Machine Learning Department
School of Computer Science
Carnegie Mellon University

Decision Trees (Part II)

Matt Gormley
Lecture 3
September 5, 2018

Q&A

Q: In our medical diagnosis example, suppose two of our doctors (i.e. experts) disagree about whether (+) or not (-) to prescribe. 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 $p(y|\mathbf{x})$

Reminders

- **Homework 1: Background**
  - **Out: Wed, Aug 29**
  - **Due: Wed, Sep 05 at 11:59pm**
  - unique policy for this assignment: we will grant (essentially) any and all extension requests
- **Homework 2: Decision Trees**
  - **Out: Wed, Sep 05**
  - **Due: Wed, Sep 19 at 11:59pm**

DECISION TREES

Decision Trees

Chalkboard

- Example: Medical Diagnosis
- Does memorization = learning?
- Decision Tree as a hypothesis
- Function approximation for DTs

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+ .
| | | | 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

Decision Trees

Chalkboard

- An Aside:
  - The Majority Vote Classifier
  - Error Rate
- Decision Tree Learning
- Information Theory primer
  - Entropy
  - (Specific) Conditional Entropy
  - Conditional Entropy
  - Information Gain / Mutual Information
- Information Gain as DT splitting criterion

Majority Vote Classifier 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

In-Class Exercise

What is the **training error** (i.e. *error rate on the training data*) of the **majority vote classifier** on this dataset?

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 by ID3?

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

In-Class Exercise

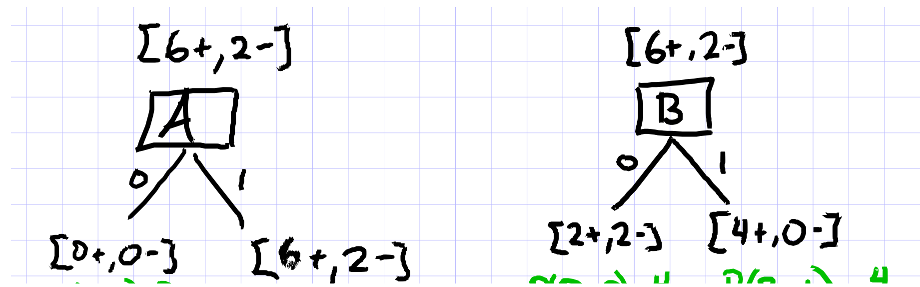
1. Which attribute would **misclassification rate** select for the next split?
2. Which attribute would **information gain** select for the next split?
3. *Justify your answers.*

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

$r(A) = 2/8$

$r(A) = 2/8$

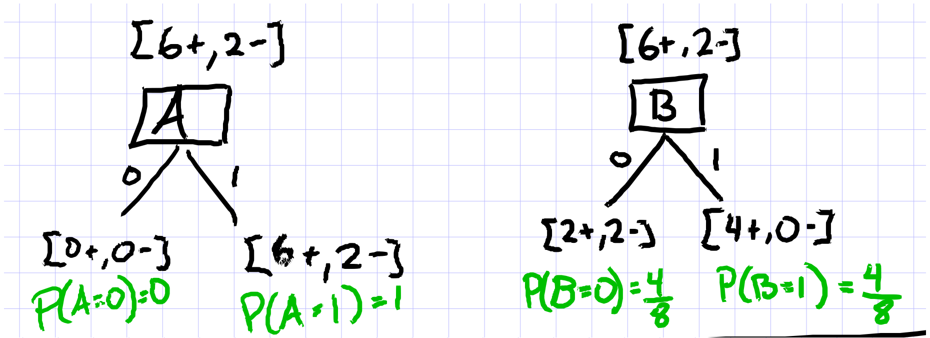
r(·) treats attributes as equally good

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



Info Gain

$H(Y) = 2/8 \log 2/8 + 6/8 \log 6/8$

A

$H(Y|A=0) = \text{"undef"}$

$H(Y|A=1) = 2/8 \log 2/8 + 6/8 \log 6/8 = H(Y)$

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

$= 0 + H(Y|A=1) = H(Y)$

$I(Y;A) = H(Y) - H(Y|B=1) = 0$

B

$H(Y|B=0) = 2/4 \log 2/4 + 2/4 \log 2/4$

$H(Y|B=1) = 0 \log 0 + 1 \log 1 = 0$

$H(Y|B) = 4/8\,(0) + 4/8\,(H(Y|B=0))$

$I(Y;B) = H(Y) - 4/8\,H(Y|B=0) > 0$

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?

