Decision Trees: Discussion

Machine Learning
Spring 2019

The slides are mainly from Vivek Srikumar
This lecture: Learning Decision Trees

1. **Representation**: What are decision trees?

2. **Algorithm**: Learning decision trees
   - The ID3 algorithm: A greedy heuristic

3. Some extensions
Tips and Tricks

1. Decision tree variants

2. Handling examples with missing feature values

3. Non-Boolean features

4. Avoiding overfitting
Basic Decision Tree Algorithm: ID3

**Input:**

- $S$ the set of Examples
- $Label$ is the target attribute (the prediction)
- $Attributes$ is the set of measured attributes

### ID3($S$, $Attributes$, $Label$):

1. **If** all examples have same label:
   
   Return a leaf node with the label ; if $Attributes$ empty, return a leaf node with the most common label

2. **Otherwise**

   1. Create a **Root node** for tree
   2. $A = \text{attribute in } Attributes \text{ that } best \text{ splits } S$
   3. for each possible value $v$ of that $A$ can take:
      
      1. Add a new tree branch corresponding to $A=v$
      2. Let $S_v$ be the subset of examples in $S$ with $A=v$
      3. if $S_v$ is empty:
         
         add leaf node with the most common value of $Label$ in $S$
         
         Else:
         
         below this branch add the subtree ID3($S_v$, $Attributes - \{A\}$, $Label$)

4. Return **Root node**

**why?**

For generalization at test time
Basic Decision Tree Algorithm: ID3

ID3(S, Attributes, Label):

1. If all examples have same label:
   - Return a leaf node with the label; if Attributes empty, return a leaf node with the most common label

2. Otherwise
   1. Create a Root node for tree
   2. \( A = \) attribute in Attributes that best splits S
   3. for each possible value \( v \) of that \( A \) can take:
      1. Add a new tree branch corresponding to \( A=v \)
      2. Let \( S_v \) be the subset of examples in \( S \) with \( A=v \)
      3. if \( S_v \) is empty:
         add leaf node with the most common value of Label in \( S \)
      Else:
         below this branch add the subtree ID3(\( S_v \), Attributes - \{A\}, Label)
   4. Return Root node

Input:
- \( S \) the set of Examples
- Label is the target attribute (the prediction)
- Attributes is the set of measured attributes

Why?
For generalization at test time
Information Gain

The *information gain* of an attribute $A$ is the expected reduction in entropy caused by partitioning on this attribute $S_v$: the subset of examples where the value of attribute $A$ is set to value $v$

$$Gain(S, A) = Entropy(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$S_v$: the subset of examples where the value of attribute $A$ is set to value $v$

*Choose the feature with maximum information gain!*

$$Entropy(S) = H(S) = -p_+ \log(p_+) - p_- \log(p_-)$$

(binary)

$$H(\{p_1, p_2, \ldots, p_K\}) = -\sum_{i=1}^{K} p_i \log(p_i)$$

( general)
1. Variants of information gain

Information gain is defined using entropy to measure the purity of the labels in the split datasets.

Other ways to measure purity

Example (1): The *MajorityError (ME)*, which computes:

“Suppose the tree was not grown below this node and the majority label were chosen, what would be the error?”

Suppose at some node, there are 15 + and 5 − examples. What is the *MajorityError*?
1. Variants of information gain

Information gain is defined using entropy to measure the purity of the labels in the splitted datasets.

Other ways to measure purity

Example (1): The *MajorityError (ME)*, which computes:

“Suppose the tree was not grown below this node and the majority label were chosen, what would be the error?”

Suppose at some node, there are 15 + and 5 − examples. What is the *MajorityError*? Answer: ¼
1. Variants of information gain

Information gain is defined using entropy to measure the purity of the labels in the splitted datasets.

Other ways to measure purity

Example (1): The **MajorityError (ME)**, which computes:

“The tree was not grown below this node and the majority label were chosen, what would be the error?”

