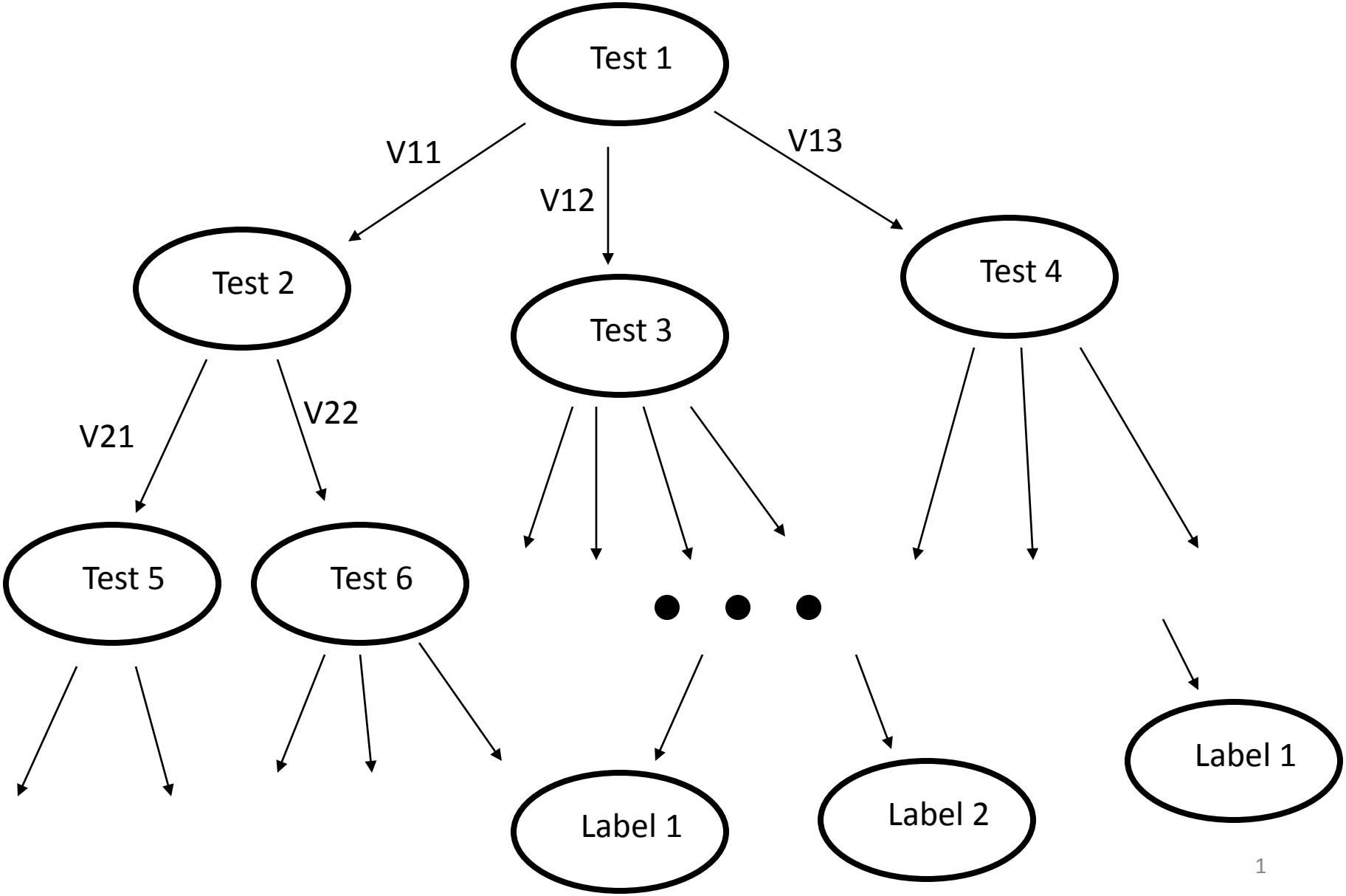
Learning decision trees

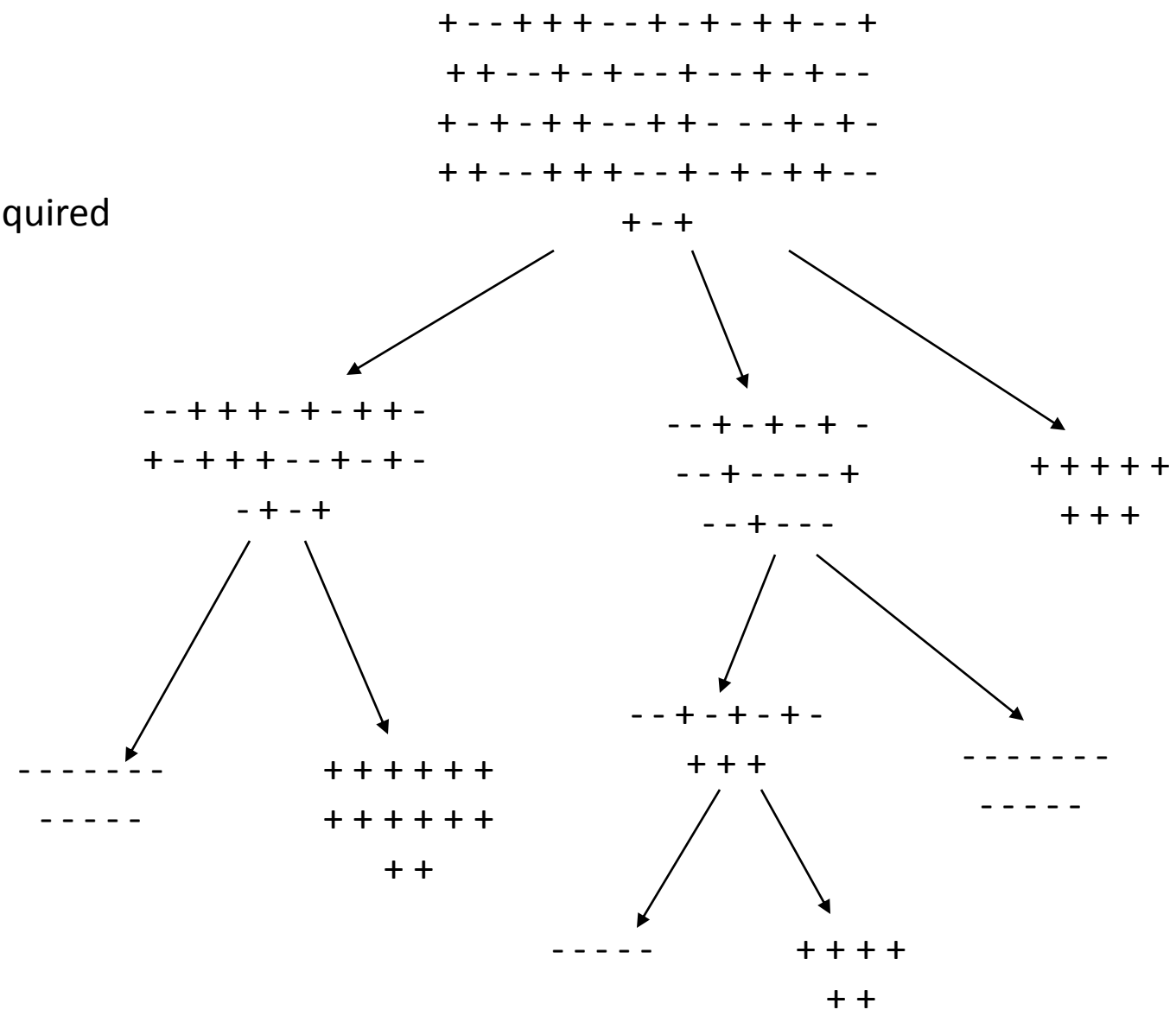


Training Data

Highly Disorganized

High Entropy

Much Information Required



Highly Organized

Low Entropy

Little Information Required

Measuring Information

What is the expected number of bits?

- 16/32 use 1 bit
- 8/32 use 2 bits
- 4 x 2/32 use 4 bits

S_7 =

A A A A A A A A A A A A A A A A	16
B B B B B B B B	8
C C D D E E F F	2 2 2 2

0.5(1) + 0.25(2) + 0.0625(4) +
0.0625(4) + 0.0625(4) + 0.0625(4)

