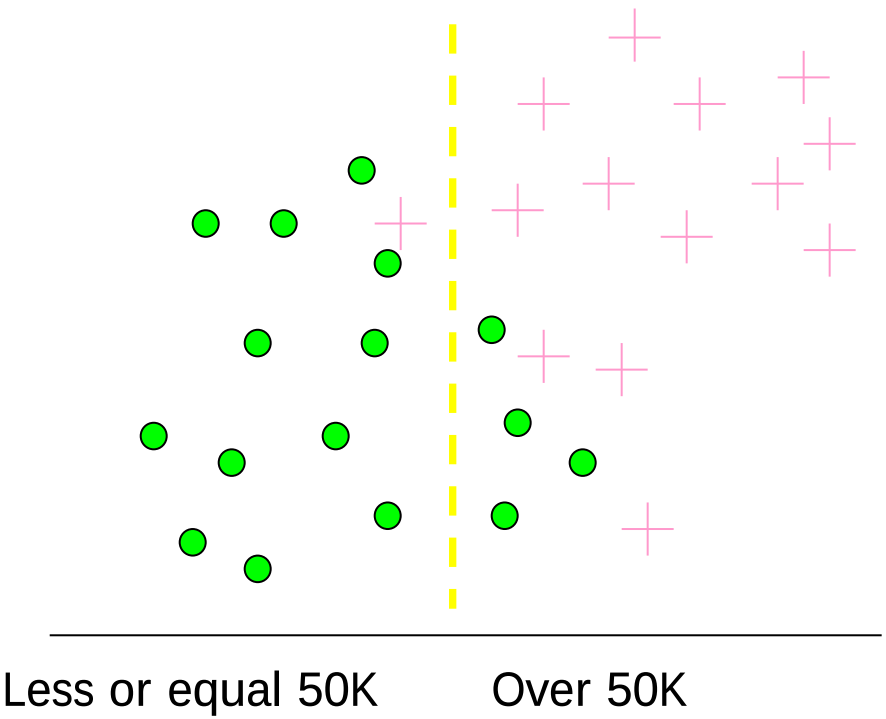
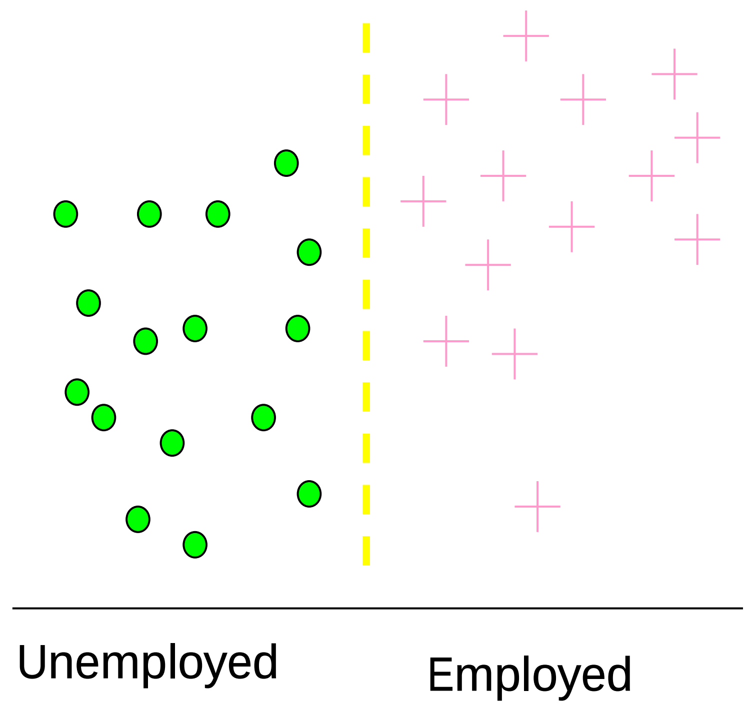
Information Gain

Which test is more informative?

