As you login

1. Rename yourself in Zoom to *pre*-pend your house number
   - e.g. "0 – Pat Virtue"

2. Open Piazza (getting ready for polls)

3. Download preview slides from course website

4. Grab something to write with/on ☺

Announcements

Assignments

- HW1
  - Due Thu, 9/10, 11:59 pm (all times will be Pittsburgh time)
  - HW1 only – Extension request form – see Piazza
- HW2
  - Out after HW1 due
  - Due Mon, 9/21, 11:59 pm

Announcements

Participation

- Earn participation points for study group sessions
- See Piazza for details

OH by appointment

- Check OH Appointment Slots link on course webpage
- Private post on Piazza with appointment request

Breakout rooms

- Video on, unmute
- Introduce yourself if you haven’t already met

Introduction to Machine Learning

Decision Trees

Instructor: Pat Virtue

Plan

Last time

- Problem formulation (notation)
- Algorithm 0-2: Memorization, majority vote, decision stump

Today

- Decision trees
  - Recursive algorithm
  - Better splitting criteria (entropy, mutual information)

Next time

- Wrap up decision trees (overfitting, continuous features)
- Nearest neighbor methods

Decision Trees

A few tools

Majority vote:

$$\hat{y} = \underset{c}{\operatorname{argmax}} \frac{N_c}{N}$$

Classification error rate:

$$ErrorRate = \frac{1}{N} \sum_i \mathbb{I}\left(y^{(i)} \neq \hat{y}^{(i)}\right)$$

What fraction did we predict incorrectly

Expected value

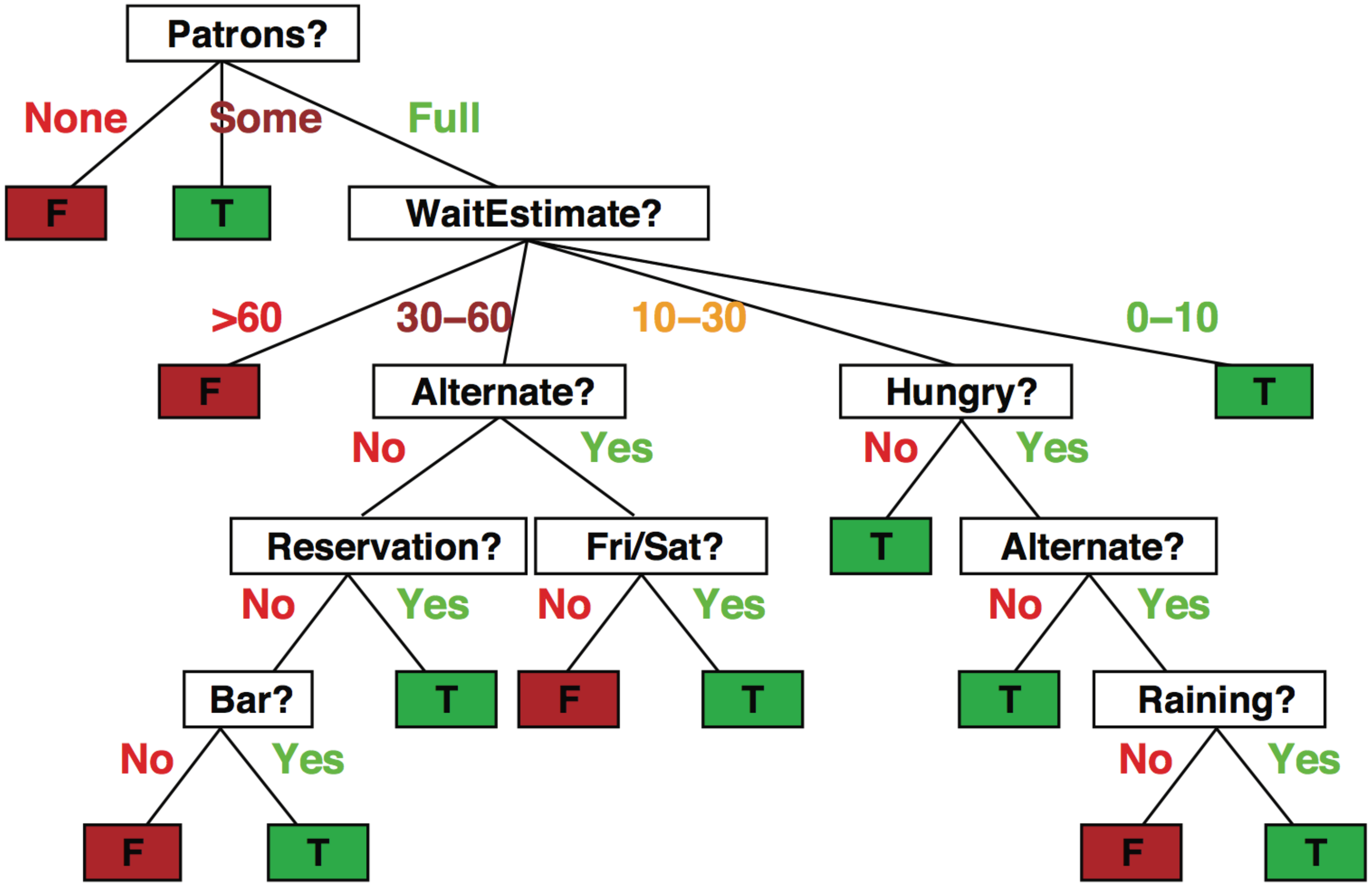
$$\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} f(x)\, P(X = x) \quad \text{or} \quad \mathbb{E}[f(X)] = \int_{\mathcal{X}} f(x)\, p(x)\, dx$$

Decision trees

Popular representation for classifiers

- Even among humans!

