决策树

统计学习方法之决策树

- 决策树的好处：可解释性强、直观；只要场景符合信息论的建模假设，模型就异常强大
- 决策树的弊端：特征通常是混乱的且取值多，并且特征之间也并不独立，有时候不能如此条理清晰的一层一层分下去
- 应用：应用在医疗类领域十分有说服力

一、熵有关的

《信息论》--- 香农：

- 熵 $H(X)$ 的大小表示分布 X 的不确定性，是分布 X 蕴含的信息量的期望，单位是bit（或者nat）
- 熵的计算公式： $H(X) = -\sum_{i=1}^{n} p_i * log(p_i)$， 其中 $p_i = P(X = i), i = 1, 2...n$
  公式中的 $-log(p_i)$ 表示信息量，信息量乘上它发生的概率再加和相当于是对信息量加权平均，也就表示了 X 这个分布所蕴含的信息量
- 为什么信息量可以表示为 $-log(p_i)$ ：
  首先，一件事情越**确定**其所蕴含的信息量就越少。
  比如太阳从东边升起发生概率为 100%，你听了之后也不会做出什么反应、得到什么信息，反之如果我说太阳从西边升起了，那这句话背后的信息可能有：地球磁场紊乱、我眼睛出问题了等等，蕴含的信息量就会很大。

  其次，信息应该具有**可加性**。
  比如我事件A有 25%的机会获得4 bit信息量，事件B有 50% 的机会获得5 bit的信息量，那么当A和B同时发生时（概率为 $25\% * 50\%$），我应该能获得 $4 + 5$ bit的信息量，也就是说概率相乘而信息量相加。

  所以，对信息量的建模应该满足：
  单调递减函数、且 $f(x_1 * x_2) = f(x_1) + f(x_2)$ ，所以 $-log(x)$ 就被选中了。
- 条件熵 $H(X)$ 表示在已知 X 的情况下 的不确定性，$H(X) = \sum_{i=1}^{n} p_i * H(X = x_i)$
- 当熵（条件熵）的概率是由数据估计得到时，其又被称为经验熵（条件经验熵）
- 信息增益 $g(,) = H() - H()$ 表示由于知晓了A之后D的不确定性降低的程度
- 信息增益比 $g(,) = {}_{H()-H()H()}$ ，norm了一下信息增益，因为如果仅按照信息增益的差来选择特征的话特征取值的选项越多，信息增益自然就越大，所以要除以一个A的取值的熵来约束一下
- 基尼指数 $ini(p) = \sum_{=1} p * (1 - p)$ ，也是用来衡量分布 p 的不确定性的，他的真实含义是：从总体中随机抽取两个样本，这两个样本属于不同种类的概率

二、剪枝

- 为什么 () 能够表征一颗决策树的预测误差?
  因为 () 本身是叶子节点的熵之和，决策树的叶子节点是“分完类”的确定节点，按道理来说其熵值应该低（因为它有很强的确定性，信息量应该含有的很少），所以可以用它来表示预测误差。但对于训练数据剪枝来说，他肯定是0，所以后面要再加一个 树的复杂程度来惩罚一下这个式子