CSC 411: Lecture 06 - Decision Trees

Ethan Fetaya, James Lucas and Emad Andrews

University of Toronto
Today

- **Decision Trees**
  - entropy
  - information gain
Another Classification Idea

- We learned about linear classification (e.g., logistic regression), and nearest neighbors. Any other idea?

- **Pick an attribute**, do a simple test

- Conditioned on a choice, pick another attribute, do another test

- In the leaves, assign a class with **majority vote**

- Do other branches as well
Another Classification Idea

- Gives *axes aligned decision boundaries*
Decision Tree: Example

width > 6.5cm?

Yes

height > 9.5cm?

Yes

height > 6.0cm?

Yes

No
Decision Tree: Classification

Test example

- **width > 6.5cm?**
  - Yes
  - **height > 9.5cm?**
    - Yes
    - **height > 9.5cm?**
      - Yes
      - Lemon
    - No
      - Orange
  - No
    - **height > 6.0cm?**
      - Yes
      - Lemon
    - No
      - Orange
Example with Discrete Inputs

What if the attributes are discrete?

<table>
<thead>
<tr>
<th>Example</th>
<th>Alt</th>
<th>Bar</th>
<th>Fri</th>
<th>Hungry</th>
<th>Patrons</th>
<th>Price</th>
<th>Raining</th>
<th>Reservations</th>
<th>Type</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>x₁</strong></td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Some</td>
<td>$$$</td>
<td>No</td>
<td>Yes</td>
<td>French</td>
<td>0–10</td>
</tr>
<tr>
<td><strong>x₂</strong></td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Full</td>
<td>$</td>
<td>No</td>
<td>No</td>
<td>Thai</td>
<td>30–60</td>
</tr>
<tr>
<td><strong>x₃</strong></td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Some</td>
<td>$</td>
<td>No</td>
<td>No</td>
<td>Burger</td>
<td>0–10</td>
</tr>
<tr>
<td><strong>x₄</strong></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Full</td>
<td>$</td>
<td>Yes</td>
<td>No</td>
<td>Thai</td>
<td>10–30</td>
</tr>
<tr>
<td><strong>x₅</strong></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Full</td>
<td>$$$</td>
<td>No</td>
<td>Yes</td>
<td>French</td>
<td>&gt;60</td>
</tr>
<tr>
<td><strong>x₆</strong></td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Some</td>
<td>$$</td>
<td>Yes</td>
<td>Yes</td>
<td>Italian</td>
<td>0–10</td>
</tr>
<tr>
<td><strong>x₇</strong></td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>None</td>
<td>$</td>
<td>Yes</td>
<td>No</td>
<td>Burger</td>
<td>0–10</td>
</tr>
<tr>
<td><strong>x₈</strong></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Some</td>
<td>$$</td>
<td>Yes</td>
<td>Yes</td>
<td>Thai</td>
<td>0–10</td>
</tr>
<tr>
<td><strong>x₉</strong></td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Full</td>
<td>$</td>
<td>Yes</td>
<td>No</td>
<td>Burger</td>
<td>&gt;60</td>
</tr>
<tr>
<td><strong>x₁₀</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Full</td>
<td>$$$</td>
<td>No</td>
<td>Yes</td>
<td>Italian</td>
<td>10–30</td>
</tr>
<tr>
<td><strong>x₁₁</strong></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>None</td>
<td>$</td>
<td>No</td>
<td>No</td>
<td>Thai</td>
<td>0–10</td>
</tr>
<tr>
<td><strong>x₁₂</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Full</td>
<td>$</td>
<td>No</td>
<td>No</td>
<td>Burger</td>
<td>30–60</td>
</tr>
</tbody>
</table>

