10-601 Introduction to Machine Learning

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Overfitting

+

k-Nearest Neighbors

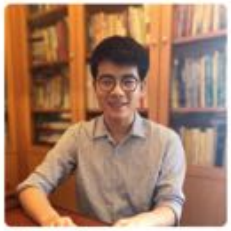
Matt Gormley
Lecture 4
Jan. 27, 2020

Course Staff

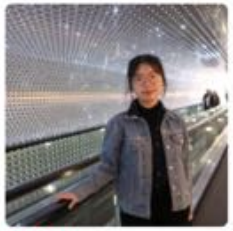

David Xu


Kelly Shi


Ayushi Sood

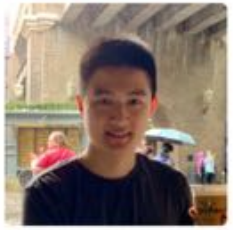

Mike Chen


Eu Jing, Chua


Zola Liu


Sankalp Patro

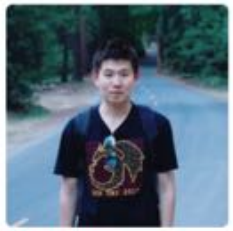

Filipp Shelobolin


Scott Liu

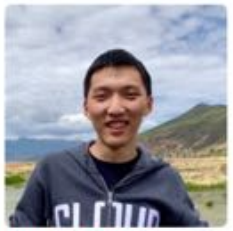

Kyle Chin


Vishal Baskar


Yufei Wang


Shaotong


Quentin Cheng


Hongyi Zhang


Yiming Wen


Sana Lakdawala


Ani Chowdhury


Matt Gormley


Brynn Edmunds


Vinay Kadi


Hanyue Chai

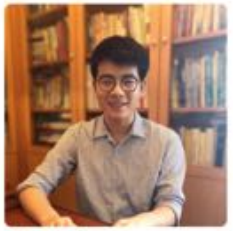

Zee Almusa

Course Staff

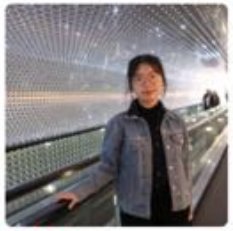

David Xu

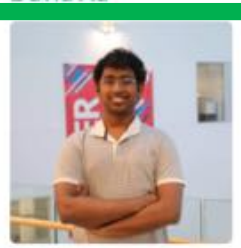

Kelly Shi


Ayushi Sood

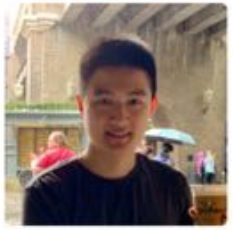

Mike Chen


Eu Jing, Chua


Zola Liu


Sankalp Patro


Filipp Shelobolin


Scott Liu

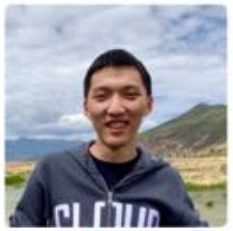

Kyle Chin


Vishal Baskar


Yufei Wang


Shaotong


Quentin Cheng


Hongyi Zhang

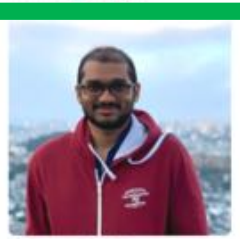

Yiming Wen


Sana Lakdawala


Ani Chowdhury


Matt Gormley


Brynn Edmunds


Vinay Kadi


Hanyue Chai

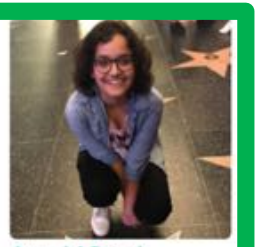

Zee Almusa

Team A

Course Staff

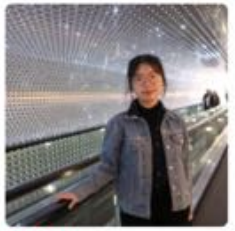

David Xu


Kelly Shi


Ayushi Sood


Mike Chen

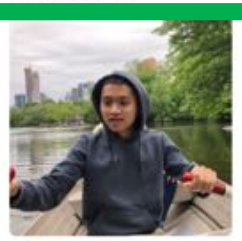

Eu Jing, Chua


Zola Liu


Sankalp Patro

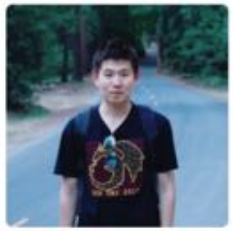

Filipp Shelobolin


Scott Liu

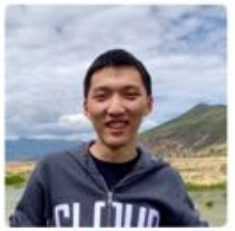

Kyle Chin


Vishal Baskar


Yufei Wang


Shaotong


Quentin Cheng


Hongyi Zhang


Yiming Wen


Sana Lakdawala


Ani Chowdhury


Matt Gormley


Brynn Edmunds


Vinay Kadi

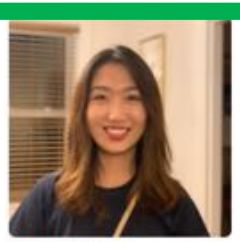

Hanyue Chai


Zee Almusa

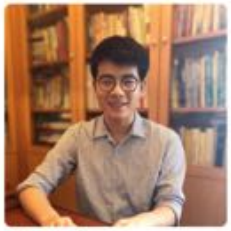
Team B

Course Staff

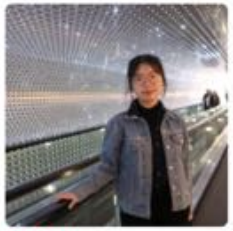

David Xu


Kelly Shi


Ayushi Sood


Mike Chen


Eu Jing, Chua


Zola Liu


Sankalp Patro

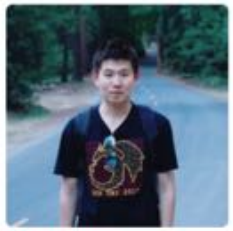

Filipp Shelobolin


Scott Liu


Kyle Chin

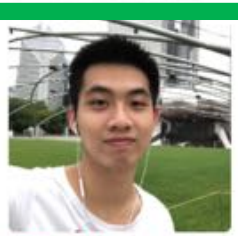

Vishal Baskar


Yufei Wang


Shaotong


Quentin Cheng


Hongyi Zhang


Yiming Wen


Sana Lakdawala

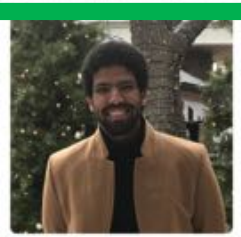

Ani Chowdhury


Matt Gormley


Brynn Edmunds


Vinay Kadi


Hanyue Chai


Zee Almusa

Team C

Course Staff


David Xu


Kelly Shi


Ayushi Sood

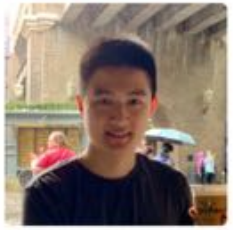

Mike Chen

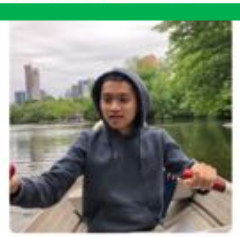

Eu Jing, Chua


Zola Liu


Sankalp Patro


Filipp Shelobolin


Scott Liu


Kyle Chin


Vishal Baskar


Yufei Wang


Shaotong


Quentin Cheng

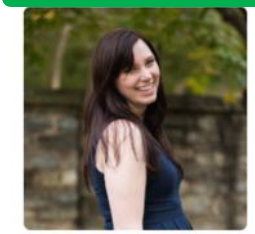

Hongyi Zhang


Yiming Wen


Sana Lakdawala


Ani Chowdhury


Matt Gormley


Brynn Edmunds


Vinay Kadi


Hanyue Chai


Zee Almusa

Team D

Course Staff

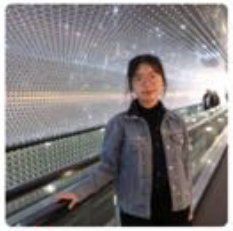

David Xu


Kelly Shi


Ayushi Sood

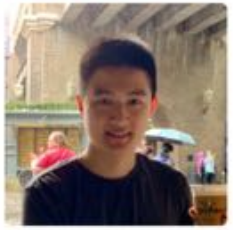

Mike Chen


Eu Jing, Chua


Zola Liu


Sankalp Patro

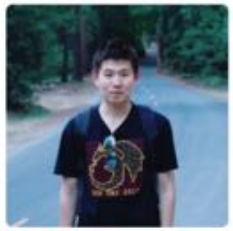

Filipp Shelobolin


Scott Liu

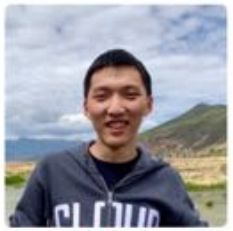

Kyle Chin


Vishal Baskar


Yufei Wang


Shaotong


Quentin Cheng


