# 10-701: Introduction to Machine Learning Lecture 2 – Decision Trees

Henry Chai

1/22/24

# Front Matter

- Announcements:
  - HW1 will be released on Wednesday (1/24), due 2/2 at 11:59 PM
  - Recitations will be held on Fridays, at the same time and place as lecture
    - HW1 recitation this Friday (1/26)
  - Office hours will start this Wednesday (1/24)
- Recommended Readings:
  - Mitchell, Chapter 3: Decision Tree Learning
  - Daumé III, Chapter 1: Decision Trees

# Schedule

Lectures are the primary mode of content delivery in this course. Attending lectures is highly recommended; there will be regular in-class activities and polls which will constitute a small portion of your final grade. Engaging in these real-time activities can greatly improve your understanding of the material. Lectures will be recorded and made available to you after the fact. However, the primary purpose of these recordings is to allow you to refer back to the content; watching recordings in lieu of attending lectures is not encouraged.

| Date | Topic | Slides | Readings/Resources |
|---|---|---|---|
| Wed, 1/17 | Introduction: Notation & Problem Formulation | Lecture 1 (Inked) | |
| Mon, 1/22 | Decision Trees | Lecture 2 (Pre-class) | Mitchell, Chapter 3<br>Daumé III, Chapter 1 |
| Wed, 1/24 | KNNs & Model Selection | | |
| Mon, 1/29 | Linear Regression | | |
| Wed, 1/31 | MLE/MAP | | |
| Mon, 2/5 | Naïve Bayes | | |

# Lecture Schedule

# Recall: Our second Machine Learning Classifier

- Alright, let's actually (try to) extract a pattern from the data

| $x_1$ Family History | $x_2$ Resting Blood Pressure | $x_3$ Cholesterol | $y$ Heart Disease? |
|---|---|---|---|
| Yes | Low | Normal | No |
| No | Medium | Normal | No |
| No | Low | Abnormal | Yes |
| Yes | Medium | Normal | Yes |
| Yes | High | Abnormal | Yes |

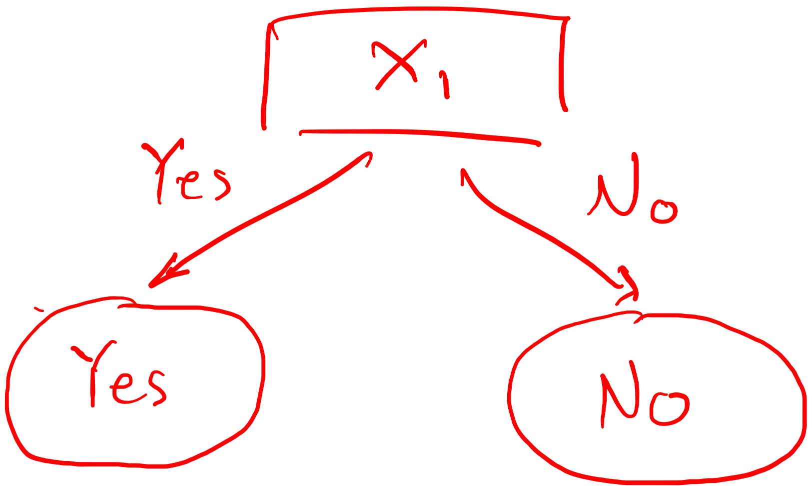
- Decision stump on $x_1$:

$$h(\boldsymbol{x}') = h(x_1', \dots, x_D') = \begin{cases} \text{“Yes” if } x_1' = \text{“Yes”} \\ \text{“No” otherwise} \end{cases}$$

# Recall: Our second Machine Learning Classifier

- Alright, let's actually (try to) extract a pattern from the data

| $x_1$ Family History | $x_2$ Resting Blood Pressure | $x_3$ Cholesterol | $y$ Heart Disease? |
|---|---|---|---|
| Yes | Low | Normal | No |
| No | Medium | Normal | No |
| No | Low | Abnormal | Yes |
| Yes | Medium | Normal | Yes |
| Yes | High | Abnormal | Yes |



# Decision Stumps: Questions

1. How can we pick which feature to split on?

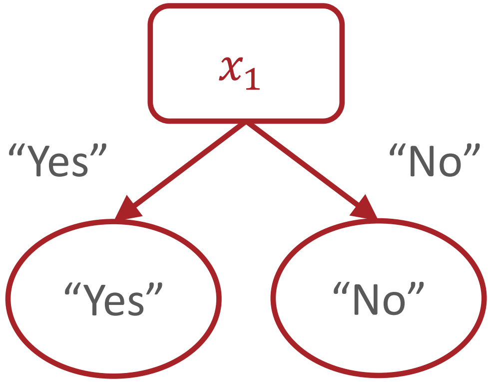
2. Why stop at just one feature?

# Splitting Criterion

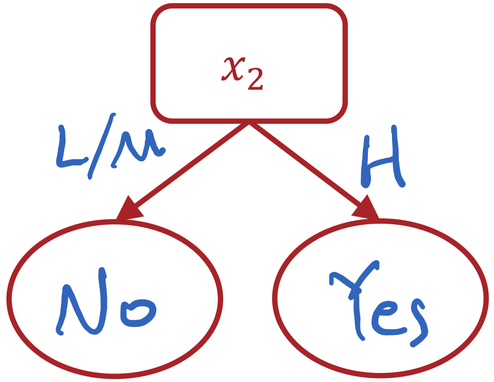
- A **splitting criterion** is a function that measures how good or useful splitting on a particular feature is *for a specified dataset*
- Idea: use the feature that optimizes the splitting criterion for our decision stump.

# Training error rate as a Splitting Criterion

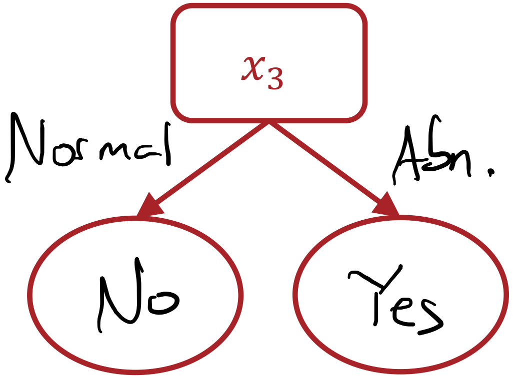
| $x_1$ Family History | $x_2$ Resting Blood Pressure | $x_3$ Cholesterol | $y$ Heart Disease? |
|---|---|---|---|
| Yes | Low | Normal | No |
| No | Medium | Normal | No |
| No | Low | Abnormal | Yes |
| Yes | Medium | Normal | Yes |
| Yes | High | Abnormal | Yes |




Training error rate: 2/5




Training error rate: 2/5

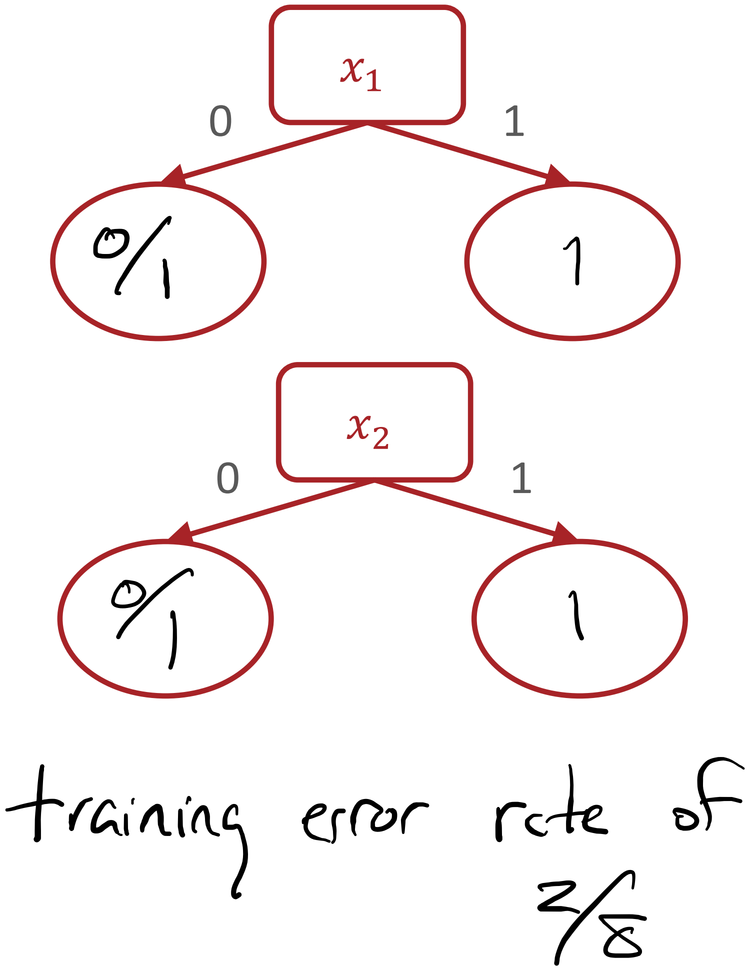



Training error rate: 1/5

# Training error rate as a Splitting Criterion?

| $x_1$ | $x_2$ | $y$ |
|---|---|---|
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |

- Which feature would you split on using training error rate as the splitting criterion?



