# Data Science

## Decision Tree Learning

# OBJECTIVE

- WHAT IS A DECISION TREE? HOW DO WE USE IT?
- HOW TO BUILD A DECISION TREE
- ADVANTAGES & DISADVANTAGES
- TUNING PARAMETERS
- BUILDING A DECISION TREE IN PYTHON

# OBJECTIVE

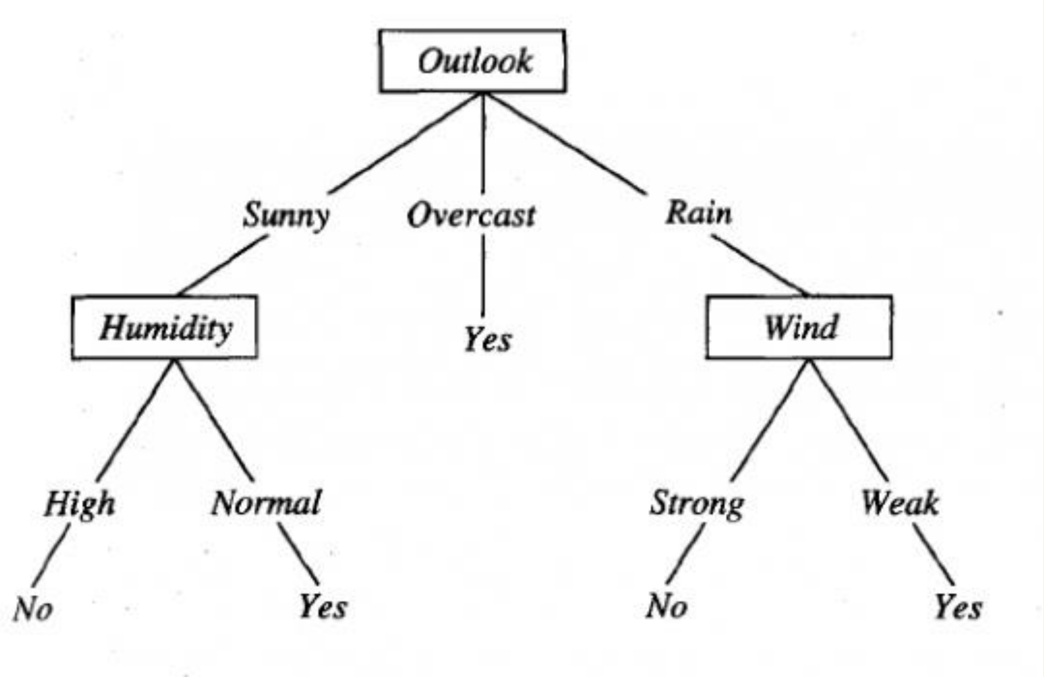
- **WHAT IS A DECISION TREE? HOW DO WE USE IT?**
- HOW TO BUILD A DECISION TREE
- ADVANTAGES & DISADVANTAGES
- TUNING PARAMETERS
- BUILDING A DECISION TREE IN PYTHON

# WHAT IS A DECISION TREE?

- Supervised learning algorithm
  - Regression: Numerical
  - Classification: Categorical
- Builds a tree by repeatedly splitting the dataset
- Easy to interpret
- Variable importance and selection
- Sets the foundation for state of the art ensemble techniques

