Machine Learning 2007: Lecture 4

Instructor: Tim van Erven (Tim.van.Erven@cwi.nl)
Website: `www.cwi.nl/˜erven/teaching/0708/ml/`

September 27, 2007

Overview

- **Organisational Matters**
- An Unbiased Hypothesis Space for List-Then-Eliminate?
- Math: Directed Graphs and Trees
- Decision Trees for Classification
  - Hypothesis Space: Decision Trees
  - Method: ID3
- Math: Probability Distributions

Organisational Matters

Course Organisation:

- Biweekly exercises: you get a full week instead of 5 days.
- Exercise 2 available this evening.
- Grades for Exercise 1 available this week.

Study Guide:

- You don't have to know the details of the Candiate-Elimination algorithm, just that it does the same thing as the List-Then-Eliminate algorithm.
- But sections 2.6 and 2.7 of Mitchell are very important! Just replace each occurrence of Candiate-Elimination by List-Then-Eliminate when reading them.

This Lecture versus Mitchell:

- Decision trees are in Mitchell, but I will discuss the underlying mathematics in much more detail.

Overview

- Organisational Matters
- **An Unbiased Hypothesis Space for List-Then-Eliminate?**
- Math: Directed Graphs and Trees
- Decision Trees for Classification
  - Hypothesis Space: Decision Trees
  - Method: ID3
- Math: Probability Distributions

List-Then-Eliminate *Algorithm*

Description:

- List-Then-Eliminate finds the set, VersionSpace, of all hypotheses that are consistent with all the training data.
- It can only classify a new feature vector $\mathbf{x}$ if all the hypotheses in VersionSpace agree.

Hypothesis Space:

$$\begin{aligned}\mathcal{H} = \{&\langle ?, ?, ?, ?, ?, ? \rangle, \langle \text{Sunny}, ?, ?, ?, ?, ? \rangle, \\ &\langle \text{Cloudy}, ?, ?, ?, ?, ? \rangle, \dots, \langle \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset \rangle\}\end{aligned}$$

- Has a very strong **representation bias**: Only 973 out of $2^{96} \approx 10^{29}$ possible hypotheses can be represented.

An Unbiased Hypothesis Space

All Possible Hypotheses:

Why not take all possible hypotheses as a hypothesis space for LIST-THEN-ELIMINATE?

$$\mathcal{H} = \{h | h \text{ is a function from } \mathcal{X} \text{ to } \mathcal{Y}\},$$

where

- $\mathcal{X} =$ set of possible feature vectors,
- $\mathcal{Y} =$ set of possible labels,
- $|\mathcal{H}| = |\mathcal{Y}|^{|\mathcal{X}|} = 2^{96}$.

An Unbiased Hypothesis Space

All Possible Hypotheses:

Why not take all possible hypotheses as a hypothesis space for LIST-THEN-ELIMINATE?

$$\mathcal{H} = \{h | h \text{ is a function from } \mathcal{X} \text{ to } \mathcal{Y}\},$$

where

- $\mathcal{X} =$ set of possible feature vectors,
- $\mathcal{Y} =$ set of possible labels,
- $|\mathcal{H}| = |\mathcal{Y}|^{|\mathcal{X}|} = 2^{96}$.

Classifying a New Feature Vector:

- Given: data $D = \begin{pmatrix} y_1 \\ \mathbf{x}_1 \end{pmatrix}, \ldots, \begin{pmatrix} y_n \\ \mathbf{x}_n \end{pmatrix}$.
- What happens if we try to classify a new feature vector $\mathbf{x}_{n+1}$?

Classifying New Instances

For any hypothesis $h \in \mathcal{H}$, there exists a $h' \in \mathcal{H}$ such that

$$\begin{aligned} h(\mathbf{x}) &\neq h'(\mathbf{x}) \quad \text{if } \mathbf{x} = \mathbf{x_{n+1}}, \\ h(\mathbf{x}) &= h'(\mathbf{x}) \quad \text{for any other } \mathbf{x}. \end{aligned}$$

Classifying New Instances

For any hypothesis $h \in \mathcal{H}$, there exists a $h' \in \mathcal{H}$ such that

$$\begin{aligned} h(\mathbf{x}) &\neq h'(\mathbf{x}) && \text{if } \mathbf{x} = \mathbf{x_{n+1}}, \\ h(\mathbf{x}) &= h'(\mathbf{x}) && \text{for any other } \mathbf{x}. \end{aligned}$$

Consequence:

- Suppose $\mathbf{x}_{n+1}$ does not occur in D.
- Then for every $h \in$ VersionSpace, there exists an alternative $h' \in$ VersionSpace that disagrees on the label of $\mathbf{x}_{n+1}$:

$$h(\mathbf{x}_{n+1}) \neq h'(\mathbf{x}_{n+1})$$

Conclusion:

In an unbiased hypothesis space, the LIST-THEN-ELIMINATE algorithm **cannot generalise** at all. Bias is unavoidable!

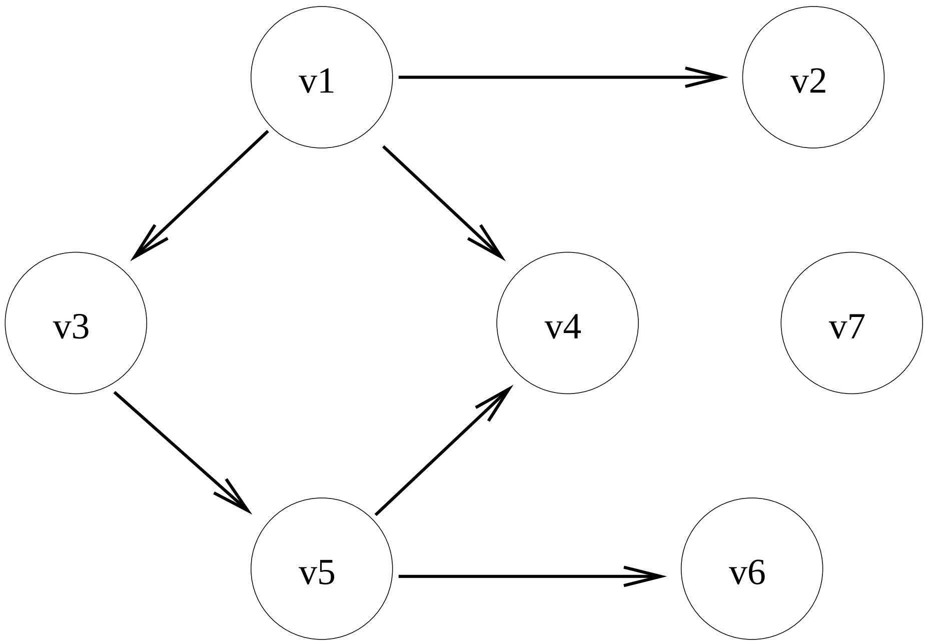
Overview

- Organisational Matters
- An Unbiased Hypothesis Space for List-Then-Eliminate?
- **Math: Directed Graphs and Trees**
- Decision Trees for Classification
  - Hypothesis Space: Decision Trees
  - Method: ID3
- Math: Probability Distributions