Split over whether Balance exceeds 50K

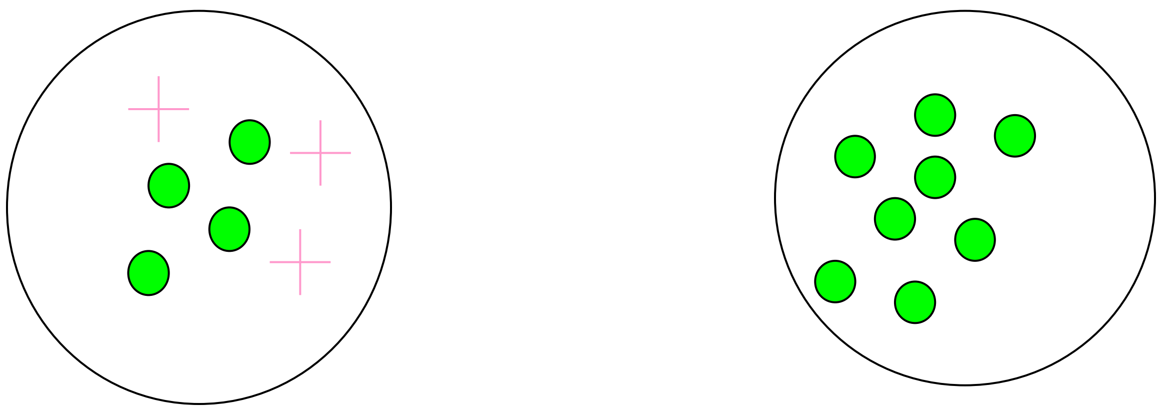


Split over whether applicant is employed



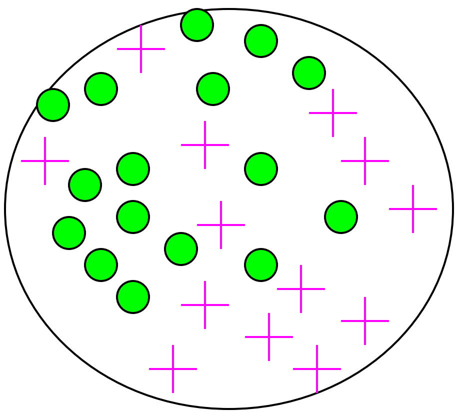
Impurity/Entropy (informal)

- Measures the level of **impurity** in a group of examples

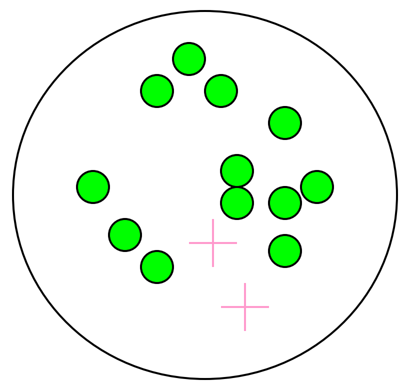


Impurity

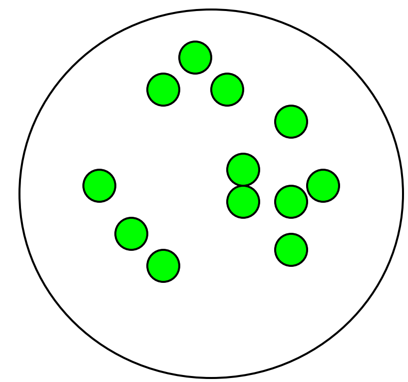
Very impure group



Less impure

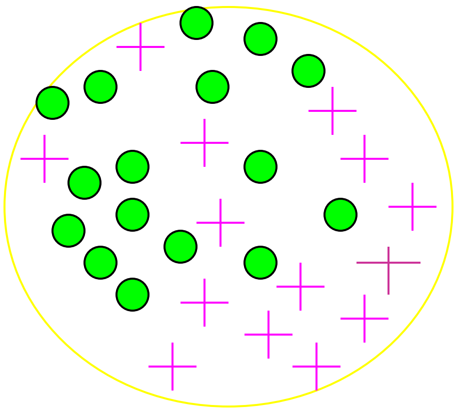


Minimum impurity



Entropy: a common way to measure impurity

- Entropy = $\sum_i - p_i \log_2 p_i$



p_i is the probability of class i

Compute it as the proportion of class i in the set.

16/30 are green circles; 14/30 are pink crosses
$\log_2(16/30) = -.9$; $\log_2(14/30) = -1.1$
Entropy = -(16/30)(-.9) –(14/30)(-1.1) = .99

- Entropy comes from information theory. The higher the entropy the more the information content.

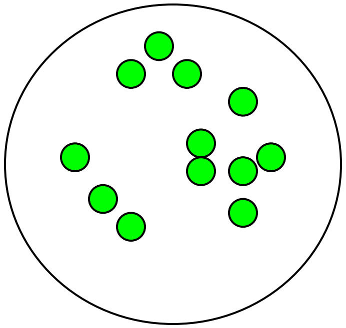
What does that mean for learning from examples?

2-Class Cases:

- What is the entropy of a group in which all examples belong to the same class?
  - entropy = $- 1 \log_2 1 = 0$

not a good training set for learning

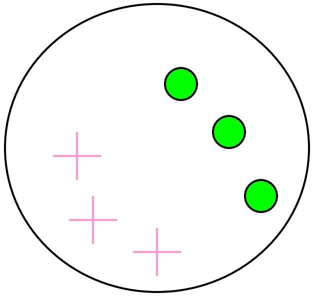
Minimum impurity



- What is the entropy of a group with 50% in either class?
  - entropy = $-0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$

good training set for learning

Maximum impurity



Information Gain

- We want to determine which attribute in a given set of training feature vectors is most useful for discriminating between the classes to be learned.

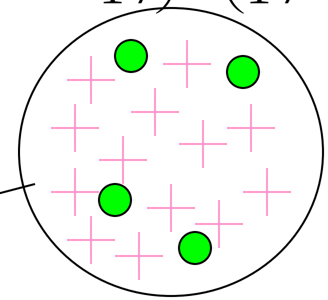
- Information gain tells us how important a given attribute of the feature vectors is.

- We will use it to decide the ordering of attributes in the nodes of a decision tree.

Calculating Information Gain

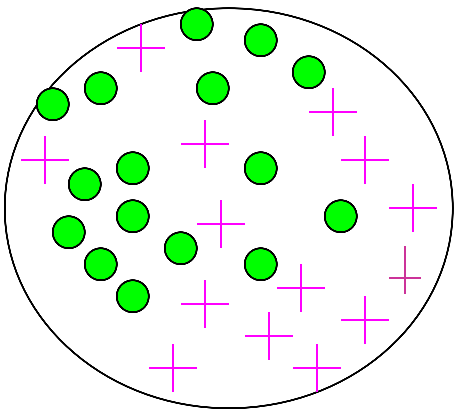
Information Gain = entropy(parent) – [average entropy(children)]

child entropy $-\left(\frac{13}{17}\cdot\log_2\frac{13}{17}\right)-\left(\frac{4}{17}\cdot\log_2\frac{4}{17}\right)=0.787$

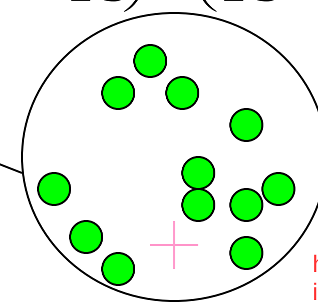


17 instances

Entire population (30 instances)



child entropy $-\left(\frac{1}{13}\cdot\log_2\frac{1}{13}\right)-\left(\frac{12}{13}\cdot\log_2\frac{12}{13}\right)=0.391$



13 instances

hh: only one is different, entropy is very small---> lead to huge information gain

parent entropy $-\left(\frac{14}{30}\cdot\log_2\frac{14}{30}\right)-\left(\frac{16}{30}\cdot\log_2\frac{16}{30}\right)=0.996$

(Weighted) Average Entropy of Children = $\left(\frac{17}{30}\cdot 0.787\right)+\left(\frac{13}{30}\cdot 0.391\right)=0.615$

Information Gain = 0.996 - 0.615 = 0.38 for this split

Entropy-Based Automatic Decision Tree Construction

Training Set S

$x_1=(f_{11},f_{12},\ldots f_{1m})$

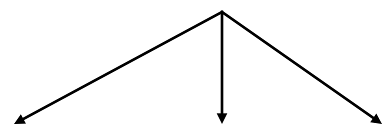
$x_2=(f_{21},f_{22}, \quad f_{2m})$

.

.

$x_n=(f_{n1},f_{22}, \quad f_{2m})$

Node 1
What feature
should be used?



What values?

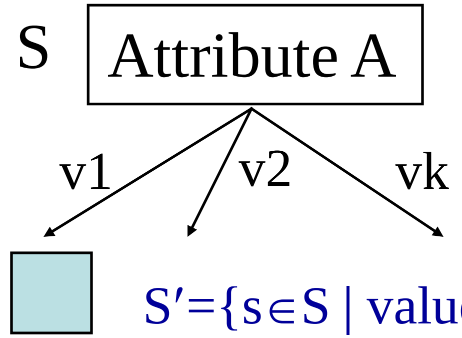
Quinlan suggested information gain in his ID3 system and later the gain ratio, both based on entropy.

Using Information Gain to Construct a Decision Tree

1

Choose the attribute A with highest information gain for the full training set at the root of the tree.

Full Training Set S



2

Construct child nodes for each value of A. Each has an associated subset of vectors in which A has a particular value.

Set S ′

$S' = \{ s \in S \mid value(A) = v1 \}$

3

repeat recursively till when?

Simple Example

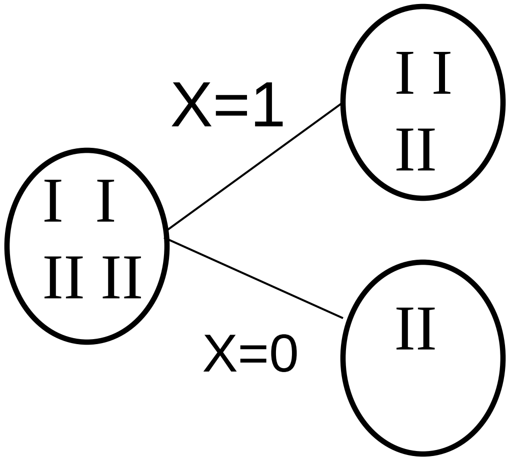
Training Set: 3 features and 2 classes

X	Y	Z	C
1	1	1	I
1	1	0	I
0	0	1	II
1	0	0	II

How would you distinguish class I from class II?

X	Y	Z	C
1	1	1	I
1	1	0	I
0	0	1	II
1	0	0	II

Split on attribute X



If X is the best attribute,
this node would be further split.

$$E_{child1} = -(1/3)\log_2(1/3) - (2/3)\log_2(2/3)$$
$$= .5284 + .39$$
$$= .9184$$

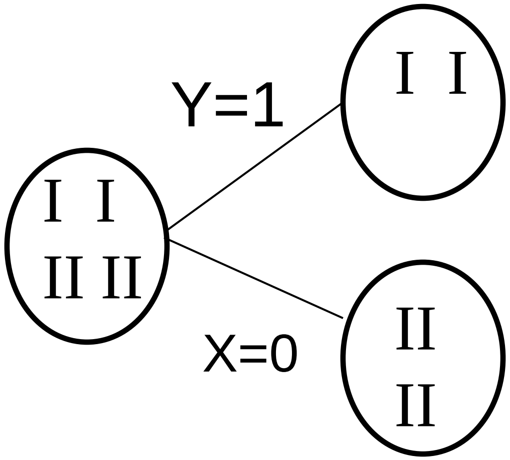
$$E_{child2} = 0$$

$E_{parent} = 1$

GAIN = 1 – (3/4)(.9184) – (1/4)(0) = .3112

X	Y	Z	C
1	1	1	I
1	1	0	I
0	0	1	II
1	0	0	II

Split on attribute Y



$E_{child1} = 0$

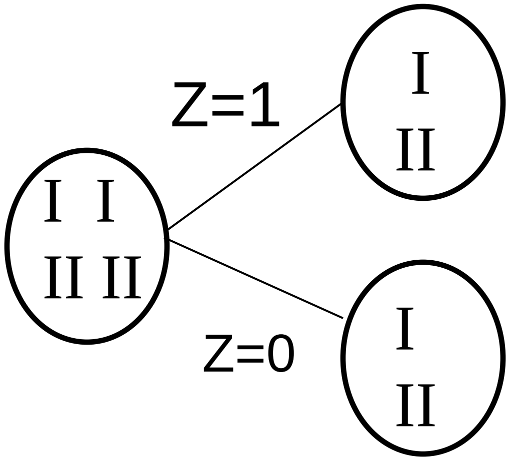
$E_{child2} = 0$

$E_{parent} = 1$

GAIN = 1 –(1/2) 0 – (1/2)0 = 1; BEST ONE

X	Y	Z	C
1	1	1	I
1	1	0	I
0	0	1	II
1	0	0	II

Split on attribute Z



$E_{child1} = 1$

$E_{child2} = 1$

$E_{parent} = 1$

GAIN = 1 – (1/2)(1) – (1/2)(1) = 0   ie. NO GAIN; WORST