# WHAT IS A DECISION TREE?

## Will there be a tennis match today?

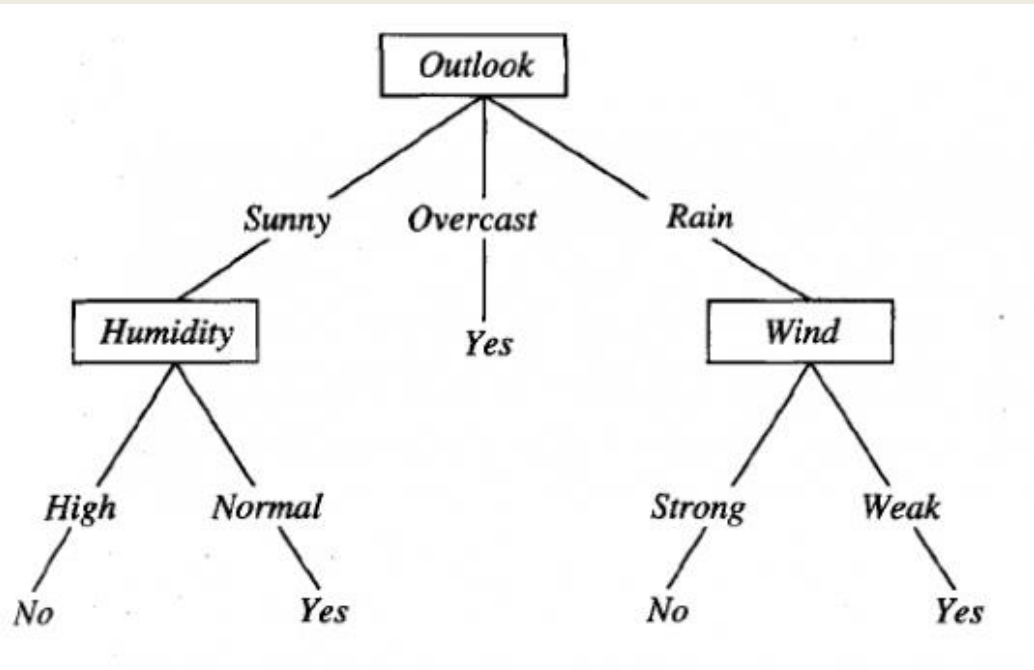



Quinlan 1986

- For new observations follow tree to leaf nodes
- Leaf nodes show the prediction
- Important variables are higher in the tree

# WHAT IS A DECISION TREE?

## Will there be a tennis match today?



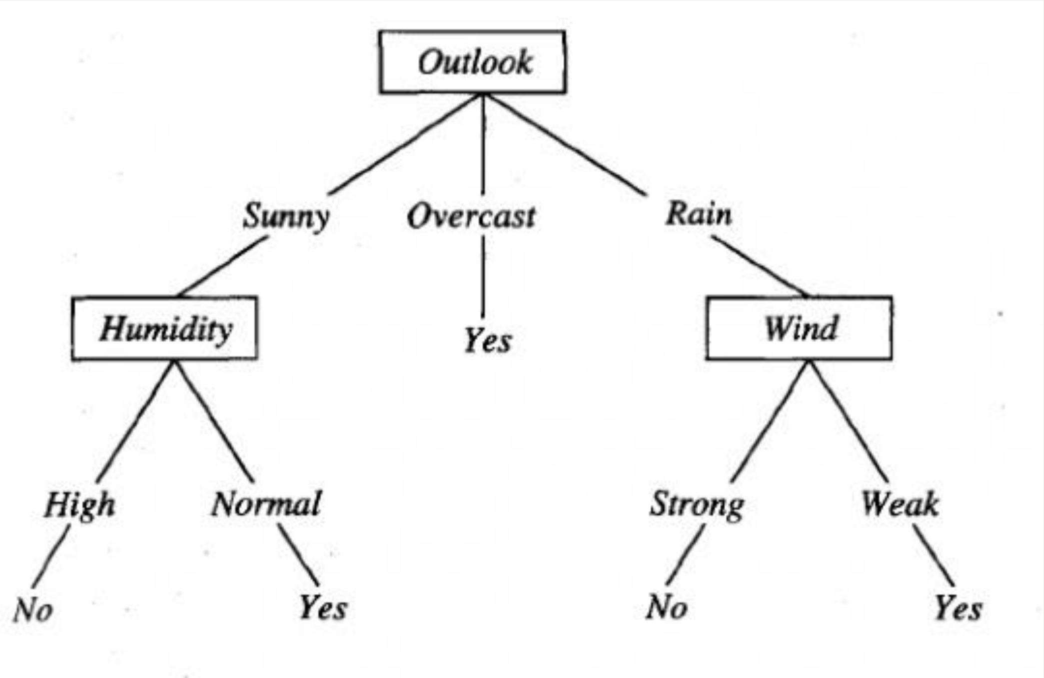

Quinlan 1986

- For new observations follow tree to leaf nodes
- Leaf nodes show the prediction
- Important variables are higher in the tree

Q1: The outlook is overcast. Will there be a tennis match?