Directed Graphs

A directed graph G is an ordered pair $G = (V, E)$, where

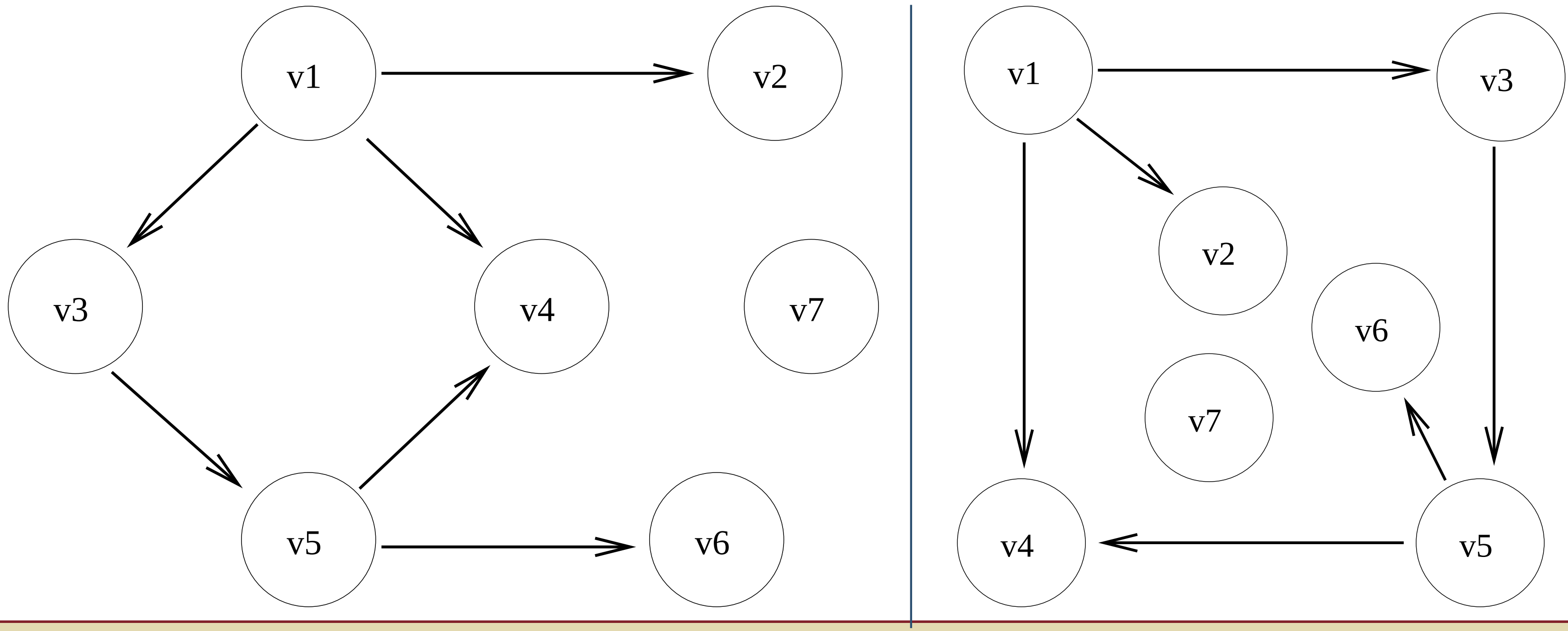
- $V = \{v_1, \ldots, v_m\}$ is a set of **vertices/nodes**;
- $E = \{e_1, \ldots, e_n\}$ is a set of **directed edges** between the vertices in V.
- Each directed edge e from vertex u to vertex v is an ordered pair $e = (u, v)$.
- I can draw the same directed graph in different ways.



Directed Graphs

A directed graph G is an ordered pair $G = (V, E)$, where

- $V = \{v_1, \ldots, v_m\}$ is a set of **vertices/nodes**;
- $E = \{e_1, \ldots, e_n\}$ is a set of **directed edges** between the vertices in V.
- Each directed edge e from vertex u to vertex v is an ordered pair $e = (u, v)$.
- I can draw the same directed graph in different ways.

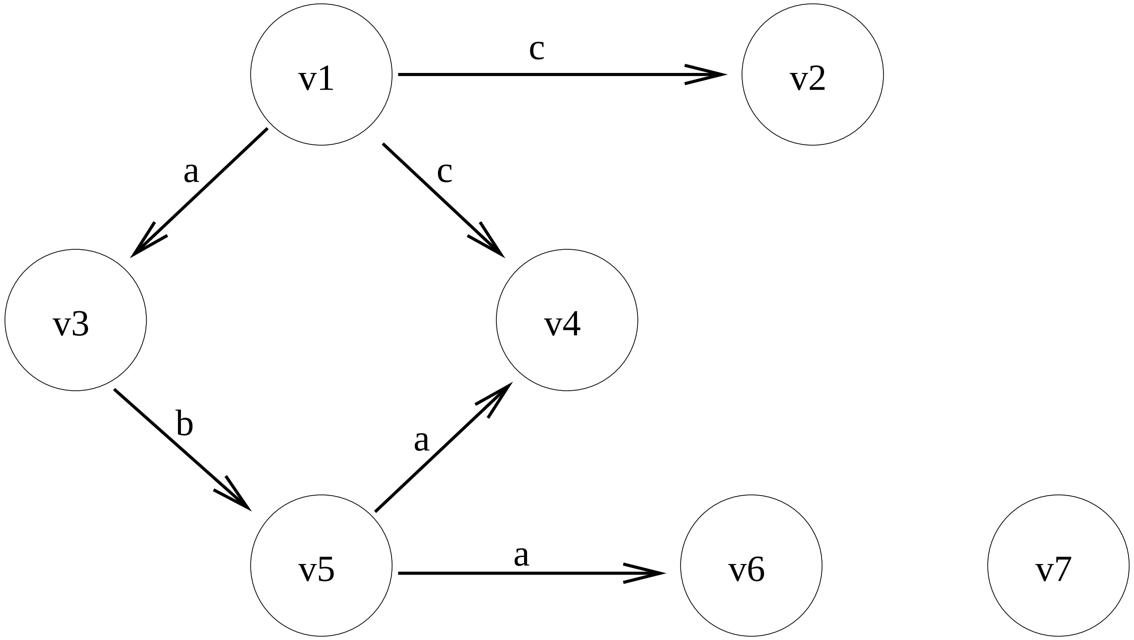


Directed Graphs with Edge Labels

- We can also **label edges** with labels from some set of possible labels L. Now $G = (V, E, L)$.
- Each directed edge e with label $l \in L$ from vertex u to vertex v is an ordered pair $e = (u, l, v)$.

