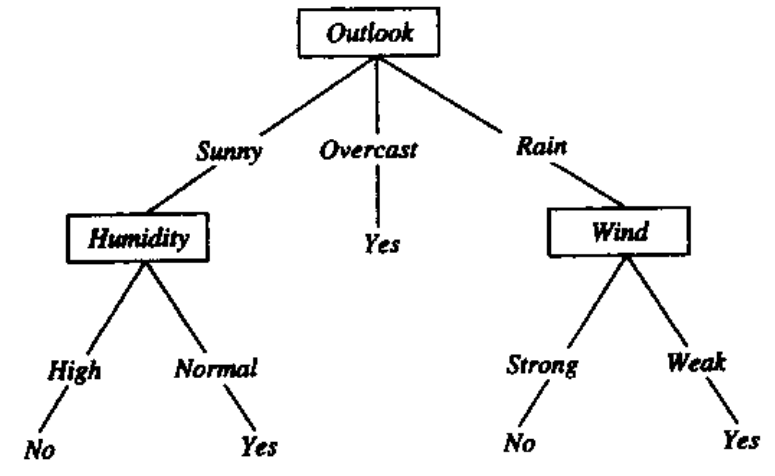
本周主要任务：学习机器学习之决策树其一

决策树：

决策树(decision tree)是一种基本的分类与回归方法，也是一种逼近离散值目标函数的方法，在这个方法中学习得到的函数表示为一棵决策树。而且学习得到的决策树也能在被表示为多个if-then的规则。

下面是一棵决策树：



决策树的表示方法：

决策树把实例从根结点排列（sort）到某个叶子结点来分类实例，叶子结点对应的就是该实例所属的分类。树上的每一个结点说明了对该实例某个属性（attribute）的测试，并且该结点的每一个后继分支对应于该属性的一个可能值。分类实例的方法就是从这棵树的根结点开始，测试这个结点的属性，然后按照给定该实例的该属性的值对应的树枝向下移动。然后以新结点为根重复这个过程。

通常情况下，决策树表示的是实例的属性值约束的合取（逻辑与）的析取式（逻辑或）。从树根到叶子的每一条路径都对应一组属性测试的合取，树本身对应这些合取的析取。

决策树可以转化为if-then规则，将决策树转换成if-then规则的过程是这样的：由决策树的根结点(root node)到叶结点(leaf node)的每一条路径构建一条规则；路径上内部结点的特征对应着规则的条件，而叶结点的类对应着规则的结论。决策树的路径或其对应的if-then规则集合具有一个重要的性质：互斥并且完备。这就是说，每一个实例都被一条路径或一条规则所覆盖，而且只被一条路径或一条规则所覆盖。这里所覆盖是指实例的特征与路径上的特征一致或实例满足规则的条件。

使用决策树来预测的过程：

收集数据：可以使用任何方法。比如想构建一个相亲系统，我们可以从媒婆那里，或者通过参访相亲对象获取数据。根据他们考虑的因素和最终的选择结果，就可以得到一些供我们利用的数据了。