# WHAT IS A DECISION TREE?

## Will there be a tennis match today?



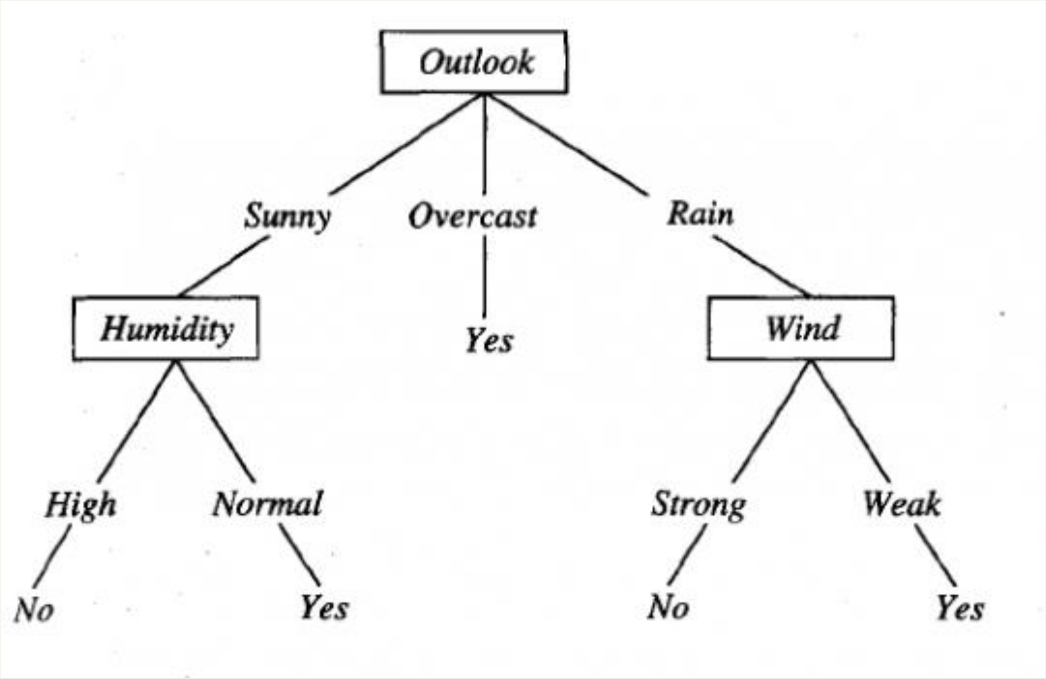

Quinlan 1986

- For new observations follow tree to leaf nodes
- Leaf nodes show the prediction
- Important variables are higher in the tree

Q1: The outlook is overcast. Will there be a tennis match?
A1: Yes

# WHAT IS A DECISION TREE?

## Will there be a tennis match today?




Quinlan 1986

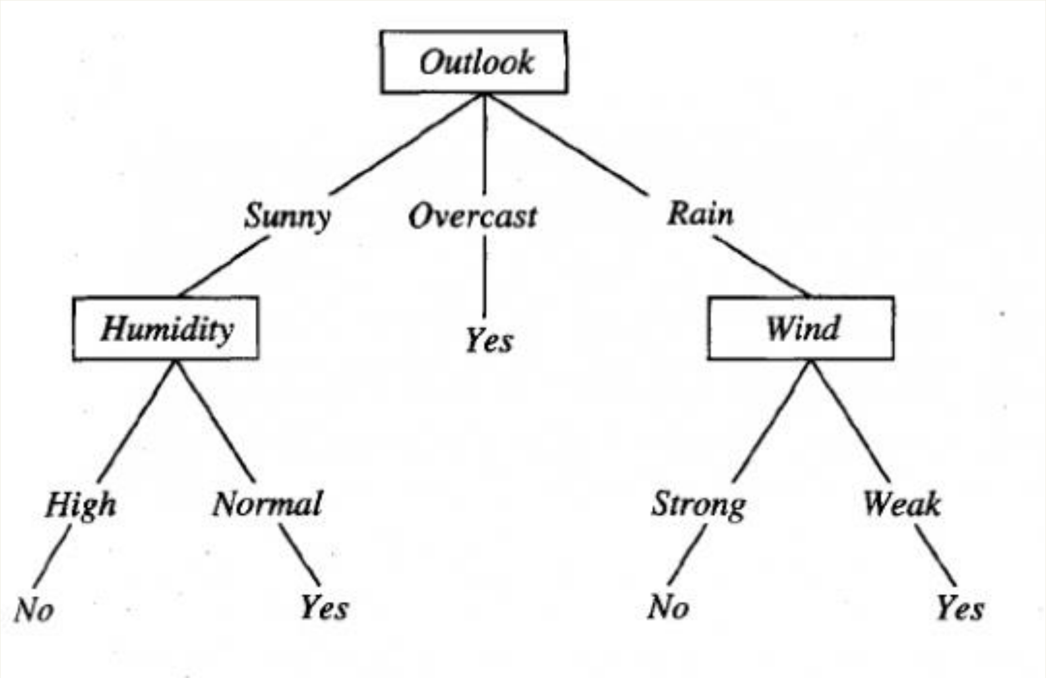
- For new observations follow tree to leaf nodes
- Leaf nodes show the prediction
- Important variables are higher in the tree

