研究生算法课课堂笔记

上课日期：2016 年 10 月 13 日　　第(1)节课

组长学号及姓名：白剑刚 1601111321

组员学号及姓名：郭天宇 1601111323 李逸峰 1601111328

注意：请提交 Word 格式文档

一、　内容概要

本节课所学内容包括以下几点:

1. 回顾了决策树(Decision Tree)中决策属性选择的三种标准.
2. 说明了决策树和逻辑斯蒂回归(Logistic Regression)两者决策边界(Decision Boundary)的区别.
3. 讲解了决策树在数值型(Numerical),离散有序型(Discrete but ordered)和分类型(Categorical)三种不同属性类型的数据上划分时的处理细节.
4. 讲解了属性值分支较多导致决策树泛化能力弱,属性值获取有成本(Attributes with Costs)以及属性值缺失(Missing Value)三个问题的处理方法.

二、　详细内容

1. 决策属性选择的三种标准
   a) 熵(Entropy):选择能够使以某属性划分以后的数据熵下降最大的这个属性作为决策树划分的决策属性.

   b) 基尼不纯度(Gini Impurity):基尼不纯度指有放回(with replacement)地随机在划分出的数据集合内抽取两个样本,这两个样本属于不同类别的概率.如果某数据集合中有 k 类样本,每类样本出现的概率是$p_i(i=1,\dots,k)$,则该集合的基尼不纯度可以用以下公式刻画:

$$GI=\sum_{i=1}^{k}p_i(1-p_i)=1-\sum_{i=1}^{k}p_i^2$$

   基尼不纯度在数据集中只含有一类样本时取最小值 0,在数据集 k 类样本数目均相等时取值最大,为$1-\frac{1}{k}$.

   在用基尼不纯度作为决策树选取决策属性的标准时,同样需要选择能够让划分后的数据基尼不纯度减少最大的属性作为决策属性.

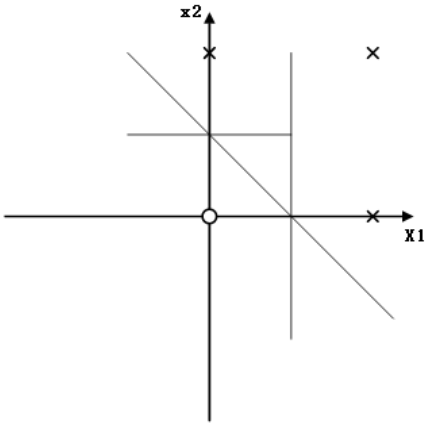
   c) 误分类率(Miss-classification Rate):在这里,我们将每一个数据集合中样本数量不是最多的类别都称为误分类类别,误分类类别的样本占样本总体的比率称为误分类率.假设数据集中一共有 k 类样本,误分类率在数据集是同一类(即 k=1)时误分类率最小为 0,数据集中每一类拥有的样本数目均相同时误分类率最大为$1-\frac{1}{k}$.同样的,在用误分类率作为决策树选取决策属性的标准时,同样需要选择能够让划分后的数据误分类率减少最大的属性作为决策属性.

- 这三个评判标准都从不同方面刻画了数据的混乱度,数据集内数据越不纯,则这三个指标会越高.

2. 决策边界(Decision Boundary)

a) 决策树和逻辑斯蒂回归的决策边界

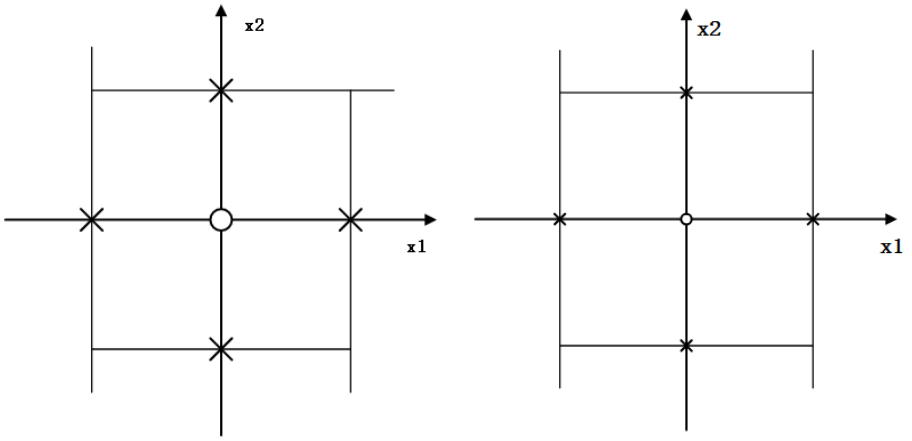
由于决策树每次划分的依据是决策属性的取值,导致决策树的决策边界是与属性的坐标轴平行的。而逻辑斯蒂回归中,我们需要通过已有数据学习线性分类平面$y = w_0 + w^T x$,这个决策平面的目的是尽可能分割开不同类型的样本,决策边界一般不会与坐标轴平行。



