研究生算法课课堂笔记

上课日期： 2016 年 10 月 10 日　　　　　　第(1)节课

组长学号及姓名：张静斌 1601111296

组员学号及姓名：郑培凯 1601214447

组员学号及姓名：秦晓冉 1601214517

一、 内容概要

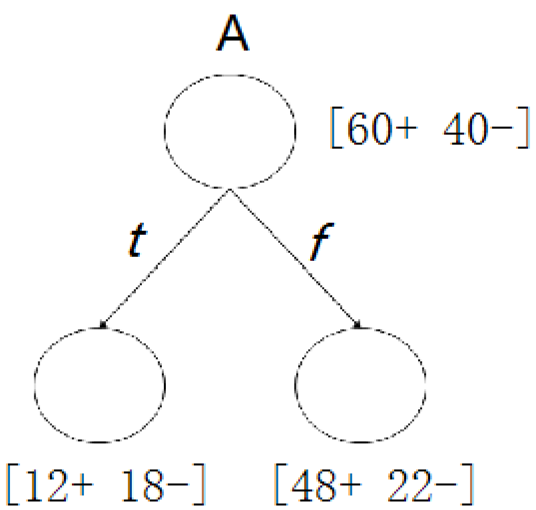
本节课内容主要包括以下几点：

1. 决策树构建过程回顾，提出量化属性划分好坏的三种方法：Information Gain, Gini Impurity 和 Misclassification Rate
2. bias – variance 理论

二、 详细内容

1. 量化属性划分好坏的方法一：Information Gain

问题：有一个样本集合 S，包括 60 个正样本和 40 个负样本，记作 S=*[60+,40-]*，对于当前节点，如何选择使得样本区分度最大的属性呢？



如上图所示，对于当前节点利用属性 A 进行划分，当判别条件 $A=t$ 时有 12 个正样本和 18 个负样本，当判别条件 $A=f$ 时有 48 个正样本和 22 个负样本。按照属性 A 划分的信息增益(Information Gain)为：

$$
\begin{aligned}
Gain(S,A) &= H(S) - \sum_{v\in(t,f)} \frac{|S_v|}{|S|} H(S_v) \\
&= H(S) - 0.3H(S_t) - 0.7H\left(S_f\right) \\
&= H(0.6,0.4) - 0.3H(0.4,0.6) - 0.7H\left(\frac{24}{35},\frac{11}{35}\right) \\
&= 0.0511
\end{aligned}
$$

一个属性 A 相对于样本集合 S 的**信息增益 *Gain(S,A)*计算公式**：

$$
Gain(S,A) = H(S) - \sum_{v\in Value(A)} \frac{|S_v|}{|S|} H(S_v)
$$

构建决策树时，我们选择信息增益最大的属性作为当前节点的划分属性。由于在计算信息增益的时候，无论选择哪个属性，被减数 *H(S)*是不变的，因此只需要使得减号后面的条件熵最小即可。决策树中用到的信息增益 *Gain(S,A)*和 information theory 中的互信息 *I(S;A)*的概念是相同的。

Information Gain 不是量化属性划分好坏的唯一方法。下面介绍另外两种方法，核心思想都是衡量划分后样本的不纯度。

2. 方法二：Gini Impurity（基尼不纯度）或者叫 Gini Index（基尼指数）

定义： Gini Index 是指在样本集中随机且有放回的取两个样本，这两个样本不是同一类的概率。Gini Index 越大，样本越不纯。

问题 1：现有一个样本集，其中包括 *k* 个已知类，每个类的概率为$\boldsymbol{P_i}$，随机有放回的取两个样本，不是同一类的概率是多少？

针对 *k=2* 的情况，两类的概率分别为P以及$(1-P)$。第一种计算方法，可以第一次选取概率为P的类，第二次选取概率为$(1-P)$的类，或者反过来第一次选取$(1-P)$，第二次选取P，那么总的$Gini\ Index = 2P(1-P)$；第二种计算方法，可以排除掉两次选取相同类的概率，$Gini\ Index = 1 - P^2 - (1-P)^2$。