= 0.5 + 0.5 + 0.25 + 0.25 + 0.25 + 0.25

= 2

FOR	SAY
A	1
B	01
C	0000
D	0001
E	0010
F	0011

$$H(S) = \sum_{v \in Labels} -\Pr(v) \cdot \log_2(\Pr(v))$$

Information Gain

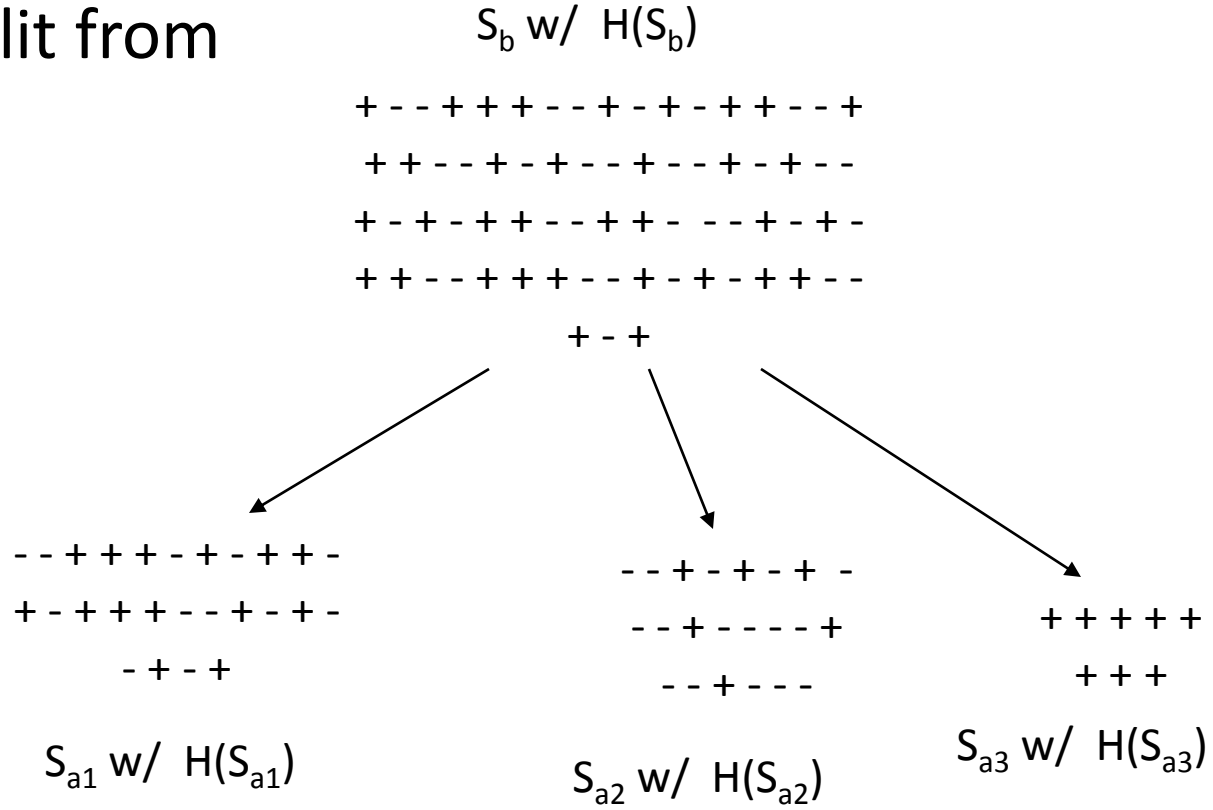
Subtract Information required after split from before

Information required:

Before $H(S_b)$

After $Pr(S_{a1}) \cdot H(S_{a1}) + Pr(S_{a2}) \cdot H(S_{a2}) + Pr(S_{a3}) \cdot H(S_{a3})$

Estimate probabilities using sample counts



Information Gain = $$\mathbf{H}(S_b) - \sum_i \mathbf{H}(S_{ai}) \frac{|S_{ai}|}{|S_b|}$$

Choosing the Most Useful Test

- Estimate information gain for each test

Information Gain = $\mathbf{H}(S_b) - \sum_i \mathbf{H}(S_{ai}) \frac{|S_{ai}|}{|S_b|}$

- Choose the highest

Example

- Restaurant example in text
- Tennis example

Will I Play Tennis?

- Features:
  - Outlook Sun, Overcast, Rain
  - Temp. Hot, Mild, Cool
  - Humidity High, Normal, Low
  - Wind Strong, Weak
  - Label +, -
- Features are evaluated in the morning
- Tennis is played in the afternoon

Training Set

1.	S H H W	-
2.	S H H S	-
3.	O H H W	+
4.	R M H W	+
5.	R C N W	+
6.	R C N S	-
7.	O C N S	+
8.	S M H W	-
9.	S C N W	+
10.	R M N W	+
11.	S M N S	+
12.	O M H S	+
13.	O H N W	+
14.	R M H S	-

Outlook: S, O, R

Temp: H, M, C

Humidity: H, N, L

Wind: S, W

9 + 5 -

$H(9/14) = -9/14 \log_2(9/14) - 5/14 \log_2(5/14)$

≈ 0.94

From N_+,N_- to H(P)