# Splitting Criterion

- A **splitting criterion** is a function that measures how good or useful splitting on a particular feature is *for a specified dataset*
- Insight: use the feature that optimizes the splitting criterion for our decision stump.
- Potential splitting criteria:
  - Training error rate (minimize)
  - Gini impurity (minimize) $\rightarrow$ CART algorithm
  - Mutual information (maximize) $\rightarrow$ ID3 algorithm

# Splitting Criterion

- A **splitting criterion** is a function that measures how good or useful splitting on a particular feature is *for a specified dataset*
- Insight: use the feature that optimizes the splitting criterion for our decision stump.
- Potential splitting criteria:
  - Training error rate (minimize)
  - Gini impurity (minimize) $\rightarrow$ CART algorithm
  - Mutual information (maximize) $\rightarrow$ ID3 algorithm

# Entropy

- Entropy of a (discrete) random variable $X$ that takes on values in $\mathcal{X}$:

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

# Entropy

- Entropy of a collection of values $S$:

$|\cdot|$ = cardinality of a collection

$$H(S) = -\sum_{v \in V(S)} \frac{|S_v|}{|S|} \log_2\left(\frac{|S_v|}{|S|}\right)$$

where $V(S)$ is the set of unique values in $S$

$S_v$ is the collection of elements in $S$ with value $v$

- If all the elements in $S$ are the same, then

$$H(S) = -\frac{N}{N}\log_2\frac{N}{N}$$

$$= -1 \log_2(1) = 0$$

# Entropy

- Entropy of a collection of values $S$:

$$H(S) = -\sum_{v \in V(S)} \frac{|S_v|}{|S|} \log_2\left(\frac{|S_v|}{|S|}\right)$$

where $V(S)$ is the set of unique values in $S$

$S_v$ is the collection of elements in $S$ with value $v$

- If $S$ is split fifty-fifty between two values, then

$$H(S) = -\frac{(N/2)}{N}\log_2\left(\frac{N/2}{N}\right) - \frac{(N/2)}{N}\log_2\left(\frac{N/2}{N}\right)$$

$$= -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = \frac{1}{2} + \frac{1}{2} = 1$$

# Mutual Information

- Mutual information between two random variables $X$ and $Y$ describes how much clarity about the value of one variable is gained by observing the other

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

# Mutual Information

- Mutual information can be used to compute how much information or clarity a particular feature provides about the label

$$I(Y; x_d) = H(Y) - \sum_{v \in V(x_d)} (f_v)\left(H(Y_{x_d=v})\right)$$

where $x_d$ is a feature

$Y$ is the collection of all labels

$V(x_d)$ is the set of unique values of $x_d$

→ $f_v$ is the fraction of inputs where $x_d = v$

$Y_{x_d=v}$ is the collection of labels where $x_d = v$

# Mutual Information: Example

| $x_d$ | $y$ |
|---|---|
| 1 | 1 |
| 1 | 1 |
| 0 | 0 |
| 0 | 0 |

2/4 (rows with $x_d = 1$); 2/4 (rows with $x_d = 0$)

$$I(Y; x_d) = H(Y) - \sum_{v \in V(x_d)} f_v H(Y_{x_d = v})$$

$$= 1 - \frac{1}{2} H(Y_{x_d=0}) - \frac{1}{2} H(Y_{x_d=1})$$

$$= 1 - \frac{1}{2}(0) - \frac{1}{2}(0) = 1$$

# Mutual Information: Example

| $x_d$ | $y$ |
|---|---|
| 1 | 1 |
| 0 | 1 |
| 1 | 0 |
| 0 | 0 |

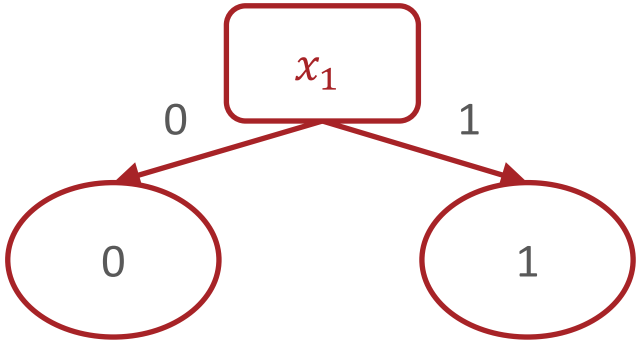
$$I(Y; x_d) = 1 - \frac{1}{2} H(Y_{x_d=0}) - \frac{1}{2} H(Y_{x_d=1})$$

$$= 1 - \frac{1}{2}(1) - \frac{1}{2}(1) = 0$$

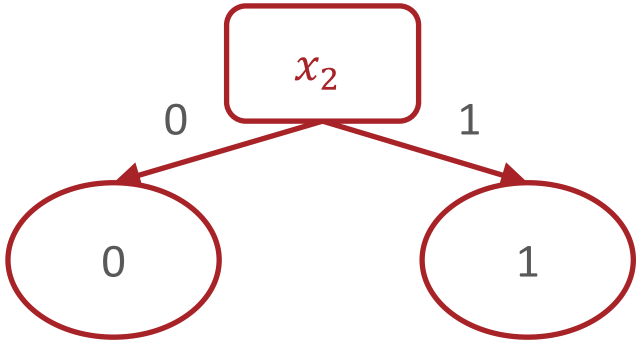
# Mutual Information as a Splitting Criterion

| $x_1$ | $x_2$ | $y$ |
|---|---|---|
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |

- Which feature would you split on using mutual information as the splitting criterion?




Mutual Information: 0




Mutual Information: $\left(-\frac{2}{8}\log_2\frac{2}{8}-\frac{6}{8}\log_2\frac{6}{8}\right)-\left(\frac{1}{2}(1)-\frac{1}{2}(0)\right)\approx 0.31$

# Decision Stumps: Questions

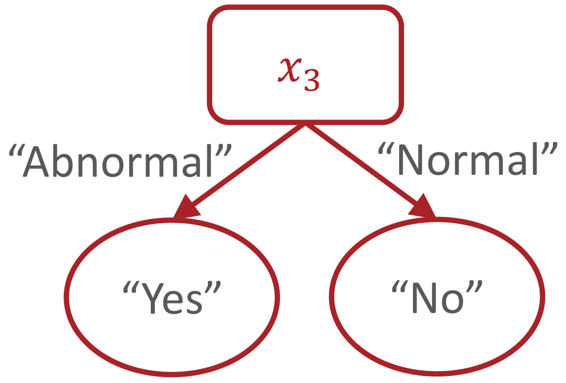
1. How can we pick which feature to split on?