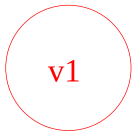
Example:

Let $L = \{a, b, c\}$.

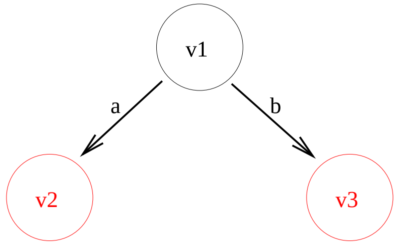


Tree Examples

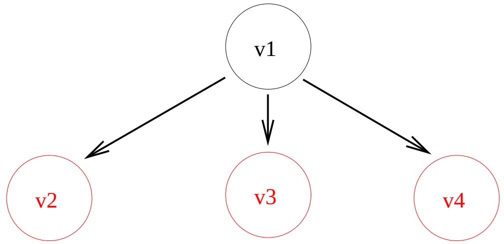
Example 1:



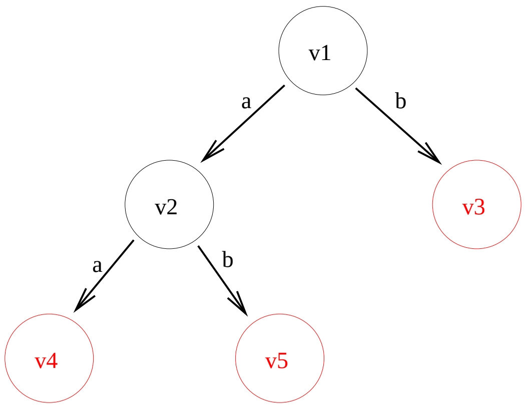
Example 2:



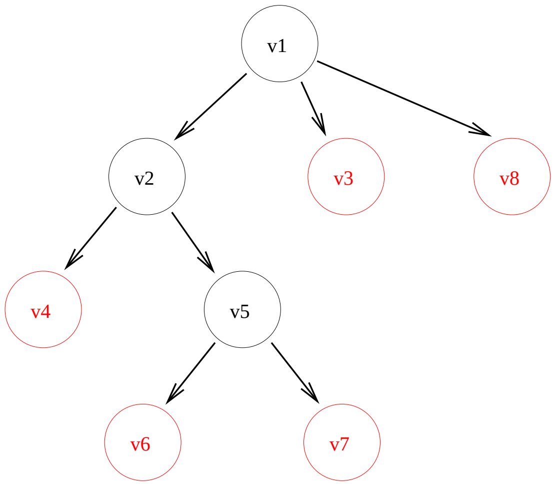
Example 3:



Example 4:



Example 5:

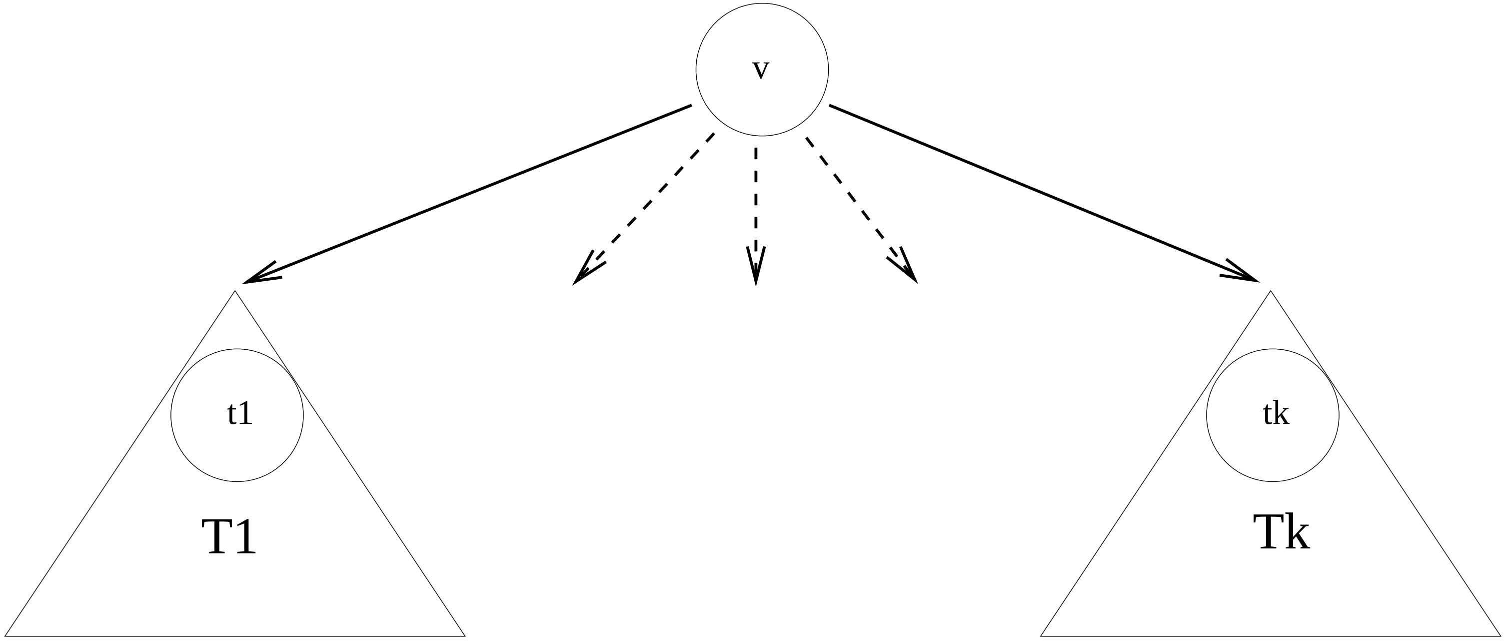


- In all examples the root of the tree is v_1.
- The nodes without outgoing edges (shown in red) are called **leaves**.
- The other nodes are called **internal** nodes.

Directed Trees

A directed graph is a **(directed) tree** $T = (V, E)$ with root $v \in V$ if and only if either:

1. v is the only node: $T = (\{v\}, \emptyset)$, or
2. - $T_1, \ldots, T_k$ are trees with roots $t_1, \ldots, t_k$,
   - v, $T_1, \ldots, T_k$ have no nodes in common, and
   - T looks like:



Properties of Trees

Let T be a (directed) tree.

- If T contains an edge $e = (u, v)$ from node u to node v, then
  - u is called the **parent** of v,
  - v is called the **child** of u.

Properties of Trees

Let T be a (directed) tree.