Entropy of a *distribution* H(P)

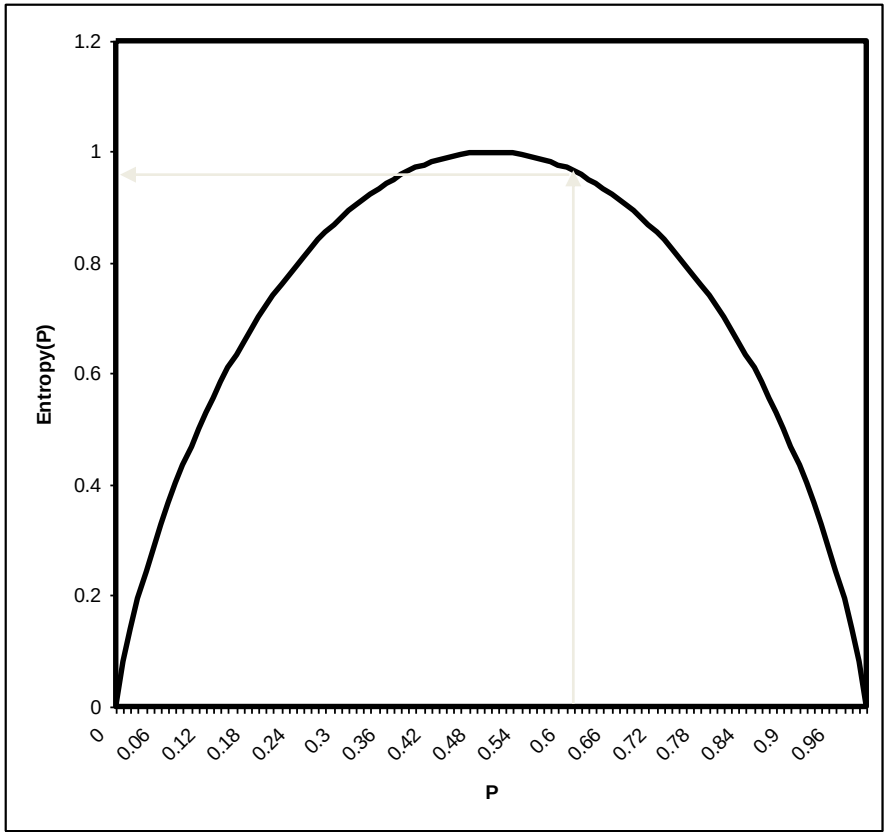
For Binomial:

$P = N_+ / (N_+ + N_-)$

$-P \log_2(P) – (1-P) \log_2(1-P)$

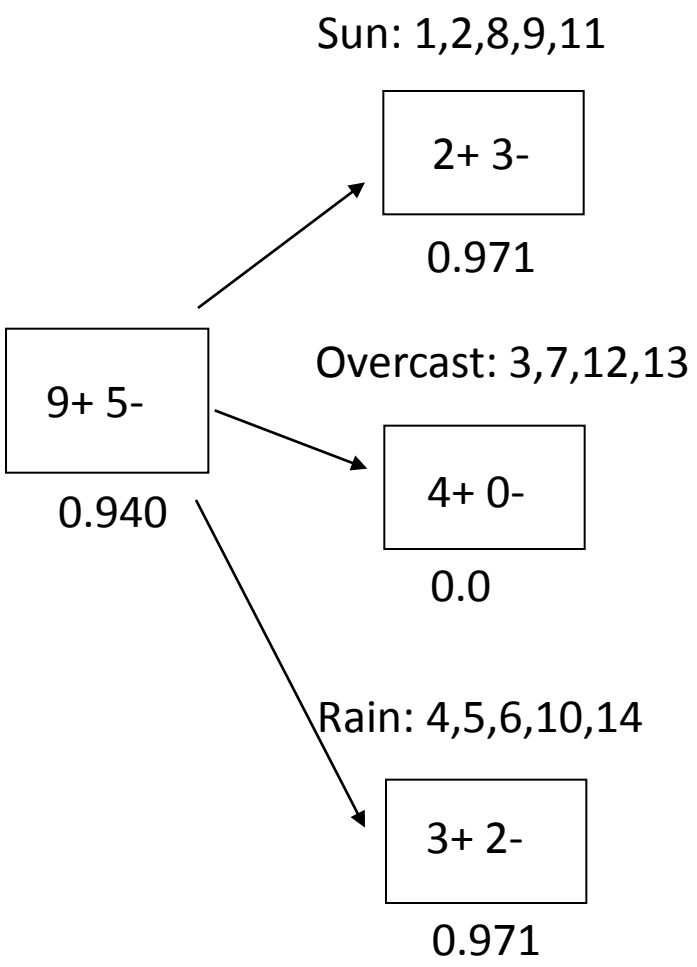
H(9/14) = Entropy(0.64)