**Goal:**

<table>
<thead>
<tr>
<th>Goal</th>
<th>WillWait</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>y₁</strong></td>
<td>Yes</td>
</tr>
<tr>
<td><strong>y₂</strong></td>
<td>No</td>
</tr>
<tr>
<td><strong>y₃</strong></td>
<td>Yes</td>
</tr>
<tr>
<td><strong>y₄</strong></td>
<td>Yes</td>
</tr>
<tr>
<td><strong>y₅</strong></td>
<td>No</td>
</tr>
<tr>
<td><strong>y₆</strong></td>
<td>Yes</td>
</tr>
<tr>
<td><strong>y₇</strong></td>
<td>No</td>
</tr>
<tr>
<td><strong>y₈</strong></td>
<td>No</td>
</tr>
<tr>
<td><strong>y₉</strong></td>
<td>No</td>
</tr>
<tr>
<td><strong>y₁₀</strong></td>
<td>No</td>
</tr>
<tr>
<td><strong>y₁₁</strong></td>
<td>No</td>
</tr>
<tr>
<td><strong>y₁₂</strong></td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Attributes:**

1. Alternate: whether there is a suitable alternative restaurant nearby.
2. Bar: whether the restaurant has a comfortable bar area to wait in.
3. Fri/Sat: true on Fridays and Saturdays.
4. Hungry: whether we are hungry.
5. Patrons: how many people are in the restaurant (values are None, Some, and Full).
6. Price: the restaurant’s price range ($, $$, $$$).
7. Raining: whether it is raining outside.
8. Reservation: whether we made a reservation.
9. Type: the kind of restaurant (French, Italian, Thai or Burger).
10. WaitEstimate: the wait estimated by the host (0-10 minutes, 10-30, 30-60, >60).
Decision Tree: Example with Discrete Inputs

- The tree to decide whether to wait (T) or not (F)
**Decision Trees**

- **Internal nodes** test attributes
- Branching is determined by attribute value
- Leaf nodes are outputs (class assignments)
Decision Tree: Algorithm

- Choose an attribute on which to descend at each level
- Condition on earlier (higher) choices
- Generally, restrict only one dimension at a time
- Declare an output value when you get to the bottom
- In the orange/lemon example, we only split each dimension once, but that is not required
Decision Tree: Classification and Regression

- Each path from root to a leaf defines a region $R_m$ of input space
- Let $\{(x^{(m_1)}, t^{(m_1)}), \ldots, (x^{(m_k)}, t^{(m_k)})\}$ be the training examples that fall into $R_m$

**Classification tree:**
- discrete output
- leaf value $y^m$ typically set to the most common value in $\{t^{(m_1)}, \ldots, t^{(m_k)}\}$

**Regression tree:**
- continuous output
- leaf value $y^m$ typically set to the mean value in $\{t^{(m_1)}, \ldots, t^{(m_k)}\}$

Note: We will only talk about classification
Expressiveness

- **Discrete-input, discrete-output case:**
  - Decision trees can express any function of the input attributes
  - E.g., for Boolean functions, truth table row $\rightarrow$ path to leaf:

```
A  B  A xor B
F  F  F
F  T  T
T  F  T
T  T  F
```

- **Continuous-input, continuous-output case:**
  - Can approximate any function arbitrarily closely
  - Trivially, there is a consistent decision tree for any training set w/ one path to leaf for each example (unless $f$ nondeterministic in $x$) but it probably won’t generalize to new examples

Need some kind of regularization to ensure more **compact** decision trees

[Slide credit: S. Russell]
How do we Learn a Decision Tree?

- How do we construct a useful decision tree?
Learning Decision Trees

Learning the simplest (smallest) decision tree is an NP complete problem [if you are interested, check: Hyafil & Rivest’76]

- Resort to a greedy heuristic:
  - Start from an empty decision tree
  - Split on next best attribute
  - Recurse

- What is best attribute?

- How about accuracy?
Choosing a Good Split

- Why isn’t accuracy a good measure?

- Is this split good? Zero accuracy gain.

- We will use intuition from information theory

Idea: Use counts at leaves to define probability distributions, so we can measure uncertainty
Choosing a Good Split

- Which attribute is better to split on, $X_1$ or $X_2$?
  - Deterministic: good (all are true or false; just one class in the leaf)
  - Uniform distribution: bad (all classes in leaf equally probable)
  - What about distributions in between?

Note: Let’s take a slight detour and remember concepts from information theory

[Slide credit: D. Sontag]
We Flip Two Different Coins

Sequence 1:
0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 ... ?

Sequence 2:
0 1 0 1 0 1 1 1 0 1 0 0 1 1 0 1 0 1 ... ?