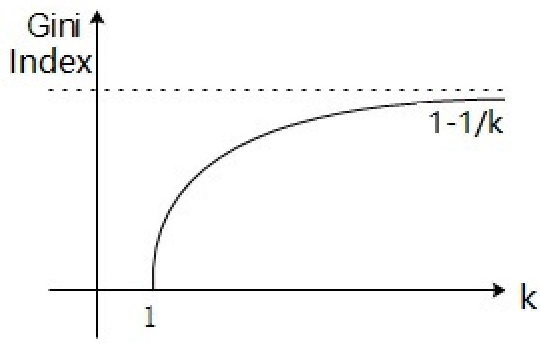
对于 k 个类的情况，也有两种求法。第一种计算方法，第一次取的样本为第 i 类的概率为P_i，第二次取的不为第 i 类的概率为$(1-P_i)$，所以$Gini\ Index = \sum_{i=1}^{k} P_i(1-P_i)$；第二种方法，排除掉两次选取相同类的概率，即$Gini\ Index = 1 -$

$\sum_{i=1}^{k} P_i^2$。

问题 2：针对 k 类样本，Gini Index 的取值范围为多少？

针对 k=2 的情况，最小值为 0，即当所有样本都为同一类时。通过对式子 $Gini\ Index = 2P(1-P)$求导，当 P=0.5 时式子取得最大值 0.5。此时取值范围为 [0,0.5]。

对于普通情况，根据 Gini Index 的定义，其最小值为 0，即样本只有一类，两次选取的结果不为同一类的概率为 0。当$k \neq 1$时，Gini Index 取得最大值的情况为各类概率相同，$P_i = {}^{1}\!/_{k}$，即样本分布越均匀 Gini Index 越大。根据问题 1 的第二种方法，两次均选取同一类的概率为$\sum_{i=1}^{k} P_i^2 = k\left({}^{1}\!/_{k}\right)^2 = {}^{1}\!/_{k}$，其对立事件$Gini\ Index = 1 - {}^{1}\!/_{k}$，当$k \to \infty$时，$Gini\ Index \to 1$。因此，$Gini\ Index \in [0, 1)$。其函数图像如下图。



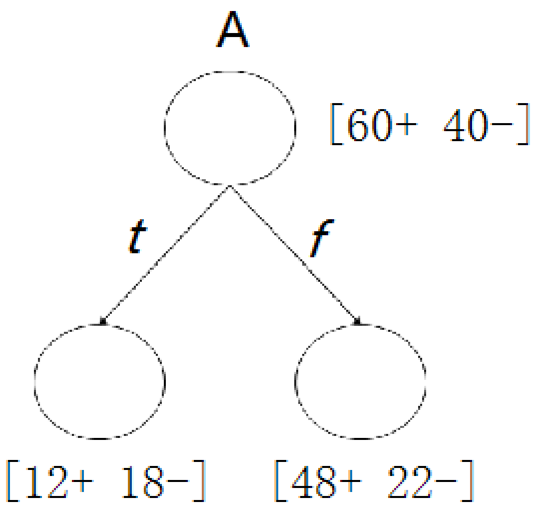
利用 Gini Index(GI)选择属性的 ***Gini(S,A)*计算公式：**

$$Gini(S, A) = GI(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} GI(S_v)$$

构建决策树时，我们选择 Gini Index 降低程度最大的属性作为当前节点的划分属性。

3. 方法三：Misclassification Rate

现在思考一下，如何用最简单直观的方法判断样本的不纯度？不需要太多的数学理论，假设我们就根据节点上样本个数最多的那一类进行预测，比如二分类问题，若正样本个数多，则预测为正，此时负样本个数为 classification error；反之亦然。



继续考虑方法一中提出的问题，一个样本集合 S，包括 60 个正样本和 40 个负样本，对于当前节点利用属性 A 进行划分，当判别条件 *A*=*t* 时有 12 个正样本和 18 个负样本，当判别条件 *A*=*f* 时有 48 个正样本和 22 个负样本。划分前，classification error = 40，划分后，当 *A*=*t* 时预测为负，当 *A*=*f* 时预测为正，classification error = 12+22=34。

