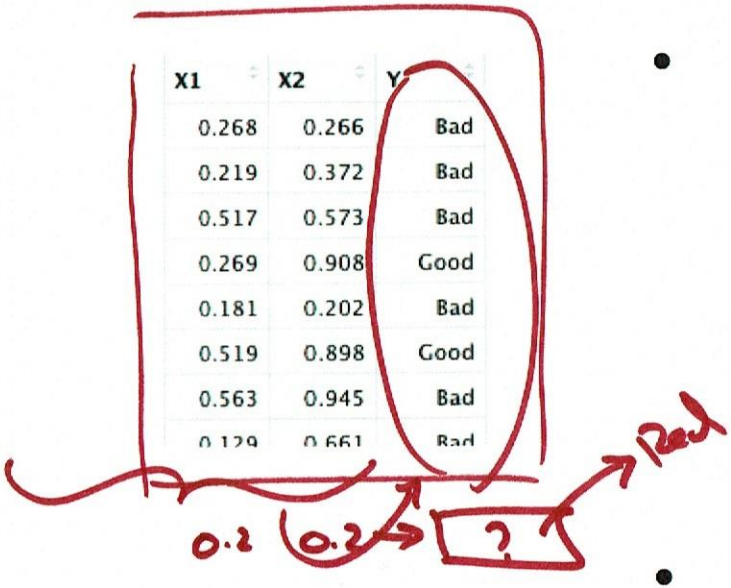
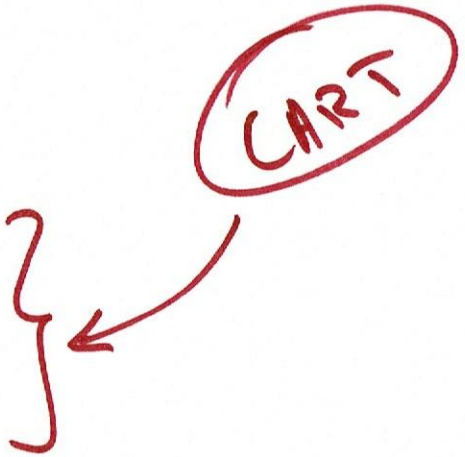
Classification and Regression

- Decision Trees can be used for both



X1	X2	Y
0.268	0.266	Bad
0.219	0.372	Bad
0.517	0.573	Bad
0.269	0.908	Good
0.181	0.202	Bad
0.519	0.898	Good
0.563	0.945	Bad
0.129	0.661	Bad

