10-315
Introduction to ML

Decision Trees

Instructor: Pat Virtue

Today

Decision trees

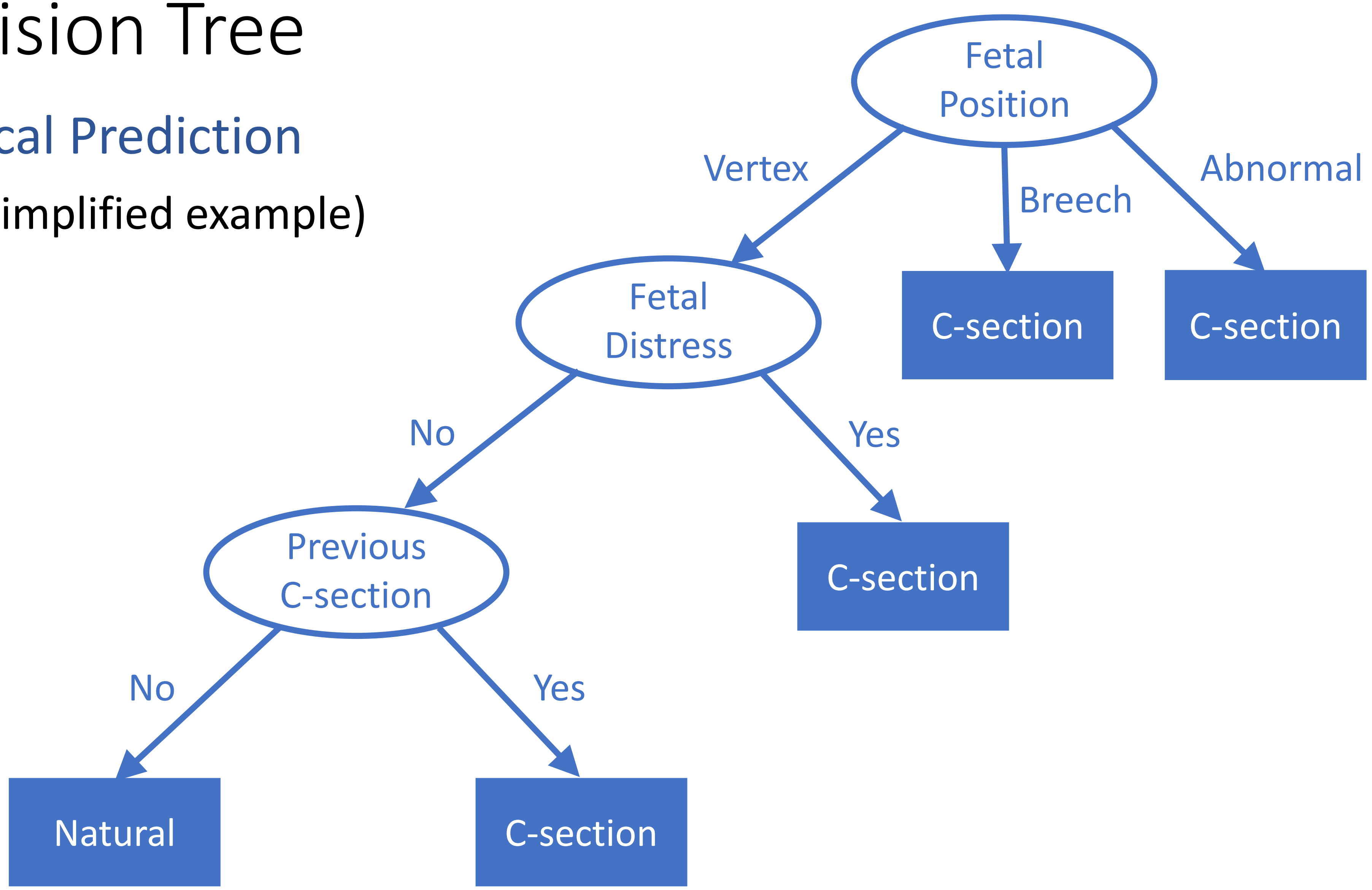
K-Nearest Neighbor

Model Selection

Decision Tree

Medical Prediction

(Oversimplified example)



Decision Trees

Iris

A few tools

N_0 N_1 N_2

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
2	5.9	3.0	5.1	1.8

Majority vote:

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

$\frac{3}{7}, \frac{3}{7}, \frac{1}{7}$

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

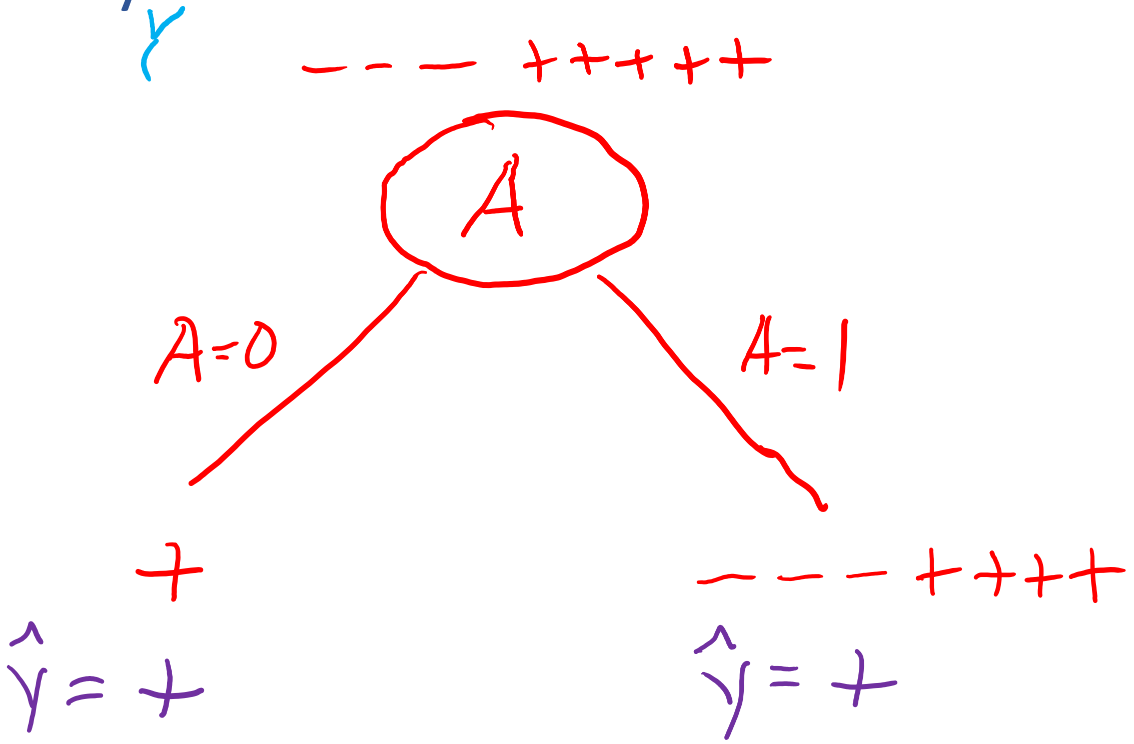
$$\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} f(x)\, P(X = x) \quad \text{or} \quad \mathbb{E}[f(X)] = \int_{\mathcal{X}} f(x)\, p(x)\, dx$$

P(Y | X)

Decision Stumps

Split data based on a single attribute

Majority vote at leaves



Dataset:

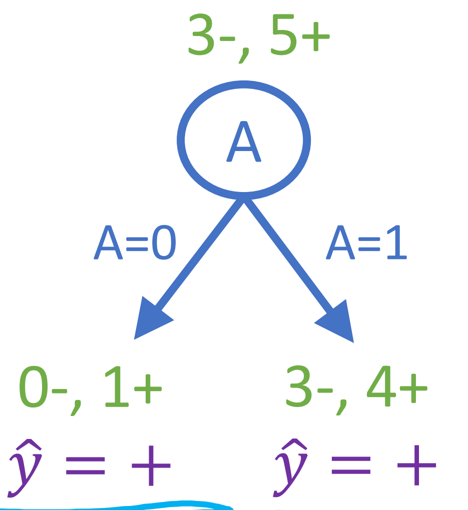
Output Y, Attributes A, B, C

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

Decision Stumps

Split data based on a single attribute

Majority vote at leaves



Error: (0 + 3) / 8

= 3/8

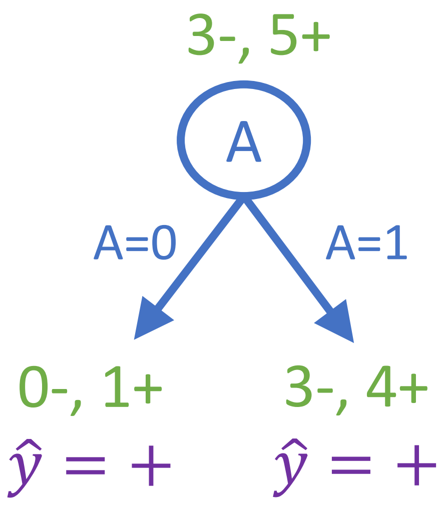
Dataset:

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

Poll 1

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



Error: (0 + 3) / 8

= 3/8

Dataset:

Output Y, Attributes A, B, C

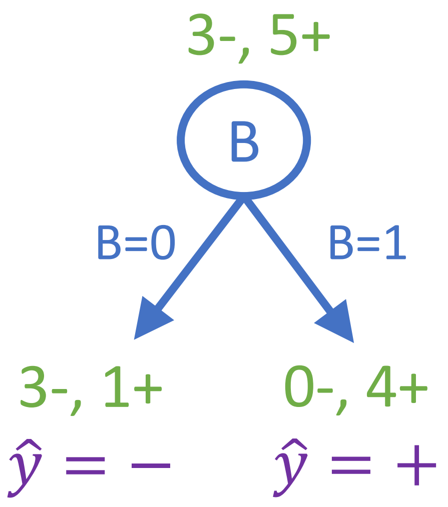
Y	A	B	C
-	1	0	0
-	1	0	1
-	1	0	0
+	0	0	1
+	1	1	0
+	1	1	1
+	1	1	0
+	1	1	1

Slide credit: CMU MLD Matt Gormley

Poll 1

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

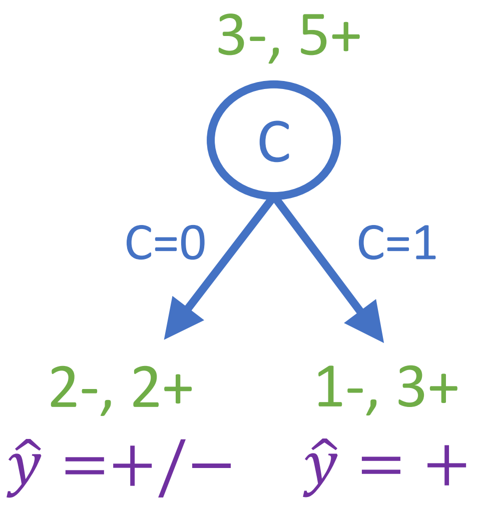
Y	A	B	C
-	1	0	0
-	1	0	1
-	1	0	0
+	0	0	1
+	1	1	0
+	1	1	1
+	1	1	0
+	1	1	1