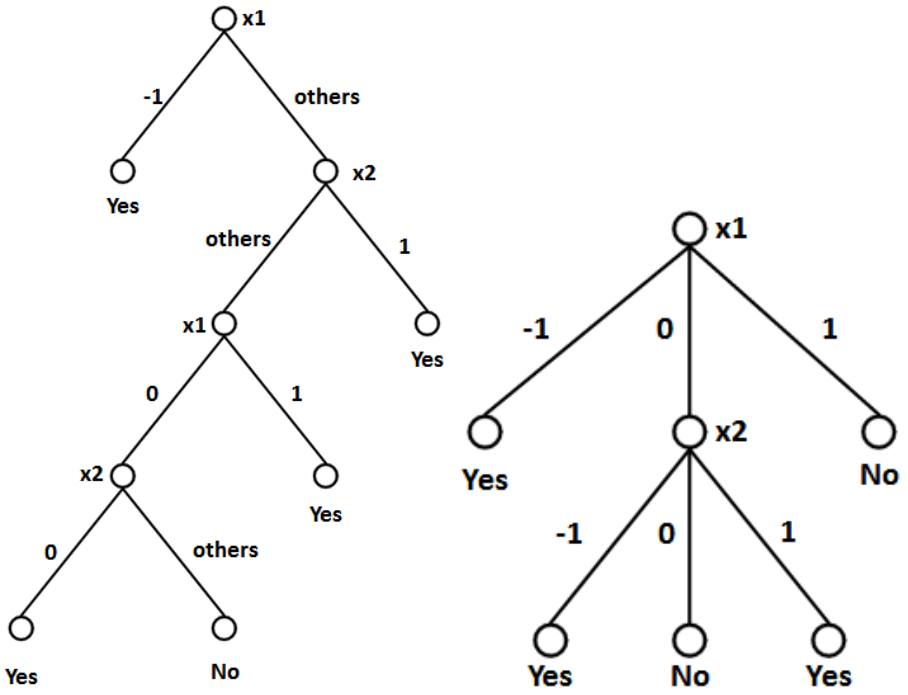
上图的二属性分类中,决策树形成的决策边界为与坐标轴平行的两条直线;而采用逻辑斯蒂回归则会产生一条斜线作为决策边界.

b) 二叉决策树和多叉决策树的决策边界

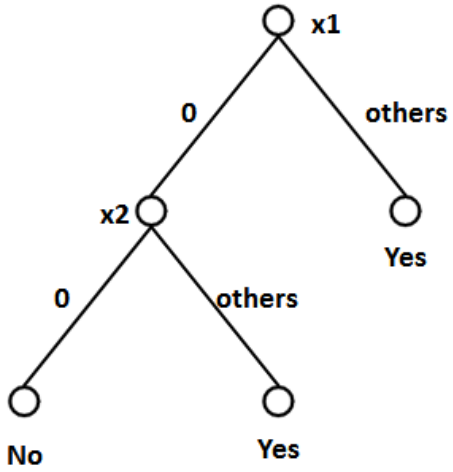
二叉决策树在每一次决策中都会对数据集做出一次切分,会产生一个决策边界,如下图左;多叉决策树在每一次决策中都会对数据集做出至少一次切分,会产生至少一个决策边界,如下图右.（下面这两个图应该在叉子和圆圈之间切一刀，而不是压在叉子上面切一刀。）