利用 classification error(CE)选择属性的 ***MR(S,A)*计算公式**：

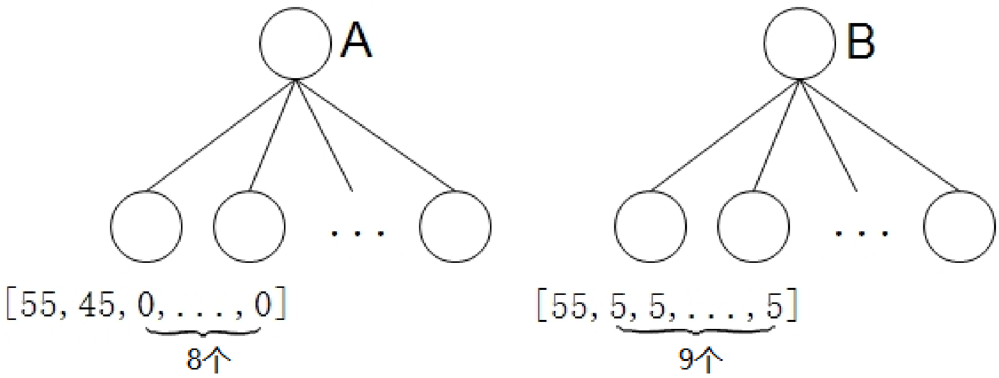
$$MR(S,A) = CE(S) - \sum_{v \in Value(A)} CE(S_v)$$

构建决策树时，我们选择 classification error 减小程度最大的属性作为当前节点的划分属性。

4. 三种方法的对比与讨论

三种方法在实际应用中的对比

实际应用中，一般采用前两种方法，即计算熵降低和 Gini Index 降低，而第三种方法 Misclassification Rate 用的不多，原因是 Misclassification Rate 在多分类问题上判断不准确，下面举例说明。



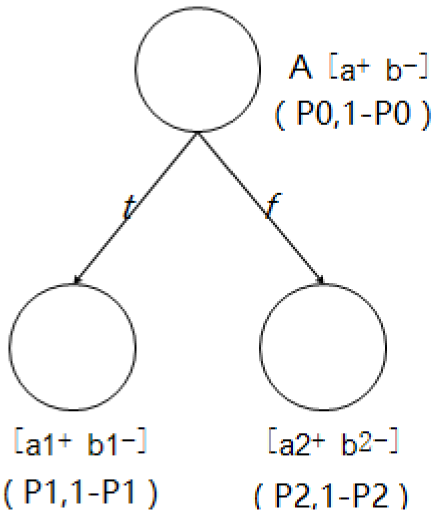
如上图所示，一个 *k* 分类问题*(k=10)*，对一个样本集合 S 分别利用属性 A 和属性 B 进行划分，划分之后，两棵决策树的叶节点除了第一个叶节点的样本分布不同，即 S_a=[55,45,0,0,0,0,0,0,0,0]，S_b=[55,5,5,5,5,5,5,5,5,5]，其余叶节点的样本分布均相同。根据方法一，由于样本集合 S 以及其余叶节点都相同，只需要考虑第一个叶节点的熵值，$\mathrm{H}(S_a) = 0.55\log\frac{1}{0.55} + 0.45\log\frac{1}{0.45}$，$\mathrm{H}(S_b) = 0.55\log\frac{1}{0.55} + 0.45\log\frac{1}{0.05}$,可以看出$\mathrm{H}(S_a) < \mathrm{H}(S_b)$,即属性 A 的信息增益较大，应该选择属性 A。然而根据方法三，属性 A 和 B 划分之后第一个叶节点的 classification error 相等为 45。从理论上分析，第三种方法只考虑了样本个数多的那一类，而前两种方法会考虑所有类别的样本分布。

划分后，熵、Gini Index、classification error 的值是否会变大的讨论

✧ **Information Gain** 利用某一属性划分之后，熵不会变大。

原因是决策树中信息增益 *Gain(S,A)*与互信息 *I(S;A)*的概念是相同的，互信息大于等于零（互信息小于零的情况是给定特殊属性值 a，但是所有属性值的熵加权平均后的互信息一定大于等于零），因此信息增益也大于等于零。特别地，熵有可能不变，例如当划分后各子节点的分布与原节点的分布相同时。

✧ **Gini Impurity** 利用某一属性划分之后，Gini Index 不会变大。分析过程如下。



首先考虑 *k=2* 时二叉决策树的情况。如上图所示，一个样本集合 S，S=*[a+, b-]*，则当前节点的概率分布为$P_0 = \frac{a}{|S|}$和$1 - P_0$。对于当前节点利用属性 A 进行划分，当判别条件 *A=t* 时有 S_t=*[a_1+, b_1-]*，概率分布为$P_1 = \frac{a_1}{|S_t|}$和$1 - P_1$；当判别条件 *A=f* 时有 S_f=*[a_2+, b_2-]*，概率分布为$P_2 = \frac{a_2}{|S_f|}$和$1 - P_2$。有$a = a_1 + a_2$且$b = b_1 + b_2$。$w = \frac{|S_t|}{S}$,为取值为 *A=t* 时的权重。根据 *Gini(S,A)*计算公式有：