- If T contains an edge $e = (u, v)$ from node u to node v, then
  - u is called the **parent** of v,
  - v is called the **child** of u.

Number of Parents:

- Each node has exactly one parent, except for the **root**, which has no parents.

Properties of Trees

Let T be a (directed) tree.

- If T contains an edge $e = (u, v)$ from node u to node v, then
  - u is called the **parent** of v,
  - v is called the **child** of u.

Number of Parents:

- Each node has exactly one parent, except for the **root**, which has no parents.

Number of Children:

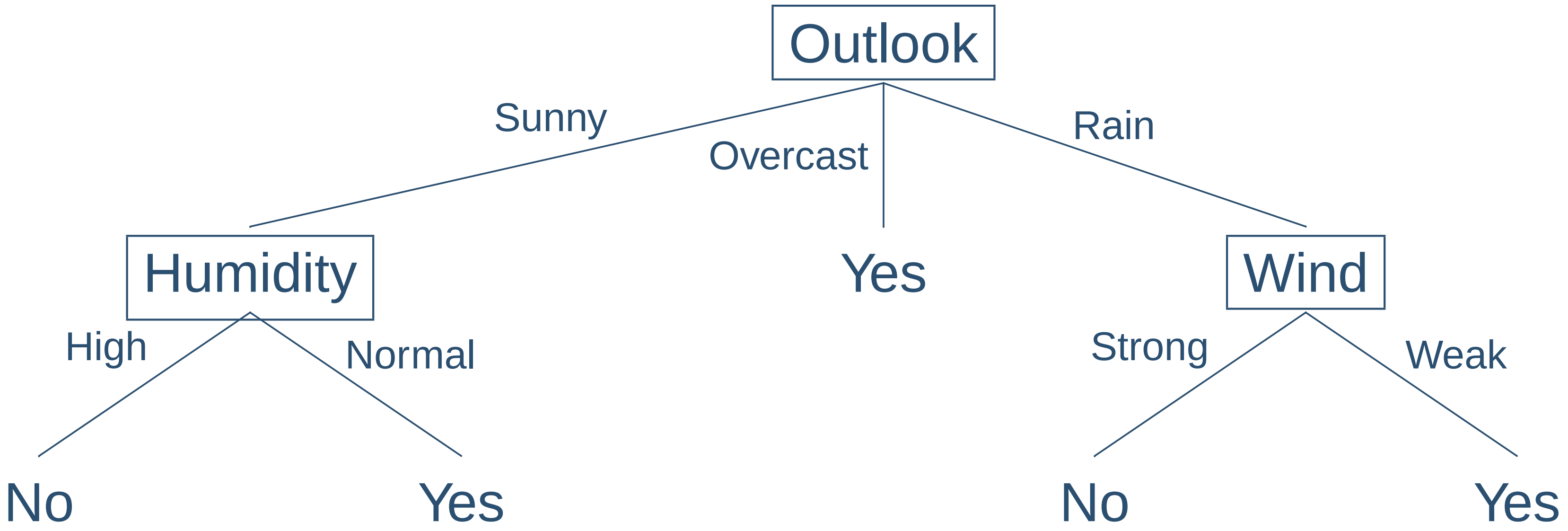
- Each node may have any (finite) number of children.
- The **leaves** are the nodes without children.
- The **internal** nodes have at least one child.

Overview

- Organisational Matters
- An Unbiased Hypothesis Space for List-Then-Eliminate?
- Math: Directed Graphs and Trees
- Decision Trees for Classification
  - **Hypothesis Space: Decision Trees**
  - Method: ID3
- Math: Probability Distributions

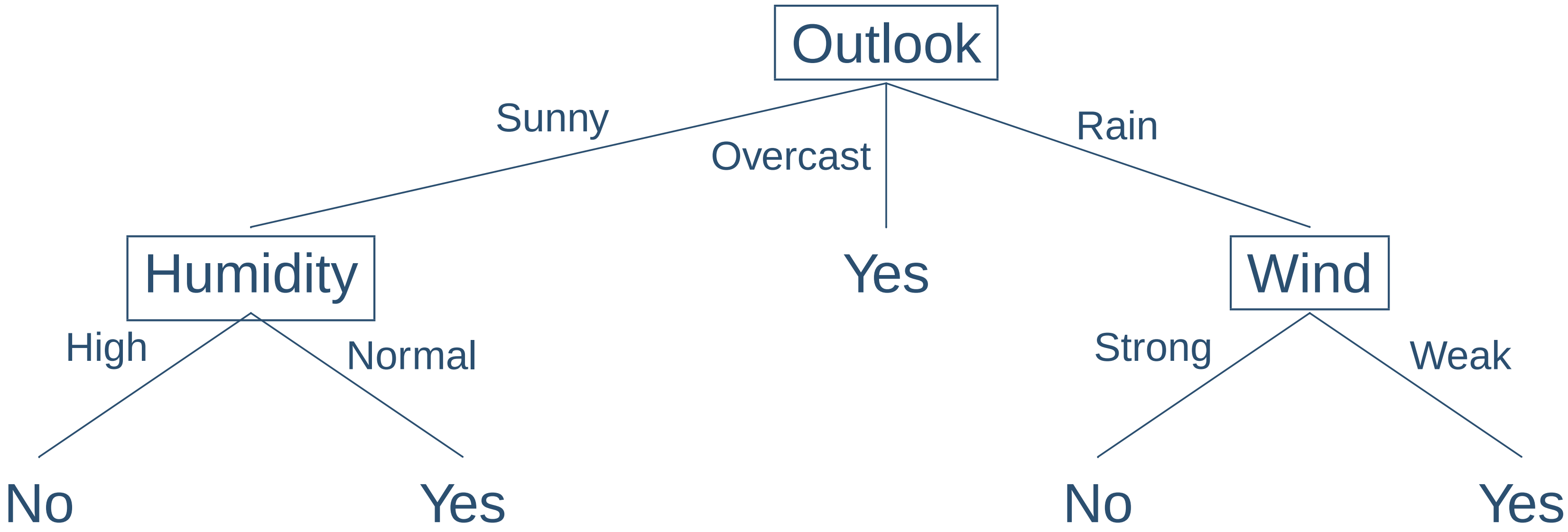
Decision Trees: Hypothesis Space

Decision Tree:



Decision Trees: Hypothesis Space

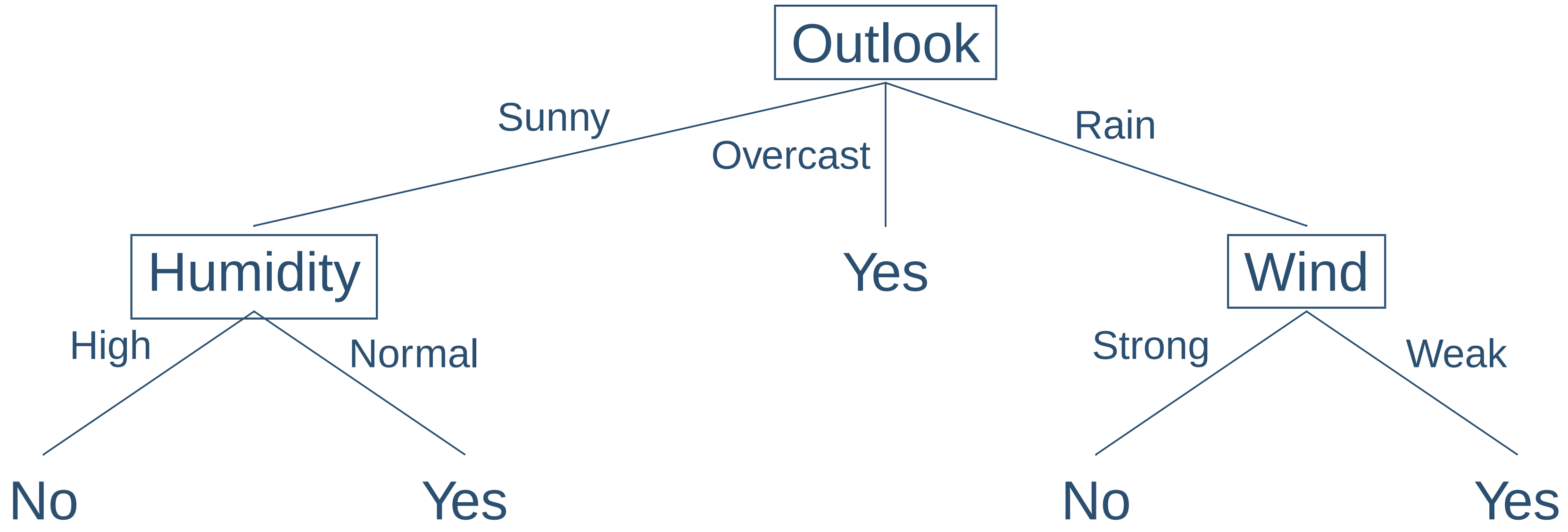
Decision Tree:



Part of tree	Interpretation	Example
Internal node	Attribute	$\boxed{\text{Outlook}}$
Leaf node	Class label	Yes
Edge label	Attribute value	Sunny

Decision Trees: Hypothesis Space

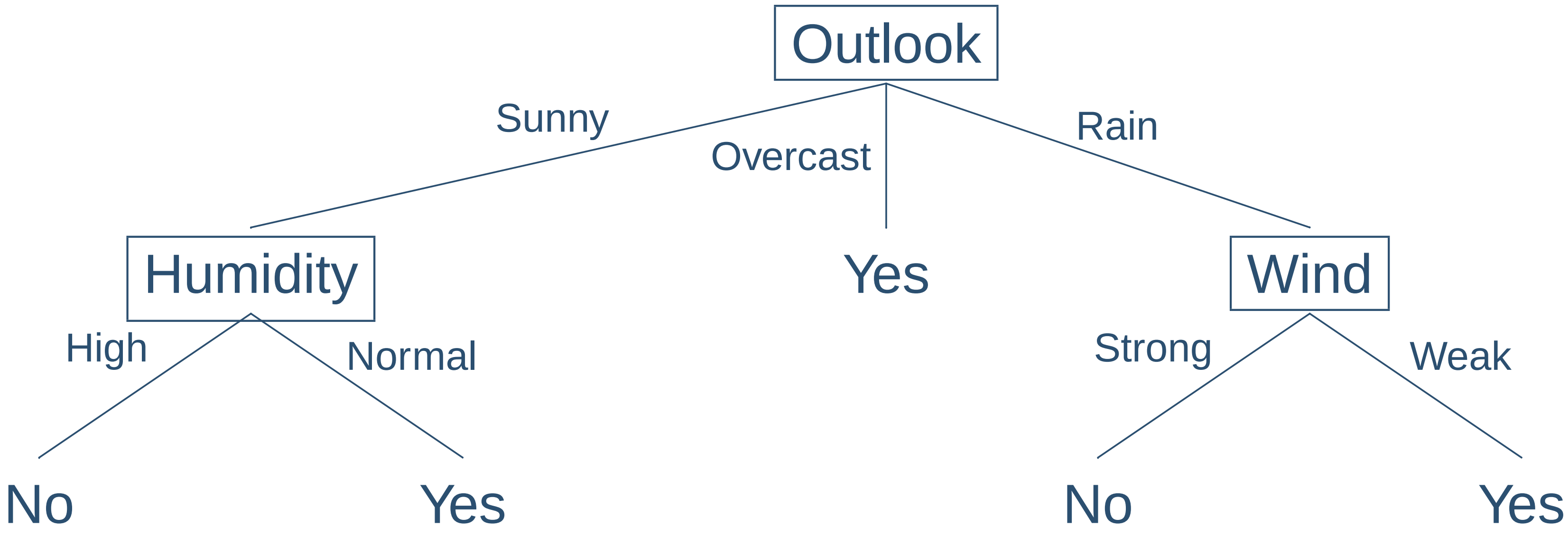
Decision Tree:



Part of tree	Interpretation	Example
Internal node	Attribute	$\boxed{\text{Outlook}}$
Leaf node	Class label	Yes
Edge label	Attribute value	Sunny

- Mitchell does not draw the arrows. They all point downwards.
- $\mathcal{H}$ is the set of all possible decision trees.