I've just arrived at a restaurant: should I stay (and wait for a table) or go elsewhere?



Slide credit: ai.berkeley.edu

Decision trees

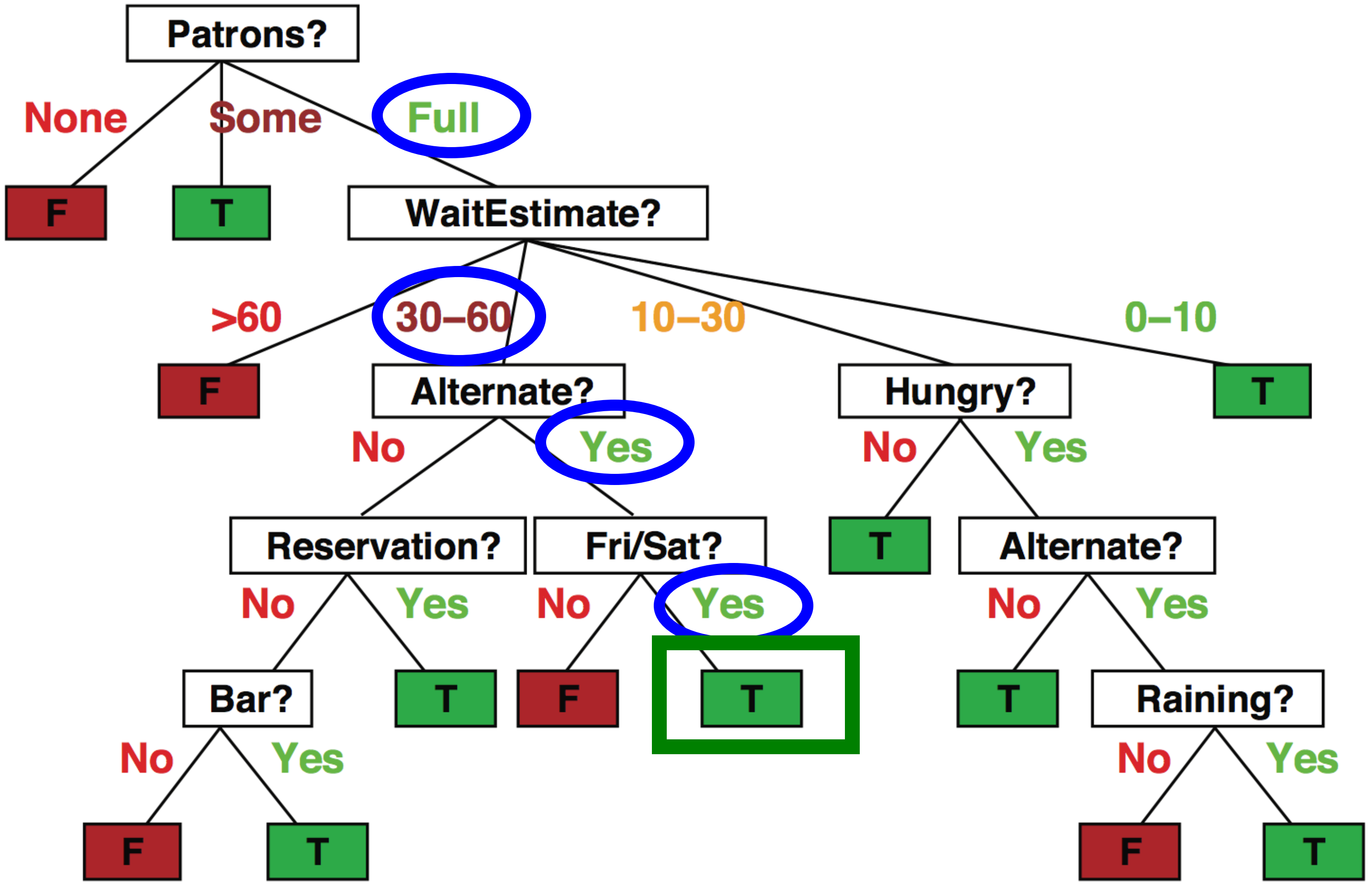
It’s Friday night and you’re hungry

You arrive at your favorite cheap but really cool happening burger place

It’s full and you have no reservation but there is a bar

The host estimates a 45 minute wait

There are alternatives nearby but it’s raining outside



Decision tree ***partitions*** the input space, assigns a label to each partition

Slide credit: ai.berkeley.edu

Problem Formulation

Medical Prediction

Y Outcome	X_1 Fetal Position	X_2 Fetal Distress	X_3 Previous C-sec
Natural	Vertex	N	N
C-section	Breech	N	N
Natural	Vertex	Y	Y
C-section	Vertex	N	Y
Natural	Abnormal	N	N