0.940



Outlook Gain = 0.246

1.	S H H W	-
2.	S H H S	-
3.	O H H W	+
4.	R M H W	+
5.	R C N W	+
6.	R C N S	-
7.	O C N S	+
8.	S M H W	-
9.	S C N W	+
10.	R M N W	+
11.	S M N S	+
12.	O M H S	+
13.	O H N W	+
14.	R M H S	-



Information After:

0.971 * 5/14 +

0.0 * 4/14 +

0.971 * 5/14

= 0.694

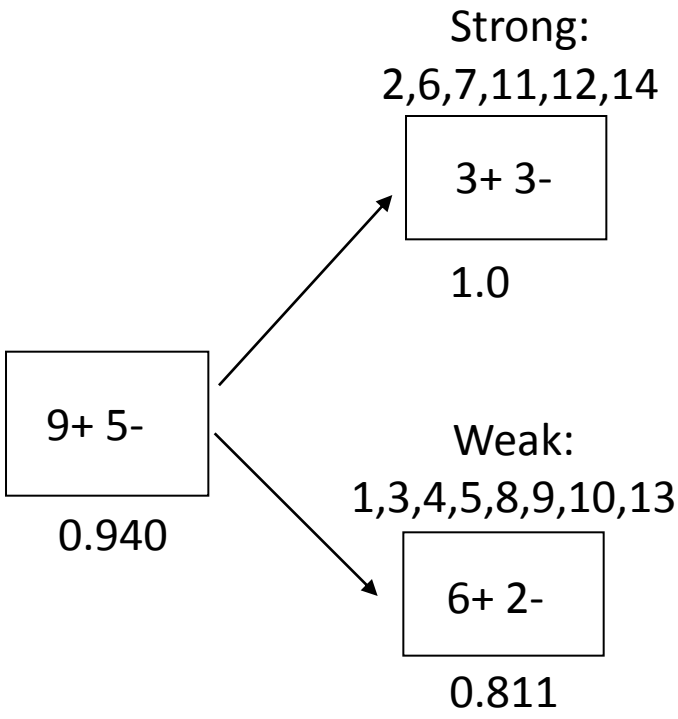
Information Gain:

0.940 – 0.694

=0.246

Wind Gain = 0.048

1.	S H H W	-
2.	S H H S	-
3.	O H H W	+
4.	R M H W	+
5.	R C N W	+
6.	R C N S	-
7.	O C N S	+
8.	S M H W	-
9.	S C N W	+
10.	R M N W	+
11.	S M N S	+
12.	O M H S	+
13.	O H N W	+
14.	R M H S	-



Information After:

1.0 * 6/14 +

0.811 * 8/14

= 0.892

Information Gain:

0.940 – 0.892

=0.048

Thought question: What is H(Strong) ?

Confidence in H(P)

Entropy H(P) for Binomial:

$P = N_+ / (N_+ + N_-)$

$-P \log_2(P) - (1-P) \log_2(1-P)$

H(3/6) = Entropy(0.5)