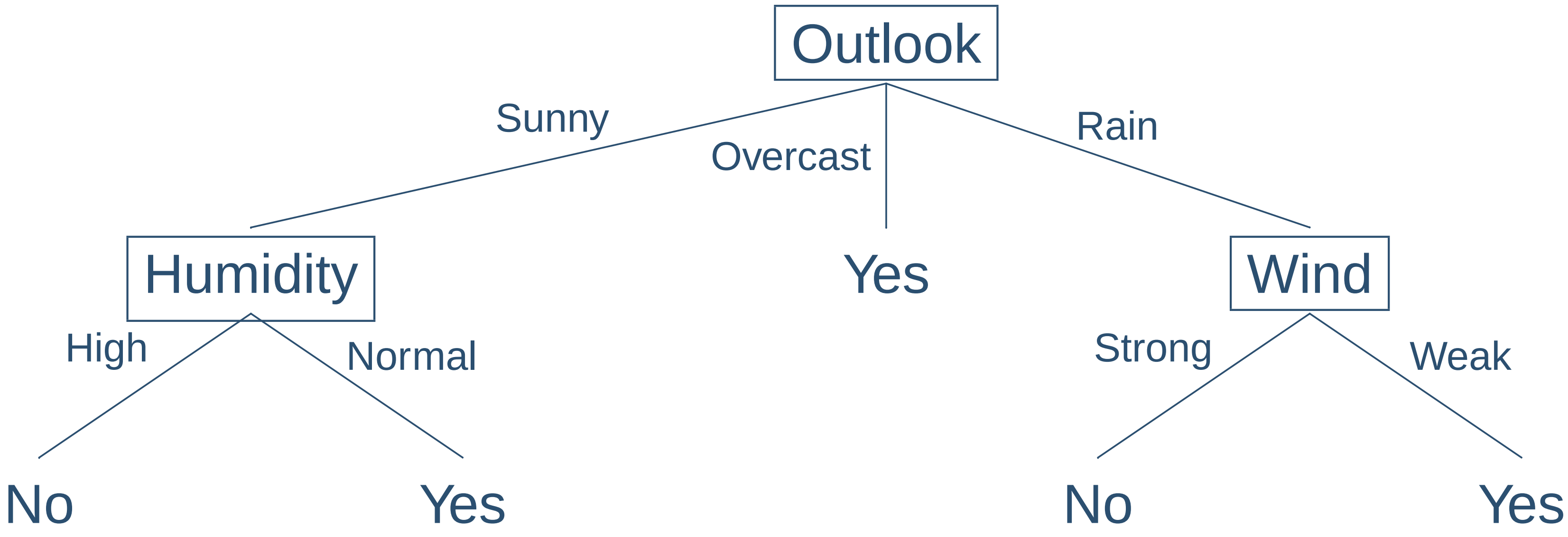
Decision Trees: Classification Examples



Classify by sorting down the tree:

x				y
Outlook	**Temperature**	**Humidity**	**Wind**	**PlayTennis**
Sunny	Hot	High	Weak	
Sunny	Hot	High	Strong	
Overcast	Hot	High	Weak	
Rain	Mild	High	Weak	
Rain	Cool	Normal	Weak	
Rain	Cool	Normal	Strong	

Decision Trees: Classification Examples

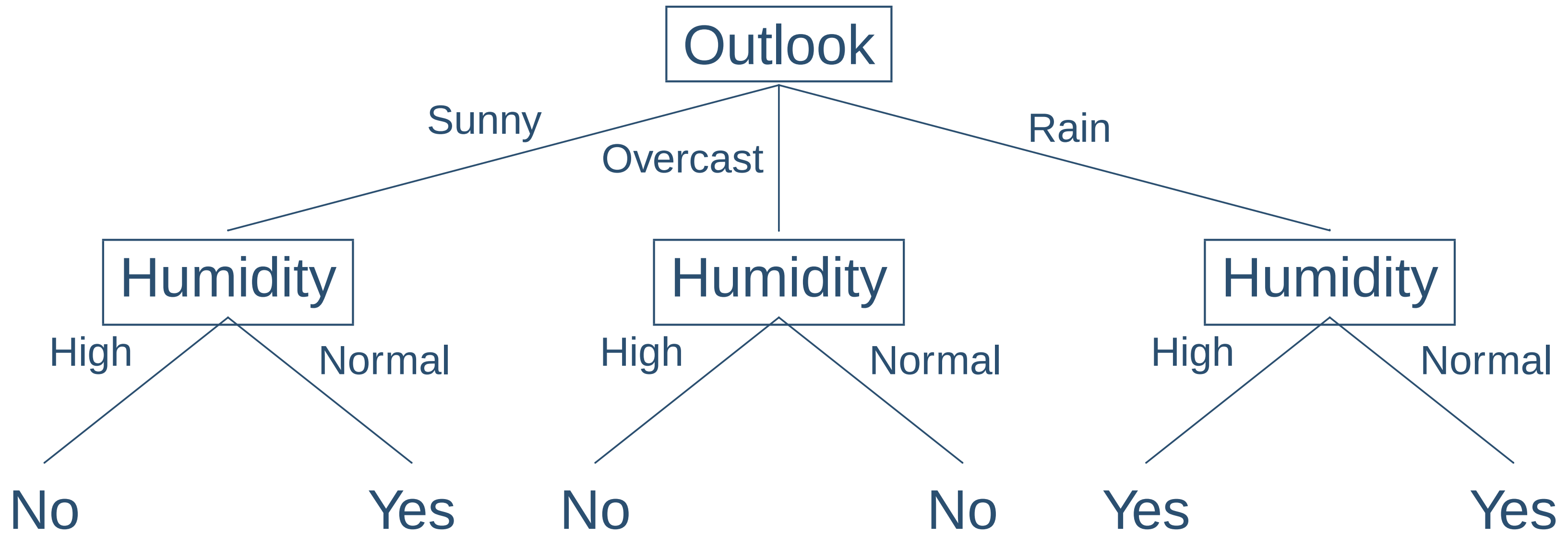


Classify by sorting down the tree:

	x			y
Outlook	**Temperature**	**Humidity**	**Wind**	**PlayTennis**
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No

Unbiased Hypothesis Space

Consider the **full tree** for the attributes Outlook and Humidity:



- By changing the labels at the leaves of the tree, we can describe **any** hypothesis about Outlook and Humidity.
- We can do the same thing for all attributes: No representation bias!
- But the size of the full tree blows up exponentially in the number of attributes.

Overview

- Organisational Matters
- An Unbiased Hypothesis Space for LIST-THEN-ELIMINATE?
- Math: Directed Graphs and Trees
- Decision Trees for Classification
  - Hypothesis Space: Decision Trees
  - **Method: ID3**
- Math: Probability Distributions

The ID3 *Algorithm*

General:

- Learns a decision tree from data.
- Hence does classification.

Main Ideas:

1. Start by selecting a root attribute for the tree.
2. Then grow the tree by adding more and more attributes to it.
3. Stop growing the tree when it is consistent with all the data.

The ID3 *Algorithm*

General:

- Learns a decision tree from data.
- Hence does classification.

Main Ideas:

1. Start by selecting a root attribute for the tree.
2. Then grow the tree by adding more and more attributes to it.
3. Stop growing the tree when it is consistent with all the data.