   Mutual information

2. Why stop at just one feature?

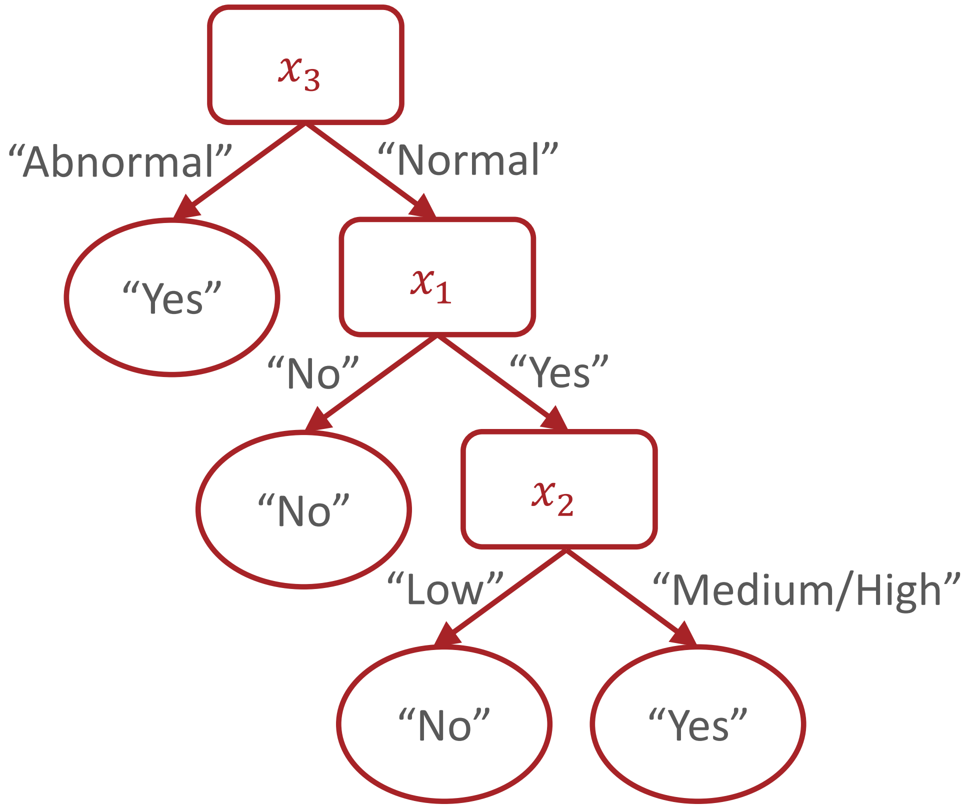
# From Decision Stump …

| $x_1$ Family History | $x_2$ Resting Blood Pressure | $x_3$ Cholesterol | $y$ Heart Disease? |
|---|---|---|---|
| Yes | Low | Normal | No |
| No | Medium | Normal | No |
| No | Low | Abnormal | Yes |
| Yes | Medium | Normal | Yes |
| Yes | High | Abnormal | Yes |



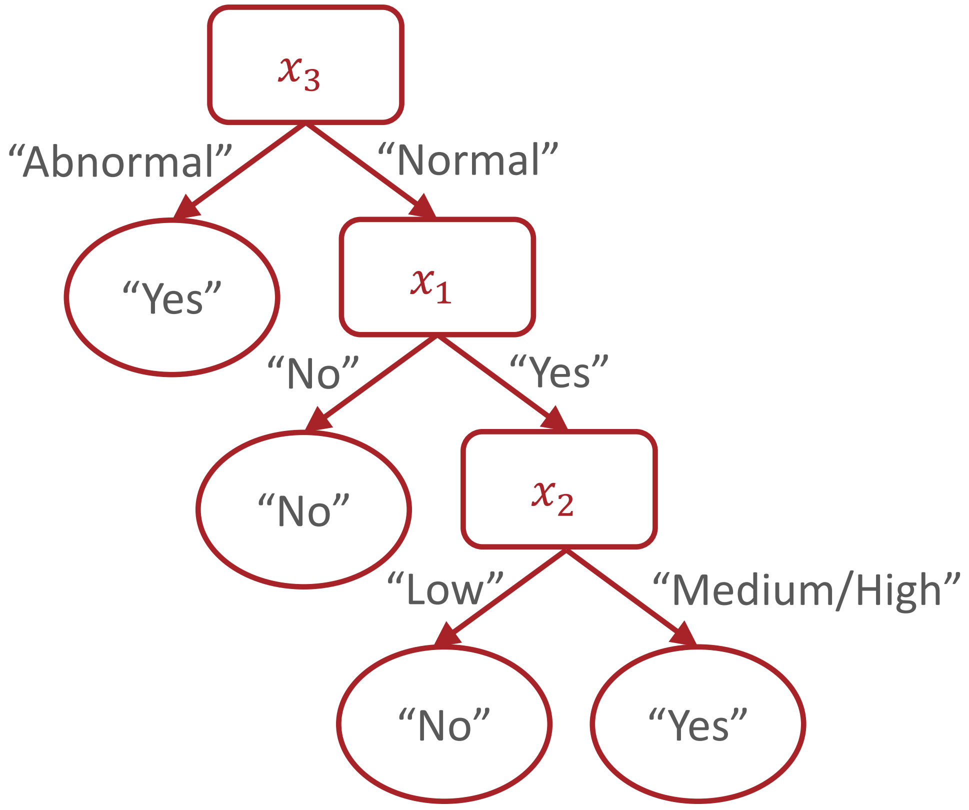
# From Decision Stump to Decision Tree

| $x_1$ Family History | $x_2$ Resting Blood Pressure | $x_3$ Cholesterol | $y$ Heart Disease? |
|---|---|---|---|
| Yes | Low | Normal | No |
| No | Medium | Normal | No |
| No | Low | Abnormal | Yes |
| Yes | Medium | Normal | Yes |
| Yes | High | Abnormal | Yes |



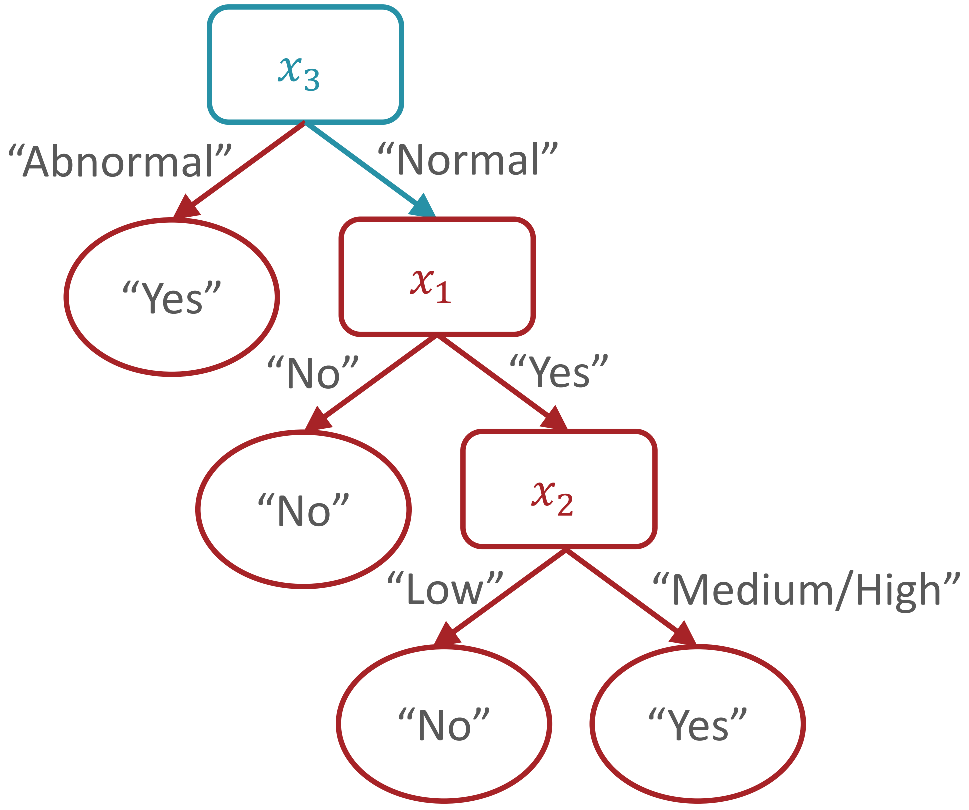
# From Decision Stump to Decision Tree

| $x_1$ Family History | $x_2$ Resting Blood Pressure | $x_3$ Cholesterol | $y$ Heart Disease? |
|---|---|---|---|
| Yes | Low | Normal | No |
| No | Medium | Normal | No |
| No | Low | Abnormal | Yes |
| Yes | Medium | Normal | Yes |
| Yes | High | Abnormal | Yes |
| | | | |
| No | High | Normal | No |



