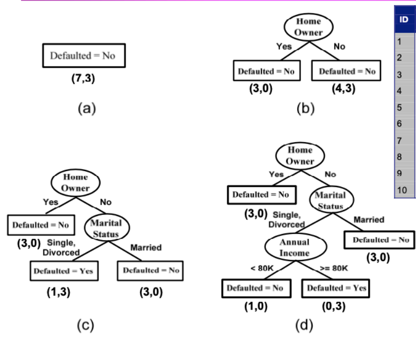
decision trees

Hunt's Algorithm



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

impurity measures = determining best split, with highest gain lowest impurity

gain = P (impurity before split) − M (impurity after split)

↳ gini index = $Gini\ Index = 1 - \sum_{i=0}^{c-1} p_i(t)^2$ Where $p_i(t)$ is the frequency of class i at node t, and c is the total number of classes

- max = $1-\frac{1}{c}$ → when all records are equally distrubuted among all classes (between 0 and 1)
- min = 0 → when all records belong to one class

* gini index for a collection of nodes = When a node p is split into k partitions (children)

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

where, n_i = number of records at child i,
n = number of records at parent node p.

* gini index for continuous attributes =

60, 70, 75, 85, 90, 95 | 100, 120, 125, 220

Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No
Annual Income										
Sorted Values	60	70	75	85	90	95	100	120	125	220

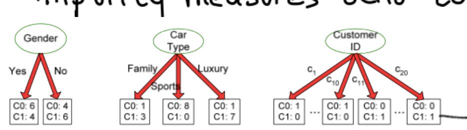
Split Positions	55		65		72		80		87		92		97		110		122		172		230	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Yes	0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0
No	0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	7	0
Gini	0.420		0.400		0.375		0.343		0.417		0.400		0.300		0.343		0.375		0.400		0.420	

↳ entrophy measure = $Entropy = -\sum_{i=0}^{c-1} p_i(t) log_2 p_i(t)$

- max = $\log_2 c$, min = 0

* information gain: $Gain_{split} = Entropy(p) - \sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)$

- impurity measures tend to prefer splits that result in large number of partitions, each being small but pure



→ costumer id has the highest gain, all are pure

gain ratio = designed to overcome the disadvantage of information gain

$$Gain\ Ratio = \frac{Gain_{split}}{Split\ Info} \qquad Split\ Info = -\sum_{i=1}^{k} \frac{n_i}{n} log_2 \frac{n_i}{n}$$

Parent Node, p is split into k partitions (children)
n_i is number of records in child node i

↳ classification error = $Error(t) = 1 - \max_i [p_i(t)]$ max = 1 − 1/c , min = 0

Comparison among Impurity Measures

For a 2-class problem: