C4.5 ALGORITHM (2) WHICH ATTRIBUTE TO CHOOSE AS A NODE

SYRACUSE UNIVERSITY
School of Information Studies

DETERMINE THE BEST ATTRIBUTE FOR SPLITTING

Information gain (IG):

A statistical measure that measures how well a given attribute separates the training examples according to their target classification (Mitchell, 1990)

DETERMINE THE BEST ATTRIBUTE FOR SPLITTING

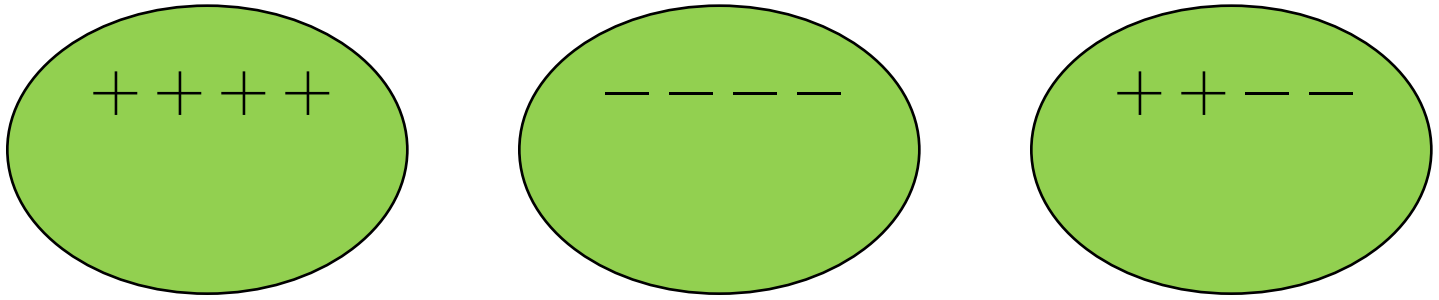
Entropy

To measure the impurity of a data set

Given a collection S, which contains positive (+) and negative (-) examples, p_i is the probability that an example belongs to Class i

Entropy(S) = $- p_+ \log_2 p_+ - p_- \log_2 p_-$

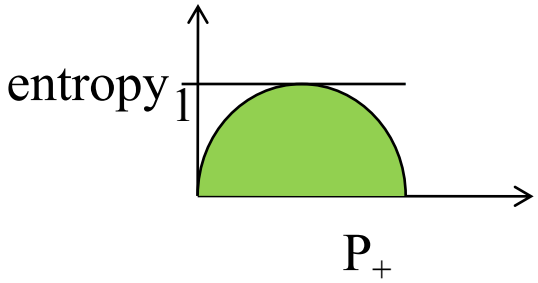
What is the entropy for each of the following collections?



DETERMINE THE BEST ATTRIBUTE FOR SPLITTING

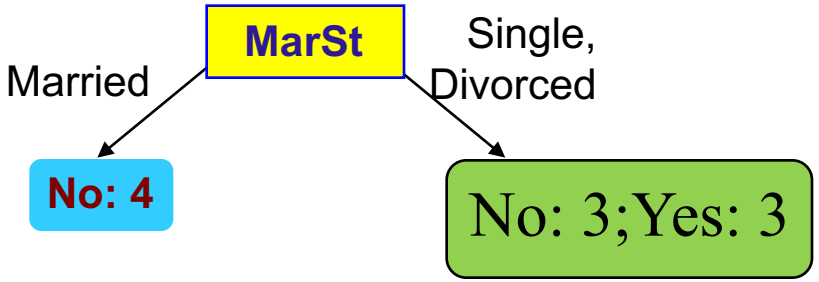
Entropy

- A measure that characterizes the impurity of a collection of examples
- Given a collection S, which contains positive (+) and negative (-) examples, p_i is the probability that an example belongs to Class i
- Entropy(S) = $- p_+ \log_2 p_+ - p_- \log_2 p_-$
- A collection of half-positive examples and half-negative examples
  - Entropy(S) = 1
- A collection of all positive examples or all negative examples
  - Entropy(S) = 0



INFORMATION GAIN: HOW MUCH IMPROVEMENT TOWARD PURITY?

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (class)
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**

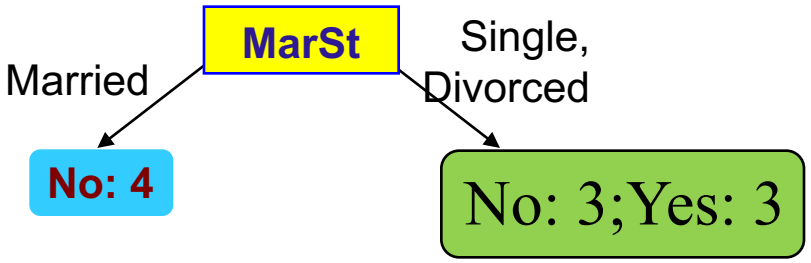


$$Gain(S, A) = Entropy(S) - \sum_{v \ Values(A)} \frac{|S_v|}{|S|} \cdot Entropy(S_v)$$

The expected reduction in entropy caused by knowing the value of attribute A

INFORMATION GAIN: HOW MUCH IMPROVEMENT TOWARD PURITY?

	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
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**



Entropy(S) = -0.7*log2(0.7)-0.3*log2(0.3) = 0.88

Entropy(S1) = 0
Entropy(S2) = 1

IG = 0.88-(0.4*0+0.6*1) = 0.28

Repeat this calculation to find the attribute that provides the highest IG.

WHICH ATTRIBUTE SHOULD BE THE FIRST NODE?

Calculate the information gain (IG) for each attribute; choose the one with the highest IG.

WHAT'S THE NEXT STEP?

Repeat the IG calculation for every subset generated from the last step …

… until all nodes are "pure" with all positive examples or all negative examples; these are all leaf nodes