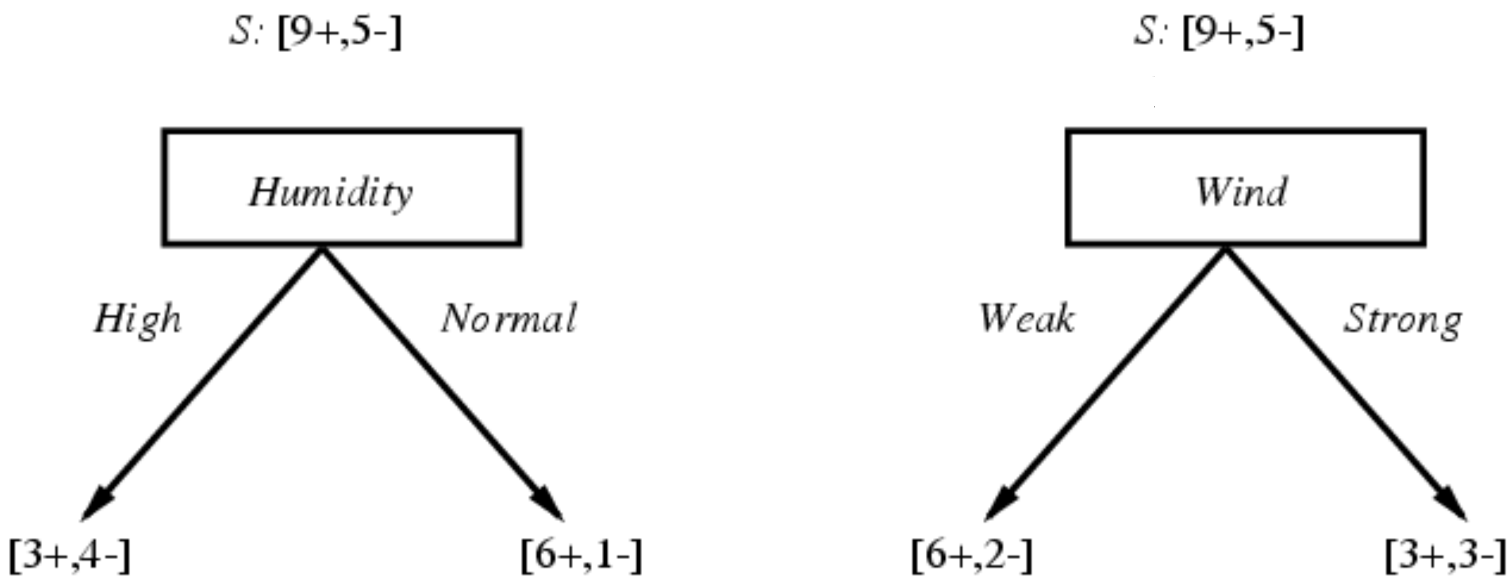


Figure from Tom Mitchell

Tennis Example

Test your understanding

Which attribute yields the best classifier?