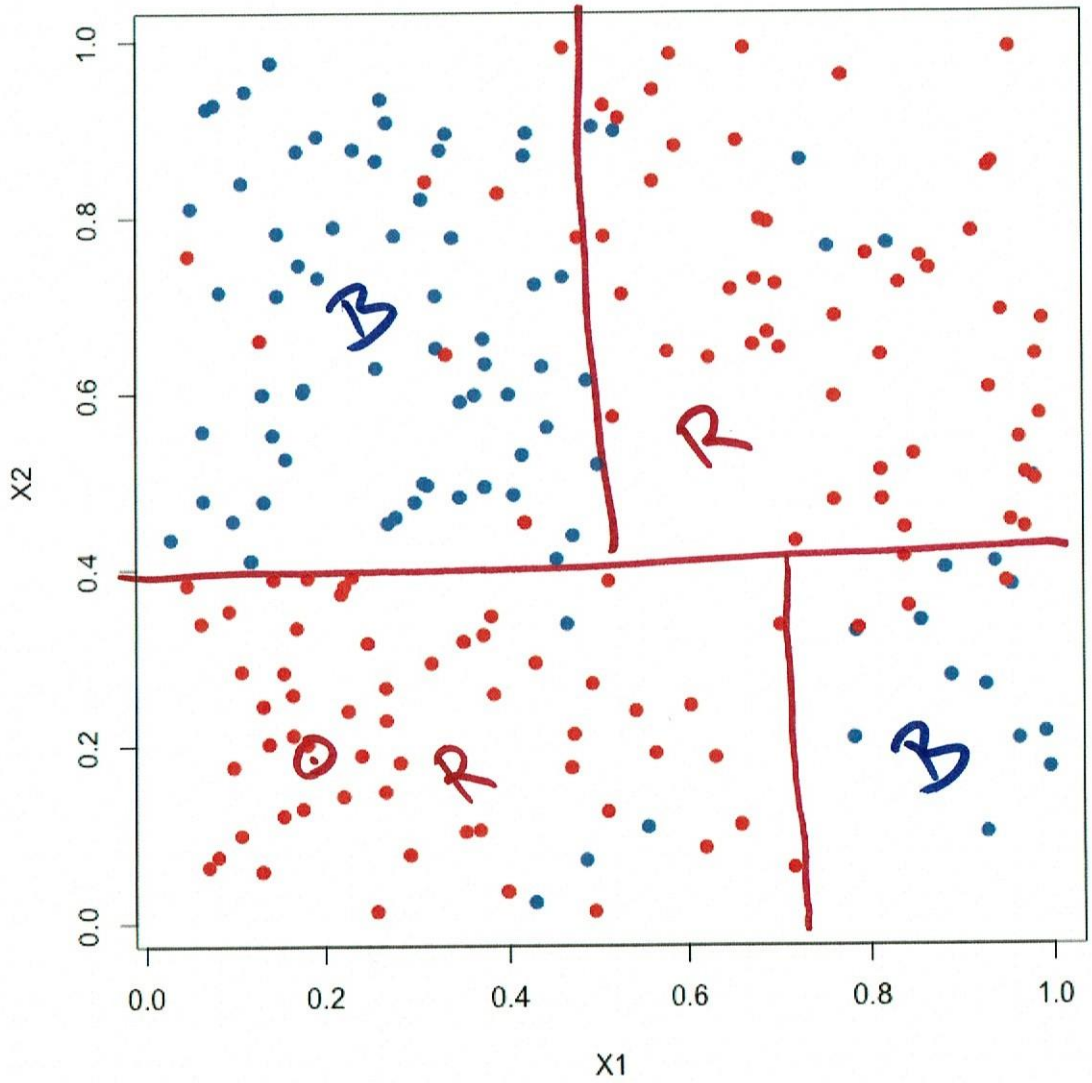
- Classification
  - Spam / not Spam
  - Admit to ICU /not
  - Lend money / deny
  - Intrusion detections

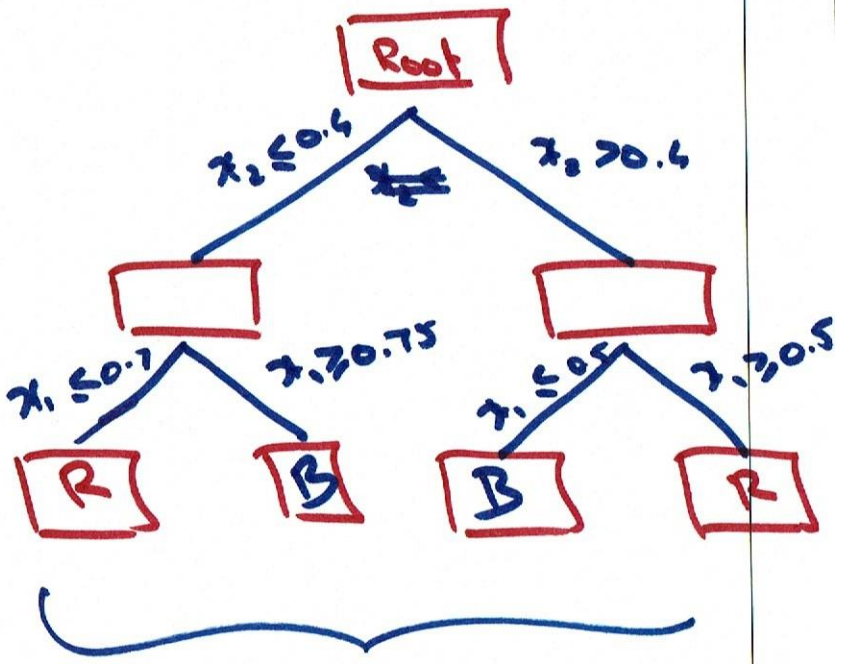


- Regression
  - Predict stock returns
  - Pricing a house or a car
  - Weather predictions (temp, rain fall etc)
  - Economic growth predictions
  - Predicting sports scores