准备数据：收集完的数据，我们要进行整理，将这些所有收集的信息按照一定规则整理出来，并排版，方便我们进行后续处理。

分析数据：可以使用任何方法，决策树构造完成之后，我们可以检查决策树图形是否符合预期。

训练算法：这个过程也就是构造决策树，同样也可以说是决策树学习，就是构造一个决策树的数据结构。

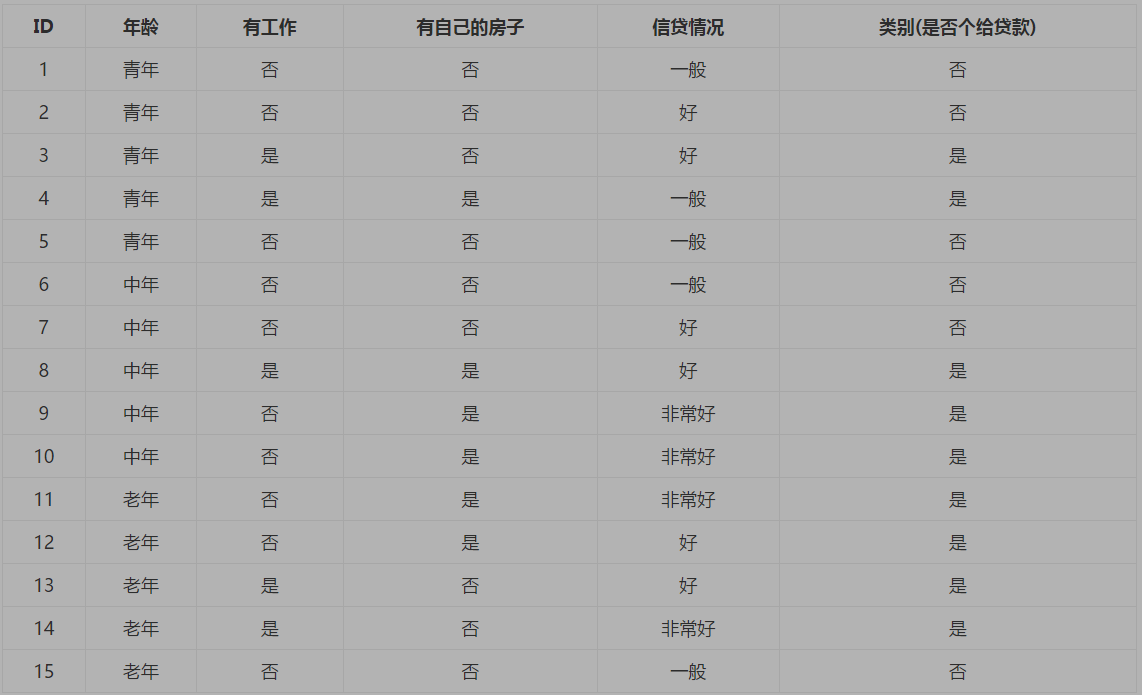
测试算法：使用经验树计算错误率。当错误率达到了可接收范围，这个决策树就可以投放使用了。

使用算法：此步骤可以使用适用于任何监督学习算法，而使用决策树可以更好地理解数据的内在含义。

决策树学习适用问题：

1. 实例由属性-值对来表示
2. 目标函数具有离散的输出值
3. 可能需要析取的描述
4. 训练数据可能包含错误
5. 训练数据可以包含缺少属性值的实例

以下的例子用下面的数据集：



如何构建一棵决策树：

这一过程可以概括为3个步骤：特征选择、决策树的生成和决策树的修剪。

特征选择：

特征选择通常选择具有分类能力的特征。通常特征选择的标准是信息增益(information gain)或信息增益比。 特征选择就是决定用哪个特征来划分特征空间，即可以选择A属性为根，也可以选择B属性，那么信息熵增益就可以决定这一标准。

关于信息增益：

在划分数据集之前之后信息发生的变化成为信息增益，知道如何计算信息增益，我们就可以计算每个特征值划分数据集获得的信息增益，获得信息增益最高的特征就是最好的选择。

香农熵：

集合信息的度量方式成为香农熵或者简称为熵(entropy)[名字来源于信息论之父克劳德·香农]

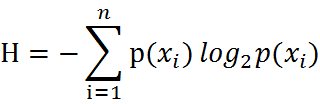
下面说下熵的计算：

  熵定义为信息的期望值。在信息论与概率统计中，熵是表示随机变量不确定性的度量。如果待分类的事务可能划分在多个分类之中，则符号xi（表示一种分类）的信息定义为

IMG_256

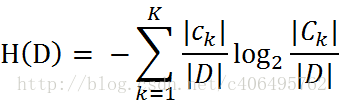
    其中p(xi)是选择该分类的概率。这是前辈的定义。上述式中的对数以2为底，也可以e为底(自然对数)。

    通过上式，我们可以得到所有类别的信息。为了计算熵，我们需要计算所有类别所有可能值包含的信息期望值(数学期望)，通过下面的公式得到：

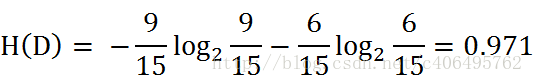


    期中n是分类的数目。熵越大，随机变量的不确定性就越大。

    当熵中的概率由数据估计(特别是最大似然估计)得到时，所对应的熵称为经验熵(empirical entropy)。什么叫由数据估计？比如有10个数据，一共有两个类别，A类和B类。其中有7个数据属于A类，则该A类的概率即为十分之七。其中有3个数据属于B类，则该B类的概率即为十分之三。浅显的解释就是，这概率是我们根据数据数出来的。我们定义贷款申请样本数据表中的数据为训练数据集D，则训练数据集D的经验熵为H(D)，|D|表示其样本容量，及样本个数。设有K个类Ck，k = 1,2,3,···,K，|Ck|为属于类Ck的样本个数，这经验熵公式可以写为：