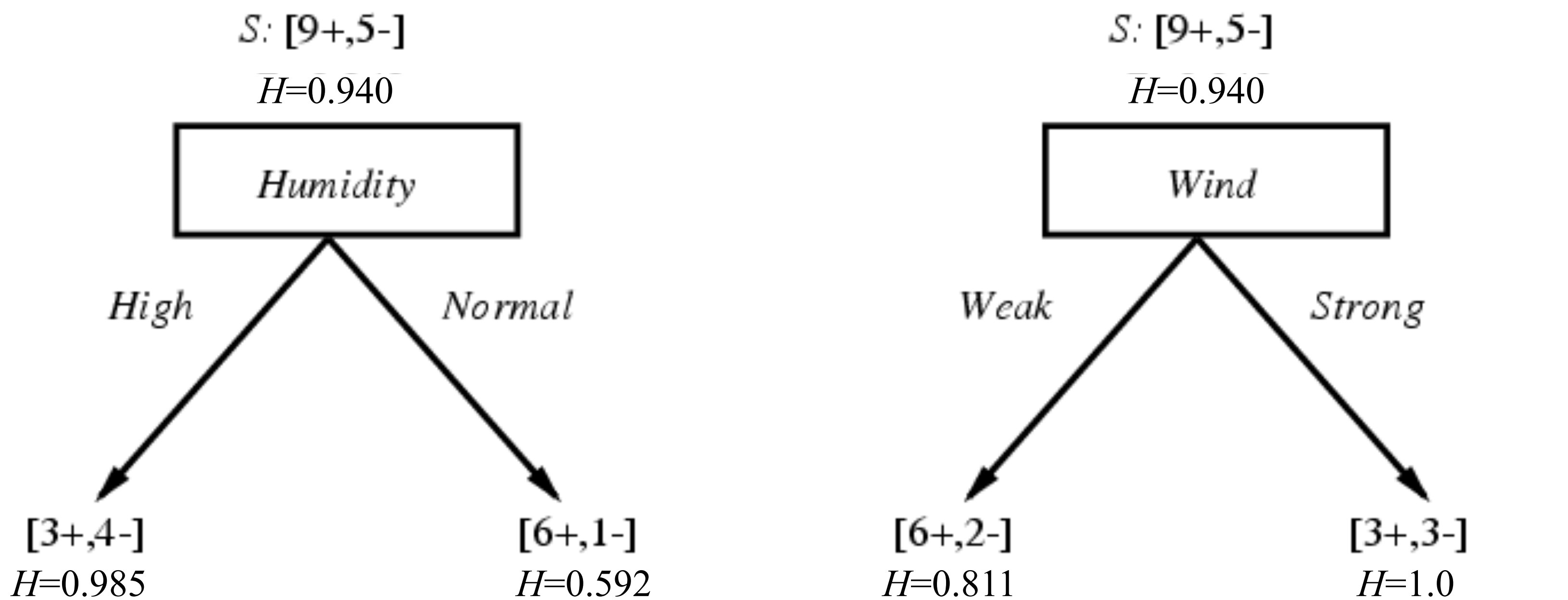
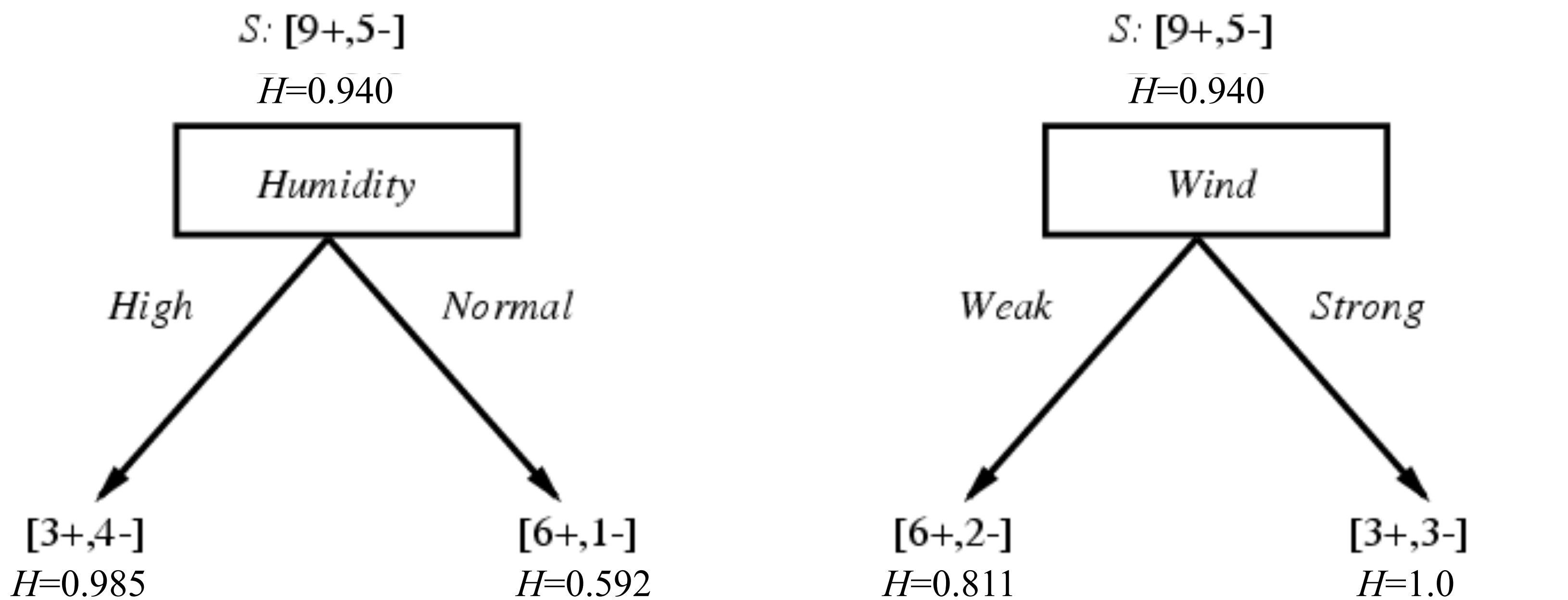


Figure from Tom Mitchell

Tennis Example

Test your understanding

Which attribute yields the best classifier?