X1	X2	Y
0.268	0.266	64.41
0.219	0.372	28.08
0.517	0.573	95.76
0.269	0.908	15.84
0.181	0.202	41.83
0.519	0.898	25.20
0.563	0.945	9.44
0.129	0.661	82.77

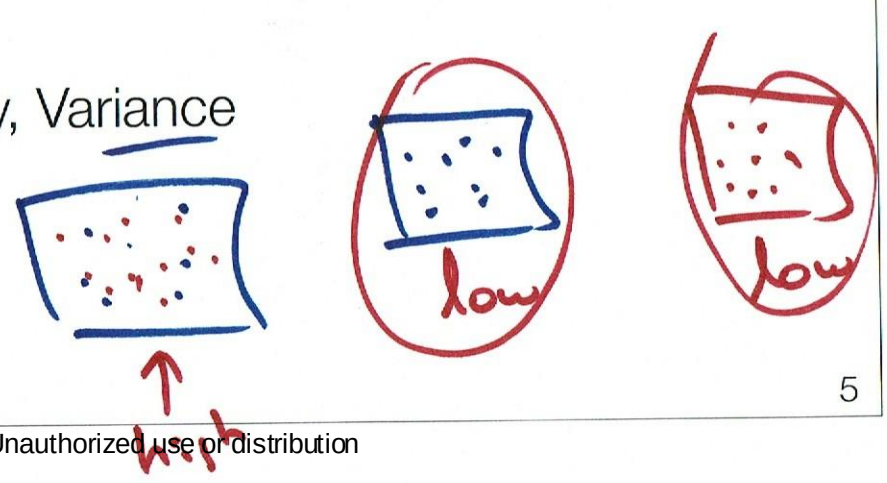
Visualizing Classification as a Tree





Metrics

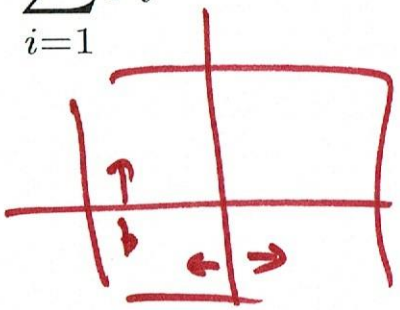
- Algorithms for constructing decision trees usually work top-down, by choosing a variable at each step that best splits the set of items.
- Different algorithms use different metrics for measuring "best"
- These metrics measure how similar a region or a node is. They are said to measure the impurity of a region.
- Larger these impurity metrics the larger the "dissimilarity" of a nodes/regions data.
- Examples: Gini impurity, Entropy, Variance