Some Notation:

- The data $D = \begin{pmatrix} y_1 \\ \mathbf{x}_1 \end{pmatrix}, \ldots, \begin{pmatrix} y_n \\ \mathbf{x}_n \end{pmatrix}$
- $A =$ the set of features/attributes that may be used to grow the decision tree. (For example, $A = \{2, 5, 6\}$ represents that we may use attributes x_2, x_5 and x_6 to grow the tree.)
- $D_{a,v} = \left\{ \begin{pmatrix} y_i \\ \mathbf{x}_i \end{pmatrix} \mid \mathbf{x}_i \text{ has value } v \text{ for attribute } x_a \right\}$

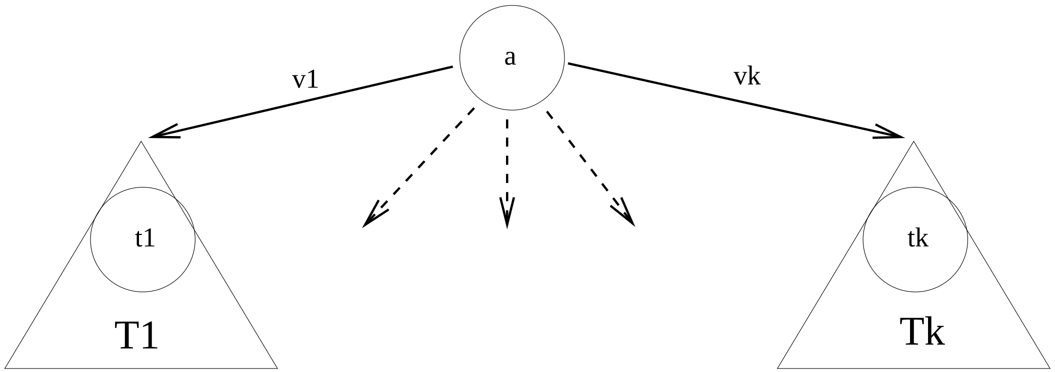
The ID3 *Algorithm*

$D =$ data; $D_{a,v} =$ data such that $\mathbf{x}$ has value v for attribute x_a;
$A =$ set of available features/attributes

ID3(D, A)

$z =$ the most common label y in D
if y is the same for all examples in D or $A = \emptyset$ **then**
    return $T = (\{z\}, \emptyset)$

Select the best[1] attribute $a \in A$ with values $v_1, \ldots, v_k$.
$T_i = \begin{cases} (\{z\}, \emptyset) & \text{if } D_{a,v_i} = \emptyset \\ \text{ID3}(D_{a,v_i}, A \setminus \{a\}) & \text{otherwise} \end{cases}$
return



[1] To be defined later

A First Discussion of ID3

- ID3 does not have a representation bias, because decision trees provide an unbiased hypothesis space. So where does the bias come in?

A First Discussion of ID3

- ID3 does not have a representation bias, because decision trees provide an unbiased hypothesis space. So where does the bias come in?
- It prefers shorter decision trees! This is called a **preference bias**.

A First Discussion of ID3

- ID3 does not have a representation bias, because decision trees provide an unbiased hypothesis space. So where does the bias come in?
- It prefers shorter decision trees! This is called a **preference bias**.
- **Not completely robust** against noise/errors in the data, because it always finds a decision tree that is consistent with all training data. (Maybe a much smaller tree exists that only makes a single mistake!)
- Next week we will see an extension, C4.5, which addresses this problem.

A First Discussion of ID3

- ID3 does not have a representation bias, because decision trees provide an unbiased hypothesis space. So where does the bias come in?
- It prefers shorter decision trees! This is called a **preference bias**.
- **Not completely robust** against noise/errors in the data, because it always finds a decision tree that is consistent with all training data. (Maybe a much smaller tree exists that only makes a single mistake!)
- Next week we will see an extension, C4.5, which addresses this problem.
- Not suitable if features/attributes can take **infinitely many values** (e.g. all real numbers): infinite number of children for the corresponding node in the decision tree.

Overview

- Organisational Matters
- An Unbiased Hypothesis Space for LIST-THEN-ELIMINATE?
- Math: Directed Graphs and Trees
- Decision Trees for Classification
  - Hypothesis Space: Decision Trees
  - Method: ID3
- **Math: Probability Distributions**

Probability Distributions

- The **sample space** $\Omega = \{\omega_1, \ldots, \omega_k\}$ is the set of all possible outcomes of an experiment.
- An **event** $\mathcal{E} \subseteq \Omega$ is a (sub)set of possible outcomes.

Probability Distributions

- The **sample space** $\Omega = \{\omega_1, \ldots, \omega_k\}$ is the set of all possible outcomes of an experiment.
- An **event** $\mathcal{E} \subseteq \Omega$ is a (sub)set of possible outcomes.
- A **(probability) mass function** $p(\omega_i)$ assigns a weight to each *outcome* $\omega_i \in \Omega$ such that:
  - $0 \leq p(\omega_i) \leq 1$
  - $p(\omega_1) + \ldots + p(\omega_k) = 1$

Probability Distributions

- The **sample space** $\Omega = \{\omega_1, \ldots, \omega_k\}$ is the set of all possible outcomes of an experiment.
- An **event** $\mathcal{E} \subseteq \Omega$ is a (sub)set of possible outcomes.
- A **(probability) mass function** $p(\omega_i)$ assigns a weight to each *outcome* $\omega_i \in \Omega$ such that:
  - $0 \leq p(\omega_i) \leq 1$
  - $p(\omega_1) + \ldots + p(\omega_k) = 1$
- Any mass function $p(\omega_i)$ defines a **(probability) distribution** $P(\mathcal{E})$, which assigns a probability to each *event* $\mathcal{E} \subseteq \Omega$:

$$P(\mathcal{E}) = \sum_{\{i|\omega_i \in \mathcal{E}\}} p(\omega_i)$$

Probability Distributions

- The **sample space** $\Omega = \{\omega_1, \ldots, \omega_k\}$ is the set of all possible outcomes of an experiment.
- An **event** $\mathcal{E} \subseteq \Omega$ is a (sub)set of possible outcomes.
- A **(probability) mass function** $p(\omega_i)$ assigns a weight to each *outcome* $\omega_i \in \Omega$ such that:
  - $0 \leq p(\omega_i) \leq 1$
  - $p(\omega_1) + \ldots + p(\omega_k) = 1$
- Any mass function $p(\omega_i)$ defines a **(probability) distribution** $P(\mathcal{E})$, which assigns a probability to each *event* $\mathcal{E} \subseteq \Omega$:

$$P(\mathcal{E}) = \sum_{\{i|\omega_i \in \mathcal{E}\}} p(\omega_i)$$

- **Frequentist** interpretation of $P(\mathcal{E})$: If we perform the experiment n times, then the **relative frequency** of observing an outcome $\omega_i \in \mathcal{E}$ goes to $P(\mathcal{E})$ as $n \to \infty$.

Examples of Probability Distributions

Example 1:
Suppose $\Omega = \{a, b, c\}$ **and** $p(a) = p(b) = p(c) = 1/3$**.**

- Then $P(\{a\}) = P(\{b\}) = P(\{c\}) = 1/3$,
- $P(\{a, b\}) = p(a) + p(b) = 2/3$,
- $P(\emptyset) = P(\{\}) = 0$,
- $P(\Omega) = P(\{a, b, c\}) = p(a) + p(b) + p(c) = 1$.