Q1: The outlook is overcast. Will there be a tennis match?
A1: Yes

Q2: Sunny outlook with normal humidity. Tennis Match?

# WHAT IS A DECISION TREE?

## Will there be a tennis match today?




Quinlan 1986

- For new observations follow tree to leaf nodes
- Leaf nodes show the prediction
- Important variables are higher in the tree

Q1: The outlook is overcast. Will there be a tennis match?
A1: Yes

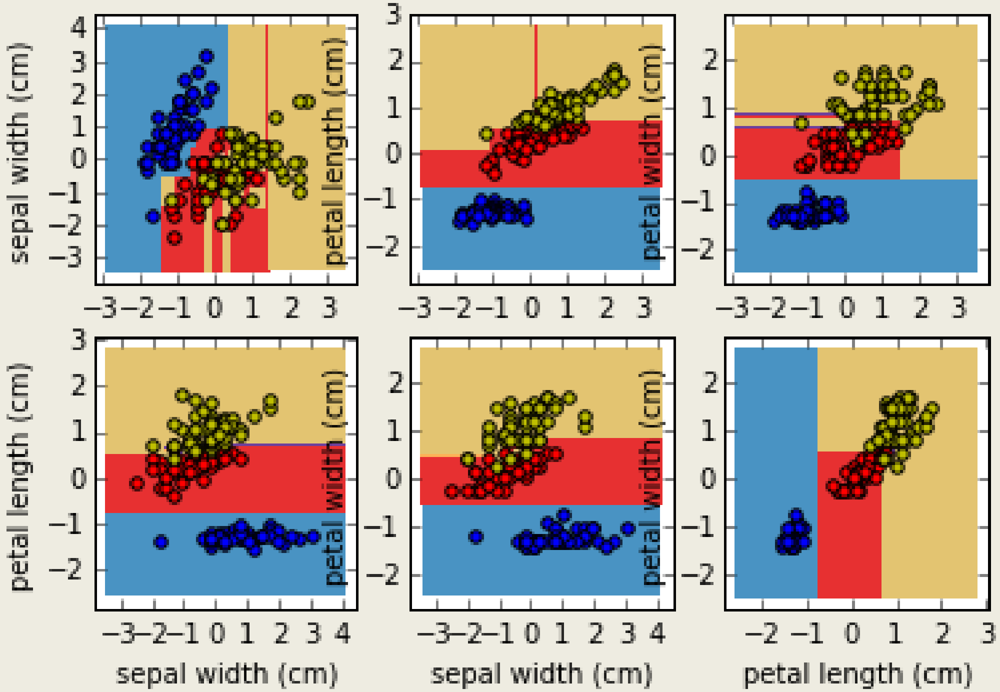
Q2: Sunny outlook with normal humidity. Tennis Match?
A2: Yes

# WHAT IS A DECISION TREE?

What is this doing?