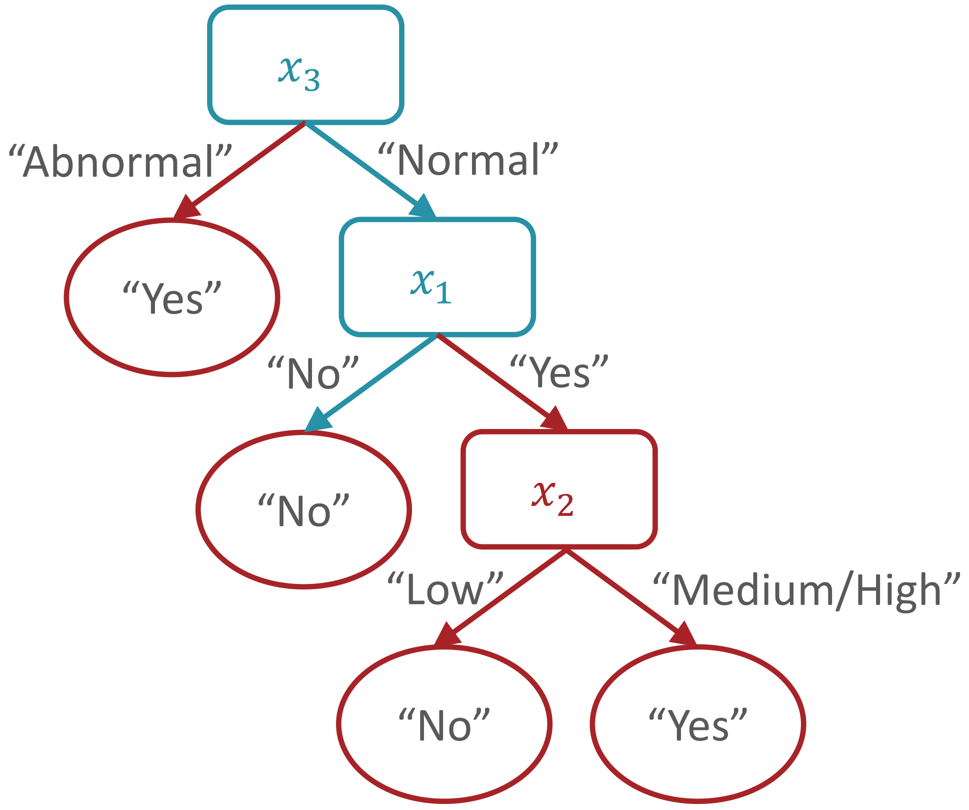
# From Decision Stump to Decision Tree

| $x_1$ Family History | $x_2$ Resting Blood Pressure | $x_3$ Cholesterol | $y$ Heart Disease? |
|---|---|---|---|
| Yes | Low | Normal | No |
| No | Medium | Normal | No |
| No | Low | Abnormal | Yes |
| Yes | Medium | Normal | Yes |
| Yes | High | Abnormal | Yes |
| No | High | Normal | No |



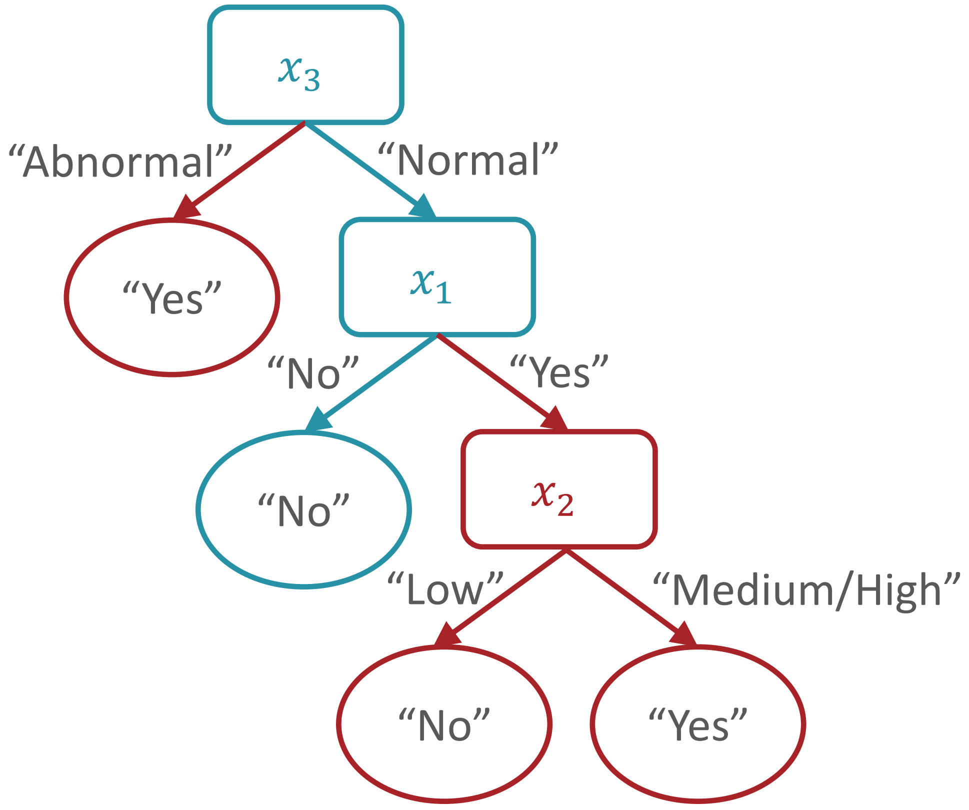
# From Decision Stump to Decision Tree

| $x_1$ Family History | $x_2$ Resting Blood Pressure | $x_3$ Cholesterol | $y$ Heart Disease? |
|---|---|---|---|
| Yes | Low | Normal | No |
| No | Medium | Normal | No |
| No | Low | Abnormal | Yes |
| Yes | Medium | Normal | Yes |
| Yes | High | Abnormal | Yes |
| No | High | Normal | No |



# From Decision Stump to Decision Tree

| $x_1$ Family History | $x_2$ Resting Blood Pressure | $x_3$ Cholesterol | $y$ Heart Disease? |
|---|---|---|---|
| Yes | Low | Normal | No |
| No | Medium | Normal | No |
| No | Low | Abnormal | Yes |
| Yes | Medium | Normal | Yes |
| Yes | High | Abnormal | Yes |
| No | High | Normal | No |



# Decision Tree Prediction: Pseudocode

$x' = [x'_1, x'_2, \ldots, x'_D]$

```
def predict(x′):
# walk from the root node to a leaf node

while (true):
    if current_node is internal:
        check the associated feature, x_d
        go down branch corresponding to x'_d
    else (current_node is a leaf):
        return label associated with current_node
```

# Decision Tree Learning: Pseudocode

```
def train(𝒟):
    store root = tree_recurse(𝒟)
def tree_recurse(𝒟′):
    q = new node()
    base case – if (SOME CONDITION):
    recursion – else:
        find the best feature to split on, x_d
        q.split = x_d
        for v in V(x_d), all possible values of x_d:
            D_v = {(x^(n), y^(n)) ∈ D′ | x_d^(n) = v}
            q.children(v) = tree_recurse(D_v)
    return q
```

# Decision Tree Learning: Pseudocode

```
def train(𝒟):
    store root = tree_recurse(𝒟)
def tree_recurse(𝒟′):
    q = new node()
    base case – if (all labels in D′ are the same
        OR if the tree is too deep
        OR if the label set has low enough entropy
        OR D′ is empty
        OR all feature vectors in D′ are
                                identical)
    recursion – else:
    return q
```

# Decision Tree: Example (Iteratively)

- How is Henry getting to work?
- Label: mode of transportation
  - $y \in \mathcal{Y} = \{\text{Bike}, \text{Drive}, \text{Bus}\}$
- Features: 4 categorial features
  - Is it raining? $x_1 \in \{\text{Rain}, \text{No Rain}\}$
  - When am I leaving (relative to rush hour)? $x_2 \in \{\text{Before}, \text{During}, \text{After}\}$
  - What am I bringing? $x_3 \in \{\text{Backpack}, \text{Lunchbox}, \text{Both}\}$
  - Am I tired? $x_4 \in \{\text{Tired}, \text{Not Tired}\}$

# Data

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | Before | Both | Tired | Drive |
| Rain | During | Both | Not Tired | Bus |
| Rain | During | Both | Tired | Drive |
| Rain | After | Backpack | Not Tired | Bus |
| Rain | After | Backpack | Tired | Bus |
| Rain | After | Lunchbox | Tired | Drive |
| No Rain | Before | Backpack | Tired | Bike |
| No Rain | Before | Lunchbox | Not Tired | Bus |
| No Rain | Before | Lunchbox | Tired | Drive |
| No Rain | During | Backpack | Not Tired | Bus |
| No Rain | During | Both | Tired | Drive |
| No Rain | After | Backpack | Not Tired | Bike |
| No Rain | After | Backpack | Tired | Bike |
| No Rain | After | Both | Not Tired | Bus |
| No Rain | After | Both | Tired | Drive |
| No Rain | After | Lunchbox | Not Tired | Bus |

Which feature would we split on first using mutual information as the splitting criterion?

