决策树

信息熵

- 用于度量样本集合纯度最常用的一种指标
- 假定样本集合D中的第k类样本所占比例为p_k,则D的信息熵定义为

$$Ent(D) = -\sum_{k=1}^{|y|} p_k log_2 p_k$$

信息增益

- 描述分支结点的影响,第v个分支结点包含了D中所有在属性上的取值为a(v)的样本，记为D(v)

$$Gain(D, a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$$

增益率

- 考虑到分支结点越多,信息增益就越容易大，但是分支结点数量多会造成时间上的开销，所以将分支结点的取值数目和信息增益综合考虑
- 增益率定义为

$$Gain_ratio(D, a) = \frac{Gain(D, a)}{IV(a)}$$

- IV(a)称为属性a的固有值,属性a的可能取值数目越多，IV(a)的值通常会越大

$$IV(a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} log_2 \frac{|D^v|}{|D|}$$

基尼指数

- CART决策树使用基尼指数来选择划分属性，用来度量数据集的纯度
- 属性a的基尼指数定义为

$$Gini(D) = \sum_{k=1}^{|y|} \sum_{k'!=k} p_k p_{k'} = 1 - \sum_{k=1}^{|y|} p_k^2$$

剪枝

- 预剪枝
  - 在划分结点的时候比较该结点下分支是否有助于验证集精度的提升，如果不能就不划分了

- 后剪枝
  - 在建完整棵树之后，自底向上判断如果将当前节点改为叶子结点是否可以提升精度，如果可以，则将其改为叶子节点
- 区别
  - 后剪枝的欠拟合风险更小，泛化性能更好，但是训练时间开销会更大

连续

- 需要连续而非离散的值的时候，我们需要找到一个数字t，将不超过t的和超过t的划分为两种类型，形成离散的效果

$$Gain(D,a) = max_{t\in T_a} Gain(D,a,t)$$
$$= max_{t\in T_a} Ent(D) - \sum_{\lambda\in\{-,+\}} \frac{|D_t^\lambda|}{|D|} Ent(D_t^\lambda)$$