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = [x_1, x_2, x_3]^T$$

$x_1 \in \{Vertex, Breech, Abn\}$
$x_2 \in \{Y, N\}$
$x_3 \in \{Y, N\}$

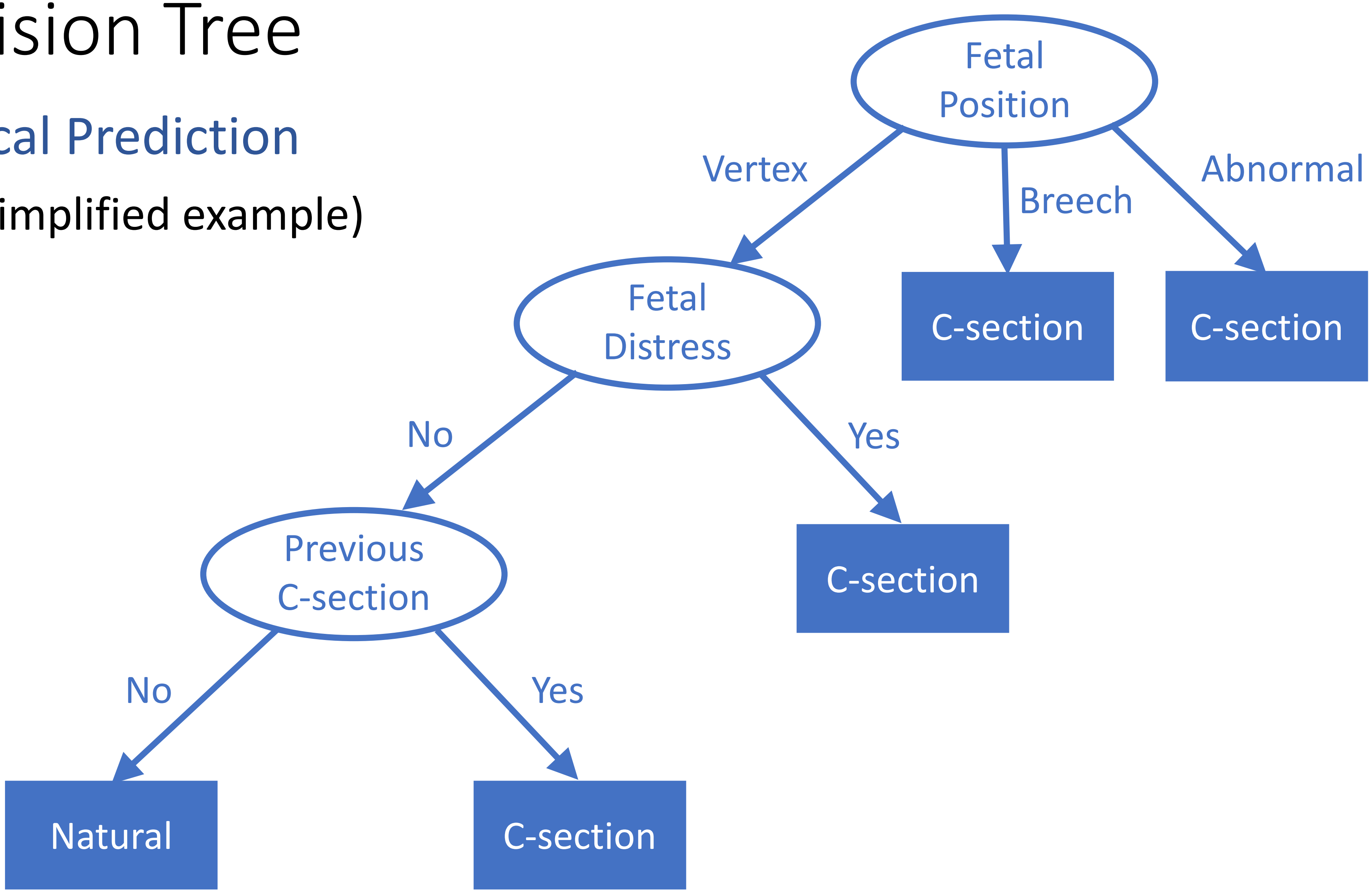
$y \in \{Csection, Natural\}$

$\hat{y} = h(\boldsymbol{x})$

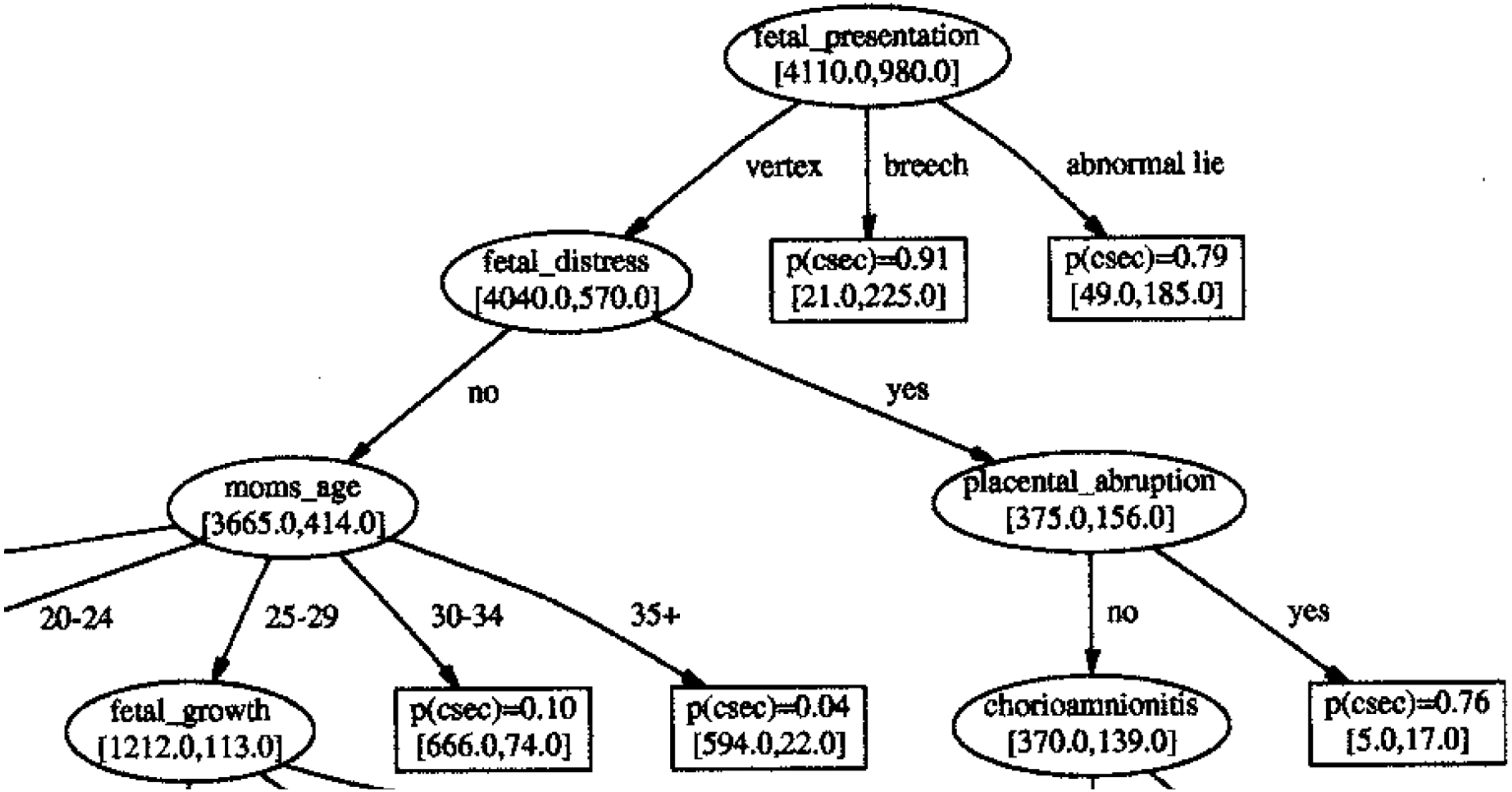
Decision Tree

Medical Prediction

(Oversimplified example)



Tree to Predict C-Section Risk



Sims, C.J., Meyn, L., Caruana, R., Rao, R.B., Mitchell, T. and Krohn, M.
American journal of obstetrics and gynecology, *2000*

Tree to Predict C-Section Risk

Learned from medical records of 1000 women