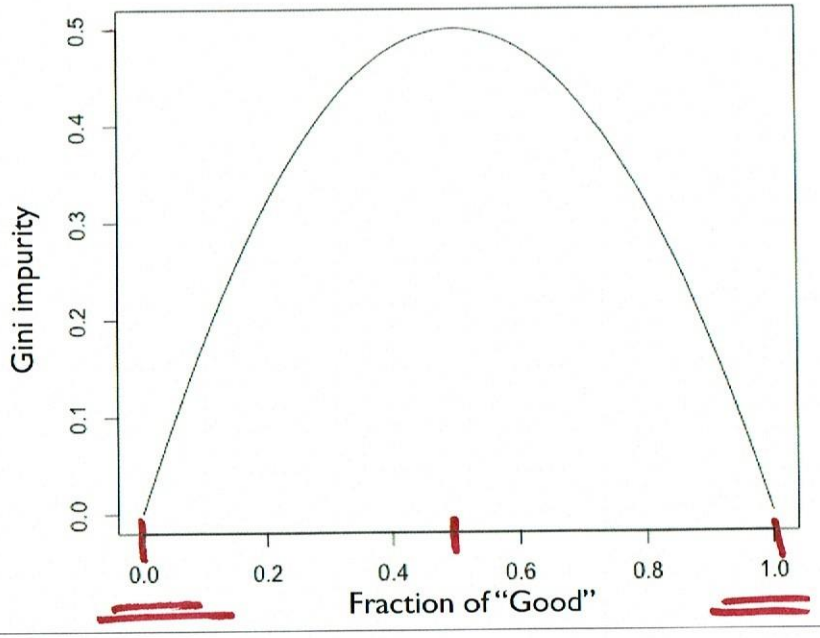


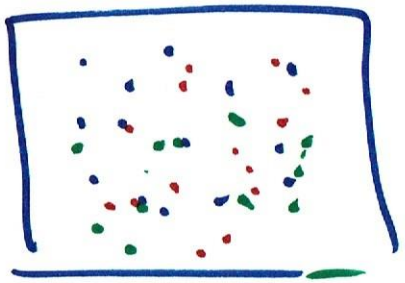
Gini impurity

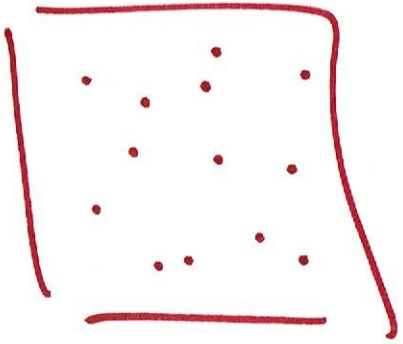
- Used by the CART
- Is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.
- Can be computed by summing the probability of an item with label i being chosen (p_i), times the probability of a mistake ($1 - p_i$) in categorizing that item.
- Simplifying gives, the Gini impurity of a set:

$$1 - \sum_{i=1}^{J} p_i^2$$









$p_1 \rightarrow ① \Rightarrow p_1(1-p_1) \leftarrow p_1p_2 + p_1p_3$

$p_2 \rightarrow ② \Rightarrow p_2(1-p_2) \leftarrow p_2p_3 + p_2p_1$

$p_3 \rightarrow ③ \Rightarrow p_3(1-p_3) \leftarrow p_3p_1 + p_3p_2$

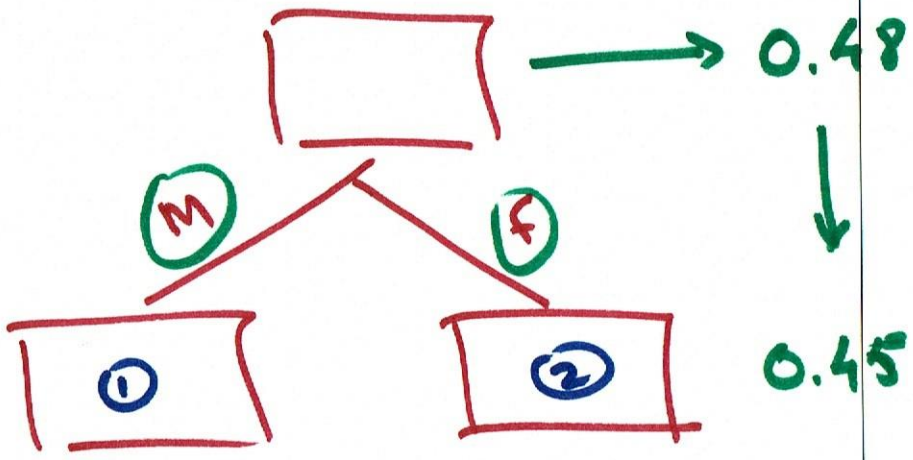
$\Downarrow$

$\sum p_i(1-p_i)$

$\sum p_i - \sum p_i^2 \Rightarrow 1 - \sum p_i^2$

CART: An Example

Cust_ID	Gender	Occupation	Age	Target
1	M	Sal	22	1
2	M	Sal	22	0
3	M	Self-Emp	23	1
4	M	Self-Emp	23	0
5	M	Self-Emp	24	1
6	M	Self-Emp	24	0
7	F	Sal	25	1
8	F	Sal	25	0
9	F	Sal	26	0
10	F	Self-Emp	26	0