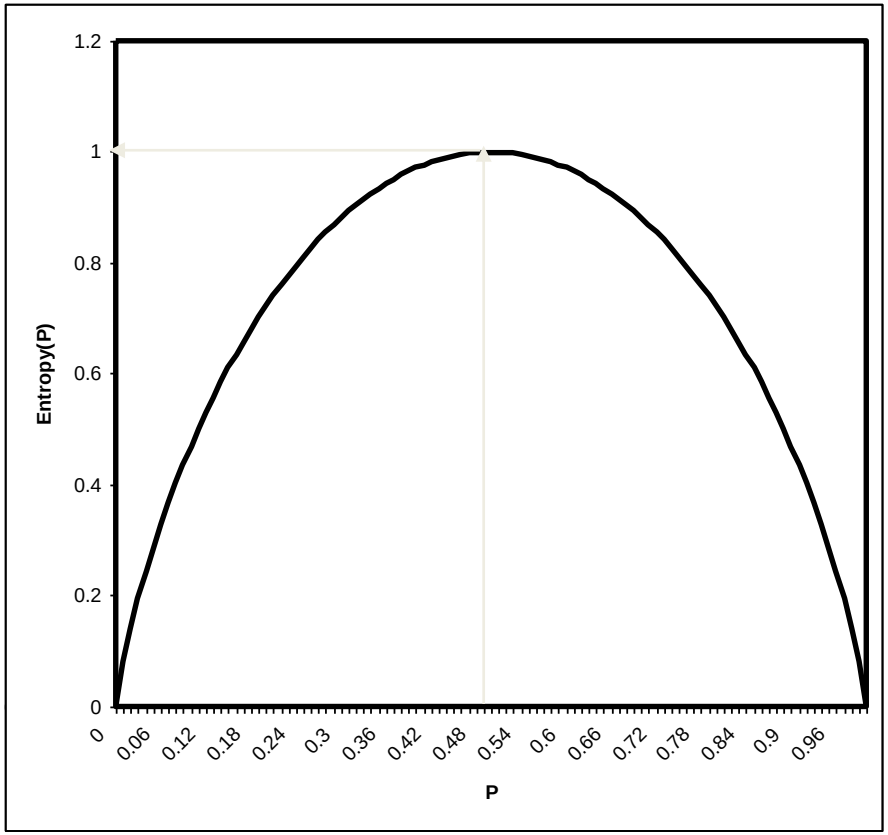
= 1.0000000000000

= H(Strong)

?????

Binomial / Beta conjugate distributi

“Bayesian averaging”



Information Gain

- Outlook 0.25
- Temperature 0.03
- Humidity 0.15
- Wind 0.05

Outlook provides greatest local gain

Split on Outlook

S H H W	-	R C N W	+	S C N W	+	O H N W	+
S H H S	-	R C N S	-	R M N W	+	R M H S	-
O H H W	+	O C N S	+	S M N S	+		
R M H W	+	S M H W	-	O M H S	+		

