10-301/601 Introduction to Machine Learning

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Decision Trees (Part II)

Matt Gormley & Henry Chai
Lecture 3
Sep. 8, 2021

Q&A

Q: In our medical diagnosis example, suppose two of our doctors (i.e. experts) disagree about whether (+) or not (-) the patient is sick. How would the decision tree represent this situation?

A: Today we will define decision trees that predict a single class by a majority vote at the leaf. More generally, the leaf could provide a probability distribution over output classes $p(y|\mathbf{x})$

Q&A

Q: What's this new collaboration policy for HW2?

A:

- For each programming assignment, you will be **randomly assigned** to a homework group of 3. Homework groups will be different for each programming assignment.
- Within that homework group, you will be able to collaborate more fully than before. Specifically, you are permitted to **show your code (either in-person or via Zoom screen share) to members of your homework group**.
- **Honor policy:** you may **not** write code while someone else's code is shared with you. That is, you are not permitted to copy down someone else's code. Though you're certainly welcome to take **mental** notes, and learn from the design decisions they've made.
- All discussion and screen sharing between homework group members must happen either in person or on Zoom using your Andrew accounts.

Q&A

Q: How do these In-Class Polls work?

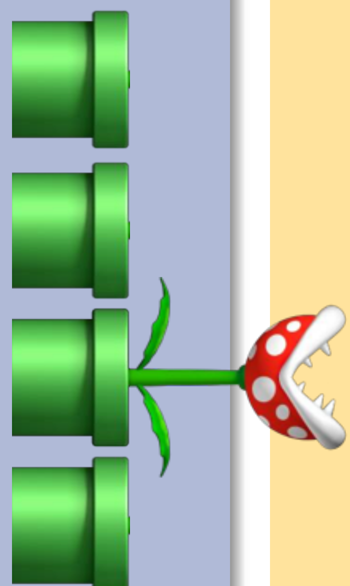
A:
- Sign into **Poll Everywhere** (link from Schedule page, http://mlcourse.org/schedule.html) using **Andrew Email**
- Answer **during lecture for full credit,** or within 24 hours for half credit
- Avoid the **toxic option** which gives negative points!
- 8 "free poll points" but can't use more than 3 free polls consecutively. All the questions for one lecture are worth 1 point total.
- Submit a **poll card** if and only if you do not have a smartphone/tablet

First In-Class Poll

Question:

Which of the following did you bring to class today?

A. Smartphone
B. Flip phone
C. ~~Pay phone~~
D. No phone



Answer:

When poll is active, respond at **pollev.com/10301601polls**

Question 1

A

B

C

D

Reminders

- **Homework 1: Background**
  - **Out: Wed, Sep 1 (2nd lecture)**
  - **Due: Wed, Sep 8 at 11:59pm**
  - unique policy for this assignment: we will grant (essentially) any and all extension requests
- **Homework 2: Decision Trees**
  - **Out: Wed, Sep. 8**
  - **Due: Wed, Sep. 20 at 11:59pm**

MAKING PREDICTIONS WITH A DECISION TREES

Decision Trees

Whiteboard

- Example Decision Tree as a hypothesis
- Defining h(x) for a decision tree
- Paper Decision Tree
  - Question 1: Given a fully specified tree, how do we make a prediction on a unseen (unlabeled) instance?
  - Question 2: Given a tree structure (i.e. all the splits), how do we learn the labels of the leaf nodes from (labeled) data?
  - Question 3: If we change the tree structure (i.e. add a split), does that change the predictions we make?
  - Question 4: Given just labeled data, how do we learn a tree structure? (NEXT SECTION)

Tree to Predict C-Section Risk

Learned from medical records of 1000 women (Sims et al., 2000)