它们对应的决策树分别是(左边图中最下面的分叉 x2 应该是 0 和 1，右边图的最右边节点应该是 Yes)：



实际上依照奥卡姆剃刀原理,课上所画的上图左侧决策树可以有更简洁的形式如下:（这个说法不对，每个属性如果只能切一刀的话，无法产生下图所示的分类面）



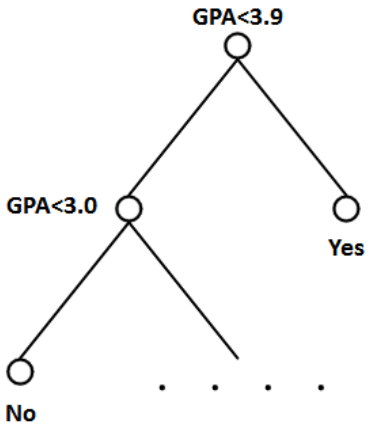
3. 不同属性类型的决策树划分
   a) 数值型(Numerical)属性
   属性取值为连续型变量的称为数值型属性.数值型属性通常会在label 发生变化的地方取边界值作为最优划分点.例如:

温度(℃)	是否打球
6.4	-
12.8	-
23.3	+
27.1	+
32.0	-
35.5	-

以温度作为属性,打球与否作为 label,我们会选择在 12.8-23.3 以及 27.1-32.0 之间选择划分点决策是否去打球.

- 问:多叉决策树中,同一个离散型属性会不会在某个样本的决策路径上出现两次以上?数值型变量呢?

  答:在多叉决策树中,由于使用离散型属性做切分后,每一个分支上的所有数据该属性值相同,所以之后不会再采用该属性作为决策属性,即同一个属性不会在一个分支上出现两次。而数值型属性在本质上是一个随机变量,每一个切分点都会产生一个数值型变量在这个切分点上的新的属性,数值型属性会在某个样本的决策路径上出现一次以上,但它们实际上都是不同的属性,只是共用了一个随机变量的符号.
  例如在招生问题中,可能会产生如下决策树,数值型变量GPA在某些样本的决策路径上出现了不止一次（左分支是 Yes，右分支是 No）:



- 问:二叉决策树中,同一个离散型属性会不会在某个样本的决策路径上出现两次以上?

  答:有可能出现两次以上.二叉决策树中,离散型属性一次只可以针对一个属性值进行划分,本质上与数值型属性并无区别.

b) 离散有序型(Discrete but ordered)属性
属性值离散但具有某种意义上的比较关系的属性被称为离散有序型属性.比如成绩={优,良,中,差},机考做出的题目数={0,1,2,…}等这些属性取值离散但有比较关系,属于离散有序属性.

c) 分类型(Categorical)属性
属性值离散但没有比较关系的属性被称为分类型属性.比如天气={阴,晴,雨,雪,霾},专业={计算机,电气工程,物理,…}等这些属性的取值并没有比较意义,属于分类型属性.
对于这类属性，我们采用 one hot encoding 的方式进行处理，将每一个属性值的取与不取视为一个新的属性的 1 与 0.这样我们会产生原属性可取值个数维度的 one hot encoding 属性向量，其中只有原

属性取值对应的新属性值一处为 1，剩下的新属性值为 0. 进行 one hot encoding 有可能导致决策树的深度变大, 因为每次切一刀时只能把一个可能的属性取值分出来。

4. 实践中可能遇到的问题及解决方法

a) 存在属性取值较多且样本在各属性上分布均匀

在实际分类问题中, 存在属性取值较多, 导致数据集被切分成很多小的数据集, 这些小数据集上样本数目较少, 往往数据分类不纯度较低, 这样的属性很可能被选择作为决策属性. 但是这种属性有可能与具体的分类问题无关, 在整个分类问题上泛化能力往往并不显著. 比如在招生问题中, 用身份证号作为属性可以将N个学生样本分为N个子数据集, 每个数据集上样本纯度为 1, 但是身份证号并不可能作为招生与否的依据, 仅仅是在形式上使决策树做出了一次较好的划分.

为了规避这一问题, 我们采用$GainRaito$作为选择决策属性的标准. 定义S为待划分数据集, A为决策属性，$GainRatio$为

$$GainRatio(S,A) = \frac{InfoGain(S,A)}{SplitInfo(S,A)}$$

其中,

$$InfoGain(\mathrm{S},\mathrm{A}) = Entropy(S) - \sum_{v \in value(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$SplitInfo(S,A) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} log\left(\frac{|S_i|}{|S|}\right)$$

$InfoGain$是上节课讲授的内容, 刻画了在本次划分后不纯度下降的大小, 在此不再赘述.

$SplitInfo$是具有熵的形式, 在属性A只有一个取值时$SplitInfo$取最小值 0, 在属性A的k个取值中都有相同个数的样本时取值最大, 为$log(\mathrm{k})$. 可以看出, $SplitInfo$在属性A有较多取值, 且在每个取值上样本分布均匀时较大. 这样会导致$GainRaito$降低, 从而导致该属性被选择的概率降低.