Sunny

S H H W	-
S H H S	-
S M H W	-
S C N W	+
S M N S	+

Overcast

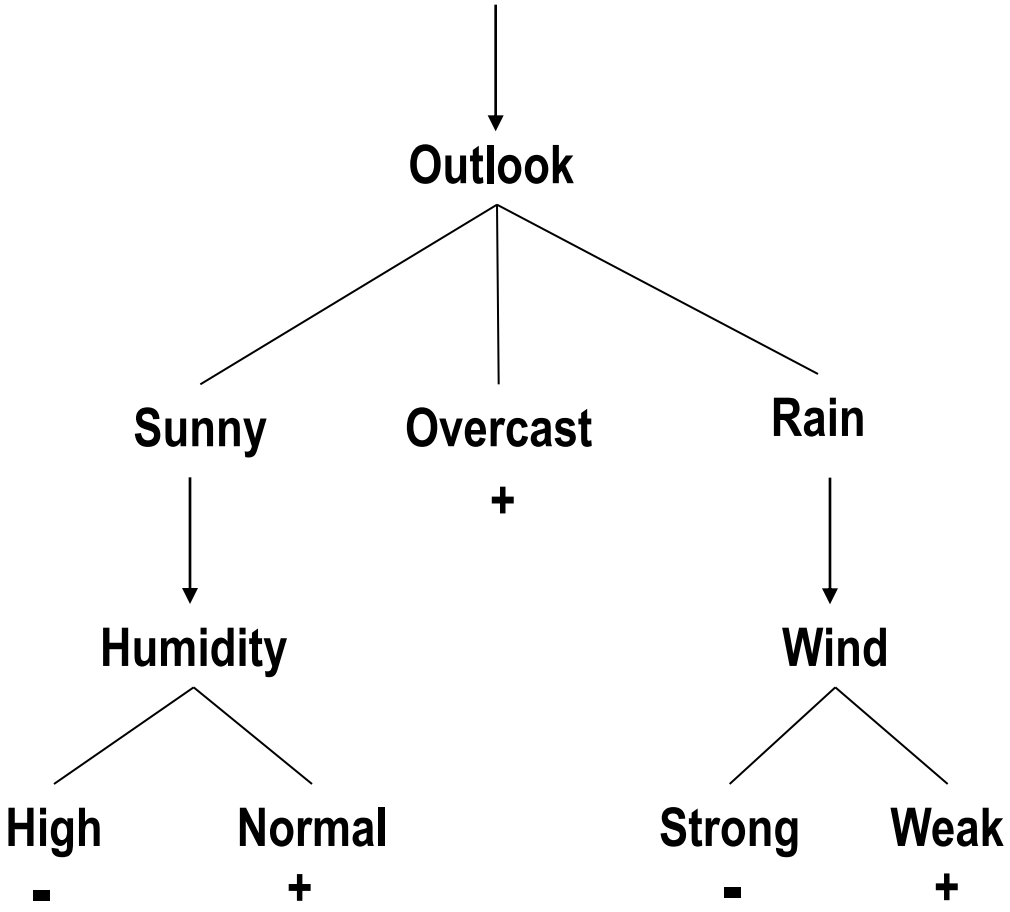
O H H W	+
O C N S	+
O M H S	+
O H N W	+

Rain

R M H W	+
R C N W	+
R C N S	-
R M N W	+
R M H S	-

Now recur on each smaller set

Final Decision Tree



Suppose under Sunny we split on Outlook (again) instead of Humidity?

What can we say about entropy as we measure additional features?

Extension I: Continuous Attributes

- Discretize into ranges
  - big, medium, small
  - Use care and prior understanding

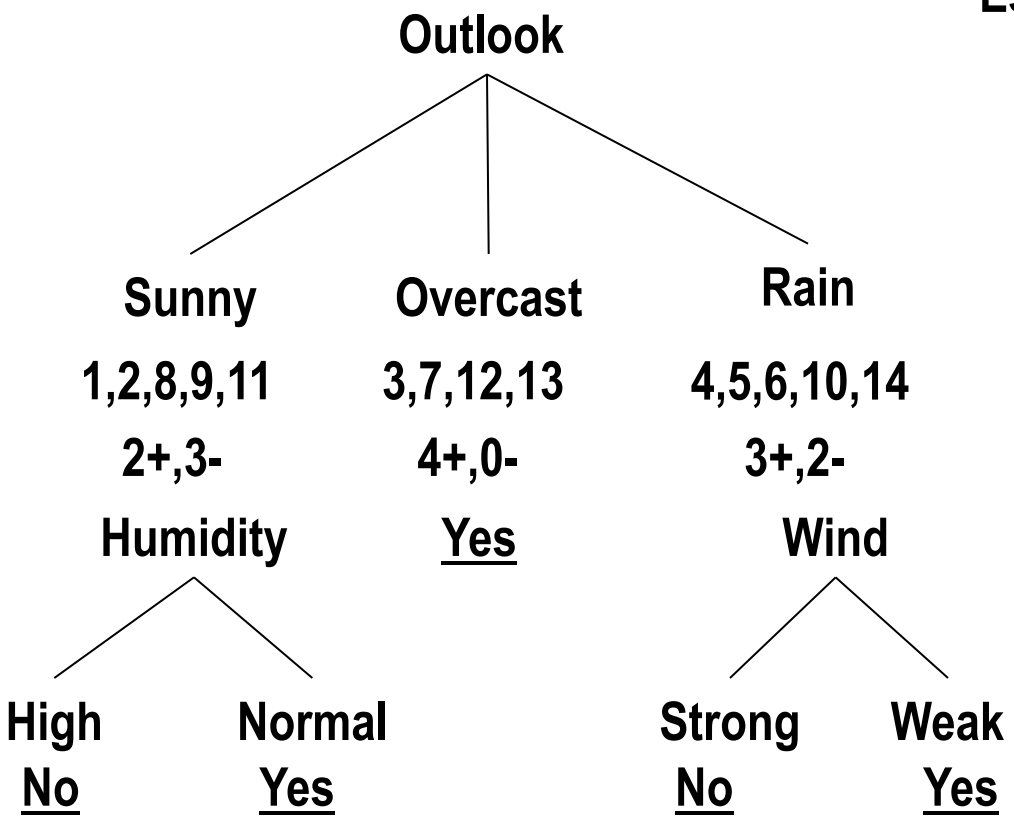
- Split by introducing thresholds
  - Form of test: A < c
  - Partition into A < c and A >= c.
  - Calculate information
  - How to find the split point with the highest gain?
    - For each continuous feature A:
    - Sort examples by A
    - Evaluate each mid-point as a possible threshold
    - Real parameter but finite interesting distinctions
    - Thought question: why use the mid-point?

Extension II: Missing Attributes

- E.g., Medical tests not yet run
- Training:
  - Information Gain splitting on attribute ‘a’
  - In some of the examples ‘a’ is not given
- Testing:
  - Classify an example x
  - But x is missing the value of ‘a’

Missing Attributes

Outlook = ???, Temp = Hot, Humidity = Normal, Wind = Strong, label = ??



Estimate Outlook:

Blend by labels

1/3 Yes + 1/3 Yes +1/3 No = Yes

Blend by estimated probability / counts (both test & label)

Does the Learning Algorithm

- Always halt?
- Yield an optimal tree?
- Yield a “good” tree?