Slide credit: CMU MLD Matt Gormley

Poll 1

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

Problem Formulation

Medical Prediction

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

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

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

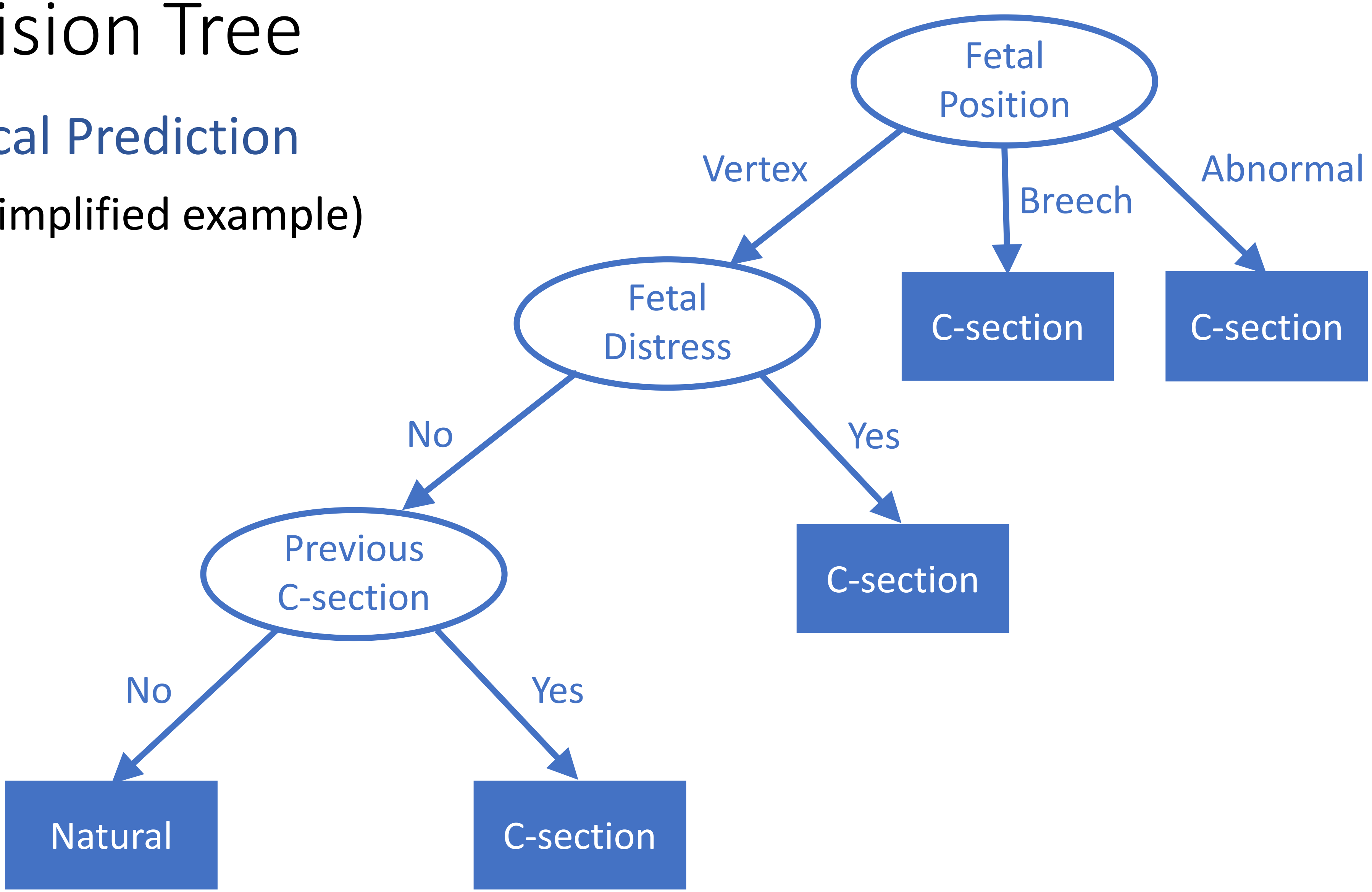
$y \in \{Csection, Natural\}$