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | Before | Both | Tired | Drive |
| Rain | During | Both | Not Tired | Bus |
| Rain | During | Both | Tired | Drive |
| Rain | After | Backpack | Not Tired | Bus |
| Rain | After | Backpack | Tired | Bus |
| Rain | After | Lunchbox | Tired | Drive |
| No Rain | Before | Backpack | Tired | Bike |
| No Rain | Before | Lunchbox | Not Tired | Bus |
| No Rain | Before | Lunchbox | Tired | Drive |
| No Rain | During | Backpack | Not Tired | Bus |
| No Rain | During | Both | Tired | Drive |
| No Rain | After | Backpack | Not Tired | Bike |
| No Rain | After | Backpack | Tired | Bike |
| No Rain | After | Both | Not Tired | Bus |
| No Rain | After | Both | Tired | Drive |
| No Rain | After | Lunchbox | Not Tired | Bus |

Recall:

$$H(S) = -\sum_{v \in V(S)} \frac{|S_v|}{|S|} \log_2\left(\frac{|S_v|}{|S|}\right)$$

$H(Y)$

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | Before | Both | Tired | Drive |
| Rain | During | Both | Not Tired | Bus |
| Rain | During | Both | Tired | Drive |
| Rain | After | Backpack | Not Tired | Bus |
| Rain | After | Backpack | Tired | Bus |
| Rain | After | Lunchbox | Tired | Drive |
| No Rain | Before | Backpack | Tired | Bike |
| No Rain | Before | Lunchbox | Not Tired | Bus |
| No Rain | Before | Lunchbox | Tired | Drive |
| No Rain | During | Backpack | Not Tired | Bus |
| No Rain | During | Both | Tired | Drive |
| No Rain | After | Backpack | Not Tired | Bike |
| No Rain | After | Backpack | Tired | Bike |
| No Rain | After | Both | Not Tired | Bus |
| No Rain | After | Both | Tired | Drive |
| No Rain | After | Lunchbox | Not Tired | Bus |

Recall:

$$H(S) = -\sum_{v \in V(S)} \frac{|S_v|}{|S|} \log_2\left(\frac{|S_v|}{|S|}\right)$$

$$H(Y) = -\frac{3}{16}\log_2\frac{3}{16} - \frac{7}{16}\log_2\frac{7}{16} - \frac{6}{16}\log_2\frac{6}{16}$$

$$\approx 1.5052$$

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | Before | Both | Tired | Drive |
| Rain | During | Both | Not Tired | Bus |
| Rain | During | Both | Tired | Drive |
| Rain | After | Backpack | Not Tired | Bus |
| Rain | After | Backpack | Tired | Bus |
| Rain | After | Lunchbox | Tired | Drive |
| No Rain | Before | Backpack | Tired | Bike |
| No Rain | Before | Lunchbox | Not Tired | Bus |
| No Rain | Before | Lunchbox | Tired | Drive |
| No Rain | During | Backpack | Not Tired | Bus |
| No Rain | During | Both | Tired | Drive |
| No Rain | After | Backpack | Not Tired | Bike |
| No Rain | After | Backpack | Tired | Bike |
| No Rain | After | Both | Not Tired | Bus |
| No Rain | After | Both | Tired | Drive |
| No Rain | After | Lunchbox | Not Tired | Bus |

Recall: $I(x_d; Y) = H(Y)$

$$- \sum_{v \in V(x_d)} (f_v)\left(H(Y_{x_d=v})\right)$$

$I(x_1, Y) =$

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | Before | Both | Tired | Drive |
| Rain | During | Both | Not Tired | Bus |
| Rain | During | Both | Tired | Drive |
| Rain | After | Backpack | Not Tired | Bus |
| Rain | After | Backpack | Tired | Bus |
| Rain | After | Lunchbox | Tired | Drive |
| No Rain | Before | Backpack | Tired | Bike |
| No Rain | Before | Lunchbox | Not Tired | Bus |
| No Rain | Before | Lunchbox | Tired | Drive |
| No Rain | During | Backpack | Not Tired | Bus |
| No Rain | During | Both | Tired | Drive |
| No Rain | After | Backpack | Not Tired | Bike |
| No Rain | After | Backpack | Tired | Bike |
| No Rain | After | Both | Not Tired | Bus |
| No Rain | After | Both | Tired | Drive |
| No Rain | After | Lunchbox | Not Tired | Bus |

Recall: $I(x_d; Y) = H(Y)$

$$- \sum_{v \in V(x_d)} (f_v)\left(H\left(Y_{x_d=v}\right)\right)$$

$I(x_1, Y) \approx 1.5052$

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | Before | Both | Tired | Drive |
| Rain | During | Both | Not Tired | Bus |
| Rain | During | Both | Tired | Drive |
| Rain | After | Backpack | Not Tired | Bus |
| Rain | After | Backpack | Tired | Bus |
| Rain | After | Lunchbox | Tired | Drive |
| No Rain | Before | Backpack | Tired | Bike |
| No Rain | Before | Lunchbox | Not Tired | Bus |
| No Rain | Before | Lunchbox | Tired | Drive |
| No Rain | During | Backpack | Not Tired | Bus |
| No Rain | During | Both | Tired | Drive |
| No Rain | After | Backpack | Not Tired | Bike |
| No Rain | After | Backpack | Tired | Bike |
| No Rain | After | Both | Not Tired | Bus |
| No Rain | After | Both | Tired | Drive |
| No Rain | After | Lunchbox | Not Tired | Bus |

Recall: $I(x_d; Y) = H(Y)$

$$-\sum_{v \in V(x_d)} (f_v)\left(H\left(Y_{x_d=v}\right)\right)$$

$I(x_1, Y) \approx 1.5052$

$$-\frac{6}{16}(1)$$

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | Before | Both | Tired | Drive |
| Rain | During | Both | Not Tired | Bus |
| Rain | During | Both | Tired | Drive |
| Rain | After | Backpack | Not Tired | Bus |
| Rain | After | Backpack | Tired | Bus |
| Rain | After | Lunchbox | Tired | Drive |
| No Rain | Before | Backpack | Tired | Bike |
| No Rain | Before | Lunchbox | Not Tired | Bus |
| No Rain | Before | Lunchbox | Tired | Drive |
| No Rain | During | Backpack | Not Tired | Bus |
| No Rain | During | Both | Tired | Drive |
| No Rain | After | Backpack | Not Tired | Bike |
| No Rain | After | Backpack | Tired | Bike |
| No Rain | After | Both | Not Tired | Bus |
| No Rain | After | Both | Tired | Drive |
| No Rain | After | Lunchbox | Not Tired | Bus |

Recall: $I(x_d; Y) = H(Y)$

$$- \sum_{v \in V(x_d)} (f_v)\left(H\left(Y_{x_d=v}\right)\right)$$

$I(x_1, Y) \approx 1.5052$

$$-\frac{6}{16}(1)$$

$$-\frac{10}{16}\left(-\frac{3}{10}\log_2\frac{3}{10} - \frac{4}{10}\log_2\frac{4}{10} - \frac{3}{10}\log_2\frac{3}{10}\right)$$

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | Before | Both | Tired | Drive |
| Rain | During | Both | Not Tired | Bus |
| Rain | During | Both | Tired | Drive |
| Rain | After | Backpack | Not Tired | Bus |
| Rain | After | Backpack | Tired | Bus |
| Rain | After | Lunchbox | Tired | Drive |
| No Rain | Before | Backpack | Tired | Bike |
| No Rain | Before | Lunchbox | Not Tired | Bus |
| No Rain | Before | Lunchbox | Tired | Drive |
| No Rain | During | Backpack | Not Tired | Bus |
| No Rain | During | Both | Tired | Drive |
| No Rain | After | Backpack | Not Tired | Bike |
| No Rain | After | Backpack | Tired | Bike |
| No Rain | After | Both | Not Tired | Bus |
| No Rain | After | Both | Tired | Drive |
| No Rain | After | Lunchbox | Not Tired | Bus |

Recall: $I(x_d; Y) = H(Y)$

$$- \sum_{v \in V(x_d)} (f_v)\left(H\left(Y_{x_d=v}\right)\right)$$

$I(x_1, Y) \approx 1.5052$

$$-\frac{6}{16}(1)$$

$$-\frac{10}{16}(1.5710)$$

$\approx 0.1482$

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | Before | Both | Tired | Drive |
| Rain | During | Both | Not Tired | Bus |
| Rain | During | Both | Tired | Drive |
| Rain | After | Backpack | Not Tired | Bus |
| Rain | After | Backpack | Tired | Bus |
| Rain | After | Lunchbox | Tired | Drive |
| No Rain | Before | Backpack | Tired | Bike |
| No Rain | Before | Lunchbox | Not Tired | Bus |
| No Rain | Before | Lunchbox | Tired | Drive |
| No Rain | During | Backpack | Not Tired | Bus |
| No Rain | During | Both | Tired | Drive |
| No Rain | After | Backpack | Not Tired | Bike |
| No Rain | After | Backpack | Tired | Bike |
| No Rain | After | Both | Not Tired | Bus |
| No Rain | After | Both | Tired | Drive |
| No Rain | After | Lunchbox | Not Tired | Bus |