16 2

versus

0 1

8 10

0 1
Quantifying Uncertainty

**Entropy $H$:**

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

- $\frac{8}{9} \log_2 \frac{8}{9} - \frac{1}{9} \log_2 \frac{1}{9} \approx \frac{1}{2}$
- $\frac{4}{9} \log_2 \frac{4}{9} - \frac{5}{9} \log_2 \frac{5}{9} \approx 0.99$

- How surprised are we by a new value in the sequence?
- How much information does it convey?
Quantifying Uncertainty

\[ H(X) = - \sum_{x \in X} p(x) \log_2 p(x) \]
Entropy

- **“High Entropy”**: 
  - Variable has a uniform like distribution
  - Flat histogram
  - Values sampled from it are less predictable

- **“Low Entropy”**
  - Distribution of variable has many peaks and valleys
  - Histogram has many lows and highs
  - Values sampled from it are more predictable

[Slide credit: Vibhav Gogate]
Example: $X = \{\text{Raining, Not raining}\}$, $Y = \{\text{Cloudy, Not cloudy}\}$

<table>
<thead>
<tr>
<th></th>
<th>Cloudy</th>
<th>Not Cloudy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raining</td>
<td>24/100</td>
<td>1/100</td>
</tr>
<tr>
<td>Not Raining</td>
<td>25/100</td>
<td>50/100</td>
</tr>
</tbody>
</table>

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(x, y)$$

$$= -\frac{24}{100} \log_2 \frac{24}{100} - \frac{1}{100} \log_2 \frac{1}{100} - \frac{25}{100} \log_2 \frac{25}{100} - \frac{50}{100} \log_2 \frac{50}{100}$$

$$\approx 1.56 \text{bits}$$
Specific Conditional Entropy

- Example: \( X = \{ \text{Raining, Not raining} \} \), \( Y = \{ \text{Cloudy, Not cloudy} \} \)

<table>
<thead>
<tr>
<th></th>
<th>Cloudy</th>
<th>Not Cloudy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raining</td>
<td>24/100</td>
<td>1/100</td>
</tr>
<tr>
<td>Not Raining</td>
<td>25/100</td>
<td>50/100</td>
</tr>
</tbody>
</table>

What is the entropy of cloudiness \( Y \), given that it is raining?

\[
H(Y|X = x) = - \sum_{y \in Y} p(y|x) \log_2 p(y|x)
\]

\[
= - \frac{24}{25} \log_2 \frac{24}{25} - \frac{1}{25} \log_2 \frac{1}{25}
\]

\[
\approx 0.24 \text{bits}
\]

- We used: \( p(y|x) = \frac{p(x,y)}{p(x)} \), and \( p(x) = \sum_y p(x,y) \) (sum in a row)
## Conditional Entropy

<table>
<thead>
<tr>
<th></th>
<th>Cloudy</th>
<th>Not Cloudy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raining</td>
<td>24/100</td>
<td>1/100</td>
</tr>
<tr>
<td>Not Raining</td>
<td>25/100</td>
<td>50/100</td>
</tr>
</tbody>
</table>

- The expected conditional entropy:

\[
H(Y|X) = \sum_{x \in X} p(x) H(Y|X = x) \\
= - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(y|x)
\]
Example: \( X = \{ \text{Raining, Not raining} \} \), \( Y = \{ \text{Cloudy, Not cloudy} \} \)

<table>
<thead>
<tr>
<th></th>
<th>Cloudy</th>
<th>Not Cloudy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raining</td>
<td>24/100</td>
<td>1/100</td>
</tr>
<tr>
<td>Not Raining</td>
<td>25/100</td>
<td>50/100</td>
</tr>
</tbody>
</table>

What is the entropy of cloudiness, given the knowledge of whether or not it is raining?

\[
H(Y|X) = \sum_{x \in X} p(x)H(Y|X = x)
\]

\[
= \frac{1}{4} H(\text{cloudy|is raining}) + \frac{3}{4} H(\text{cloudy|not raining})
\]

\[
\approx 0.75 \text{ bits}
\]
Some useful properties:

- $H$ is always non-negative
- Chain rule: $H(X, Y) = H(X|Y) + H(Y) = H(Y|X) + H(X)$
- If $X$ and $Y$ independent, then $X$ doesn’t tell us anything about $Y$: $H(Y|X) = H(Y)$
- But $Y$ tells us everything about $Y$: $H(Y|Y) = 0$
- By knowing $X$, we can only decrease uncertainty about $Y$: $H(Y|X) \leq H(Y)$
Information Gain

<table>
<thead>
<tr>
<th></th>
<th>Cloudy</th>
<th>Not Cloudy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raining</td>
<td>24/100</td>
<td>1/100</td>
</tr>
<tr>
<td>Not Raining</td>
<td>25/100</td>
<td>50/100</td>
</tr>
</tbody>
</table>

How much information about cloudiness do we get by discovering whether it is raining?

\[
IG(Y|X) = H(Y) - H(Y|X) 
\approx 0.25 \text{ bits}
\]

- Also called information gain in \( Y \) due to \( X \)
- If \( X \) is completely uninformative about \( Y \): \( IG(Y|X) = 0 \)
- If \( X \) is completely informative about \( Y \): \( IG(Y|X) = H(Y) \)
- How can we use this to construct our decision tree?
Revisiting Our Original Example

- What is the information gain of this split?

Root entropy:
\[
H(Y) = -\frac{49}{149} \log_2\left(\frac{49}{149}\right) - \frac{100}{149} \log_2\left(\frac{100}{149}\right) \approx 0.91
\]

Leafs entropy:
\[
H(Y|left) = 0, \quad H(Y|right) \approx 1.
\]

\[
IG(split) \approx 0.91 - \left(\frac{1}{3} \cdot 0 + \frac{2}{3} \cdot 1\right) \approx 0.24 > 0
\]
I made the fruit data partitioning just by eyeballing it.

We can use the information gain to automate the process.

At each level, one must choose:

1. Which variable to split.
2. Possibly where to split it.

Choose them based on how much information we would gain from the decision! (choose attribute that gives the highest gain)
Decision Tree Construction Algorithm

- Simple, greedy, recursive approach, builds up tree node-by-node

1. pick an attribute to split at a non-terminal node
2. split examples into groups based on attribute value
3. for each group:
   - if no examples – return majority from parent
   - else if all examples in same class – return class
   - else loop to step 1
## Attributes:

1. **Alternate:** whether there is a suitable alternative restaurant nearby.
2. **Bar:** whether the restaurant has a comfortable bar area to wait in.
3. **Fri/Sat:** true on Fridays and Saturdays.
4. **Hungry:** whether we are hungry.
5. **Patrons:** how many people are in the restaurant (values are None, Some, and Full).
6. **Price:** the restaurant’s price range ($, $$, $$$).
7. **Raining:** whether it is raining outside.
8. **Reservation:** whether we made a reservation.
9. **Type:** the kind of restaurant (French, Italian, Thai or Burger).
10. **WaitEstimate:** the wait estimated by the host (0-10 minutes, 10-30, 30-60, >60).

---

### Input Attributes

<table>
<thead>
<tr>
<th>Example</th>
<th>Alt</th>
<th>Bar</th>
<th>Fri</th>
<th>Hun</th>
<th>Pat</th>
<th>Price</th>
<th>Rain</th>
<th>Res</th>
<th>Type</th>
<th>Est</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Some</td>
<td>$$$</td>
<td>No</td>
<td>Yes</td>
<td>French</td>
<td>0–10</td>
</tr>
<tr>
<td>x2</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Full</td>
<td>$</td>
<td>No</td>
<td>No</td>
<td>Thai</td>
<td>30–60</td>
</tr>
<tr>
<td>x3</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Some</td>
<td>$</td>
<td>No</td>
<td>No</td>
<td>Burger</td>
<td>0–10</td>
</tr>
<tr>
<td>x4</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Full</td>
<td>$</td>
<td>Yes</td>
<td>No</td>
<td>Thai</td>
<td>10–30</td>
</tr>
<tr>
<td>x5</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Full</td>
<td>$$$</td>
<td>No</td>
<td>Yes</td>
<td>French</td>
<td>&gt;60</td>
</tr>
<tr>
<td>x6</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Some</td>
<td>$$</td>
<td>Yes</td>
<td>Yes</td>
<td>Italian</td>
<td>0–10</td>
</tr>
<tr>
<td>x7</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>None</td>
<td>$</td>
<td>Yes</td>
<td>No</td>
<td>Burger</td>
<td>0–10</td>
</tr>
<tr>
<td>x8</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Some</td>
<td>$$</td>
<td>Yes</td>
<td>Yes</td>
<td>Thai</td>
<td>0–10</td>
</tr>
<tr>
<td>x9</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Full</td>
<td>$</td>
<td>Yes</td>
<td>No</td>
<td>Burger</td>
<td>&gt;60</td>
</tr>
<tr>
<td>x10</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Full</td>
<td>$$$</td>
<td>No</td>
<td>Yes</td>
<td>Italian</td>
<td>10–30</td>
</tr>
<tr>
<td>x11</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>None</td>
<td>$</td>
<td>No</td>
<td>No</td>
<td>Thai</td>
<td>0–10</td>
</tr>
<tr>
<td>x12</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Full</td>
<td>$</td>
<td>No</td>
<td>No</td>
<td>Burger</td>
<td>30–60</td>
</tr>
</tbody>
</table>