# WHAT IS A DECISION TREE?

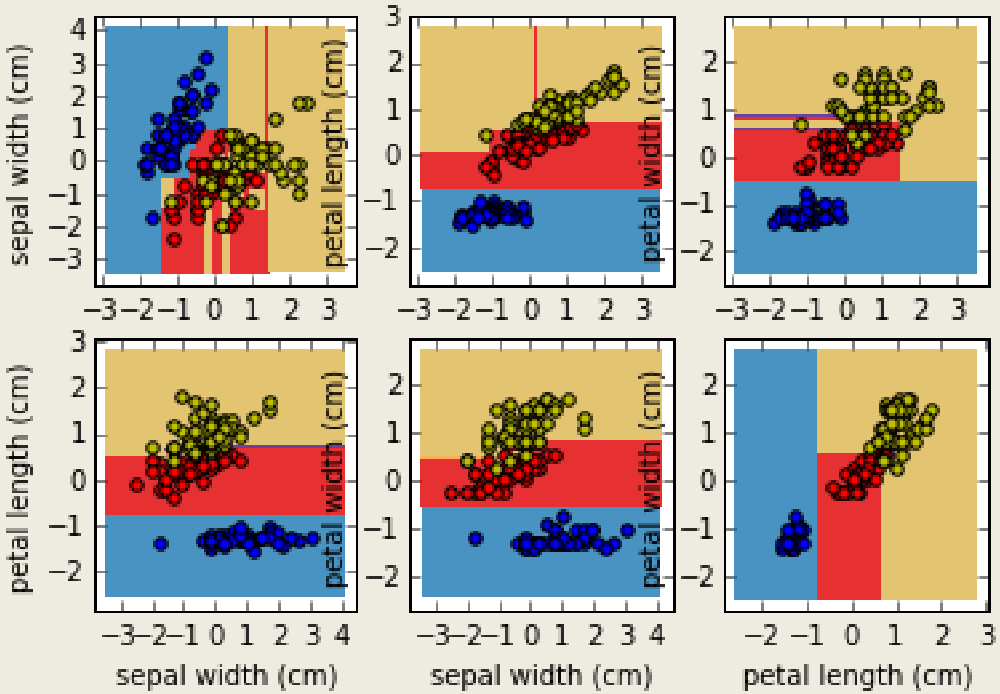
## What is this doing?




Partitioning the feature space by making axis perpendicular cuts

# WHAT IS A DECISION TREE?

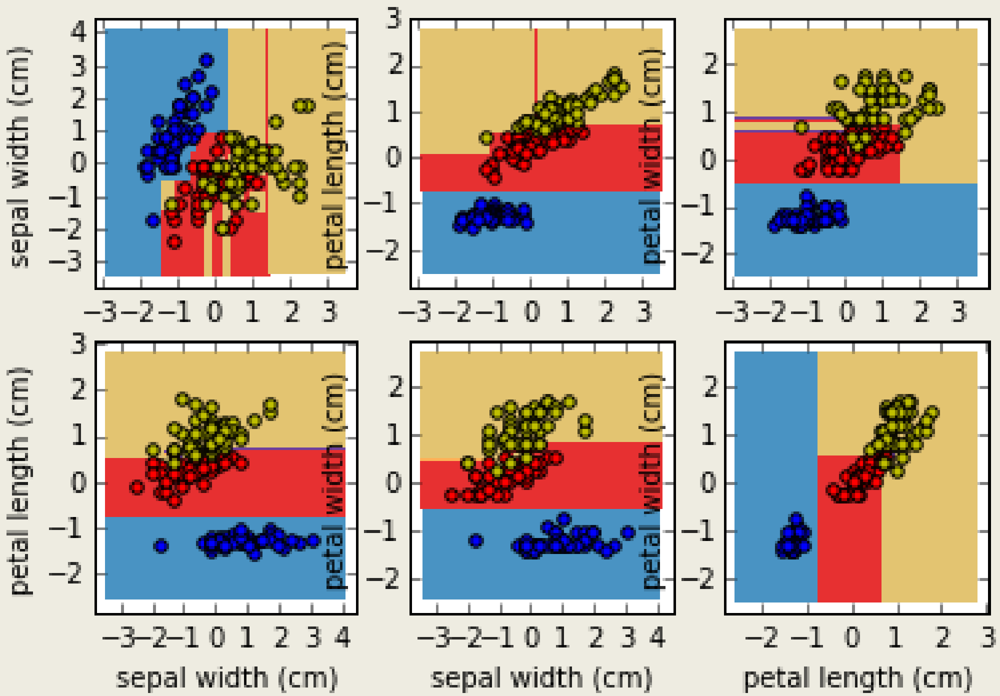
## What is this doing?




Partitioning the feature space by making axis perpendicular cuts
Q: How do we classify or regress now?

# WHAT IS A DECISION TREE?

## What is this doing?




Partitioning the feature space by making axis perpendicular cuts
Q: How do we classify or regress now?

A: Feature space is partitioned into rectangles. Take a vote or average within rectangle.

# OBJECTIVE

- WHAT IS A DECISION TREE? HOW DO WE USE IT?
- **HOW TO BUILD A DECISION TREE**
- ADVANTAGES & DISADVANTAGES
- TUNING PARAMETERS
- BUILDING A DECISION TREE IN PYTHON

# HOW TO BUILD A DECISION TREE?

- Various algorithms:
  - ID3, C4.5, CART, etc

# HOW TO BUILD A DECISION TREE?

- Various algorithms:
  - ID3, C4.5, CART, etc

Q: What questions would you ask?