Recall: $I(x_d; Y) = H(Y) - \sum_{v \in V(x_d)} (f_v)\left(H\left(Y_{x_d=v}\right)\right)$

| $I(x_d, Y)$ | |
|---|---|
| $x_1$ | 0.1482 |
| $x_2$ | 0.1302 |
| $x_3$ | 0.5358 |
| $x_4$ | 0.5576 |

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | Before | Both | Tired | Drive |
| Rain | During | Both | Not Tired | Bus |
| Rain | During | Both | Tired | Drive |
| Rain | After | Backpack | Not Tired | Bus |
| Rain | After | Backpack | Tired | Bus |
| Rain | After | Lunchbox | Tired | Drive |
| No Rain | Before | Backpack | Tired | Bike |
| No Rain | Before | Lunchbox | Not Tired | Bus |
| No Rain | Before | Lunchbox | Tired | Drive |
| No Rain | During | Backpack | Not Tired | Bus |
| No Rain | During | Both | Tired | Drive |
| No Rain | After | Backpack | Not Tired | Bike |
| No Rain | After | Backpack | Tired | Bike |
| No Rain | After | Both | Not Tired | Bus |
| No Rain | After | Both | Tired | Drive |
| No Rain | After | Lunchbox | Not Tired | Bus |

Recall: $I(x_d; Y) = H(Y) - \sum_{v \in V(x_d)} (f_v)\left(H\left(Y_{x_d=v}\right)\right)$

| $I(x_d, Y)$ | |
|---|---|
| $x_1$ | 0.1482 |
| $x_2$ | 0.1302 |
| $x_3$ | 0.5358 |
| $x_4$ | 0.5576 |

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | During | Both | Not Tired | Bus |
| Rain | After | Backpack | Not Tired | Bus |
| No Rain | Before | Lunchbox | Not Tired | Bus |
| No Rain | During | Backpack | Not Tired | Bus |
| No Rain | After | Backpack | Not Tired | Bike |
| No Rain | After | Both | Not Tired | Bus |
| No Rain | After | Lunchbox | Not Tired | Bus |
| Rain | Before | Both | Tired | Drive |
| Rain | During | Both | Tired | Drive |
| Rain | After | Backpack | Tired | Bus |
| Rain | After | Lunchbox | Tired | Drive |
| No Rain | Before | Backpack | Tired | Bike |
| No Rain | Before | Lunchbox | Tired | Drive |
| No Rain | During | Both | Tired | Drive |
| No Rain | After | Backpack | Tired | Bike |
| No Rain | After | Both | Tired | Drive |

Recall: $I(x_d; Y) = H(Y) - \sum_{v \in V(x_d)} (f_v)\left(H\left(Y_{x_d=v}\right)\right)$

| $I(x_d, Y)$ | |
|---|---|
| $x_1$ | 0.1482 |
| $x_2$ | 0.1302 |
| $x_3$ | 0.5358 |
| $x_4$ | 0.5576 |

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | During | Both | Not Tired | Bus |
| Rain | After | Backpack | Not Tired | Bus |
| No Rain | Before | Lunchbox | Not Tired | Bus |
| No Rain | During | Backpack | Not Tired | Bus |
| No Rain | After | Backpack | Not Tired | Bike |
| No Rain | After | Both | Not Tired | Bus |
| No Rain | After | Lunchbox | Not Tired | Bus |
| Rain | Before | Both | Tired | Drive |
| Rain | During | Both | Tired | Drive |
| Rain | After | Backpack | Tired | Metro |
| Rain | After | Lunchbox | Tired | Drive |
| No Rain | Before | Backpack | Tired | Bike |
| No Rain | Before | Lunchbox | Tired | Drive |
| No Rain | During | Both | Tired | Drive |
| No Rain | After | Backpack | Tired | Bike |
| No Rain | After | Both | Tired | Drive |

Recall: $I(x_d; Y) = H(Y) - \sum_{v \in V(x_d)} (f_v)\left(H(Y_{x_d=v})\right)$

| $I(x_d, Y)$ | |
|---|---|
| $x_1$ | 0.1482 |
| $x_2$ | 0.1302 |
| $x_3$ | 0.5358 |
| $x_4$ | 0.5576 |

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | During | Both | Not Tired | Bus |
| Rain | After | Backpack | Not Tired | Bus |
| No Rain | Before | Lunchbox | Not Tired | Bus |
| No Rain | During | Backpack | Not Tired | Bus |
| No Rain | After | Backpack | Not Tired | Bike |
| No Rain | After | Both | Not Tired | Bus |
| No Rain | After | Lunchbox | Not Tired | Bus |
| Rain | Before | Both | Tired | Drive |
| Rain | During | Both | Tired | Drive |
| Rain | After | Backpack | Tired | Bus |
| Rain | After | Lunchbox | Tired | Drive |
| No Rain | Before | Backpack | Tired | Bike |
| No Rain | Before | Lunchbox | Tired | Drive |
| No Rain | During | Both | Tired | Drive |
| No Rain | After | Backpack | Tired | Bike |
| No Rain | After | Both | Tired | Drive |

# Decision Tree: Example

$x_4$

Not Tired

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | During | Both | Not Tired | Bus |
| Rain | After | Backpack | Not Tired | Bus |
| No Rain | Before | Lunchbox | Not Tired | Bus |
| No Rain | During | Backpack | Not Tired | Bus |
| No Rain | After | Backpack | Not Tired | Bike |
| No Rain | After | Both | Not Tired | Bus |
| No Rain | After | Lunchbox | Not Tired | Bus |

Tired

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | Before | Both | Tired | Drive |
| Rain | During | Both | Tired | Drive |
| Rain | After | Backpack | Tired | Bus |
| Rain | After | Lunchbox | Tired | Drive |
| No Rain | Before | Backpack | Tired | Bike |
| No Rain | Before | Lunchbox | Tired | Drive |
| No Rain | During | Both | Tired | Drive |
| No Rain | After | Backpack | Tired | Bike |
| No Rain | After | Both | Tired | Drive |

$x_4$

Not Tired

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | During | Both | Not Tired | Bus |
| Rain | After | Backpack | Not Tired | Bus |
| No Rain | Before | Lunchbox | Not Tired | Bus |
| No Rain | During | Backpack | Not Tired | Bus |
| No Rain | After | Backpack | Not Tired | Bike |
| No Rain | After | Both | Not Tired | Bus |
| No Rain | After | Lunchbox | Not Tired | Bus |

Tired

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | Before | Both | Tired | Drive |
| Rain | During | Both | Tired | Drive |
| Rain | After | Backpack | Tired | Bus |
| Rain | After | Lunchbox | Tired | Drive |
| No Rain | Before | Backpack | Tired | Bike |
| No Rain | Before | Lunchbox | Tired | Drive |
| No Rain | During | Both | Tired | Drive |
| No Rain | After | Backpack | Tired | Bike |
| No Rain | After | Both | Tired | Drive |

$x_4$

Not Tired

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | During | Both | Not Tired | Bus |
| Rain | After | Backpack | Not Tired | Bus |
| No Rain | Before | Lunchbox | Not Tired | Bus |
| No Rain | During | Backpack | Not Tired | Bus |
| No Rain | After | Backpack | Not Tired | Bike |
| No Rain | After | Both | Not Tired | Bus |
| No Rain | After | Lunchbox | Not Tired | Bus |