Hongyi Zhang


Yiming Wen


Sana Lakdawala


Ani Chowdhury


Matt Gormley


Brynn Edmunds


Vinay Kadi


Hanyue Chai


Zee Almusa

Q&A

Q: When and how do we decide to stop growing trees? What if the set of values an attribute could take was really large or even infinite?

A: We'll address this question for discrete attributes today. If an attribute is real-valued, there's a clever trick that only considers O(L) splits where L = # of values the attribute takes in the training set. Can you guess what it does?

Reminders

- **Homework 2: Decision Trees**
  - **Out: Wed, Jan. 22**
  - **Due: Wed, Feb. 05 at 11:59pm**
- **Required Readings:**
  - **10601 Notation Crib Sheet**
  - **Command Line and File I/O Tutorial (check out our colab.google.com template!)**

SPLITTING CRITERIA FOR DECISION TREES

Decision Tree Learning

- *Definition*: a **splitting criterion** is a function that measures the effectiveness of splitting on a particular attribute
- Our decision tree learner **selects the "best" attribute** as the one that maximizes the splitting criterion
- Lots of options for a splitting criterion:
  - error rate (or *accuracy* if we want to pick the tree that *maximizes* the criterion)
  - Gini gain
  - Mutual information
  - random
  - …

Decision Tree Learning Example

Dataset:

Output Y, Attributes A and B

Y	A	B
-	1	0
-	1	0
+	1	0
+	1	0
+	1	1
+	1	1
+	1	1
+	1	1

In-Class Exercise

Which attribute would **error rate** select for the next split?

1. A
2. B
3. A or B (tie)
4. Neither

Decision Tree Learning Example

Dataset:

Output Y, Attributes A and B

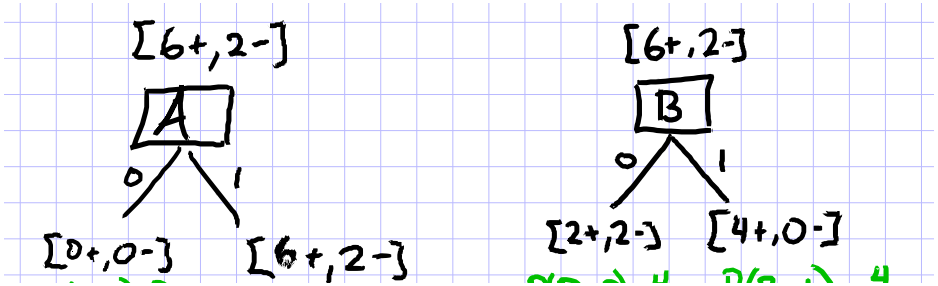
Y	A	B
-	1	0
-	1	0
+	1	0
+	1	0
+	1	1
+	1	1
+	1	1
+	1	1

Decision Tree Learning Example

Dataset:

Output Y, Attributes A and B

Y	A	B
-	1	0
-	1	0
+	1	0
+	1	0
+	1	1
+	1	1
+	1	1
+	1	1



Misclass. Rate

$r(A) = 2/8$

$r(A) = 2/8$

r(·) treats attributes as equally good

Gini Impurity

Chalkboard

- Expected Misclassification Rate:
  - Predicting a Weighted Coin with another Weighted Coin
  - Predicting a Weighted Dice Roll with another Weighted Dice Roll
- Gini Impurity
- Gini Impurity of a Bernoulli random variable
- Gini Gain as a splitting criterion

Decision Tree Learning Example

Dataset:

Output Y, Attributes A and B

Y	A	B
-	1	0
-	1	0
+	1	0
+	1	0
+	1	1
+	1	1
+	1	1
+	1	1

In-Class Exercise

Which attribute would **Gini gain** select for the next split?

1. A
2. B
3. A or B (tie)
4. Neither

Decision Tree Learning Example

Dataset:

Output Y, Attributes A and B

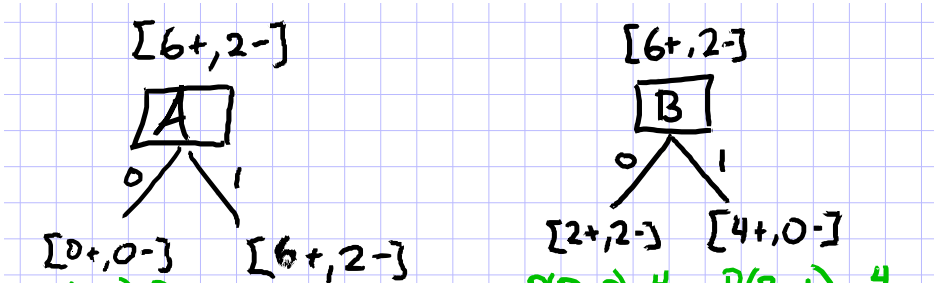
Y	A	B
-	1	0
-	1	0
+	1	0
+	1	0
+	1	1
+	1	1
+	1	1
+	1	1

Decision Tree Learning Example

Dataset:

Output Y, Attributes A and B

Y	A	B
-	1	0
-	1	0
+	1	0
+	1	0
+	1	1
+	1	1
+	1	1
+	1	1



1) $G(Y) = 1 - (6/8)^2 - (2/8)^2 = 0.375$

2) $P(A=1) = 8/8 = 1$
3) $P(A=0) = 0/8 = 0$
4) $G(Y \mid A=1) = G(Y)$
5) $G(Y \mid A=0) = \text{undef}$
6) $\text{GiniGain}(Y \mid A) =$
   $0.375 - 0(\text{undef}) - 1(0.375) = 0$

7) $P(B=1) = 4/8 = 0.5$
8) $P(B=0) = 4/8 = 0.5$
9) $G(Y \mid B=1) = 1 - (4/4)^2 - (0/4)^2 = 0$
10) $G(Y \mid B=0) = 1 - (2/4)^2 - (2/4)^2 = 0.5$
11) $\text{GiniGain}(Y \mid B) =$
   $0.375 - 0.5(0) - 0.5(0.5) = 0.125$

Mutual Information

Let X be a random variable with $X \in \mathcal{X}$.
Let Y be a random variable with $Y \in \mathcal{Y}$.