Negative examples are C-sections

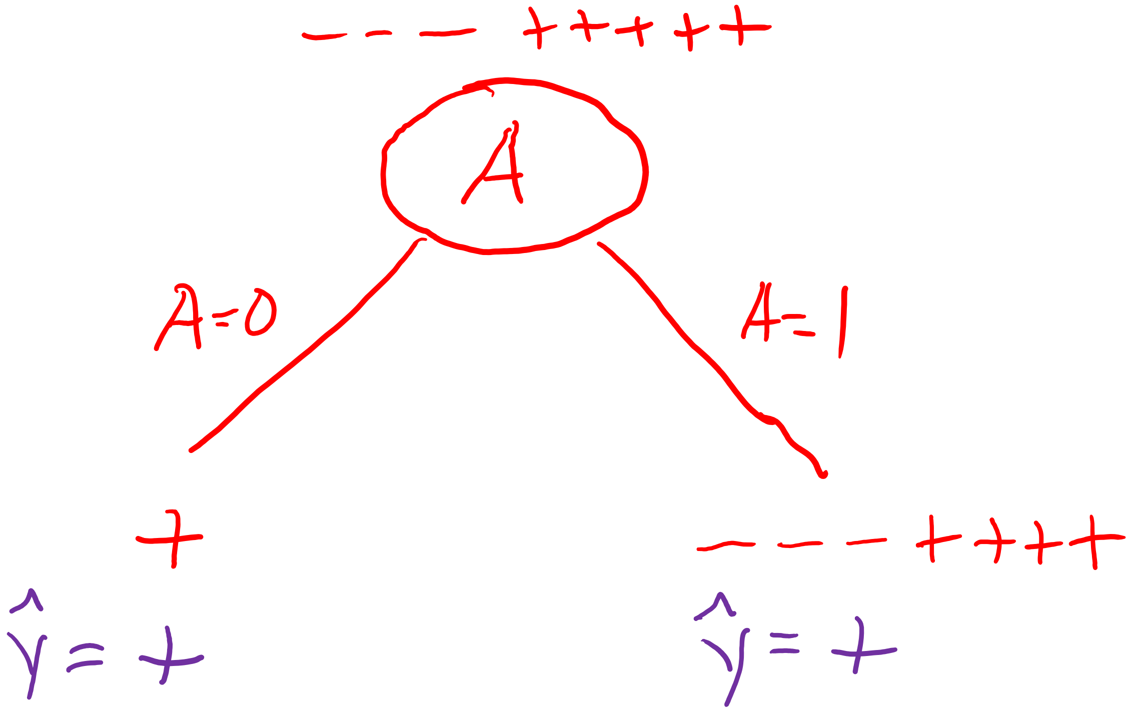
```
[833+,167-] .83+ .17-
Fetal_Presentation = 1: [822+,116-] .88+ .12-
| Previous_Csection = 0: [767+,81-] .90+ .10-
| | Primiparous = 0: [399+,13-] .97+ .03-
| | Primiparous = 1: [368+,68-] .84+ .16-
| | | Fetal_Distress = 0: [334+,47-] .88+ .12-
| | | | Birth_Weight < 3349: [201+,10.6-] .95+ .
| | | | Birth_Weight >= 3349: [133+,36.4-] .78+
| | | Fetal_Distress = 1: [34+,21-] .62+ .38-
| Previous_Csection = 1: [55+,35-] .61+ .39-
Fetal_Presentation = 2: [3+,29-] .11+ .89-
Fetal_Presentation = 3: [8+,22-] .27+ .73-
```