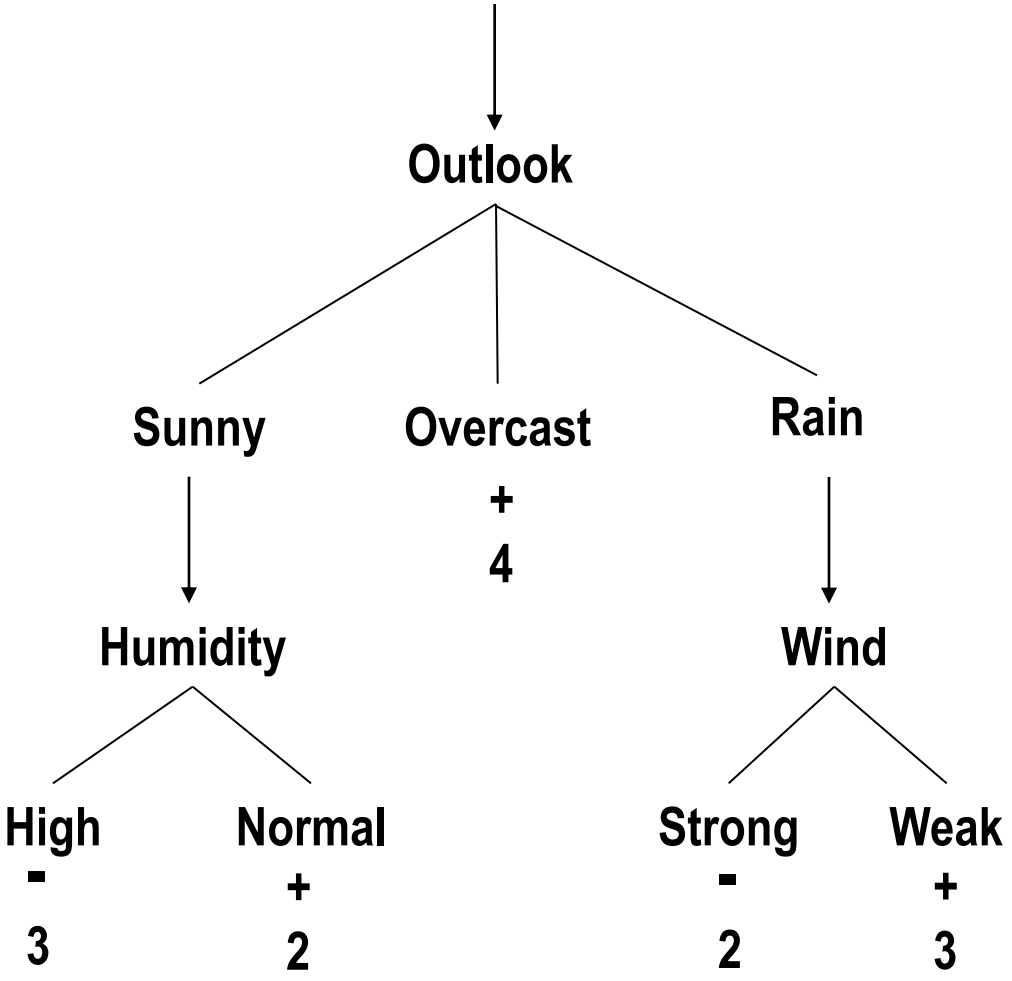
What does “good” mean?

It means performing well on the training set

NO!

It means performing well on future yet-unseen examples drawn from the same distribution

Recall Final Decision Tree



Thought Questions

Are these *actual world examples* or did someone make them up?

How could we tell?

Instead of 14 training examples, suppose we had 14,000, same pattern, same tree

Is it possible that this is the *wrong* tree?
(what would that mean?)

Overfitting
(very important & general phenomenon!)

- Concept performs
  - well on training data (drawn from X according to $\mathcal{D}$)
  - poorly on unseen examples of interest (same drawing)
- Excess flexibility in hypothesis space H
  - Finds training set pattern not in population
  - Concept selection from too little (insignificant) training data
- Confidence in selecting $c \in H$
  - Diversity of H reflects our ignorance
  - Training set Z provides information
  - Information(Z) $\geq$ Information need(H) [>> for high confidence]
- Often Learning algorithms cannot tell
- With low confidence, expected behavior of "best" concept (on Z) is poor on underlying population
- Extreme:
  - Rote learning of training data
  - No generalization