$$Gini(S, A) = GI(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} GI(S_v)$$

$$= 2P_0(1 - P_0) - 2wP_1(1 - P_1) - 2(1 - w)P_2(1 - P_2)$$

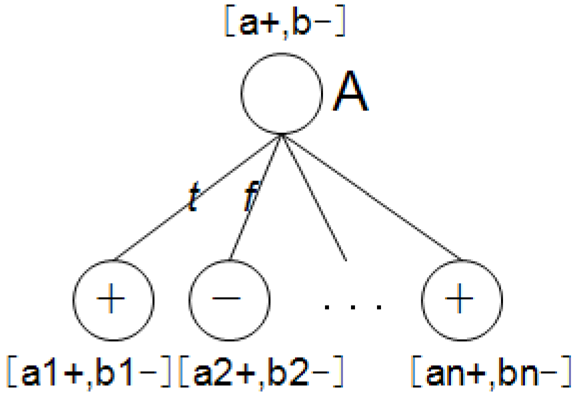
其中$wP_1 + (1 - w)P_2 = P_0$，令$P_2 = \frac{P_0}{1-w} - \frac{wP_1}{1-w}$，带入上式，由于$P_0$是定值，式子可变为关于$P_1$的二项式，可得：

$$\frac{2w^2}{1 - w}{P_1}^2 - \frac{4w}{1 - w}P_0P_1 + \frac{{P_0}^2}{1 - w} - 2{P_0}^2$$

由于$\frac{2w^2}{1-w} > 0$，当$P_1 = P_0$时，上式有最小值 0。所以$Gini(S, A) \geq 0$成立，Gini Index 不会变大。如果$|\text{value}(A)| > 2$，需要用到 Lagrange 乘子法，对约束条件$\sum w_i = 1, \sum w_i P_i = P_0$进行限定，在这里就不给予证明了。对于 *k>2* 的情况，同理。

综上所述，利用某一属性划分之后，Gini Index 不会变大。

✧ **Misclassification Rate** 利用某一属性划分之后，classification error 不会变大。分析过程如下。



首先考虑 *k=2* 的情况。如上图所示，一个样本集合 S，S=*[a+, b-](a>b)*，对于当前节点利用属性 A 进行划分，当判别条件 *A=t* 时有 S_t=*[a_1+, b_1-]*，当判别条件 *A=f* 时有 S_f=*[a_2+, b_2-]*……，有$a=\sum_{v\in Value(A)}a_v$且$b=\sum_{v\in Value(A)}b_v$。我们可以把所有叶节点分为两类 S_{v1} 和 S_{v2}，S_{v1}=*[a_{v1}+, b_{v1}-]($a_{v1}>b_{v1}$)*，S_{v2}=*[a_{v2}+, b_{v2}-]($a_{v2}<b_{v2}$)*，S_{v1} 预测为正，S_{v2} 预测为负。根据 *MR(S,A)*计算公式有：

$$MR(S,A)=CE(S)-\sum_{v\in Value(A)}CE(S_v)$$

$$=b-\sum_{v1\in Value(A)}b_{v1}-\sum_{v2\in Value(A)}a_{v2}$$

要证明

$$\mathrm{b}-\sum_{v1\in Value(A)}b_{v1}-\sum_{v2\in Value(A)}a_{v2}\geq 0$$

由于 $a_{v2}<b_{v2}$

$$\mathrm{b}-\sum_{v1\in Value(A)}b_{v1}-\sum_{v2\in Value(A)}a_{v2}\geq \mathrm{b}-\sum_{v1\in Value(A)}b_{v1}-\sum_{v2\in Value(A)}b_{v2}$$

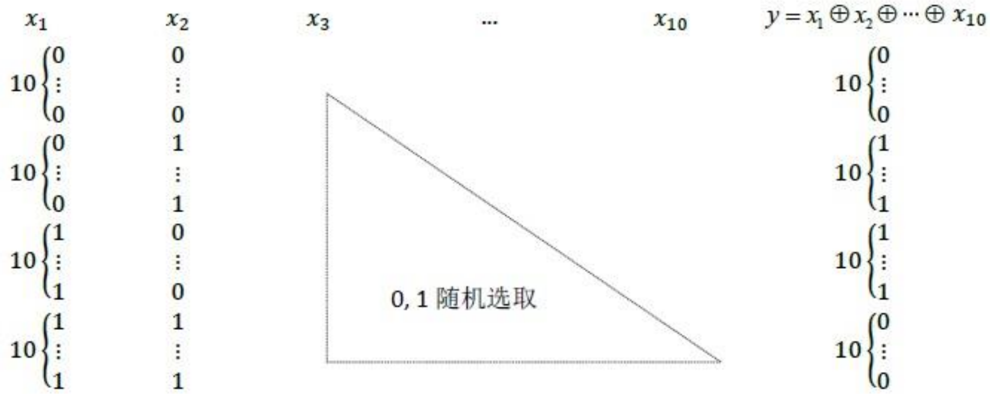
$$=b-\sum\nolimits_{v\in Value(A)}b_v=0$$