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

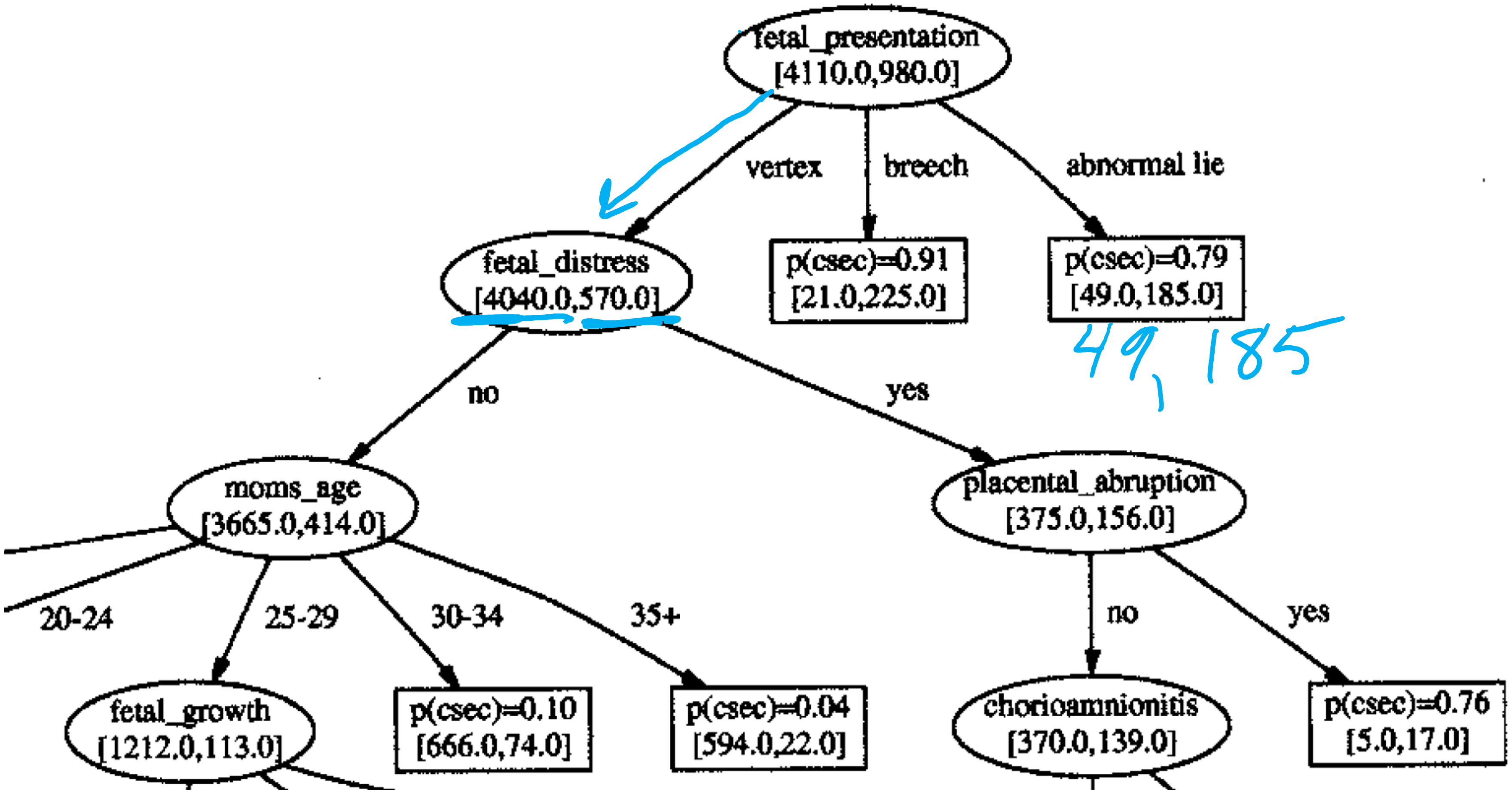
Decision Tree

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

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
```

A

pres

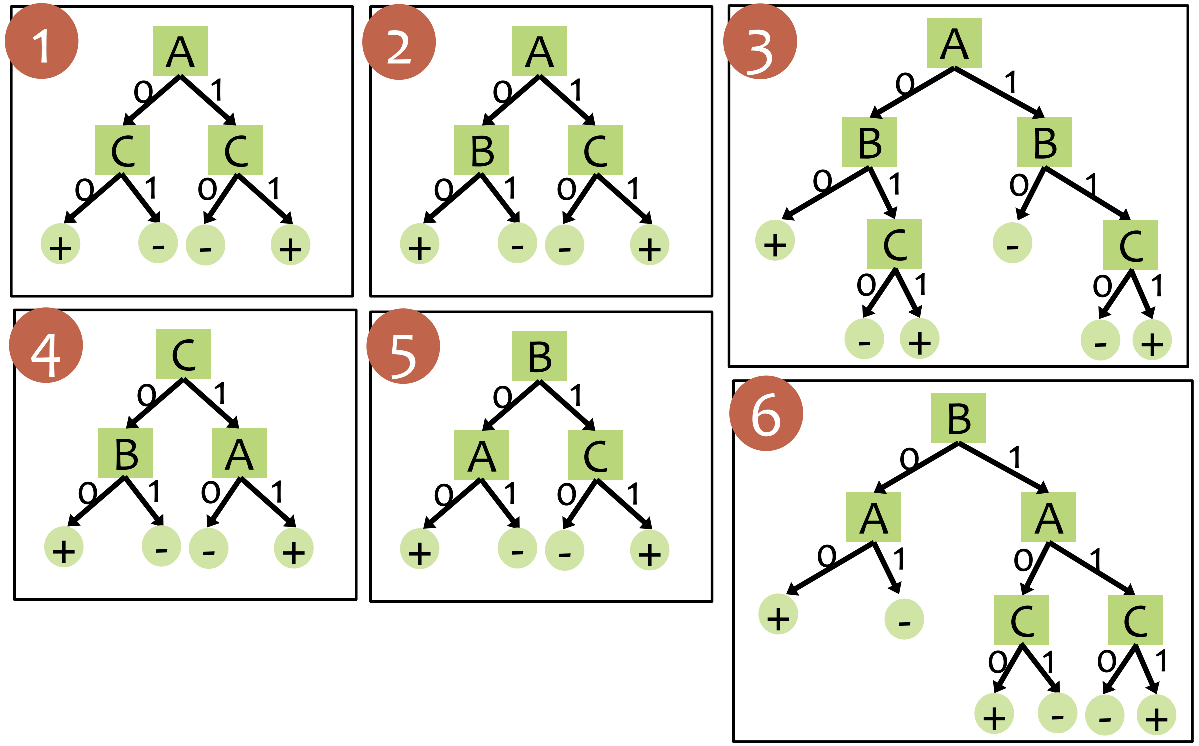
abnormal

normal

Poll 2

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