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+ .0
| | | | 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-
```

Figure from Tom Mitchell

LEARNING A DECISION TREE

Decision Trees

Whiteboard

- Decision Tree Learning

Decision Tree Learning Example

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

In-Class Exercise

Using **error rate** as the splitting criterion, what decision tree would be learned?

Decision Trees

Whiteboard

- Example of Decision Tree Learning with Error Rate as splitting criterion

SPLITTING CRITERION: ERROR RATE

Decision Tree Learning

- *Definition*: a **splitting criterion** is a function that measures the effectiveness of splitting on a particular attribute
- Our decision tree learner **selects the "best" attribute** as the one that maximizes the splitting criterion
- Lots of options for a splitting criterion:
  - error rate (or *accuracy* if we want to pick the tree that *maximizes* the criterion) 1 - error
  - Gini gain
  - Mutual information
  - random
  - …

Decision Tree Learning Example

Question 2

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

In-Class Exercise

Which attribute would **error rate** select for the next split?

1. A
2. B
3. A or B (tie)
4. ~~Neither~~ toxic

Question 2

1

2

3

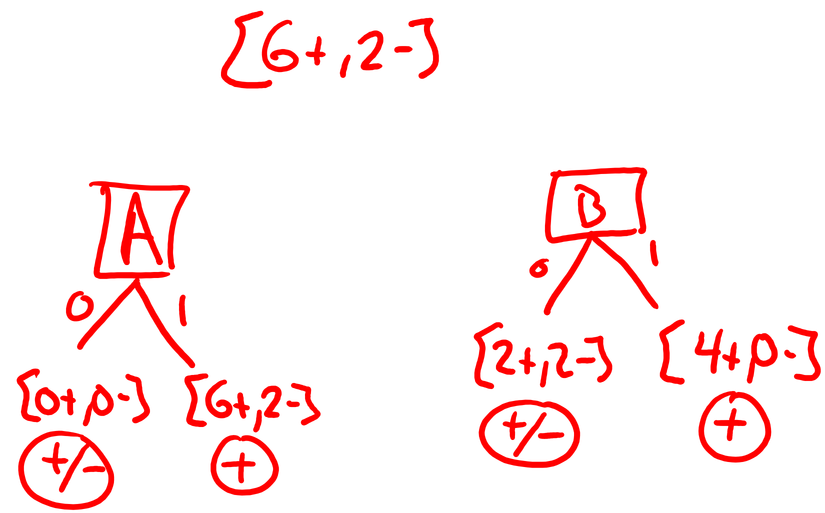
4

Decision Tree Learning Example

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



$\text{error}_A = 2/8$

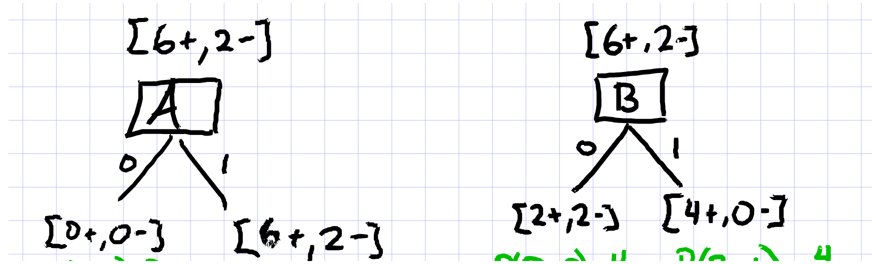
$\text{error}_B = 1/4 = 2/8$

Decision Tree Learning Example

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



Misclass. Rate

$r(A) = 2/8$

$r(A) = 2/8$

r(·) treats attributes as equally good

SPLITTING CRITERION: MUTUAL INFORMATION

Information Theory & DTs

Whiteboard

- Information Theory primer
  - Entropy
  - (Specific) Conditional Entropy
  - Conditional Entropy
  - Information Gain / Mutual Information