Slide credit: CMU MLD Tom Mitchell

Decision Stumps

Split data based on a single attribute

Majority vote at leaves



Dataset:

Output Y, Attributes A, B, C

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

How could we implement training and prediction?

Algorithm 2: Decision stump algorithm

def train(D)

① pick an attribute X_m

② Divide dataset as

$$D^{(0)} = \{(\vec{x}, y) \in D \mid X_m = 0\}$$

$$D^{(1)} = \{(\vec{x}, y) \in D \mid X_m = 1\}$$

③ two votes

$$v^{(0)} = \text{majority}(D^{(0)})$$

$$v^{(1)} = \text{majority}(D^{(1)})$$

def $h(\vec{x})$

if $x_m = 0$

return $v^{(0)}$

if $x_m = 1$

return $v^{(1)}$

Slide credit: CMU MLD Matt Gormley

Decision Stumps

Splitting on which attribute {A, B, C} creates a decision stump with the lowest training error?

Dataset:

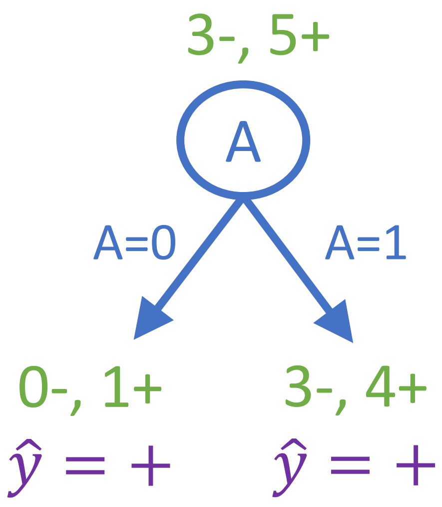
Output Y, Attributes A, B, C

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

Decision Stumps

Splitting on which attribute {A, B, C} creates a decision stump with the lowest training error?

Answer: B



Error: (0 + 3) / 8

= 3/8

Dataset:

Output Y, Attributes A, B, C

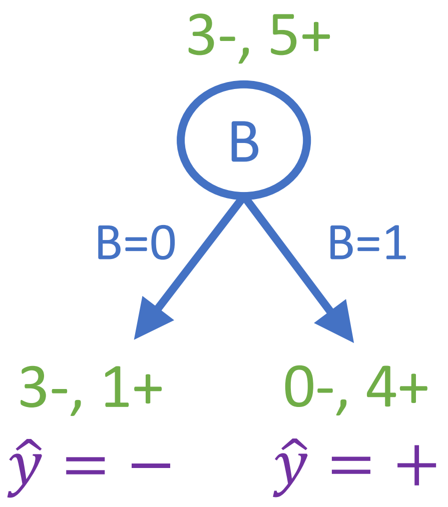
Y	A	B	C
-	1	0	0
-	1	0	1
-	1	0	0
+	0	0	1
+	1	1	0
+	1	1	1
+	1	1	0
+	1	1	1

Slide credit: CMU MLD Matt Gormley

Decision Stumps

Splitting on which attribute {A, B, C} creates a decision stump with the lowest training error?

Answer: B



Error: (1 + 0) / 8

= 1/8

Dataset:

Output Y, Attributes A, B, C

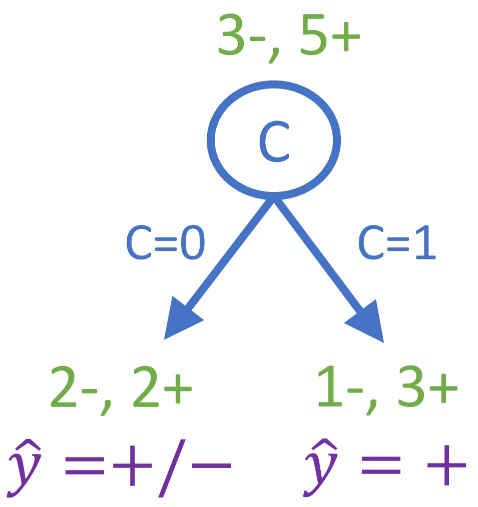
Y	A	B	C
-	1	0	0
-	1	0	1
-	1	0	0
+	0	0	1
+	1	1	0
+	1	1	1
+	1	1	0
+	1	1	1

Slide credit: CMU MLD Matt Gormley

Decision Stumps

Splitting on which attribute {A, B, C} creates a decision stump with the lowest training error?