### Goal

- \( y_1 = \text{Yes} \)
- \( y_2 = \text{No} \)
- \( y_3 = \text{Yes} \)
- \( y_4 = \text{Yes} \)
- \( y_5 = \text{No} \)
- \( y_6 = \text{Yes} \)
- \( y_7 = \text{Yes} \)
- \( y_8 = \text{Yes} \)
- \( y_9 = \text{No} \)
- \( y_{10} = \text{No} \)
- \( y_{11} = \text{No} \)
- \( y_{12} = \text{Yes} \)
Attribute Selection

\[ IG(Y) = H(Y) - H(Y|X) \]

\[ IG(type) = 1 - \left[ \frac{2}{12} H(Y|Fr.) + \frac{2}{12} H(Y|It.) + \frac{4}{12} H(Y|Thai) + \frac{4}{12} H(Y|Bur.) \right] = 0 \]

\[ IG(Patrons) = 1 - \left[ \frac{2}{12} H(0, 1) + \frac{4}{12} H(1, 0) + \frac{6}{12} H(\frac{2}{6}, \frac{4}{6}) \right] \approx 0.541 \]
Which Tree is Better?
What Makes a Good Tree?

- **Not too small:** need to handle important but possibly subtle distinctions in data
- **Not too big:**
  - Computational efficiency (avoid redundant, spurious attributes)
  - Avoid over-fitting training examples
- **Occam’s Razor:** find the simplest hypothesis (smallest tree) that fits the observations
- **Inductive bias:** small trees with informative nodes near the root
Decision Tree Miscellany

- Problems:
  - You have exponentially less data at lower levels
  - Too big of a tree can overfit the data
  - Greedy algorithms don’t necessarily yield the global optimum

- In practice, one often regularizes the construction process to try to get small but highly-informative trees
  - Pruning.
  - Minimum gain for split.
  - Minimum node size.

- Decision trees can also be used for regression on real-valued outputs, but it requires a different formalism
Comparison to k-NN

K-Nearest Neighbors
- Decision boundaries: piece-wise linear
- Test complexity: non-parametric, few parameters besides (all?) training examples

Decision Trees
- Decision boundaries: axis-aligned, tree structured
- Test complexity: attributes and splits
Applications of Decision Trees: XBox!

- Decision trees are in XBox

Applications of Decision Trees: XBox!

- Decision trees are in XBox: Classifying body parts

depth image $\rightarrow$ body parts $\rightarrow$ 3D joint proposals
Applications of Decision Trees: XBox!

- Trained on million(s) of examples
Applications of Decision Trees: XBox!

- Trained on million(s) of examples

Results:
Applications of Decision Trees

- Can express any Boolean function, but most useful when function depends critically on few attributes
- Bad on: parity, majority functions; also not well-suited to continuous attributes
- Practical Applications:
  - Tubular data
  - Flight simulator: 20 state variables; 90K examples based on expert pilot’s actions; auto-pilot tree
  - Yahoo Ranking Challenge
  - Random Forests: Microsoft Kinect Pose Estimation