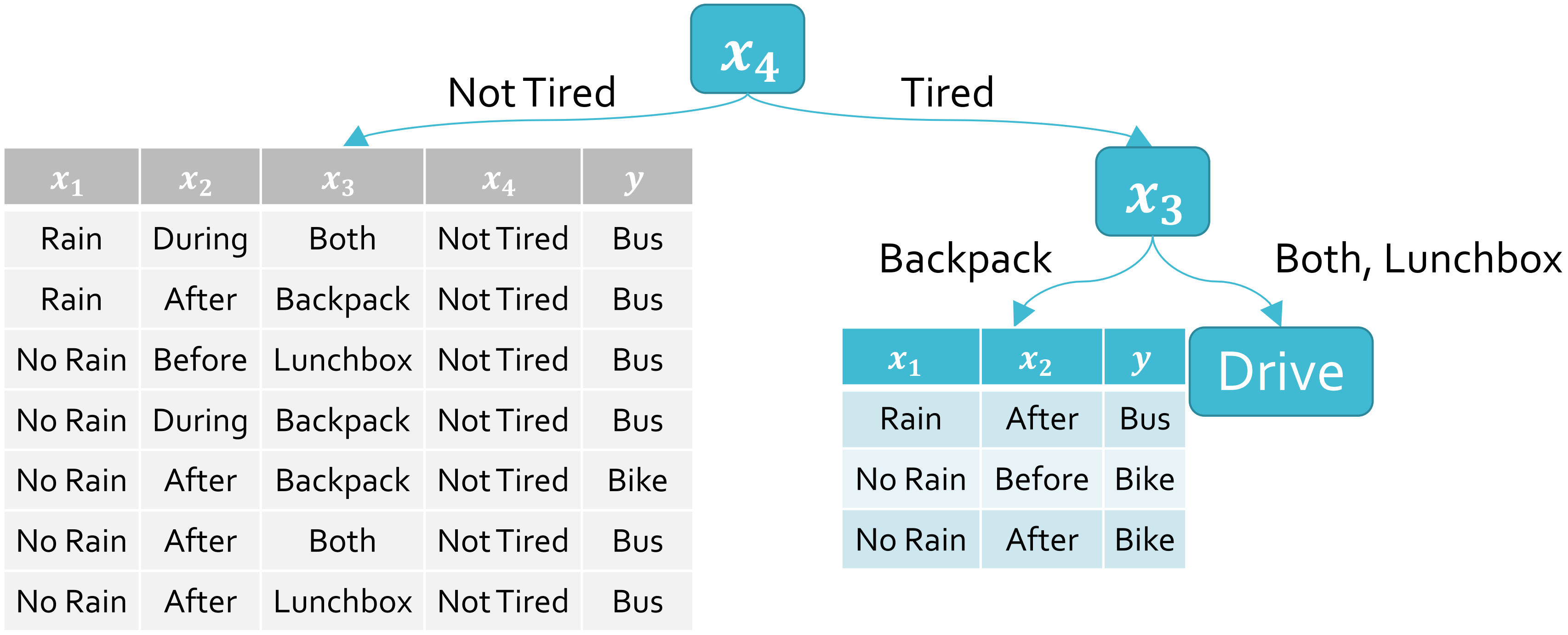
Tired

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | After | Backpack | Tired | Bus |
| No Rain | Before | Backpack | Tired | Bike |
| No Rain | After | Backpack | Tired | Bike |
| Rain | Before | Both | Tired | Drive |
| Rain | During | Both | Tired | Drive |
| No Rain | During | Both | Tired | Drive |
| No Rain | After | Both | Tired | Drive |
| Rain | After | Lunchbox | Tired | Drive |
| No Rain | Before | Lunchbox | Tired | Drive |

$$I\left(x_1, Y_{x_4=\text{Tired}}\right) \approx 0.3244$$

$$I\left(x_2, Y_{x_4=\text{Tired}}\right) \approx 0.2516$$

$$I\left(x_3, Y_{x_4=\text{Tired}}\right) \approx \mathbf{0.9183}$$



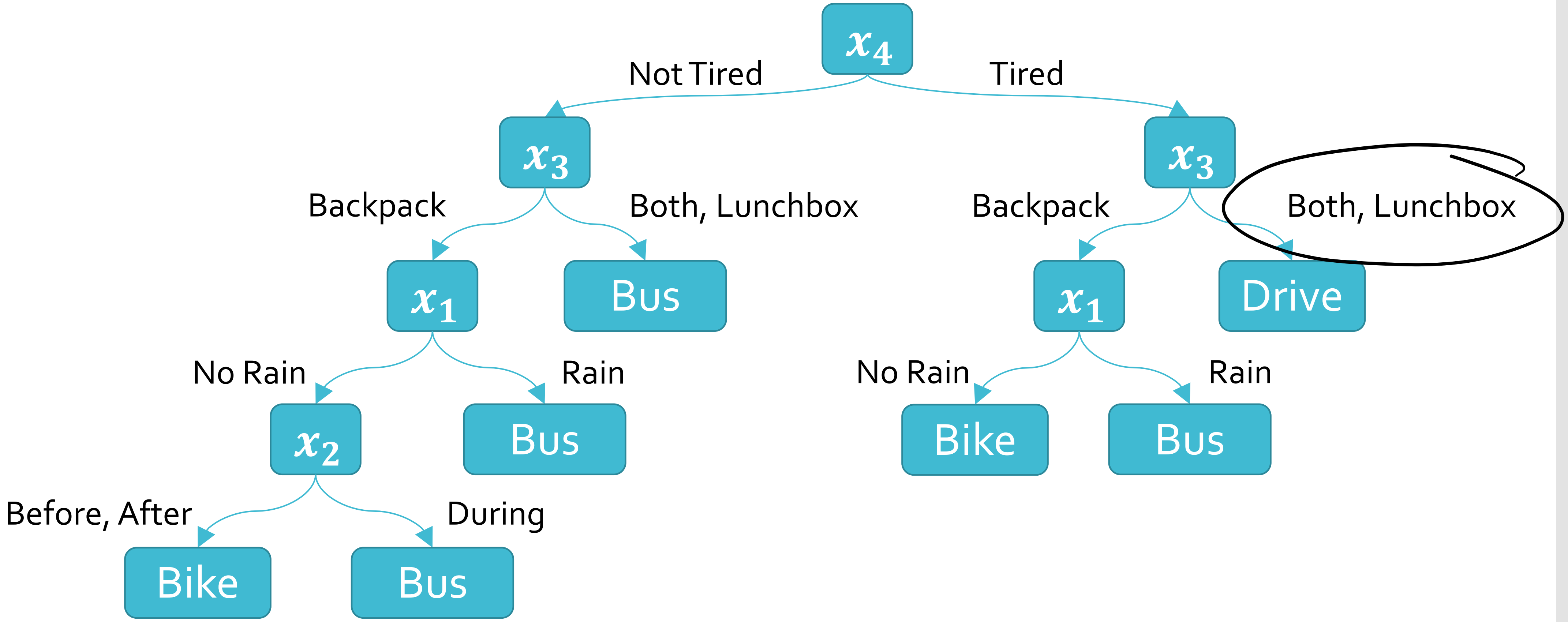

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | During | Both | Not Tired | Bus |
| Rain | After | Backpack | Not Tired | Bus |
| No Rain | Before | Lunchbox | Not Tired | Bus |
| No Rain | During | Backpack | Not Tired | Bus |
| No Rain | After | Backpack | Not Tired | Bike |
| No Rain | After | Both | Not Tired | Bus |
| No Rain | After | Lunchbox | Not Tired | Bus |

| $x_1$ | $x_2$ | $y$ |
|---|---|---|
| Rain | After | Bus |
| No Rain | Before | Bike |
| No Rain | After | Bike |

$$I\left(x_1, Y_{x_4=\text{Tired}}\right) \approx 0.3244$$

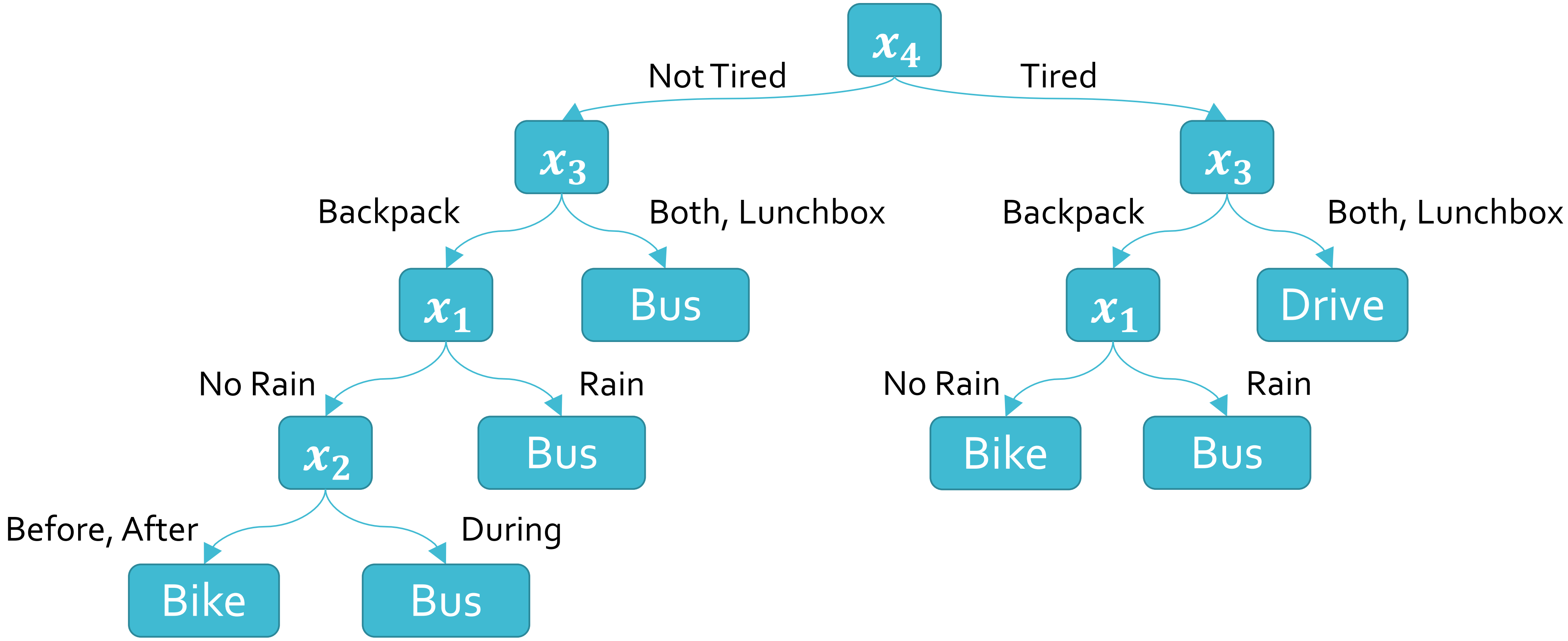
$$I\left(x_2, Y_{x_4=\text{Tired}}\right) \approx 0.2516$$