实际操作中, 我们通常用$InfoGain$选择出能够带来信息增益最大的几个属性, 然后为了防止选中取值多且样本在各取值上分布均匀的属性, 我们对这几个属性计算$GainRatio$然后选择$GainRatio$最大的属性作为本次的决策属性.

b) 属性值获取有成本

通常情况下, 属性值的获取只需要简单的采样成本, 但是某些特殊情况下属性值获取需要一定的时间, 人力以及物力, 这种情况下我们不能无视属性值的获取成本进行决策属性的选择. 为了避免获取属性值带来的高额成本, 我们需要对获取成本较高属性值的重要性在选取时进行一定的抑制. 可以用如下两个公式进行决策属性的选择, 可以降低对获取成本较高属性的选中概率:

$$\frac{InfoGain(S,A)}{Cost(A)}$$

和

$$\frac{2^{InfoGain(S,A)}-1}{(Cost(A)+1)^{w}}$$

$w \in [0,1]$代表属性成本在属性选择中的影响的重要程度.

c) 属性值缺失

在数据集中,往往存在一些样本的一些属性值缺失,无法进行准确分类的情况.对于这种情况,通常有如下的处理方法:对于大量样本中只有少量样本的属性缺失,可以抛弃这些属性值缺失的样本,不会对最后的决策树分类器产生较大影响.

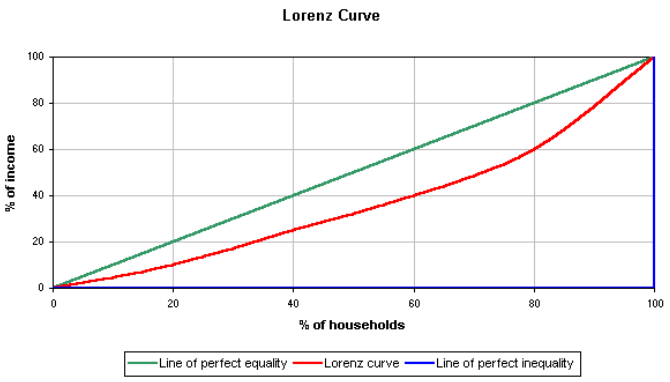
在样本数较少或较多样本发生了属性缺失时,抛弃这些属性缺失的样本会对最后的决策树产生很大影响,使它泛化能力降低.我们可以采取如下三种处理方式:

1. 将缺失属性的样本所处数据集合中,缺失属性取值最多的属性值赋予该样本的缺失属性,然后进行进一步决策.
2. 统计方法,允许样本按照所处数据集合中缺失属性的取值比例而被分入不同的数据集合,将决策树叶节点中该样本被划分成的正负类别按比例各自加和确定该样本最终的决策结果.
3. 将缺失属性视为新的属性值Unknown,然后进行进一步决策.

- 问:利用统计方法产生决策树有什么缺陷?
  答:按照统计方法产生决策树,将缺失属性视为数据集合上该属性的取值按照该属性取值的样本数占样本总数的比例组合而成,导致属性缺失的样本会依照属性取值比例的不同被部分分入子树的样本集合中去.但是在实际问题中,样本的属性缺失是由于分类结果与样本在该属性上的取值并不相关,该属性缺失的情况下也可以做出置信度较高的判断.例如在医院诊断中,病人只会做与病情相关的检查,并不会做所有的检查,医生有把握缺失的检查属性不会影响他对病人病情的诊断.在这类情况下,不应该将样本在统计意义上分开,这样反而会降低分类器泛化能力.

PS:基尼指数(Gini Index)[1]

设右图中的实际收入分配曲线(红线)和收入分配绝对平等线(绿线)之间的面积为A,和收入分配绝对不平等线(蓝线)之间的面积为B,



[1] Source: 基尼系数-维基百科 http 5%B0

则表示收入与人口之间的比例的基尼系数为$\frac{A}{A+B}$.

如果A为零,即基尼系数为0,表示收入分配完全平等(红线和绿线重叠);如果B为零,则系数为1,收入分配绝对不平等(红线和蓝线重叠).该系数可在0和1之间取任何值.收入分配越趋向平等,劳伦茨曲线的弧度越小(斜度越倾向45度),基尼系数也越小;反之,收入分配越趋向不平等,劳伦茨曲线的弧度越大,那么基尼系数也越大.