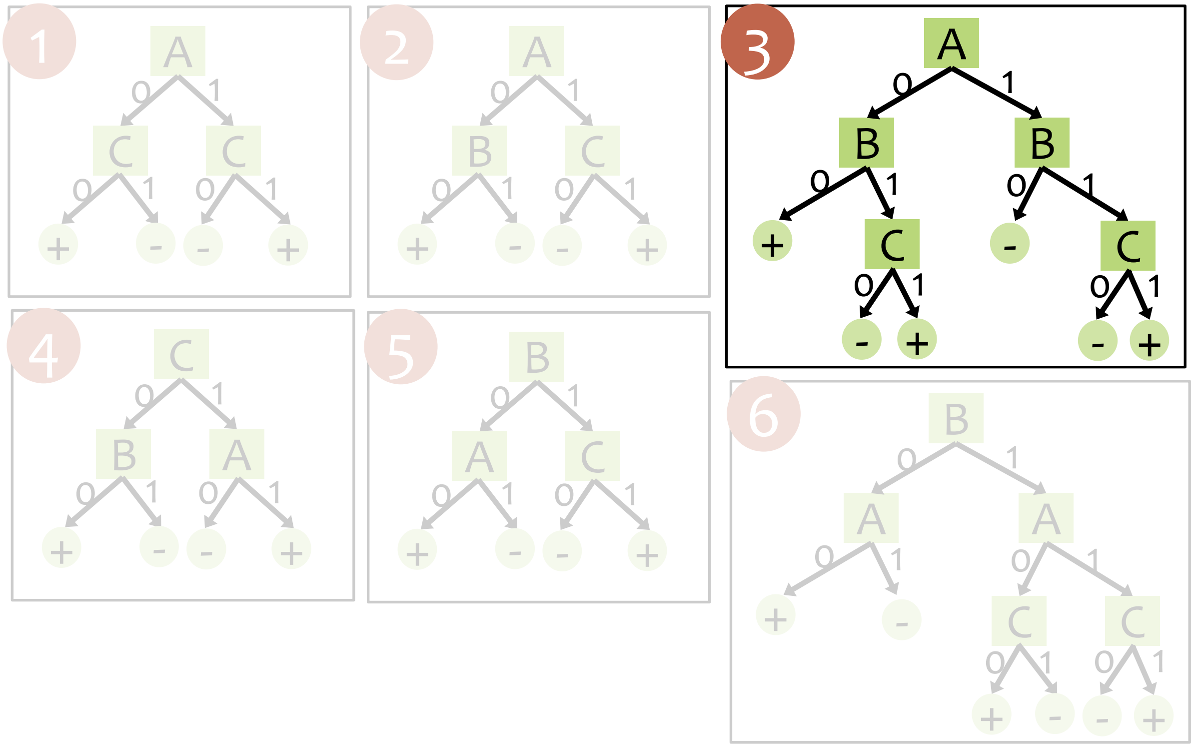
Y	A	B	C
+	0	0	0
+	0	0	1
-	0	1	0
+	0	1	1
-	1	0	0
-	1	0	1
-	1	1	0
+	1	1	1

Slide credit: CMU MLD Matt Gormley

Poll 2

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

Slide credit: CMU MLD Matt Gormley

Poll 3

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

Poll 3

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
```

→ Mutual info.

