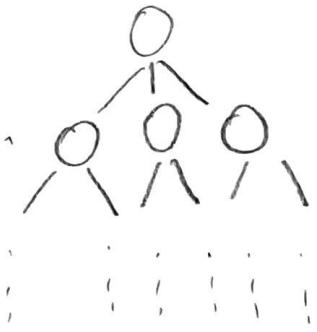
2022.02.28

Decision Tree



从一开始的input，不断通过各种问题将input分类，渐渐细化。可binary split / multi split.

- Descretion
  - 将continuous attribute 转化为descret 的几个range
  - static：在一开始一次性descretize
  - dynamic: ranges can be found by equal interval bucketing

- 如何看哪种split更好？
  最终各个split中data point的portion尽量不同。

例

	S_1	S_2	S_3
包含	1个C_1	0个C_1	7个C_1
	2个C_2	8个C_2	0个C_2

Gini index: (1个node要split时怎样更好？)

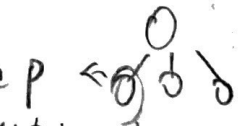
$GINI(t) = 1 - \sum_j [p(j|t)]^2$ 表示data原本的类，t表示最终split.

↓ $p(j|t)$表示node t中j的占比 例 S_1 C_1为$\frac{1}{3}$ C_2为$\frac{2}{3}$

Combine

$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$

k表示partition个数 n为node p中总records，n_i为child i中的records.



GINI split在算每一步split的好坏时决定此node如何split.

例看split S. 则算[不split的 GINI index - split S的GINI index.]
哪个split使 ← 最大, 哪个为最好.

∴ Continuous attribute中
input 先要排序

· Stop criteria
① node 中所有 records 属于同一 class
② node 中 所有records 有similar attribute values.

· Model evaluation
例 class 1 有9990 samples, class 2 有10 samples.
若model 将所有都分成class1, accuracy 为99%. problem!

Cost Matrix

		predicted Yes	predicted No
actual	Yes	C(Y|Y)	C(N|Y)
	No	C(Y|N)	C(N|N)

→ 给每个cell 给个 cost.
例

10	20
-1	50

C(i|j) 表示 actual为j, predict 成 i 的 cost.
最终算总 cost.

· Test Model

· Holdout
$\frac{2}{3}$ for training, $\frac{1}{3}$ for testing

·· cross validation
将 data分为 k个 partitions (disjoint), k-1会用来training,
求剩余1个用来test.

例：25个 classifier，每个error rate为 0.35．每个unseen record会根据 majority of ↓来分类

majority 错的 prob：$\sum_{i=13}^{25} \varepsilon^i (1-\varepsilon)^{25-i} = 0.06$．

classifier 间需 independent．

· 用GINI index 分某个 attribute (binary)

① sort the attribute values，
② 尝试所有 possible split 计算GINI　③选最小值．

例　attribute values　3　7　11
　　2　4　8　12
　　↙
　　分水岭

根据与"分水岭"大小关系来split成2组，算组算GINI