Entropy: $H(Y) = -\sum_{y \in \mathcal{Y}} P(Y = y) \log_2 P(Y = y)$

Specific Conditional Entropy: $H(Y \mid X = x) = -\sum_{y \in \mathcal{Y}} P(Y = y \mid X = x) \log_2 P(Y = y \mid X = x)$

Conditional Entropy: $H(Y \mid X) = \sum_{x \in \mathcal{X}} P(X = x) H(Y \mid X = x)$

Mutual Information: $I(Y; X) = H(Y) - H(Y|X) = H(X) - H(X|Y)$

- For a decision tree, we can use **mutual information** of the output class *Y* and some attribute *X* on which to split **as a splitting criterion**
- Given a dataset *D* of training examples, we can estimate the required probabilities as…

$$P(Y = y) = N_{Y=y}/N$$
$$P(X = x) = N_{X=x}/N$$
$$P(Y = y|X = x) = N_{Y=y,X=x}/N_{X=x}$$

where $N_{Y=y}$ is the number of examples for which $Y = y$ and so on.

Mutual Information

Let X be a random variable with $X \in \mathcal{X}$.
Let Y be a random variable with $Y \in \mathcal{Y}$.

Entropy: $H(Y) = -\sum_{y \in \mathcal{Y}} P(Y = y) \log_2 P(Y = y)$

Specific Conditional Entropy: $H(Y \mid X = x) = -\sum_{y \in \mathcal{Y}} P(Y = y \mid X = x) \log_2 P(Y = y \mid X = x)$

Conditional Entropy: $H(Y \mid X) = \sum_{x \in \mathcal{X}} P(X = x) H(Y \mid X = x)$

Mutual Information: $I(Y; X) = H(Y) - H(Y|X) = H(X) - H(X|Y)$

- **Entropy** measures the **expected # of bits** to code one random draw from X**.**
- For a decision tree, we want to **reduce the entropy of the random variable we are trying to predict**!

Conditional entropy is the expected value of specific conditional entropy
$E_{P(X=x)}[H(Y \mid X = x)]$

Informally, we say that **mutual information** is a measure of the following:
If we know X, how much does this reduce our uncertainty about Y?

Decision Tree Learning Example

Dataset:

Output Y, Attributes A and B

Y	A	B
-	1	0
-	1	0
+	1	0
+	1	0
+	1	1
+	1	1
+	1	1
+	1	1

In-Class Exercise

Which attribute would **mutual information** select for the next split?

1. A
2. B
3. A or B (tie)
4. Neither

Decision Tree Learning Example

Dataset:

Output Y, Attributes A and B

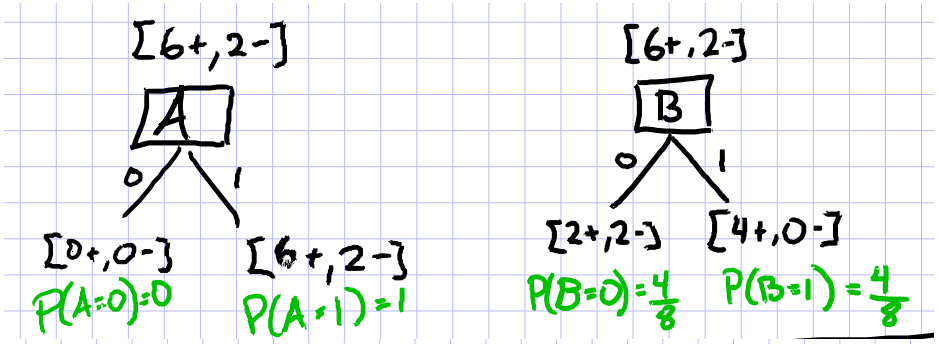
Y	A	B
-	1	0
-	1	0
+	1	0
+	1	0
+	1	1
+	1	1
+	1	1
+	1	1

Decision Tree Learning Example

Dataset:

Output Y, Attributes A and B

Y	A	B
-	1	0
-	1	0
+	1	0
+	1	0
+	1	1
+	1	1
+	1	1
+	1	1



Info Gain

$H(Y) = -\frac{2}{8}\log\frac{2}{8} - \frac{6}{8}\log\frac{6}{8}$

A:

- $H(Y|A=0)$ = "undef"
- $H(Y|A=1) = -\frac{2}{8}\log\frac{2}{8} - \frac{6}{8}\log\frac{6}{8} = H(Y)$
- $H(Y|A) = P(A=0)H(Y|A=0) + P(A=1)H(Y|A=1)$
  $= 0 + H(Y|A=1) = H(Y)$
- $I(Y;A) = H(Y) - H(Y|B=1) = 0$

B:

- $H(Y|B=0) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}$
- $H(Y|B=1) = -0\log 0 - 1\log 1 = 0$
- $H(Y|B) = \frac{4}{8}(0) + \frac{4}{8}(H(Y|B=0))$
- $I(Y;B) = H(Y) - \frac{4}{8}H(Y|B=0) > 0$

Tennis Example

Test your understanding

Dataset:

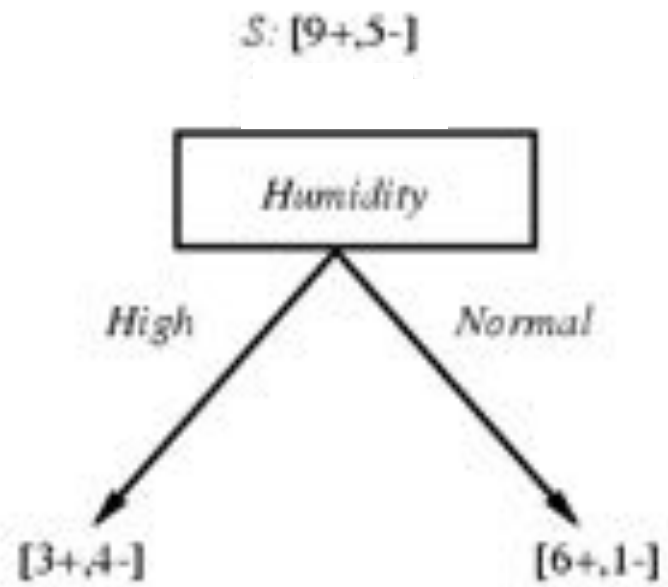
Day	Outlook	Temperature	Humidity	Wind	PlayTennis?
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Figure from Tom Mitchell

Tennis Example

Test your understanding

Which attribute yields the best classifier?