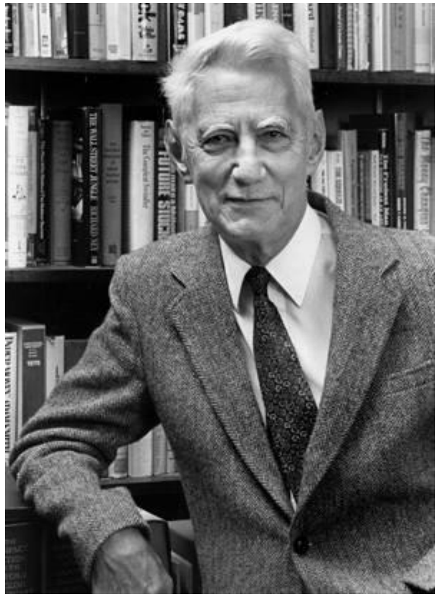
Gini impurity

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

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

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



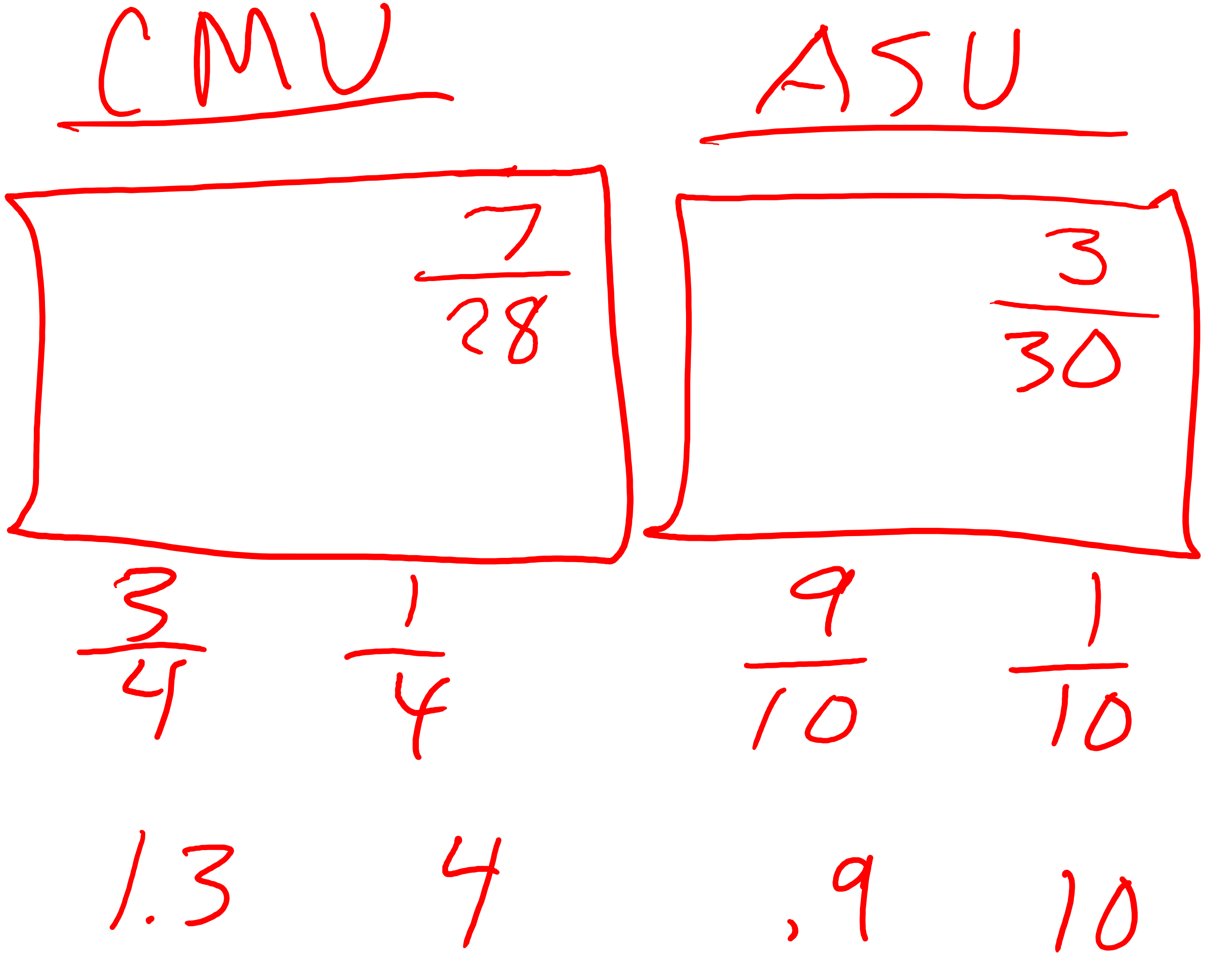
Claude Shannon (1916 – 2001), most of the work was done in Bell labs