# HOW TO BUILD A DECISION TREE?

- Various algorithms:
    - ID3, C4.5, CART, etc

<u>Questions:</u>

- What feature to use at a node?
- How many features should I use?
- How big do I make my tree?
- Are there any constraints on the leaf nodes?

# HOW TO BUILD A DECISION TREE?

- Various algorithms:
  - ID3, C4.5, CART, etc

<u>Questions:</u>

- What feature to use at a node?
- How many features should I use?
- How big do I make my tree?
- Are there any constraints on the leaf nodes?

Q: Why are these questions important?

# HOW TO BUILD A DECISION TREE?

- Various algorithms:
  - ID3, C4.5, CART, etc

Questions:

- What feature to use at a node?
- How many features should I use?
- How big do I make my tree?
- Are there any constraints on the leaf nodes?

Q: Why are these questions important?
Q: Any ideas on how to solve these problems?

# HOW TO BUILD A DECISION TREE?

<u>Questions:</u>

- What feature to use at a node?
  - ?

- How many features should I use?
  - ?
- How big do I make my tree?
  - ?
- Are there any constraints on the leaf nodes?
  - ?

# HOW TO BUILD A DECISION TREE?

Questions:

- What feature to use at a node?
  - Information Gain
  - Gini Impurity
  - Etc
- How many features should I use?
  - ?
- How big do I make my tree?
  - ?
- Are there any constraints on the leaf nodes?
  - ?

# HOW TO BUILD A DECISION TREE?

<u>Questions:</u>

- What feature to use at a node?
  - Information Gain
  - Gini Impurity
  - Etc
- How many features should I use?
  - Cross Validation
- How big do I make my tree?
  - ?
- Are there any constraints on the leaf nodes?
  - ?

# HOW TO BUILD A DECISION TREE?

<u>Questions:</u>

- What feature to use at a node?
  - Information Gain
  - Gini Impurity
  - Etc
- How many features should I use?
  - Cross Validation
- How big do I make my tree?
  - Cross Validation
- Are there any constraints on the leaf nodes?
  - ?

# HOW TO BUILD A DECISION TREE?

Questions:

- What feature to use at a node?
  - Information Gain
  - Gini Impurity
  - Etc
- How many features should I use?
  - Cross Validation
- How big do I make my tree?
  - Cross Validation
- Are there any constraints on the leaf nodes?
  - Cross Validation

# Information Gain

Entropy: Expected information contained in a message

Entropy = $-\Sigma\ p \log(p)$

# Information Gain

Entropy: Expected information contained in a message

Entropy = $-\Sigma \, p \log(p)$

Q: What is the entropy of a 90/10 coin? A 50/50 coin?

# Information Gain

Entropy: Expected information contained in a message
Entropy = $-\Sigma\, p \log(p)$

Q: What is the entropy of a 90/10 coin? A 50/50 coin?
A: 90/10 = .47 50/50 = 1.0

The higher the entropy the higher the information content. We can use this to determine which feature will give us the most information gain.

# Information Gain

Entropy: Expected information contained in a message
Entropy = $-\Sigma\ p \log(p)$

Q: What is the entropy of a 90/10 coin? A 50/50 coin?
A: 90/10 = .47 50/50 = 1.0

The higher the entropy the higher the information content. We can use this to determine which feature will give us the most information gain.

# Information Gain

Entropy: Expected information contained in a message

Entropy = $-\Sigma$ p log(p)

Q: What is the entropy of a 90/10 coin? A 50/50 coin?

A: 90/10 = .47 50/50 = 1.0

The higher the entropy the higher the information content. We can use this to determine which feature will give us the most information gain.

Information Gain = Entropy(parent) - average entropy(children)

# Information Gain

Entropy: Expected information contained in a message

Entropy = $-\Sigma \text{ p} \log(\text{p})$

Q: What is the entropy of a 90/10 coin? A 50/50 coin?

A:  90/10 = .47   50/50 = 1.0

The higher the entropy the higher the information content. We can use this to determine which feature will give us the most information gain.

Information Gain = Entropy(parent) - average entropy(children)

| X | Y | Z | C |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 1 | 1 | 0 | 1 |
| 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 |

Example:

Q: What feature is the best?

