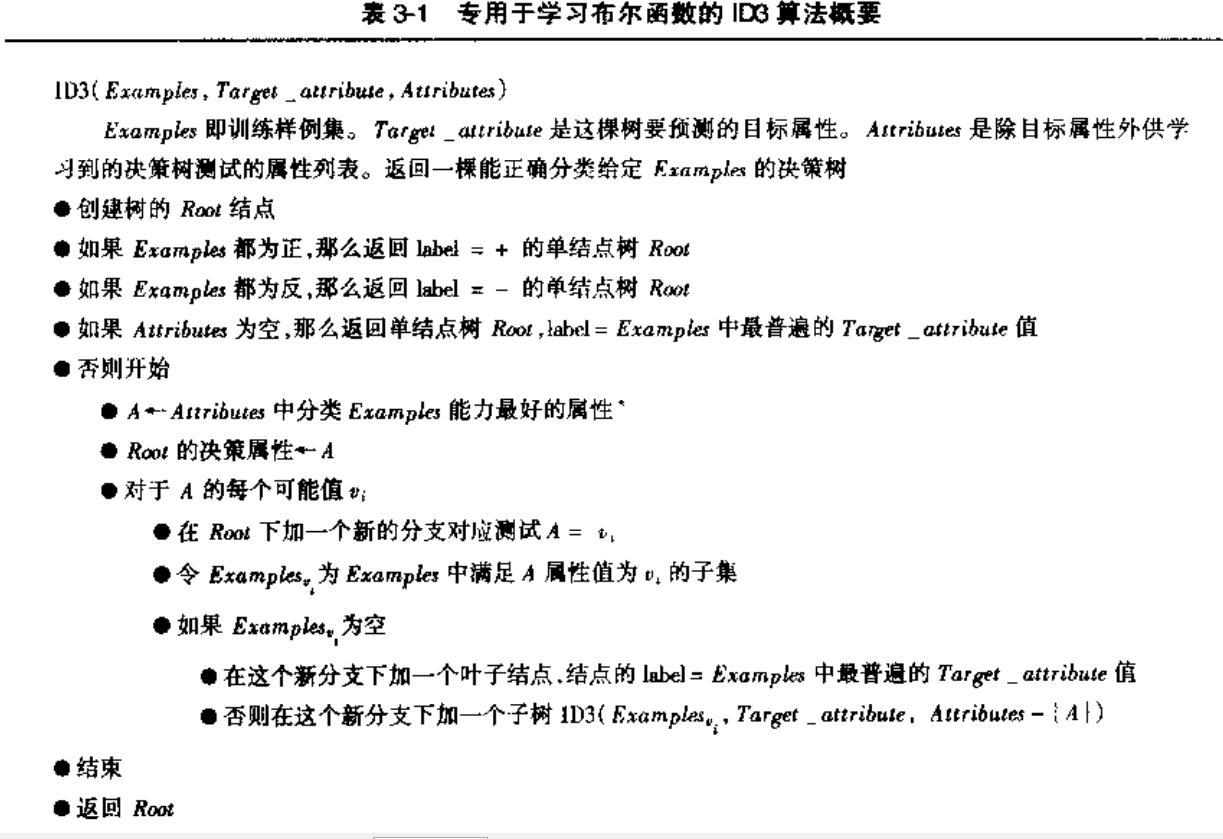
本周主要工作：机器学习之决策树其二

构建决策树的算法大多是采取自顶向下的贪婪搜索遍历所有可能的决策空间。

ID3算法

基本的ID3算法采取自顶向下构造决策树进行学习。构造过程是从“哪一个属性将在树的根结点被测试？”这个问题开始。为解决这个问题，算法使用统计测试来确定每一个实例属性单独分类训练实例的能力。分类能力最好的属性被选做树的根结点测试。然后根结点属性的每个取值可能产生一个分支，并把训练实例排列到适当的分支之下。然后重复这个过程，用每个分支结点关联的训练样例来选取在该结点被测试的最佳属性。这样就形成了对合格决策树的贪婪搜索（greed search）。算法从不回溯重新考虑之前的选择。