    根据此公式计算经验熵H(D)，分析贷款申请样本数据表中的数据。最终分类结果只有两类，即放贷和不放贷。根据表中的数据统计可知，在15个数据中，9个数据的结果为放贷，6个数据的结果为不放贷。所以数据集D的经验熵H(D)为：



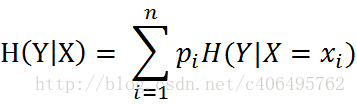
    经过计算可知，数据集D的经验熵H(D)的值为0.971。

信息增益：

如何选择特征，需要看信息增益。也就是说，信息增益是相对于特征而言的，信息增益越大，特征对最终的分类结果影响也就越大，我们就应该选择对最终分类结果影响最大的那个特征作为我们的分类特征。

条件熵：

条件熵H(Y|X)表示在已知随机变量X的条件下随机变量Y的不确定性，随机变量X给定的条件下随机变量Y的条件熵(conditional entropy) H(Y|X)，定义X给定条件下Y的条件概率分布的熵对X的数学期望：



这里，

IMG_257

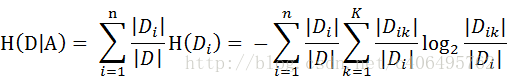
    同理，当条件熵中的概率由数据估计(特别是极大似然估计)得到时，所对应的条件熵成为条件经验熵(empirical conditional entropy)。

    明确了条件熵和经验条件熵的概念。接下来，让我们说说信息增益。前面也提到了，信息增益是相对于特征而言的。所以，特征A对训练数据集D的信息增益g(D,A)，定义为集合D的经验熵H(D)与特征A给定条件下D的经验条件熵H(D|A)之差，即

IMG_258

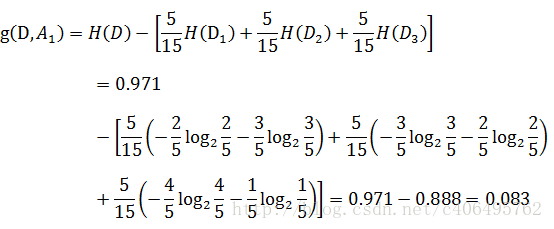
    一般地，熵H(D)与条件熵H(D|A)之差成为互信息(mutual information)。决策树学习中的信息增益等价于训练数据集中类与特征的互信息。

    设特征A有n个不同的取值{a1,a2,···,an}，根据特征A的取值将D划分为n个子集D1,D2，···,Dn，|Di|为Di的样本个数。记子集Di中属于Ck的样本的集合为Dik，即Dik = Di ∩ Ck，|Dik|为Dik的样本个数。于是经验条件熵的公式可以些为：

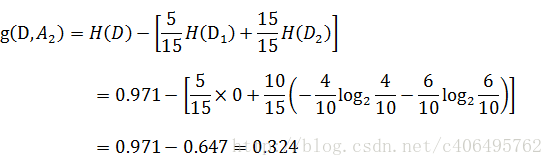


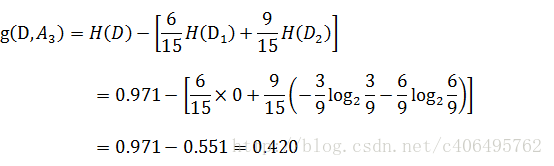
    说了这么多概念性的东西，没有听懂也没有关系，举几个例子，再回来看一下概念，就懂了。

    以贷款申请样本数据表为例进行说明。看下年龄这一列的数据，也就是特征A1，一共有三个类别，分别是：青年、中年和老年。我们只看年龄是青年的数据，年龄是青年的数据一共有5个，所以年龄是青年的数据在训练数据集出现的概率是十五分之五，也就是三分之一。同理，年龄是中年和老年的数据在训练数据集出现的概率也都是三分之一。现在我们只看年龄是青年的数据的最终得到贷款的概率为五分之二，因为在五个数据中，只有两个数据显示拿到了最终的贷款，同理，年龄是中年和老年的数据最终得到贷款的概率分别为五分之三、五分之四。所以计算年龄的信息增益，过程如下：



    同理，计算其余特征的信息增益g(D,A2)、g(D,A3)和g(D,A4)。分别为：





IMG_263

    最后，比较特征的信息增益，由于特征A3(有自己的房子)的信息增益值最大，所以选择A3作为最优特征。

参考资料：

1. Machine Learning Tom M.Mitchell著
2. CSDN博主Jack-Cui http://my.csdn.net/c406495762