Answer: B



Error: (2 + 1) / 8

= 3/8

Dataset:

Output Y, Attributes A, B, C

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

Building a Decision Tree

```
Function BuildTree(D,A)
    # D: dataset at current node, A: current set of attributes
    If empty(A) or all labels in D are the same
        # Leaf node
        class = most common class in D
    else
        # Internal node
        a ⇐ bestAttribute(D,A)
        LeftNode = BuildTree(D(a=1), A \ {a})
        RightNode = BuildTree(D(a=0), A \ {a})
    end
end
```

Building a Decision Tree

Dataset:

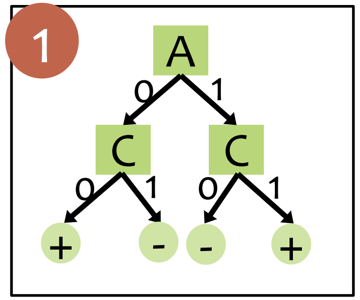
Output Y, Attributes A, B, C

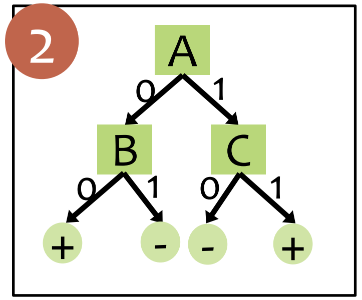
Y	A	B	C
-	1	0	0
-	1	0	1
-	1	0	0
+	0	0	1
+	1	1	0
+	1	1	1
+	1	1	0
+	1	1	1

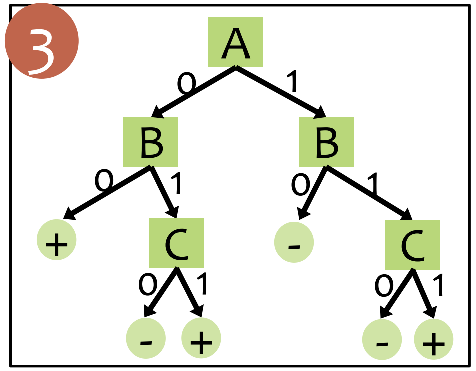
Piazza Poll 1

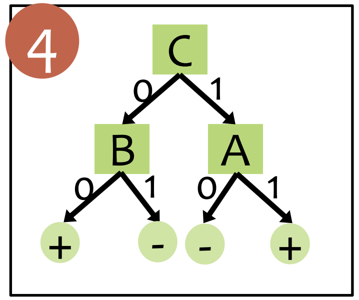
Which of the following trees would be learned by the decision tree learning algorithm using "error rate" as the splitting criterion?

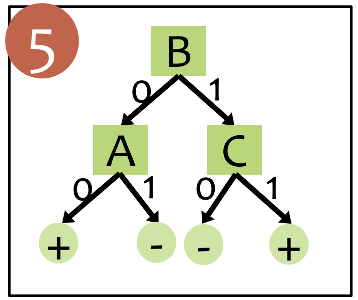
(Assume ties are broken alphabetically.)

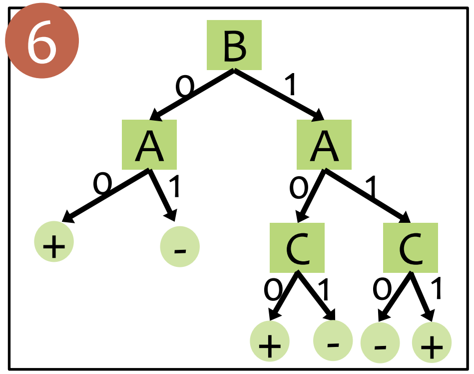












Dataset:

Output Y, Attributes A, B, C

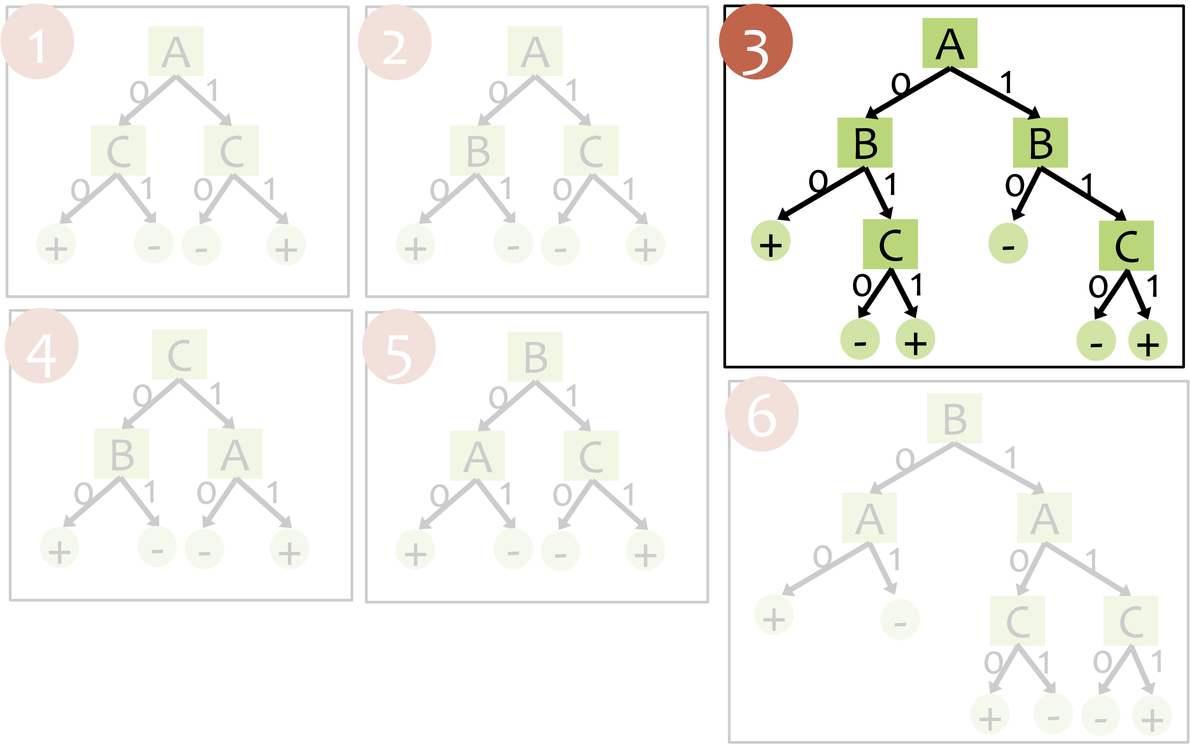
Y	A	B	C
+	0	0	0
+	0	0	1
-	0	1	0
+	0	1	1
-	1	0	0
-	1	0	1
-	1	1	0
+	1	1	1