以下是专门用于学习布尔值函数的ID3算法

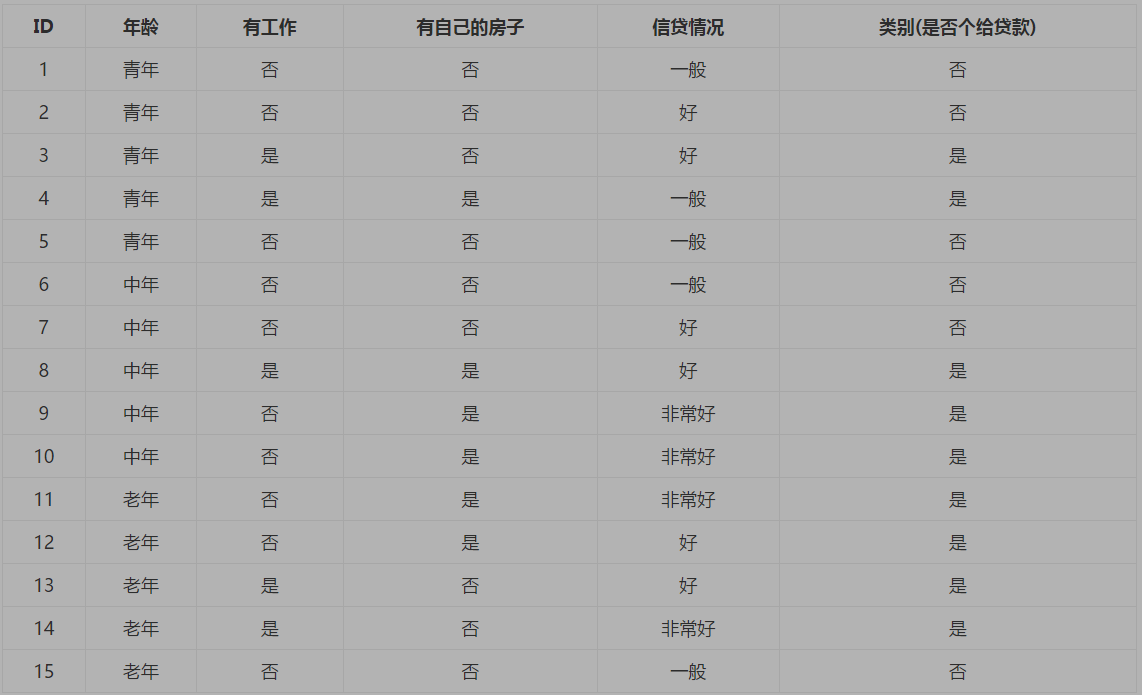


ID3算法选择根结点的原则：

 ID3算法的核心是在决策树各个结点上对应信息增益准则选择特征，递归地构建决策树。具体方法是：从根结点(root node)开始，对结点计算所有可能的特征的信息增益，选择信息增益最大的特征作为结点的特征，由该特征的不同取值建立子节点；再对子结点递归地调用以上方法，构建决策树；直道所有特征的信息增益均很小或没有特征可以选择位为止。最后得到一个决策树。ID3相当于用极大似然法进行概率模型的选择。

经验熵与条件熵的差值就是信息熵增（在第九次报告中具体说明），选取结点的准则就是选取信息增益最大的属性（体现了分类能力大）。

下面还是以第九次报告的放贷为例构建决策树：



由于特征A3(有自己的房子)的信息增益值最大，所以选择特征A3作为根结点的特征。它将训练集D划分为两个子集D1(A3取值为”是”)和D2(A3取值为”否”)。由于D1只有同一类的样本点，所以它成为一个叶结点，结点的类标记为“是”。

     对D2则需要从特征A1(年龄)，A2(有工作)和A4(信贷情况)中选择新的特征，计算各个特征的信息增益：

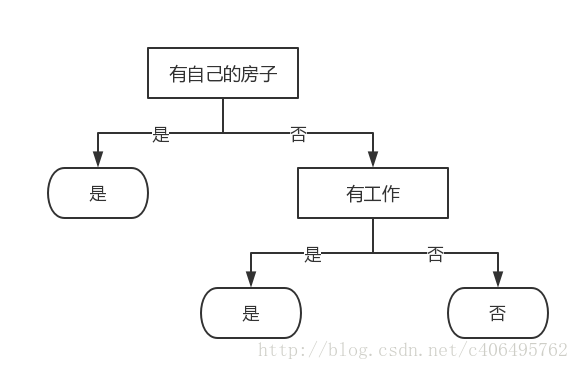
g(D2,A1) = H(D2) - H(D2 | A1) = 0.251

g(D2,A2) = H(D2) - H(D2 | A2) = 0.918 （信息增益最大，选取为结点）

g(D2,A3) = H(D2) - H(D2 | A3) = 0.474

    根据计算，选择信息增益最大的特征A2(有工作)作为结点的特征。由于A2有两个可能取值，从这一结点引出两个子结点：一个对应”是”(有工作)的子结点，包含3个样本，它们属于同一类，所以这是一个叶结点，类标记为”是”；另一个是对应”否”(无工作)的子结点，包含6个样本，它们也属于同一类，所以这也是一个叶结点，类标记为”否”。

这样子决策树就构建好了，表示为：{'有自己的房子': {0: {'有工作': {0: 'no', 1: 'yes'}}, 1: 'yes'}}，图形化表示为：



使用决策树进行分类

当决策树构建好了之后，然后就可以用来预测测试数据的结果了。

测试数据先从根开始比较，逐渐递归到叶子结点，最终的预测结果就是叶子结点的属性.。

例如测试数据：{“无房子”,”有工作”}的测试结果为：“是”。

参考资料：

1.Machine Learning Tom M.Mitchell著

2.CSDN博主Jack-Cui http://my.csdn.net/c406495762