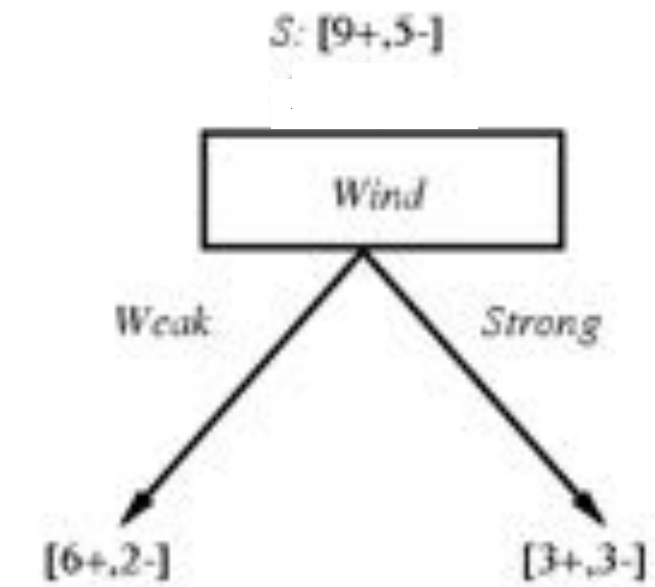
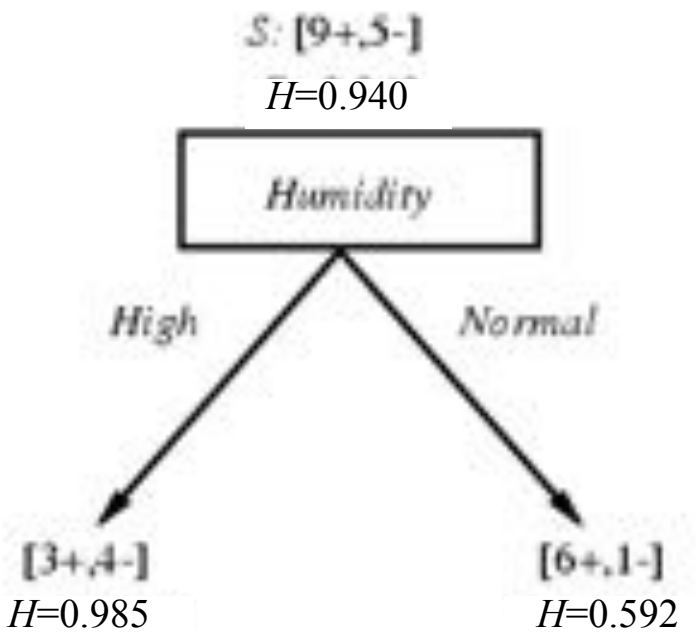


Figure from Tom Mitchell

Tennis Example

Test your understanding

Which attribute yields the best classifier?



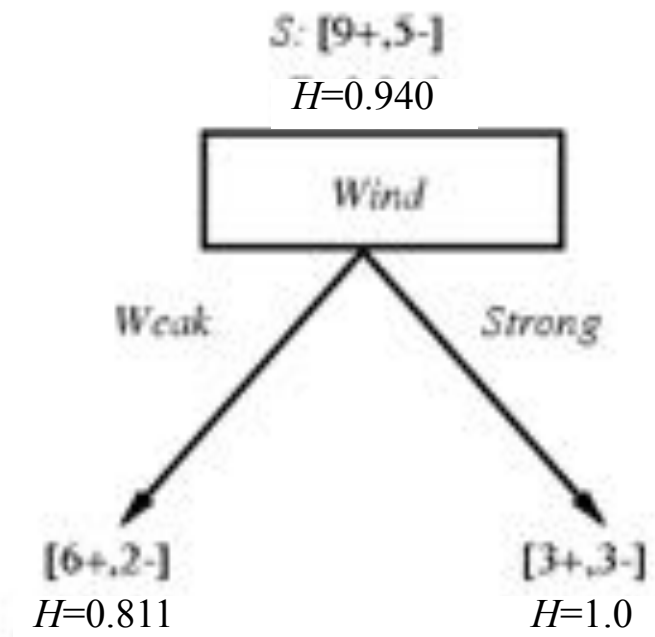


Figure from Tom Mitchell

Tennis Example

Test your understanding

Which attribute yields the best classifier?

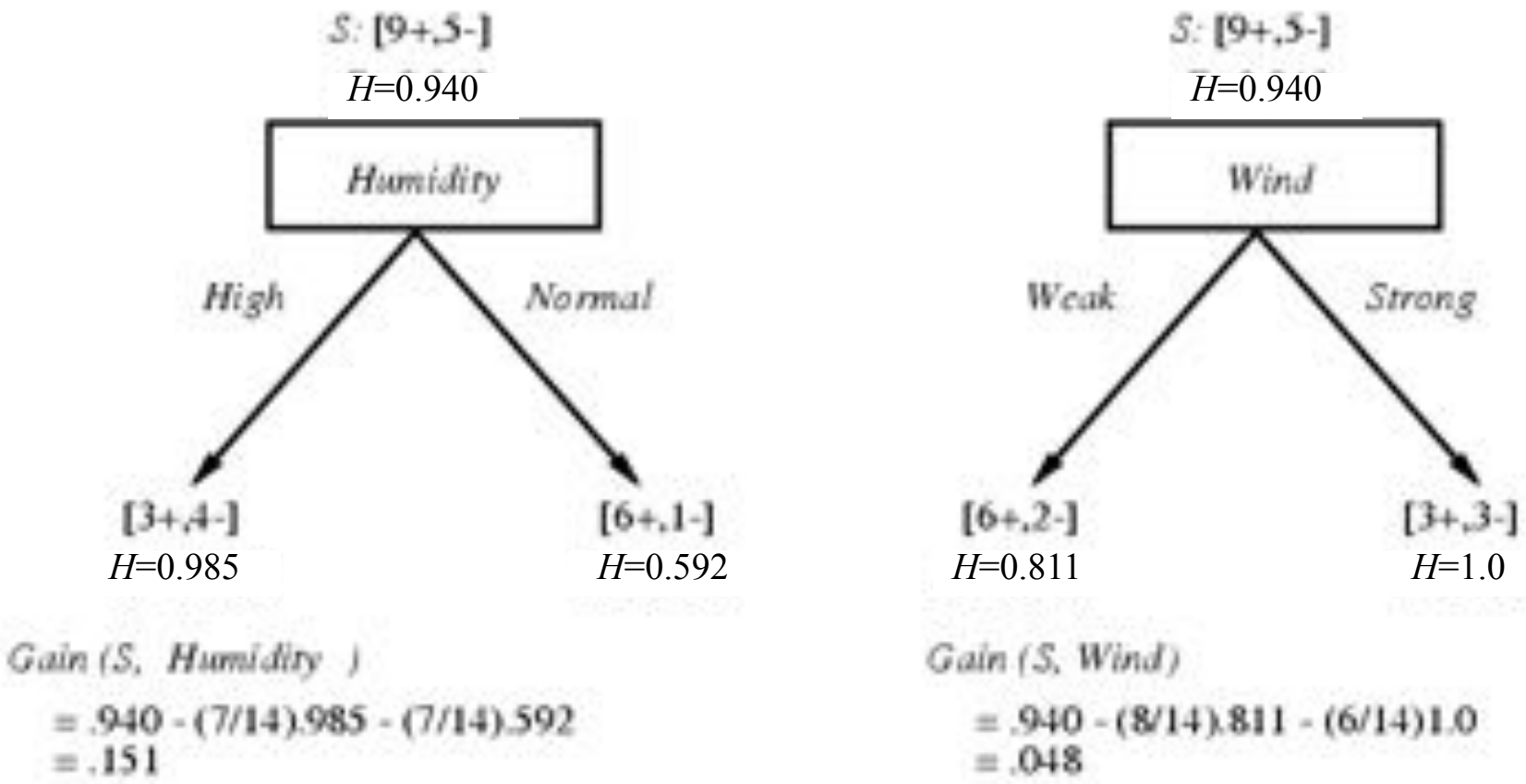
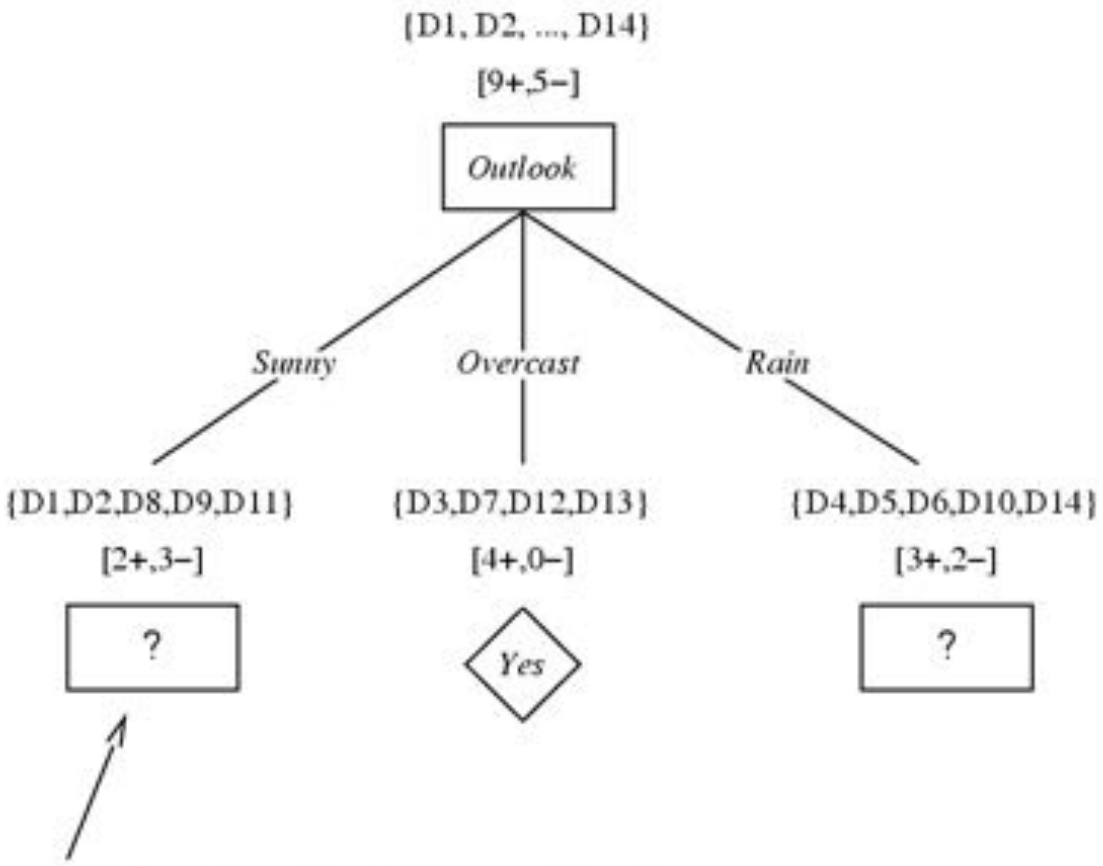