Slide credit: CMU MLD Matt Gormley

Piazza Poll 1

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.)



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

Piazza Poll 2

Which attribute {A, B} would error rate select for the next split?

1) A
2) B
3) A or B (tie)
4) I don't know

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

Piazza Poll 2

Which attribute {A, B} would error rate select for the next split?

1) A
2) B
3) A or B (tie)
4) I don't know

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

Slide credit: CMU MLD Matt Gormley

Building a Decision Tree

```
Function BuildTree(D,A)
    # D: dataset at current node, A: current set of attributes
    If empty(A) or all labels in D are the same
        # Leaf node
        class = most common class in D
    else
        # Internal node
        a ⇐ bestAttribute(D,A)
        LeftNode = BuildTree(D(a=1), A \ {a})
        RightNode = BuildTree(D(a=0), A \ {a})
    end
end
```

Entropy

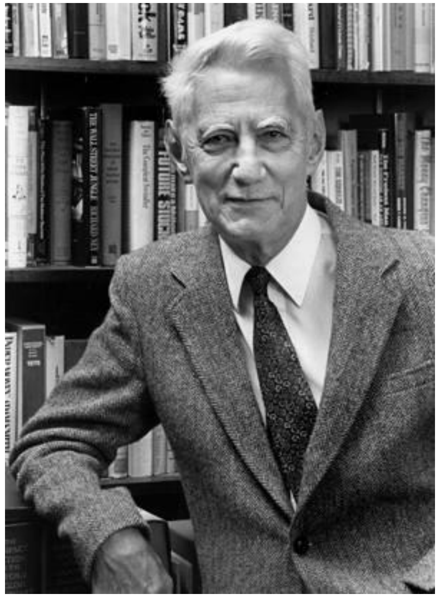
- Quantifies the amount of uncertainty associated with a specific probability distribution
- The higher the entropy, the less confident we are in the outcome

Entropy

- Quantifies the amount of uncertainty associated with a specific probability distribution
- The higher the entropy, the less confident we are in the outcome
- Definition

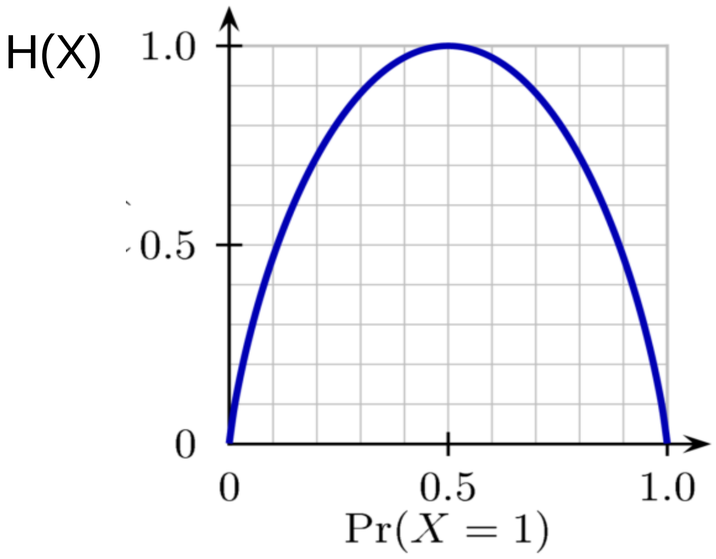
$$H(X) = \sum_{x} p(X = x) \log_2 \frac{1}{p(X = x)}$$