Suppose at some node, there are 15 + and 5 – examples. What is the **MajorityError**? **Answer: ¼**

$$\text{Gain}(S, A) = \text{ME}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{ME}(S_v)$$
1. Variants of information gain

Information gain is defined using entropy to measure the purity of the labels.

Other ways to measure purity

Example (2): The *Gini Index (GI)*, which computes:

- **Binary**:
  \[
  1 - (p_-^2 + p_+^2)
  \]

- **General**:
  \[
  1 - \sum_{t=1}^{k} p_k^2
  \]
1. Variants of information gain

Information gain is defined using entropy to measure the purity of the labels.

Other ways to measure purity

Example (2) : The *Gini Index (GI)*, which computes:

- **Binary**
  \[ \text{GI}_{\text{binary}} = 1 - (p_-^2 + p_+^2) \]

- **General**
  \[ \text{GI}_{\text{general}} = 1 - \sum_{t=1}^{k} p_{t}^2 \]

Suppose at some node, there are 15 + and 5 − examples. What is the GI?
1. Variants of information gain

Information gain is defined using entropy to measure the purity of the labels.

Other ways to measure purity

Example (2): The *Gini Index (GI)*, which computes:

\[ 1 - \left( p_-^2 + p_+^2 \right) \quad \text{binary} \]

\[ 1 - \sum_{t=1}^{k} p_k^2 \quad \text{general} \]

Suppose at some node, there are 15 + and 5 − examples. What is the GI?

\[ 1 - \left( \frac{5}{20} \right)^2 - \left( \frac{15}{20} \right)^2 = \frac{3}{8} \]
1. Variants of information gain

Information gain is defined using entropy to measure the purity of the labels.

Other ways to measure purity

Example (2): The *Gini Index (GI)*

\[
\text{Gain}(S, A) = \text{GI}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{GI}(S_v)
\]
2. Missing feature values

Suppose an example is missing the value of an attribute. What can we do at training time?

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>PlayTennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>???</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>9</td>
<td>Sunny</td>
<td>Cool</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
</tbody>
</table>
2. Missing feature values

Suppose an example is missing the value of an attribute. What can we do at training time?

“Complete the example” by

– Using the most common value of the attribute in the data
– Using the most common value of the attribute among all examples with the same output
2. Missing feature values

Suppose an example is missing the value of an attribute. What can we do at training time?

“Complete the example” by

– Using the most common value of the attribute in the data
– Using the most common value of the attribute among all examples with the same output
– Using fractional counts of the attribute values in training data
  • Eg: Outlook={5/14 Sunny, 4/14 Overcast, 5/14 Rain}
Suppose an example is missing the value of an attribute. What can we do at training time?

“Complete the example” by

- Using the most common value of the attribute in the data
- Using the most common value of the attribute among all examples with the same output
- Using fractional counts of the attribute values in training data
  
  Eg: Outlook={5/14 Sunny, 4/14 Overcast, 5/14 Rain}

<table>
<thead>
<tr>
<th></th>
<th>Mild</th>
<th>High</th>
<th>Weak</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>5/14</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
</tr>
<tr>
<td>5/14</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
</tr>
<tr>
<td>5/14</td>
<td>Rain.</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
</tr>
</tbody>
</table>
2. Missing feature values

Suppose an example is missing the value of an attribute. What can we do at training time?

“Complete the example” by

– Using the most common value of the attribute in the data
– Using the most common value of the attribute among all examples with the same output
– Using fractional counts of the attribute values in training data
  • Eg: Outlook={5/14 Sunny, 4/14 Overcast, 5/14 Rain}
  • Exercise: Will this change probability computations?
2. Missing feature values

Suppose an example is missing the value of an attribute. What can we do at training time?

“Complete the example” by

- Using the most common value of the attribute in the data
- Using the most common value of the attribute among all examples with the same output
- Using fractional counts of the attribute values
  - Eg: Outlook={5/14 Sunny, 4/14 Overcast, 5/14 Rain}
  - Exercise: Will this change probability computations?