Figure from Tom Mitchell

Tennis Example

Test your understanding



Which attribute should be tested here?

S_{sunny} = {D1,D2,D8,D9,D11}

Gain (S_{sunny}, *Humidity*) = .970 − (3/5) 0.0 − (2/5) 0.0 = .970

Gain (S_{sunny}, *Temperature*) = .970 − (2/5) 0.0 − (2/5) 1.0 − (1/5) 0.0 = .570

Gain (S_{sunny}, *Wind*) = .970 − (2/5) 1.0 − (3/5) .918 = .019

Figure from Tom Mitchell

EMPIRICAL COMPARISON OF SPLITTING CRITERIA

Experiments: Splitting Criteria

Bluntine & Niblett (1992) compared 4 criteria (random, Gini, mutual information, Marshall) on 12 datasets

Medical Diagnosis Datasets: (4 of 12)

- **hypo**: data set of 3772 examples records expert opinion on possible hypo- thyroid conditions from 29 real and discrete attributes of the patient such as sex, age, taking of relevant drugs, and hormone readings taken from drug samples.
- **breast:** The classes are reoccurrence or non-reoccurrence of breast cancer sometime after an operation. There are nine attributes giving details about the original cancer nodes, position on the breast, and age, with multi-valued discrete and real values.
- **tumor:** examples of the location of a primary tumor
- **lymph**: from the lymphography domain in oncology. The classes are normal, metastases, malignant, and fibrosis, and there are nineteen attributes giving details about the lymphatics and lymph nodes

Table 1. Properties of the data sets

Data Set	Classes	Attrs	Training Set	Test Set
hypo	4	29	1000	2772
breast	2	9	200	86
tumor	22	18	237	102
lymph	4	18	103	45
LED	10	7	200	1800
mush	2	22	200	7924
votes	2	17	200	235
votes1	2	16	200	235
iris	3	4	100	50
glass	7	9	100	114
xd6	2	10	200	400
pole	2	4	200	1647

Table from Bluntine & Niblett (1992)

Experiments: Splitting Criteria

Table 3. Error for different splitting rules (pruned trees).

Data Set	Splitting Rule			
	GINI	Info. Gain	Marsh.	Random
hypo	1.01 ± 0.29	0.95 ± 0.22	1.27 ± 0.47	7.44 ± 0.53
breast	28.66 ± 3.87	28.49 ± 4.28	27.15 ± 4.22	29.65 ± 4.97
tumor	60.88 ± 5.44	62.70 ± 3.89	61.62 ± 3.98	67.94 ± 5.68
lymph	24.44 ± 6.92	24.00 ± 6.87	24.33 ± 5.51	32.33 ± 11.25
LED	33.77 ± 3.06	32.89 ± 2.59	33.15 ± 4.02	38.18 ± 4.57
mush	1.44 ± 0.47	1.44 ± 0.47	7.31 ± 2.25	8.77 ± 4.65
votes	4.47 ± 0.95	4.57 ± 0.87	11.77 ± 3.95	12.40 ± 4.56
sound	12.79 ± 1.48	13.04 ± 1.65	15.13 ± 2.89	15.62 ± 2.73
iris	5.00 ± 3.08	4.90 ± 3.08	5.50 ± 2.59	14.20 ± 6.77
glass	39.56 ± 6.20	50.57 ± 6.73	40.53 ± 6.41	53.20 ± 5.01
sd6	22.14 ± 3.23	22.17 ± 3.36	22.06 ± 3.37	31.86 ± 3.62
pole	15.43 ± 1.51	15.47 ± 0.88	15.01 ± 1.15	26.38 ± 6.92

Key Takeaway: GINI gain and Mutual Information are statistically indistinguishable!

Info. Gain is another name for *mutual information*

Table from Bluntine & Niblett (1992)

Experiments: Splitting Criteria

Table 4. Difference and significance of error for GINI splitting rule versus others.

Data Set	Splitting Rule		
	Info. Gain	Marsh.	Random
hypo	−0.06 (0.82)	0.26 (0.99)	6.43 (1.00)
breast	−0.17 (0.23)	−1.51 (0.94)	0.99 (0.72)
tumor	1.81 (0.84)	0.74 (0.39)	7.06 (0.99)
lymph	−0.44 (0.83)	0.11 (0.05)	7.89 (0.99)
LED	0.12 (0.17)		
mush	0.00 (0.00)	5.86	
vote1			
vote		34	
iris		50	
glass		96	
xd6		97	
pole		43	

Key Takeaway: GINI gain and Mutual Information are statistically indistinguishable!