Entropy

Definition

$$H(Y) = \sum_y p(Y=y) \log_2 \frac{1}{p(Y=y)}$$

$$H(Y) = -\sum_y p(Y=y) \log_2 p(Y=y)$$



wreport

X= rain X= sun

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

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

$I(Y;A) = H(Y) - H(Y|A)$

$I(Y;B) = H(Y) - H(Y|B)$

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

Splitting with Mutual Information

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

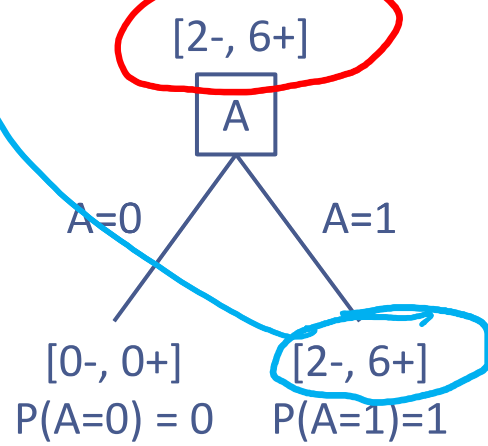
$$H(Y) = -\left[\frac{2}{8}\log_2\frac{2}{8} + \frac{6}{8}\log_2\frac{6}{8}\right]$$

$$H(Y \mid A = 0) = undefined$$

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

$$H(Y \mid A) = P(A = 0)H(Y \mid A = 0) + P(A = 1)H(Y \mid A = 1)$$
$$= 0 + H(Y \mid A = 1)$$
$$= H(Y)$$

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



Slide credit: CMU MLD Matt Gormley

Decision Tree Learning Example

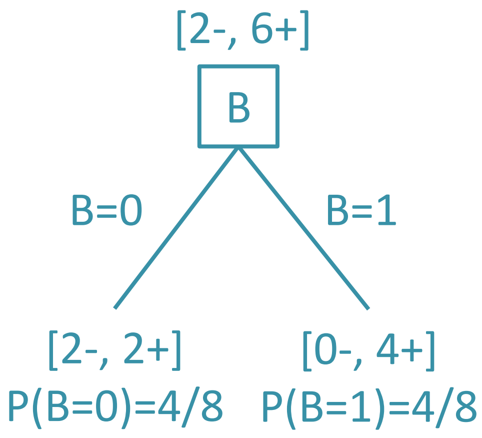
Entropy: $H(Y) = -\sum_{y \in \mathcal{Y}} P(Y = y) \log_2 P(Y = y)$

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



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

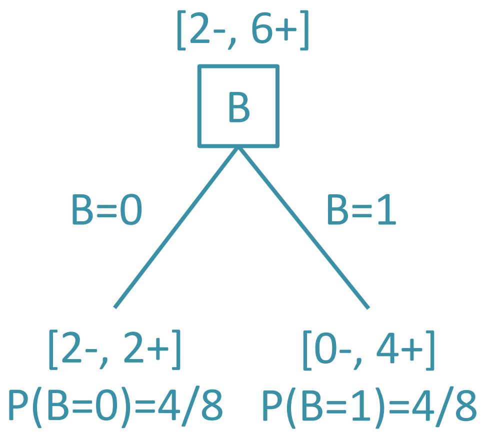
Mutual Information: $I(Y;X) = H(Y) - H(Y|X)$

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.

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