$$H(X) = -\sum_{x} p(X = x) \log_2 p(X = x)$$



Claude Shannon (1916 – 2001), most of the work was done in Bell labs

Entropy



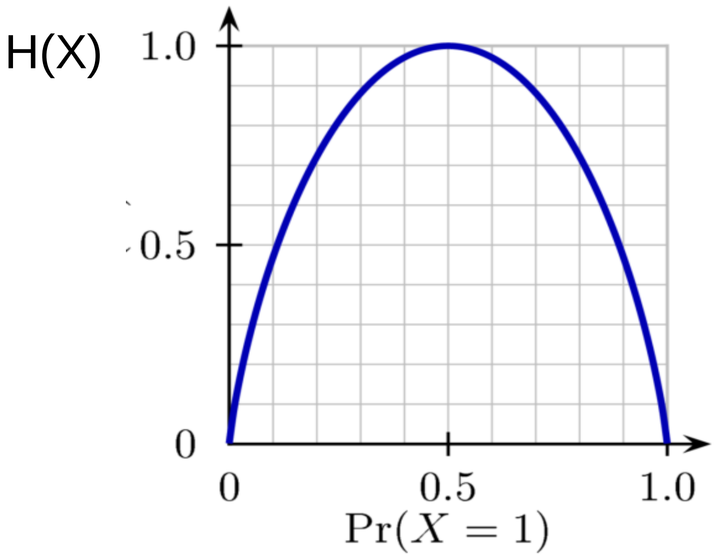
- Definition

$$H(X) = -\sum_{x} p(X = x) \log_2 p(X = x)$$

- So, if P(X=1) = 1 then

- If P(X=1) = .5 then

Entropy



- Definition

$$H(X) = -\sum_x p(X=x)\log_2 p(X=x)$$

- So, if P(X=1) = 1 then

$$H(X) = -p(x=1)\log_2 p(X=1) - p(x=0)\log_2 p(X=0)$$
$$= -1\log 1 - 0\log 0 = 0$$

- If P(X=1) = .5 then

$$H(X) = -p(x=1)\log_2 p(X=1) - p(x=0)\log_2 p(X=0)$$
$$= -.5\log_2 .5 - .5\log_2 .5 = -\log_2 .5 = 1$$

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

- 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.

Slide credit: CMU MLD Matt Gormley

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

- **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?*

Slide credit: CMU MLD Matt Gormley

Piazza Poll 3

Which attribute {A, B} would **mutual information** select for the next split?

1) A
2) B
3) A or B (tie)
4) I don't know

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

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

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

Slide credit: CMU MLD Matt Gormley

Decision Tree Learning Example

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

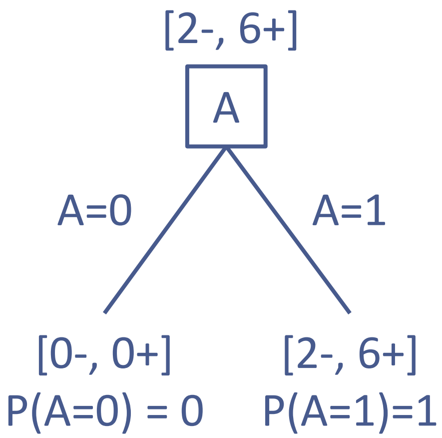
Mutual Information: $I(Y;X) = H(Y) - H(Y|X)$

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

$$H(Y) = -\left[\frac{2}{8}\log_2\frac{2}{8} + \frac{6}{8}\log_2\frac{6}{8}\right]$$

$$H(Y \mid A=0) = undefined$$
$$H(Y \mid A=1) = -\left[\frac{2}{8}\log_2\frac{2}{8} + \frac{6}{8}\log_2\frac{6}{8}\right] = H(Y)$$

$$\begin{aligned} H(Y \mid A) &= P(A=0)H(Y \mid A=0) + P(A=1)H(Y \mid A=1) \\ &= 0 + H(Y \mid A=1) \\ &= H(Y) \end{aligned}$$
$$I(Y;A) = H(Y) - H(Y \mid A) = 0$$



Slide credit: CMU MLD Matt Gormley

Decision Tree Learning Example

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

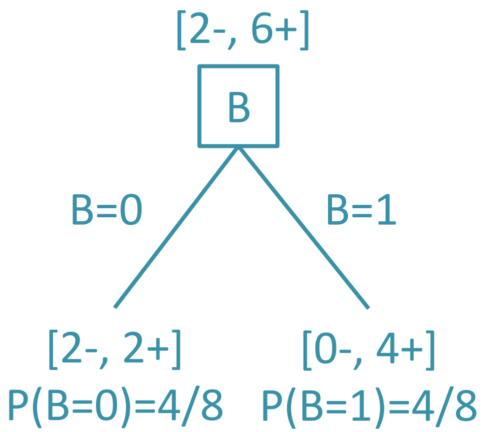
Mutual Information: $I(Y;X) = H(Y) - H(Y|X)$

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

$$H(Y) = -\left[\frac{2}{8}\log_2\frac{2}{8} + \frac{6}{8}\log_2\frac{6}{8}\right]$$

$$H(Y \mid B=0) = -\left[\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right]$$
$$H(Y \mid B=1) = -[0\log_2 0 + 1\log_2 1] = 0$$

$$H(Y \mid B) = P(B=0)H(Y \mid B=0) + P(B=1)H(Y \mid B=1)$$
$$= \frac{4}{8}H(Y \mid B=0) + \frac{4}{8} \cdot 0$$



$$I(Y;B) = H(Y) - H(Y \mid B) > 0$$

$I(Y;B)$ ends up being greater than $I(Y;A) = 0$, so we split on B

Slide credit: CMU MLD Matt Gormley

Mutual Information Notation

We use mutual information in the context of before and after a split, regardless of where that split is in the tree.

$$I(Y; X) = H(Y) - H(Y \mid X)$$