Results are of the form A.AA (B.BB) where:

1. A.AA is the **average difference in errors** between the two methods
2. B.BB is the **significance** of the difference according to a two-tailed **paired t-test**

Table from Bluntine & Niblett (1992)

INDUCTIVE BIAS (FOR DECISION TREES)

Decision Tree Learning Example

Dataset:

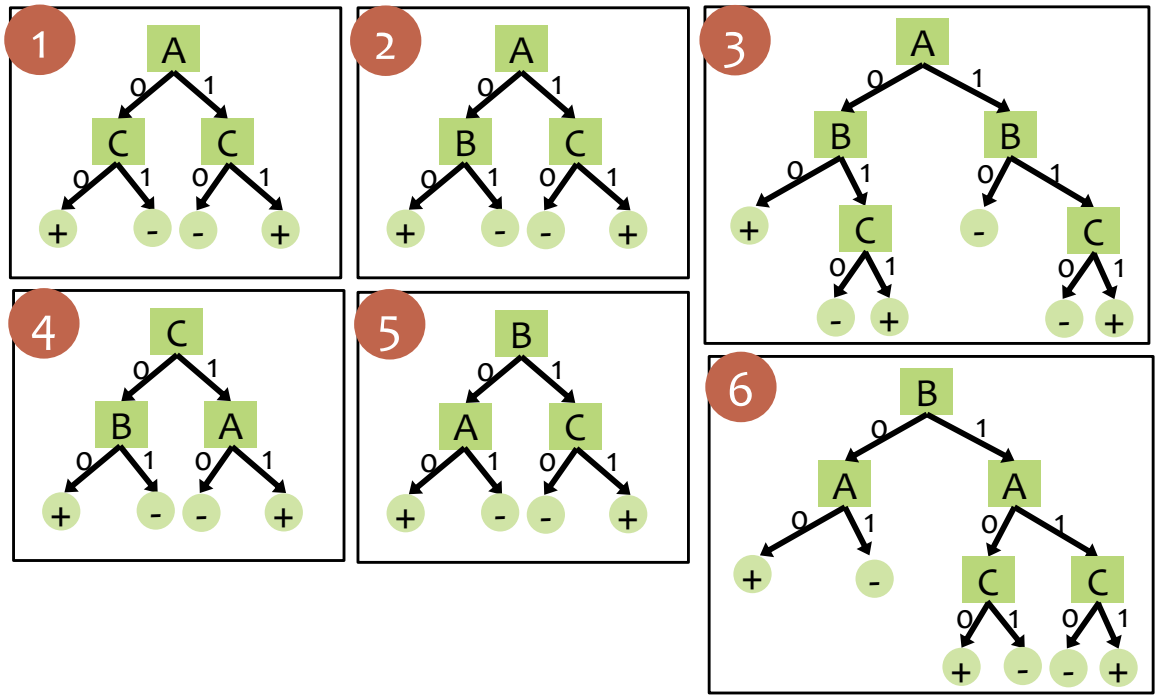
Output Y, Attributes A, B, C

Y	A	B	C
+	0	0	0
+	0	0	1
-	0	1	0
+	0	1	1
-	1	0	0
-	1	0	1
-	1	1	0
+	1	1	1

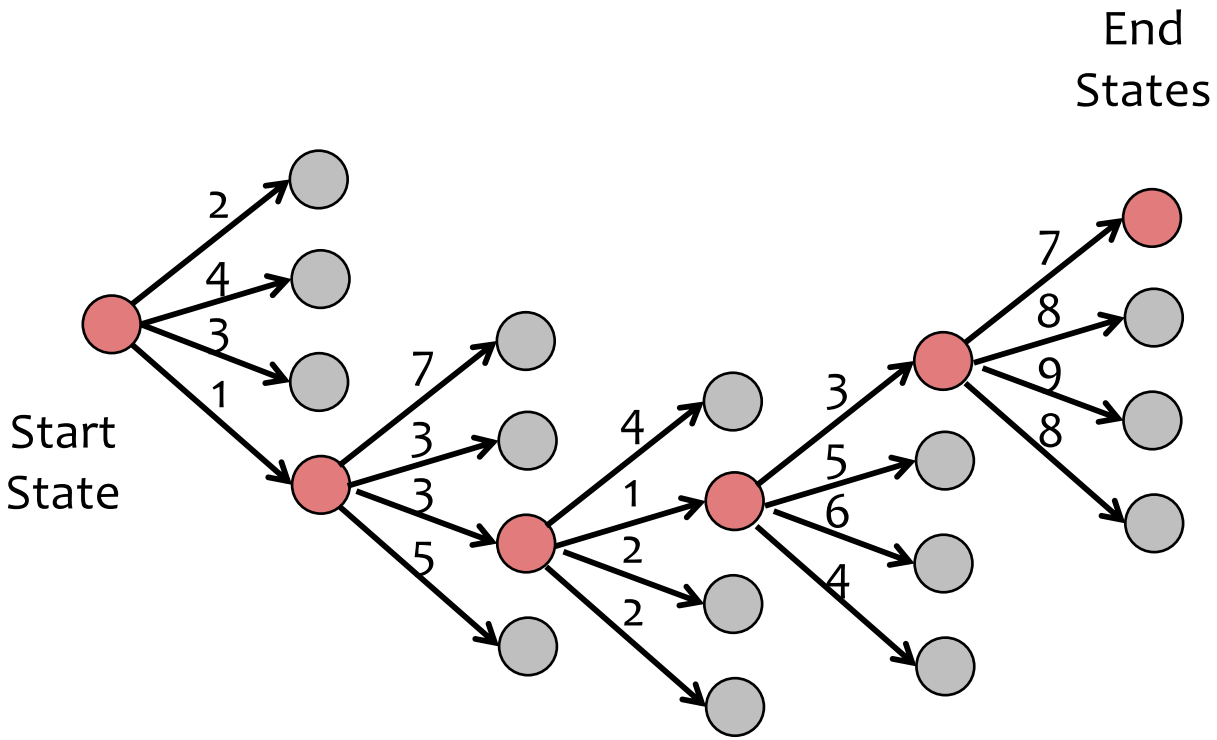
In-Class Exercise

Which of the following trees would be **learned by the the decision tree learning algorithm** using "error rate" as the splitting criterion?

(Assume ties are broken alphabetically.)



Background: Greedy Search



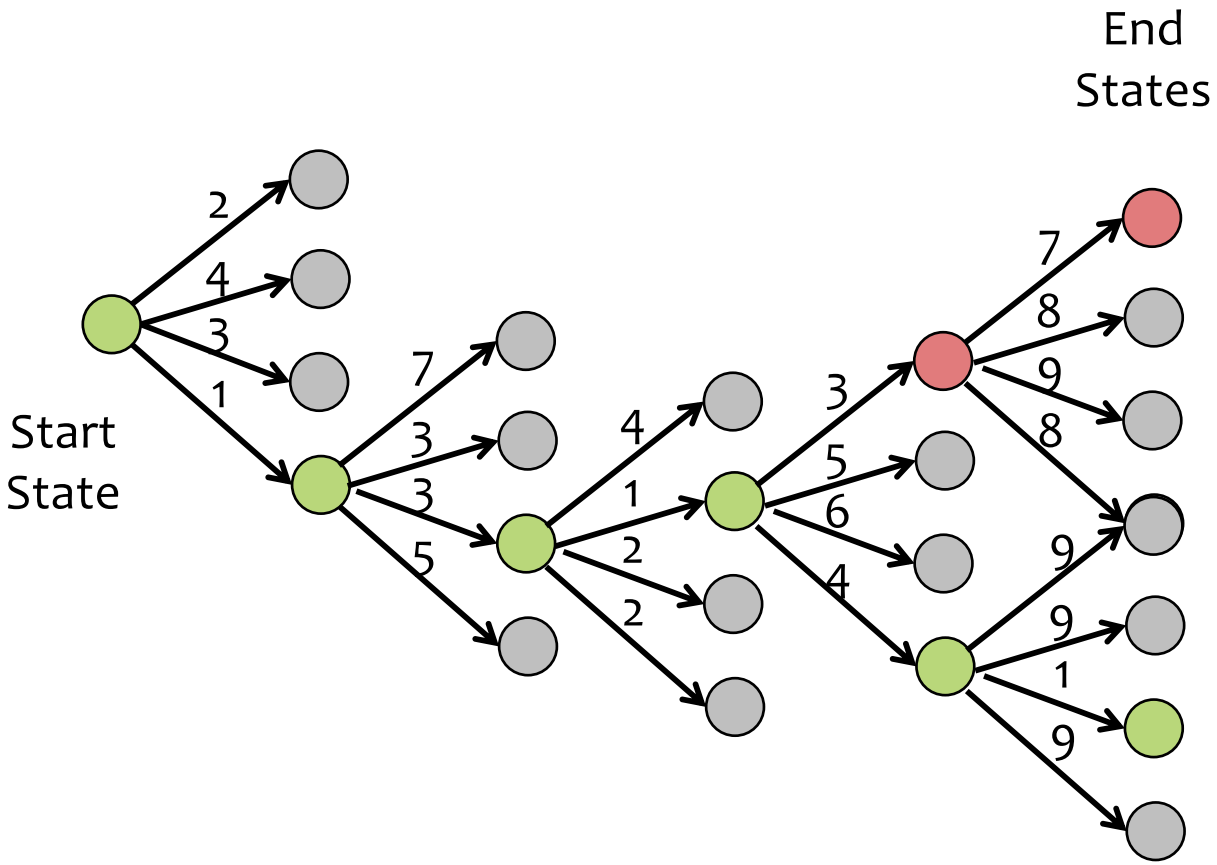
Goal:

- Search space consists of nodes and weighted edges
- Goal is to find the lowest (total) weight path from root to a leaf

Greedy Search:

- At each node, selects the edge with lowest (immediate) weight
- Heuristic method of search (i.e. does *not* necessarily find the best path)

Background: Greedy Search



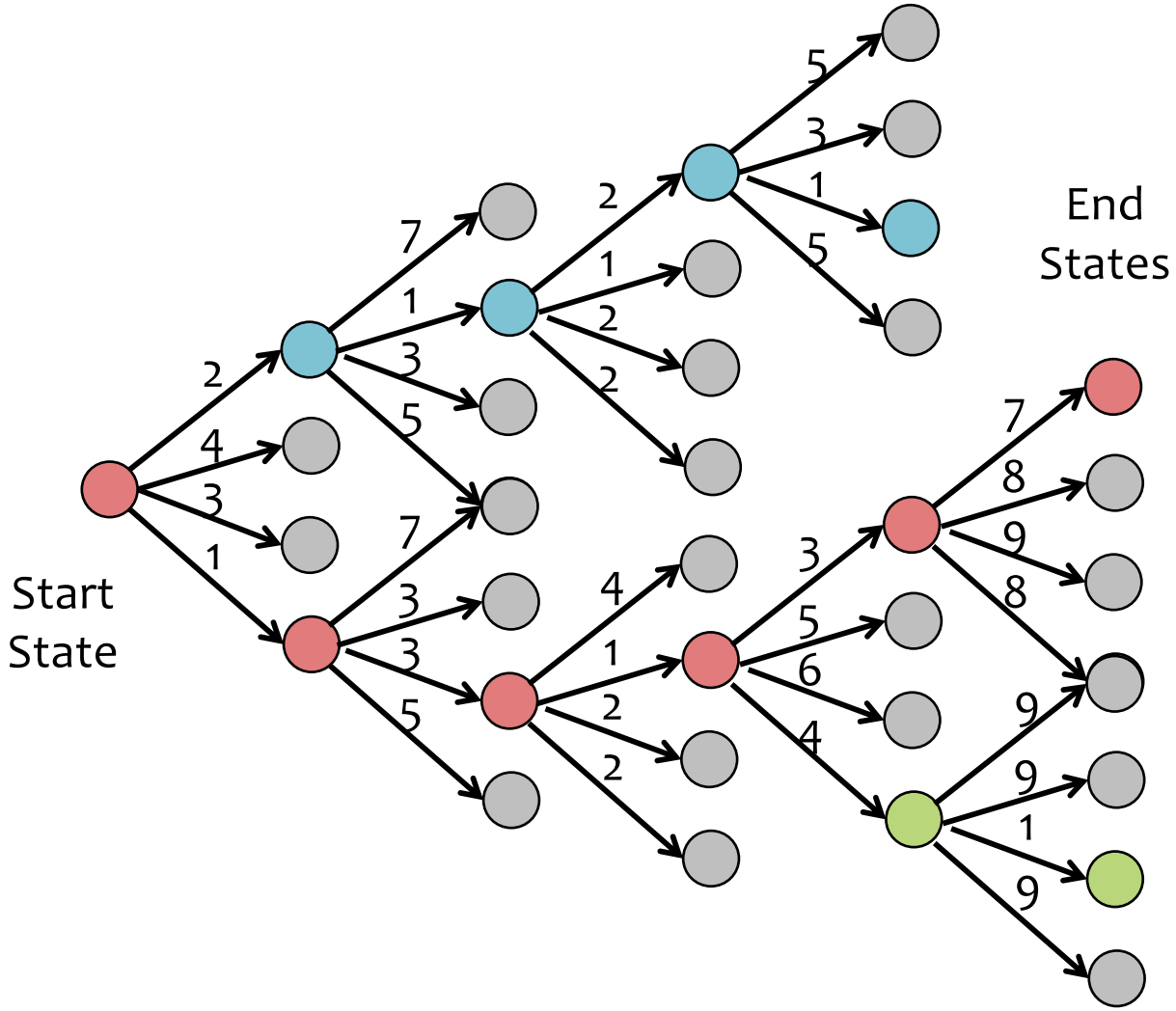
Goal:

- Search space consists of nodes and weighted edges
- Goal is to find the lowest (total) weight path from root to a leaf

Greedy Search:

- At each node, selects the edge with lowest (immediate) weight
- Heuristic method of search (i.e. does *not* necessarily find the best path)

Background: Greedy Search



Goal:

- Search space consists of nodes and weighted edges
- Goal is to find the lowest (total) weight path from root to a leaf

Greedy Search:

- At each node, selects the edge with lowest (immediate) weight
- Heuristic method of search (i.e. does *not* necessarily find the best path)

Decision Trees

Chalkboard

- Decision Tree Learning as Search

DT: Remarks

ID3 = Decision Tree Learning with Mutual Information as the splitting criterion

Question: Which tree does ID3 find?

DT: Remarks

ID3 = Decision Tree Learning with Mutual Information as the splitting criterion

Question: Which tree does ID3 find?

Definition:

We say that the **inductive bias** of a machine learning algorithm is the principal by which it generalizes to unseen examples

Inductive Bias of ID3:

Smallest tree that matches the data with high mutual information attributes near the top

Occam's Razor: (restated for ML)

Prefer the simplest hypothesis that explains the data

Decision Tree Learning Example

Dataset:

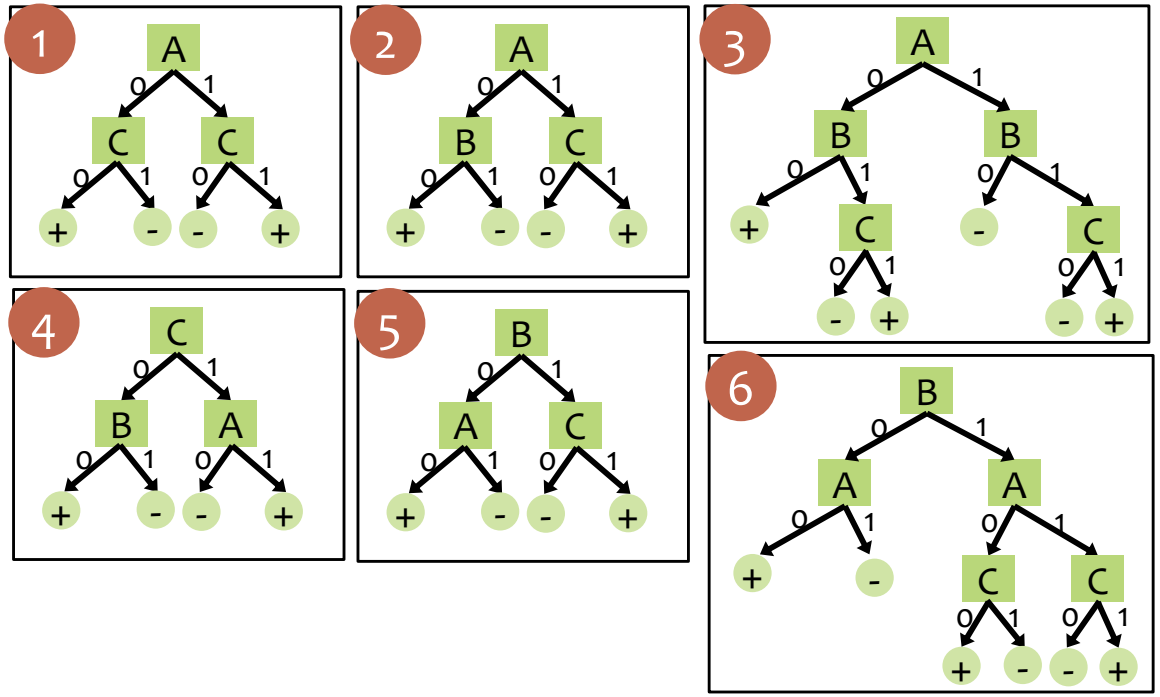
Output Y, Attributes A, B, C

Y	A	B	C
+	0	0	0
+	0	0	1
-	0	1	0
+	0	1	1
-	1	0	0
-	1	0	1
-	1	1	0
+	1	1	1

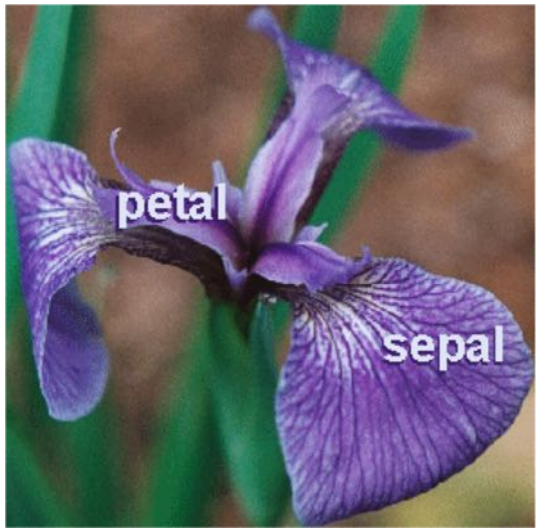
In-Class Exercise

Suppose you had an algorithm that found **the tree with lowest training error that was as small as possible (i.e. exhaustive global search)**, which tree would it return?

(Assume ties are broken by choosing the smallest.)



CLASSIFICATION


petal
sepal

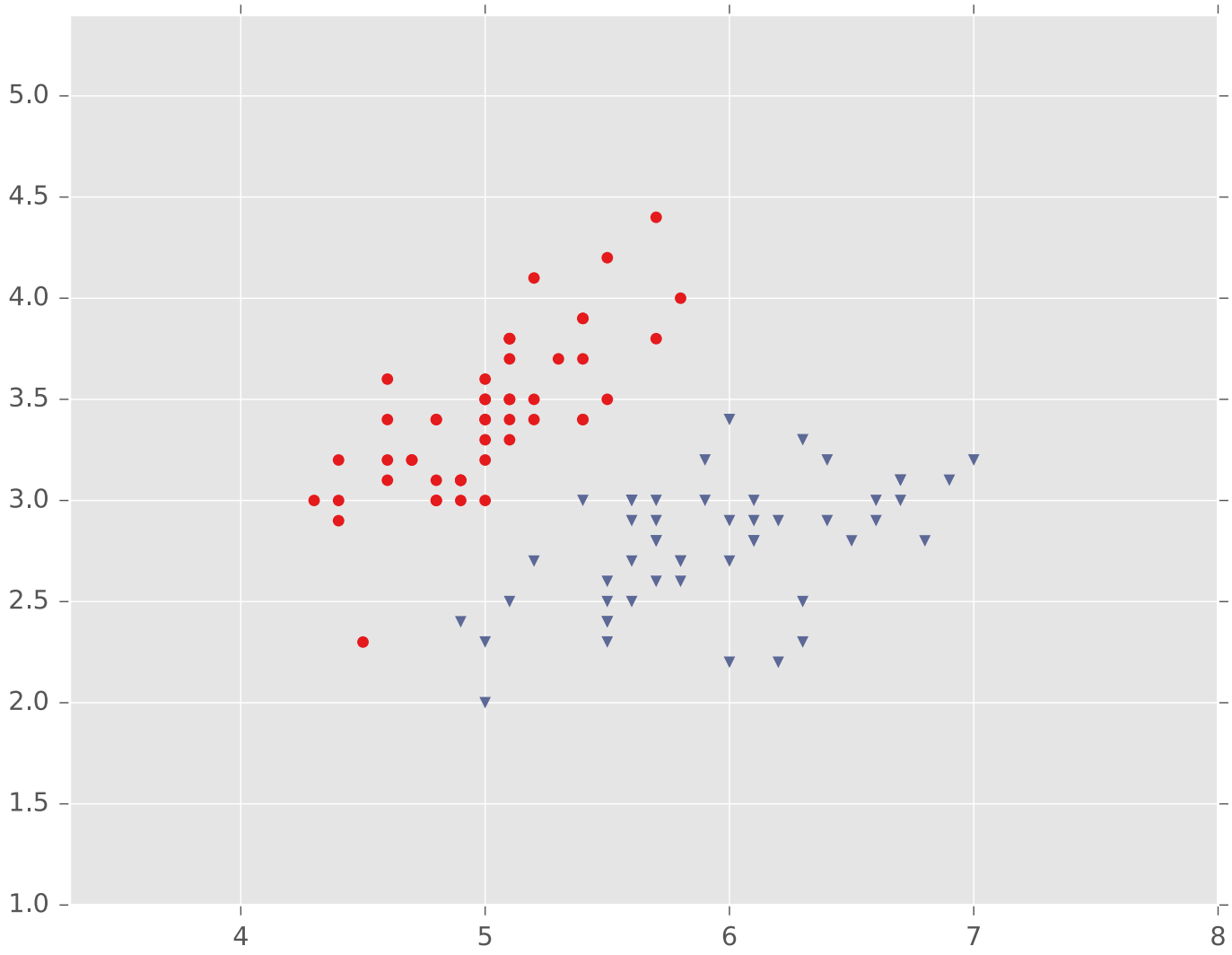
Fisher Iris Dataset

Fisher (1936) used 150 measurements of flowers from 3 different species: Iris setosa (0), Iris virginica (1), Iris versicolor (2) collected by Anderson (1936)

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
0	4.3	3.0	1.1	0.1
0	4.9	3.6	1.4	0.1
0	5.3	3.7	1.5	0.2
1	4.9	2.4	3.3	1.0
1	5.7	2.8	4.1	1.3
1	6.3	3.3	4.7	1.6
1	6.7	3.0	5.0	1.7

Full dataset: https://en.wikipedia.org/wiki/Iris_flower_data_set

Fisher Iris Dataset



K-NEAREST NEIGHBORS

Classification

Chalkboard:

- Binary classification
- 2D examples
- Decision rules / hypotheses

k-Nearest Neighbors

Chalkboard:

- Nearest Neighbor classifier
- KNN for binary classification