# Information Gain

Entropy = $-\Sigma\ p \log(p)$

Information Gain = Entropy(parent) - average entropy(children)

| X | Y | Z | C |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 1 | 1 | 0 | 1 |
| 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 |

Example:
Q: What feature is the best?

## What is information Gain?

# Information Gain

Entropy = $-\Sigma$ p log(p)

Information Gain = Entropy(parent) - average entropy(children)

| X | Y | Z | C |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 1 | 1 | 0 | 1 |
| 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 |

Example:

Q: What feature is the best?

## What is Information Gain?

## Entropy(Parent) = ?

# Information Gain

Entropy = $-\Sigma\ p \log(p)$

Information Gain = Entropy(parent) - average entropy(children)

| X | Y | Z | C |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 1 | 1 | 0 | 1 |
| 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 |

Example:

Q: What feature is the best?

## What is Information Gain?

- Entropy(Parent) = 1
- X: ?

# Information Gain

Entropy = $-\Sigma\, p \log(p)$

Information Gain = Entropy(parent) - average entropy(children)

| X | Y | Z | C |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 1 | 1 | 0 | 1 |
| 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 |

Example:

Q: What feature is the best?

## What is Information Gain?

- Entropy(Parent) = 1
- X: C1 = -(1/3)log(1/3) – (2/3)log(2/3) = .9184, C2 = 0
  - Info Gain = 1- (3/4)*0.9184 – 0 = 0.3112
- Y: ?

# Information Gain

Entropy = $-\Sigma$ p log(p)

Information Gain = Entropy(parent) - average entropy(children)

| X | Y | Z | C |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 1 | 1 | 0 | 1 |
| 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 |

Example:

Q: What feature is the best?

## What is Information Gain?

- Entropy(Parent) = 1
- X: C1 = -(1/3)log(1/3) – (2/3)log(2/3) = .9184, C2 = 0
  - Info Gain = 1- (3/4)*0.9184 – 0 = 0.3112
- Y: C1 = 0, C2 = 0
  - Info Gain = 1-0-0= 1
- Z: ?

# Information Gain

Entropy = $-\Sigma\ p \log(p)$

Information Gain = Entropy(parent) - average entropy(children)

| X | Y | Z | C |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 1 | 1 | 0 | 1 |
| 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 |

Example:

Q: What feature is the best?

## What is Information Gain?

- Entropy(Parent) = 1
- X: C1 = -(1/3)log(1/3) – (2/3)log(2/3) = .9184, C2 = 0
  - Info Gain = 1- (3/4)*0.9184 – 0 = 0.3112
- Y: C1 = 0, C2 = 0
  - Info Gain = 1-0-0= 1
- Z: C1 = 1, C2 = 1
  - Info Gain = 1 –(1/2)*1 = (1/2)*1 = 0

# Information Gain

Entropy = $-\Sigma$ p log(p)

Information Gain = Entropy(parent) - average entropy(children)

| X | Y | Z | C |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 1 | 1 | 0 | 1 |
| 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 |

Example:

Q: What feature is the best?

A: Feature Y is the best feature

## What is Information Gain?

- Entropy(Parent) = 1
- X: C1 = -(1/3)log(1/3) – (2/3)log(2/3) = .9184, C2 = 0
  - Info Gain = 1- (3/4)*0.9184 – 0 = 0.3112
- **Y: C1 = 0, C2 = 0**
  - **Info Gain = 1-0-0= 1**
- Z: C1 = 1, C2 = 1
  - Info Gain = 1 –(1/2)*1 = (1/2)*1 = 0

# OBJECTIVE

- WHAT IS A DECISION TREE? HOW DO WE USE IT?
- HOW TO BUILD A DECISION TREE
- **ADVANTAGES & DISADVANTAGES**
- TUNING PARAMETERS
- BUILDING A DECISION TREE IN PYTHON

# Advantage & Disadvantages

Advantages:

- Can handle continuous and categorical predictors
- Interpretability
- Ensembles are extremely powerful

Disadvantages

- Overfitting
- Predictive power
- High Variance

# OBJECTIVE

- WHAT IS A DECISION TREE? HOW DO WE USE IT?
- HOW TO BUILD A DECISION TREE
- ADVANTAGES & DISADVANTAGES
- **TUNING PARAMETERS**
- **BUILDING A DECISION TREE IN PYTHON**