- Information Gain as DT splitting criterion

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

- 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.

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

- **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?*

Decision Tree Learning Example

Q3:

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

In-Class Exercise

Which attribute would **mutual information** select for the next split?

1. A
2. B
3. A or B (tie)
4. Neither

When poll is active, respond at **pollev.com/10301601polls**

Question 3

1

2

3

4

Decision Tree Learning Example

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

$\boxed{A}$: 0 → {0+,0-}, 1 → [6+,2-]

$\boxed{B}$: 0 → {2+,2-}, 1 → [4+,0-]

$H(Y;D) = -\left(2/8 \log_2 2/8 + 6/8 \log_2 6/8\right)$

A: $I(Y;A;D) = H(Y;D) - \left(P(A=0)\, H(Y,D_{A=0}) + P(A=1)\, \underbrace{H(Y,D_{A=1})}_{H(Y;D)}\right)$

$P(A=0) \to 0$, $P(A=1) \to 1$

$= 0$

B: $I(Y;B;D) = H(Y;D) - \left(4/8\, H(Y;D_{B=0}) + 4/8\, H(Y;D_{B=1})\right)$

$H(Y;D_{B=0}) = -\left(2/4 \log_2 2/4 + 2/4 \log_2 2/4\right)$, $H(Y;D_{B=1}) \to 0$

> 0

Tennis Example

Test your understanding

Dataset:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis?
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Figure from Tom Mitchell

Tennis Example

Test your understanding

Which attribute yields the best classifier?

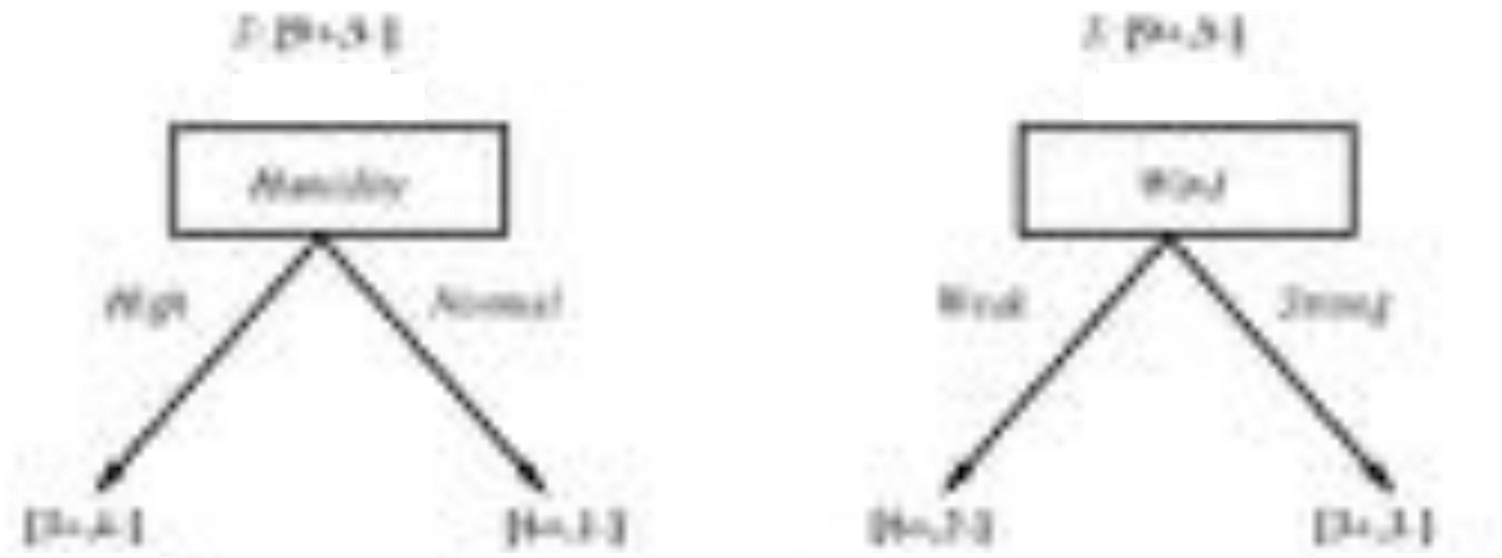


Figure from Tom Mitchell

Tennis Example

Test your understanding

Which attribute yields the best classifier?

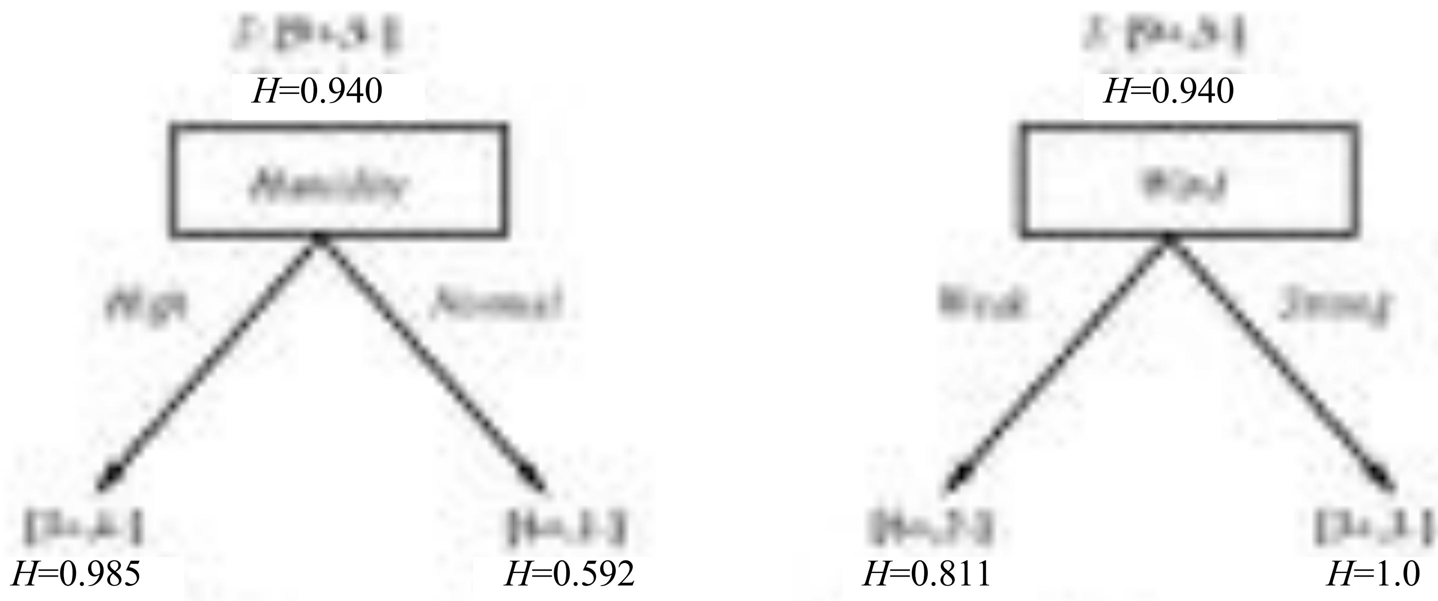


Figure from Tom Mitchell

Tennis Example

Test your understanding

Which attribute yields the best classifier?

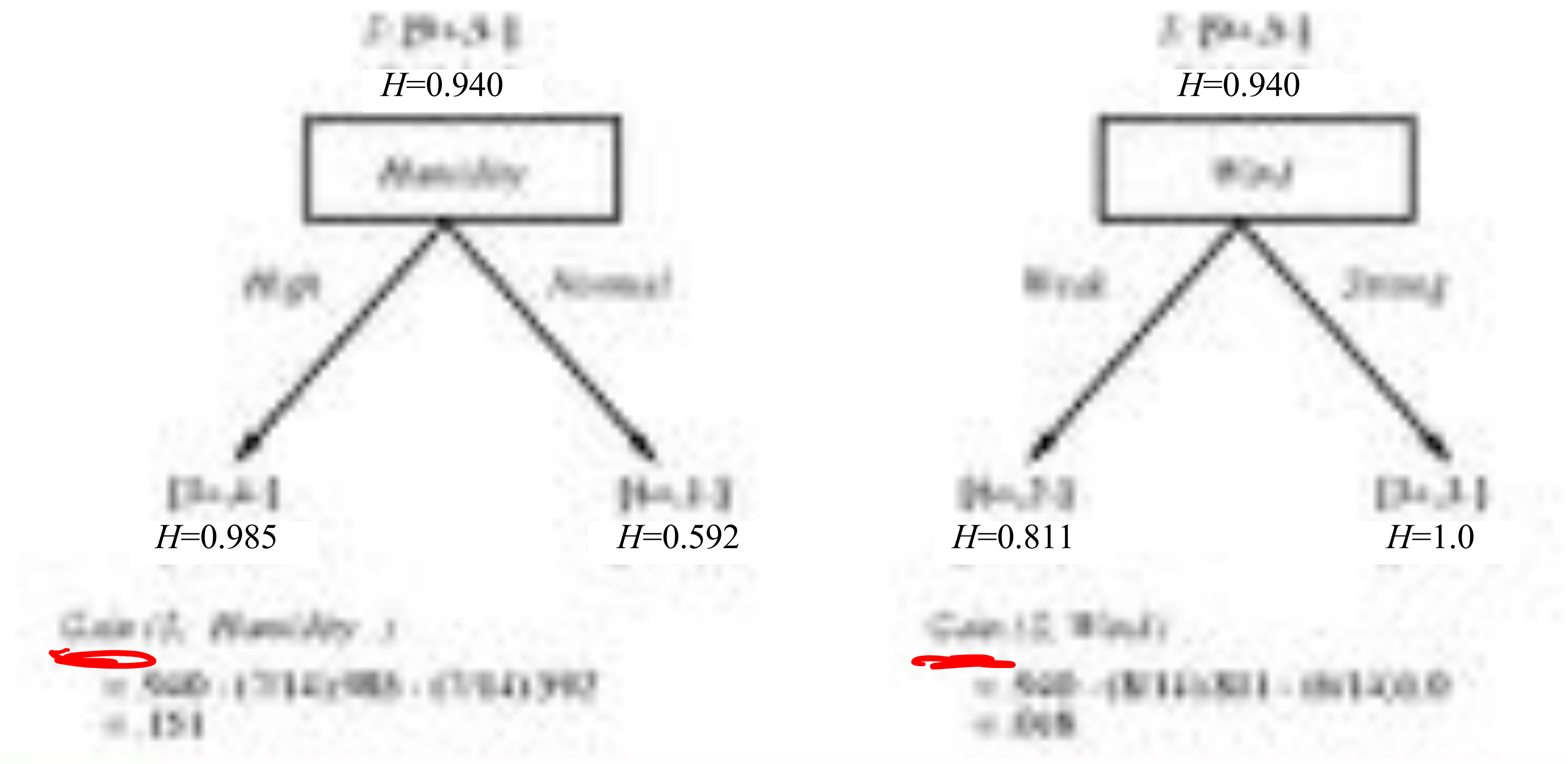


Figure from Tom Mitchell

Tennis Example

Test your understanding

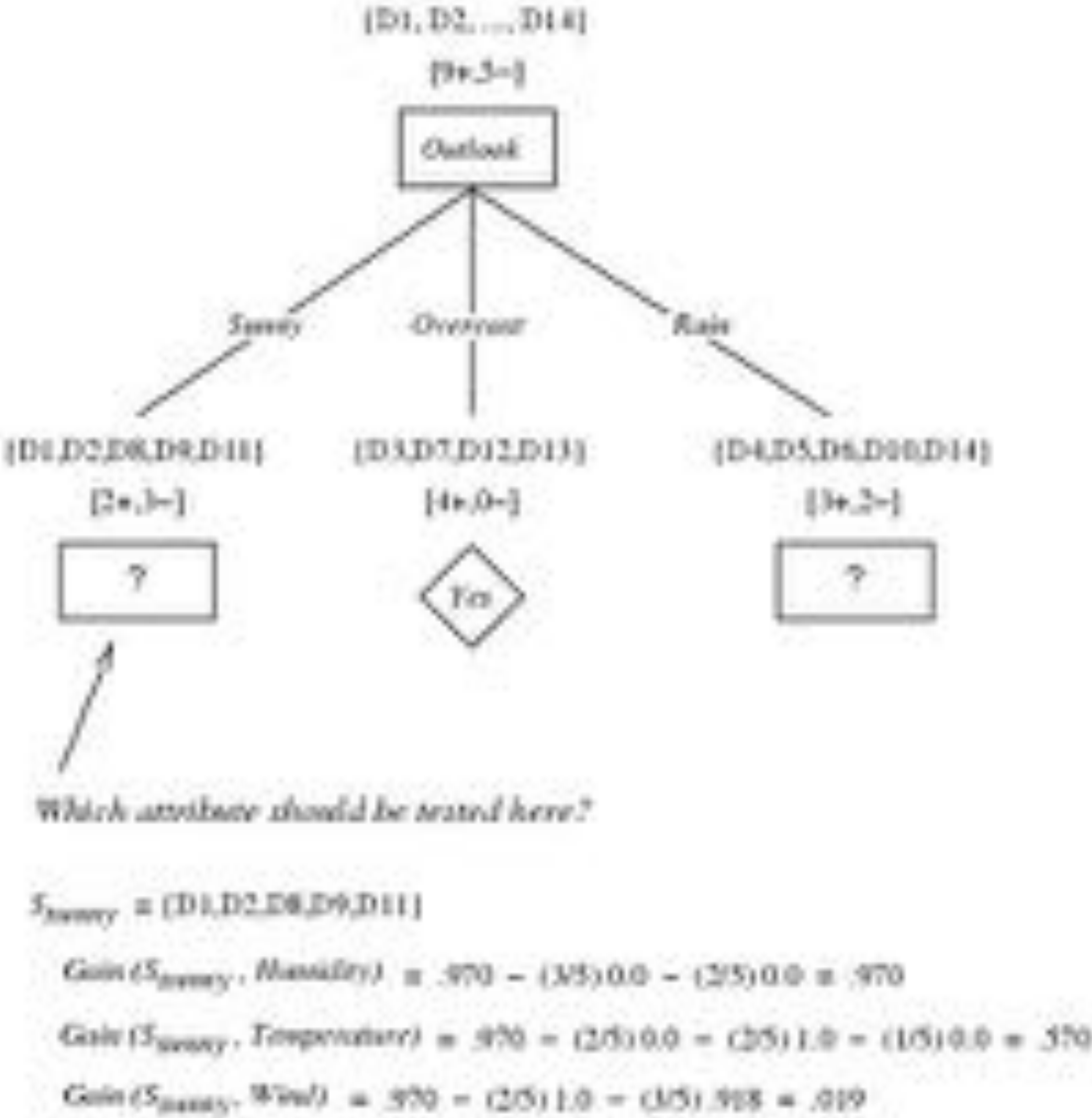


Figure from Tom Mitchell