Examples of Probability Distributions

Example 1:
Suppose $\Omega = \{a, b, c\}$ **and** $p(a) = p(b) = p(c) = 1/3$**.**

- Then $P(\{a\}) = P(\{b\}) = P(\{c\}) = 1/3$,
- $P(\{a, b\}) = p(a) + p(b) = 2/3$,
- $P(\emptyset) = P(\{\}) = 0$,
- $P(\Omega) = P(\{a, b, c\}) = p(a) + p(b) + p(c) = 1$.

Example 2:
Suppose $\Omega = \{1, 2, \ldots, 10\}$ **and** $p(i) = i/55$**.**

- Then $P(\emptyset) = 0$, $P(\Omega) = 1$,
- $P(\{3, 4, 8\}) = (3 + 4 + 8)/55 = 3/11$.

Properties of Probability Distributions

The Impossible and the Certain Event:

$P(\emptyset) = \sum_{\{i|\omega_i \in \emptyset\}} p(\omega_i) = 0 \quad P(\Omega) = 1$

Organisational Matters

List-Then-Eliminate

Directed Graphs and Trees

Hypothesis Space: Decision Trees

ID3

Probability Distributions

Properties of Probability Distributions

The Impossible and the Certain Event:

$P(\emptyset) = \sum_{\{i | \omega_i \in \emptyset\}} p(\omega_i) = 0 \quad P(\Omega) = 1$

Combining Events:

For any two events $\mathcal{E}_1, \mathcal{E}_2 \subseteq \Omega$, the

- **union** $\mathcal{E}_1 \cup \mathcal{E}_2 = \{\omega_i \mid \omega_i \in \mathcal{E}_1$ or $\omega_i \in \mathcal{E}_2\}$ and
- **intersection** $\mathcal{E}_1 \cap \mathcal{E}_2 = \{\omega_i \mid \omega_i \in \mathcal{E}_1$ and $\omega_i \in \mathcal{E}_2\}$

are also events.

Properties of Probability Distributions

The Impossible and the Certain Event:

$P(\emptyset) = \sum_{\{i|\omega_i \in \emptyset\}} p(\omega_i) = 0 \quad P(\Omega) = 1$

Combining Events:

For any two events $\mathcal{E}_1, \mathcal{E}_2 \subseteq \Omega$, the

- **union** $\mathcal{E}_1 \cup \mathcal{E}_2 = \{\omega_i \mid \omega_i \in \mathcal{E}_1$ or $\omega_i \in \mathcal{E}_2\}$ and
- **intersection** $\mathcal{E}_1 \cap \mathcal{E}_2 = \{\omega_i \mid \omega_i \in \mathcal{E}_1$ and $\omega_i \in \mathcal{E}_2\}$

are also events.

Relating the Probability of Unions and Intersections:

$$P(\mathcal{E}_1 \cup \mathcal{E}_2) = P(\mathcal{E}_1) + P(\mathcal{E}_2) - P(\mathcal{E}_1 \cap \mathcal{E}_2) \tag{1}$$

Properties of Probability Distributions

The Impossible and the Certain Event:

$P(\emptyset) = \sum_{\{i|\omega_i \in \emptyset\}} p(\omega_i) = 0 \quad P(\Omega) = 1$

Combining Events:

For any two events $\mathcal{E}_1, \mathcal{E}_2 \subseteq \Omega$, the

- **union** $\mathcal{E}_1 \cup \mathcal{E}_2 = \{\omega_i \mid \omega_i \in \mathcal{E}_1$ or $\omega_i \in \mathcal{E}_2\}$ and
- **intersection** $\mathcal{E}_1 \cap \mathcal{E}_2 = \{\omega_i \mid \omega_i \in \mathcal{E}_1$ and $\omega_i \in \mathcal{E}_2\}$

are also events.

Relating the Probability of Unions and Intersections:

$$P(\mathcal{E}_1 \cup \mathcal{E}_2) = P(\mathcal{E}_1) + P(\mathcal{E}_2) - P(\mathcal{E}_1 \cap \mathcal{E}_2) \tag{1}$$

An Event Not Happening:

- For any event $\mathcal{E}$, its **complement** $\bar{\mathcal{E}} = \{\omega_i \mid \omega_i \notin \mathcal{E}\}$ is the event describing that $\mathcal{E}$ does **not** occur.
- It follows from (1) that $P(\bar{\mathcal{E}}) = 1 - P(\mathcal{E})$.

Conditional Probability

Suppose P is a probability distribution on sample space Ω, and $\mathcal{E}_1, \mathcal{E}_2 \subseteq \Omega$ are events.

Definition:

The **conditional probability** $P(\mathcal{E}_1 \mid \mathcal{E}_2)$ of $\mathcal{E}_1$ given $\mathcal{E}_2$ is

$$P(\mathcal{E}_1 \mid \mathcal{E}_2) = \frac{P(\mathcal{E}_1 \cap \mathcal{E}_2)}{P(\mathcal{E}_2)}.$$

Example:

Let $\Omega = \{aa, ab, ba, bb\}$. Then

$$P(\{ba\} \mid \{ab, ba\}) = \frac{P(\{ba\})}{P(\{ab, ba\})}.$$

Random Variables

Let $\Omega = \{\omega_1, \ldots, \omega_k\}$ be a sample space.

Definition: A **random variable** $X(\omega_i)$ is a function from Ω to $\mathbb{R}$.

Example:

Suppose $\Omega = \{aa, ab, ba, bb\}$. Then we might define the random variable that counts the number of a's in an outcome: $X(aa) = 2$, $X(ab) = 1$, $X(ba) = 1$, $X(bb) = 0$.

Random Variables

Let $\Omega = \{\omega_1, \ldots, \omega_k\}$ be a sample space.

Definition: A **random variable** $X(\omega_i)$ is a function from Ω to $\mathbb{R}$.

Example:

Suppose $\Omega = \{aa, ab, ba, bb\}$. Then we might define the random variable that counts the number of a's in an outcome: $X(aa) = 2$, $X(ab) = 1$, $X(ba) = 1$, $X(bb) = 0$.

Probability Distribution of a Random Variable:

- Suppose P is a probability distribution on Ω.
- We define the shorthand notation:

$$P(X = x) = P\big(\{\omega_i \mid X(\omega_i) = x\}\big).$$

Example Continued:

$$P(X = 1) = P(\{ab, ba\})$$

Overview

- Organisational Matters
- An Unbiased Hypothesis Space for List-Then-Eliminate?
- Math: Directed Graphs and Trees
- Decision Trees for Classification
  - Hypothesis Space: Decision Trees
  - Method: ID3
- Math: Probability Distributions

References

- D. Wood, ”Theory of Computation,” Harper and Row, Publishers, 1987.
- A.N. Shiryaev, “Probability”, Second Edition, 1996