$$I\left(x_3, Y_{x_4=\text{Tired}}\right) \approx \mathbf{0.9183}$$

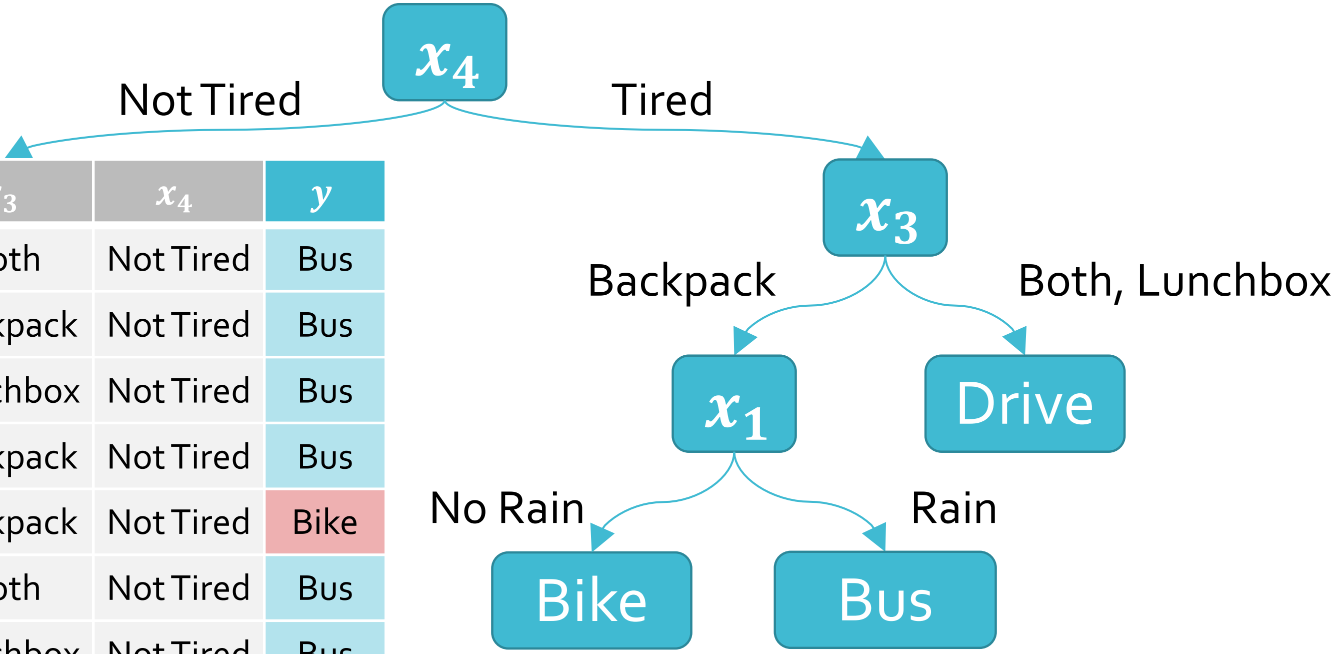

$x_4$
Not Tired
Tired
$x_3$
$x_3$
Backpack
Both, Lunchbox
Backpack
Both, Lunchbox
$x_1$
Bus
$x_1$
Drive
No Rain
Rain
No Rain
Rain
$x_2$
Bus
Bike
Bus
Before, After
During
Bike
Bus

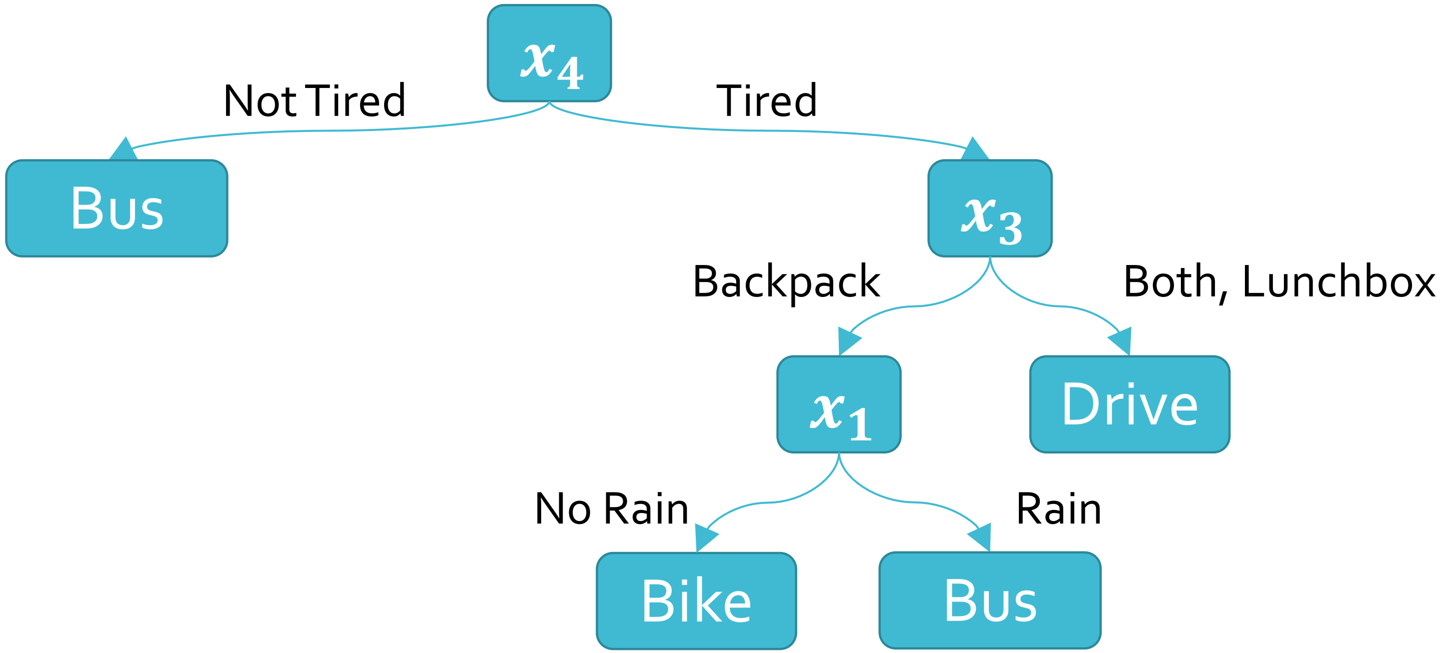
# Decision Trees: Inductive Bias

- The **inductive bias** of a machine learning algorithm is the principal by which it generalizes to unseen examples
- What is the inductive bias of the ID3 algorithm i.e., decision tree learning with mutual information maximization as the splitting criterion?
  - Try to find the shortest tree that achieves zero training error with high mutual information features at the top


$x_4$
Not Tired
Tired
$x_3$
$x_3$
Backpack
Both, Lunchbox
Backpack
Both, Lunchbox
$x_1$
Bus
$x_1$
Drive
No Rain
Rain
No Rain
Rain
$x_2$
Bus
Bike
Bus
Before, After
During
Bike
Bus

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| Rain | During | Both | Not Tired | Bus |
| Rain | After | Backpack | Not Tired | Bus |
| No Rain | Before | Lunchbox | Not Tired | Bus |
| No Rain | During | Backpack | Not Tired | Bus |
| No Rain | After | Backpack | Not Tired | Bike |
| No Rain | After | Both | Not Tired | Bus |
| No Rain | After | Lunchbox | Not Tired | Bus |






This tree only misclassifies one training data point!