所以上述不等式成立。即$MR(S,A)\geq 0$，classification error 不会变大。对于 *k>2* 的情况，同理。

综上所述，利用某一属性划分之后，classification error 不会变大。

5. 关于贪心算法构建决策树是否最优？

贪心算法不能保证最优决策树，如上节课的异或例子，但是在现实生活中这样的例子很少，所以利用贪心算法构建决策树一般可以得到比较好的解。下面是上节课举的异或例子。



有同学提出，由于 y 是所有变量 $x_1 \ldots x_{10}$ 的异或，那么直到最后一个变量 x_{10} 值确定后，y 值才能确定，所以前 9 个变量的 information gain 都是 0，直到 x_{10} 的值确定后 information gain 才变为最大。这样的想法是不正确的，需要注意的是，Information Gain 是计算出来的，不是理论推导出来的。

6. bias – variance

High-bias 模型比较简单，表达能力不强，对数据拟合程度比较低，例如线性回归。这类模型会出现欠拟合(underfitting)问题。

hard bias(representation bias)	例如线性回归。模型表达能力不强
soft bias(search bias)	例如决策树算法中选择比较矮的树。

High-variance 模型比较复杂，表达能力强，对数据拟合程度高，例如决策树和深度学习。这类模型容易出现过拟合(overfitting)问题。决策树是属于 low bias/high variance 的。

High bias 和 **High variance** 都不好。

Occam's Razor(剃刀原理)：如果有几个模型都是成立的，那么简单的模型往往是比较好的。

三、 遗留问题

我们小组针对 Gini Index 划分后会不会增高这一问题，还有另一种思路，但没有证完。希望，老师、同学们有能力的话帮忙完成，谢谢。

首先考虑 *k=2* 的情况。如上图所示，一个样本集合 S，S=*[a+, b-]*，对于当前节点利用属性 A 进行划分，当判别条件 *A=t* 时有 S_t=*[a_1+, b_1-]*，当判别条件 *A=f* 时有 S_f=*[a_2+, b_2-]*……，有$a=\sum_{v\in Value(A)} a_v$且$b=\sum_{v\in Value(A)} b_v$。根据 *Gini(S,A)* 计算公式有：

$$Gini(S,A)=GI(S)-\sum_{v\in Value(A)}\frac{|S_v|}{|S|}GI(S_v)$$

$$=2\frac{a}{|S|}\left(1-\frac{a}{|S|}\right)-\sum_{v\in Value(A)}\frac{|S_v|}{|S|}\left(2\frac{a_v}{|S_v|}\left(1-\frac{a_v}{|S_v|}\right)\right)$$

要证明

$$2\frac{a}{|S|}\left(1-\frac{a}{|S|}\right)-\sum_{v\in Value(A)}\frac{|S_v|}{|S|}\left(2\frac{a_v}{|S_v|}\left(1-\frac{a_v}{|S_v|}\right)\right)\geq 0$$

即要证明

$$a\left(1-\frac{a}{|S|}\right)-\sum_{v\in Value(A)}\left(a_v\left(1-\frac{a_v}{|S_v|}\right)\right)\geq 0$$

$$\left(a-\sum_{v\in Value(A)}a_v\right)+\left(\sum_{v\in Value(A)}\frac{a_v^2}{|S_v|}-\frac{a^2}{|S|}\right)\geq 0$$

$$\sum_{v\in Value(A)}\frac{a_v^2}{|S_v|}-\frac{a^2}{|S|}\geq 0$$

即需要证明$\sum_{v\in Value(A)}\frac{a_v^2}{|S_v|}-\frac{\left(\sum_{v\in Value(A)}a_v\right)^2}{\sum_{v\in Value(A)}|S_v|}\geq 0$，感觉可以通过放缩证明，但是很遗憾我们没有想出具体的办法，希望老师、同学们能帮忙完成。