Gain (S, Humidity)

= .940 - (7/14).985 - (7/14).592
= .151

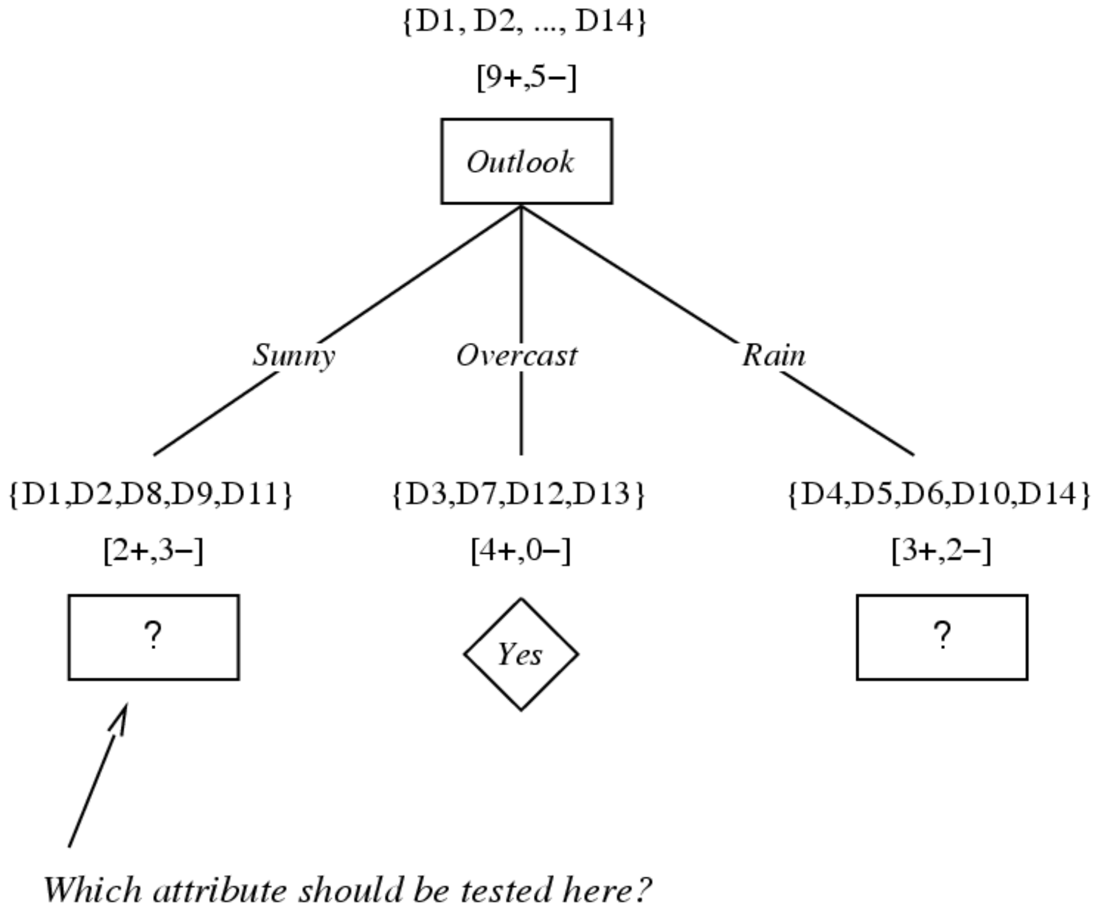
Gain (S, Wind)

= .940 - (8/14).811 - (6/14)1.0
= .048

Figure from Tom Mitchell

Tennis Example

Test your understanding



$S_{sunny} = \{D1,D2,D8,D9,D11\}$

$Gain\ (S_{sunny}, Humidity) = .970 - (3/5)\ 0.0 - (2/5)\ 0.0 = .970$

$Gain\ (S_{sunny}, Temperature) = .970 - (2/5)\ 0.0 - (2/5)\ 1.0 - (1/5)\ 0.0 = .570$

$Gain\ (S_{sunny}, Wind) = .970 - (2/5)\ 1.0 - (3/5)\ .918 = .019$

Figure from Tom Mitchell