Root node : $P_1 = 0.4$ $P_2 = 0.6$

$GI = 1 - (0.1)^2 - (0.6)^2$

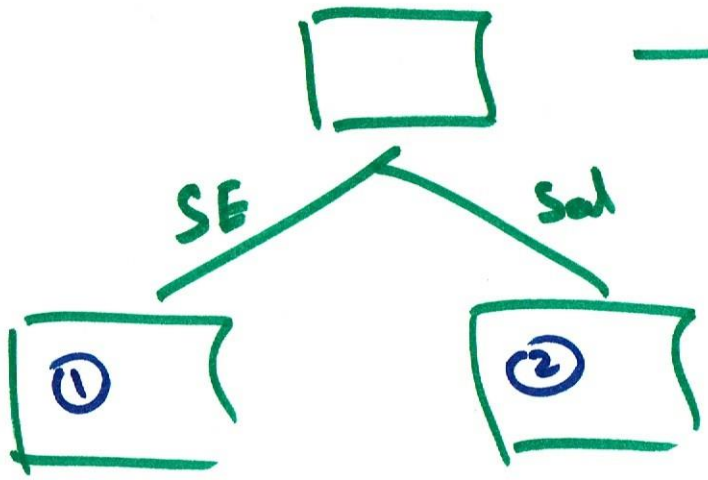
$= 0.48$

$P_1 = 0.5$
$P_2 = 0.5$
$1 - 0.5^2 - 0.5^2$
$= 0.5$

$P_1 = 0.25$
$P_2 = 0.75$
$1 - 0.25^2 - 0.75^2$
$= 0.375$

$GI = \frac{6}{10}(0.5) + \frac{4}{10}(0.375)$

$= 0.45$



0.48

① G.I = $1 - 0.4^2 - 0.6^2$

= 0.48

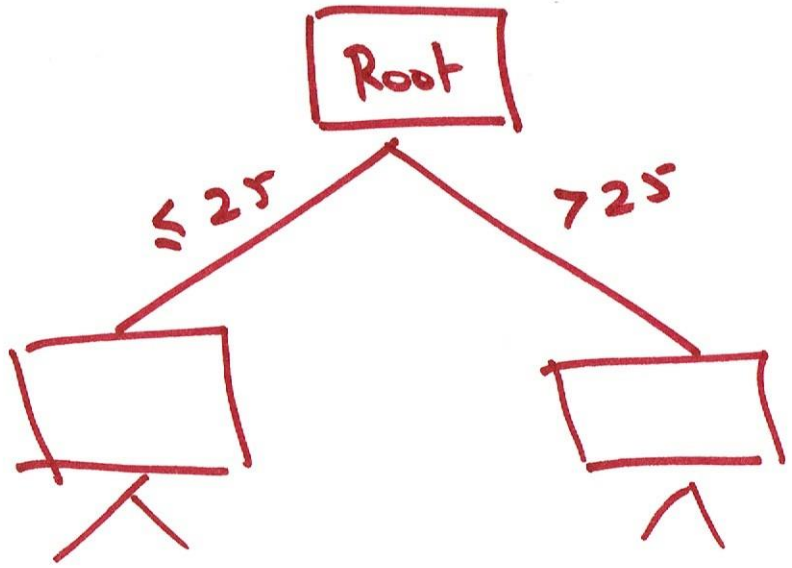
② G.I = $1 - 0.4^2 - 0.6^2$

= 0.48

$$G.I = \frac{5}{10}(0.48) + \frac{5}{10}(0.48) = 0.48$$

	Left	right	Gini Split
≤22, >22	0.5	0.47	0.48
≤23, >23	0.5	0.44	0.47
≤24, >24	0.5	0.38	0.45
≤25, >25	0.5	0	0.40

Jain ⇒ 0.08



0.48

0.40