At test time?
2. Missing feature values

Suppose an example is missing the value of an attribute. What can we do at training time?

“Complete the example” by

- Using the most common value of the attribute in the data
- Using the most common value of the attribute among all examples with the same output
- Using fractional counts of the attribute values
  - Eg: Outlook={5/14 Sunny, 4/14 Overcast, 5/14 Rain}
  - Exercise: Will this change probability computations?

At test time? Use the same method (if you use fraction counts, use the weighted majority!)
3. Non-Boolean features

- If the features can take multiple values
  - We have seen one edge per value (i.e. a multi-way split)
3. Non-Boolean features

• If the features can take multiple values
  – We have seen one edge per value (i.e. a multi-way split)
  – Another option: Make the attributes Boolean by testing for each value
    Convert Outlook=Sunny → {Outlook:Sunny=True,
                               Outlook:Overcast=False,
                               Outlook:Rain=False}
  – Or, perhaps group values into disjoint sets
3. Non-Boolean features

- If the features can take multiple values
  - We have seen one edge per value (i.e. a multi-way split)
  - Another option: Make the attributes Boolean by testing for each value
    - Convert Outlook=Sunny → \{Outlook:Sunny=True, Outlook:Overcast=False, Outlook:Rain=False\}
  - Or, perhaps group values into disjoint sets

- For numeric features, use thresholds or ranges to get Boolean/discrete alternatives
4. *Overfitting*
The “First Bit” function

• A Boolean function with n features
• Simply returns the value of the first feature, all others irrelevant

<table>
<thead>
<tr>
<th>$X_0$</th>
<th>$X_1$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

$Y = X_0$

$X_1$ is irrelevant

What is the decision tree for this function?
The “First Bit” function

- A Boolean function with n features
- Simply returns the value of the first feature, all others irrelevant

<table>
<thead>
<tr>
<th>$X_0$</th>
<th>$X_1$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

What is the decision tree for this function?
The “First Bit” function

- A Boolean function with \( n \) features
- Simply returns the value of the first feature, all others irrelevant

<table>
<thead>
<tr>
<th>( X_0 )</th>
<th>( X_1 )</th>
<th>( Y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

What is the decision tree for this function?

**Exercise**: Convince yourself that ID3 will generate this tree
The best case scenario: Perfect data

Suppose we have all $2^n$ examples for training. What will the error be on any future examples?
The best case scenario: Perfect data

Suppose we have all $2^n$ examples for training. What will the error be on any future examples?

Zero! Because we have seen every possible input!

And the decision tree can represent the function and ID3 will build a consistent tree.
Noisy data

What if the data is noisy? And we have all $2^n$ examples.

<table>
<thead>
<tr>
<th>$X_0$</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

Suppose, the outputs of both training and test sets are randomly corrupted.

Train and test sets are no longer identical.

Both have noise, possibly different.
Noisy data

What if the data is noisy? And we have all $2^n$ examples.

<table>
<thead>
<tr>
<th>$X_0$</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

Suppose, the outputs of both training and test sets are randomly corrupted.

Train and test sets are no longer identical.

Both have noise, possibly different.
E.g: Output corrupted with probability 0.25

The data is noisy. And we have all $2^n$ examples.

Test accuracy for different input sizes

The error bars are generated by running the same experiment multiple times for the same setting.
E.g: Output corrupted with probability 0.25

The data is noisy. And we have all $2^n$ examples.

**Test accuracy for different input sizes**

We can analytically compute test error in this case:

**Correct prediction:**
- $P(\text{Training example uncorrupted AND test example uncorrupted}) = 0.75 \times 0.75$
- $P(\text{Training example corrupted AND test example corrupted}) = 0.25 \times 0.25$
- $P(\text{Correct prediction}) = 0.625$

**Incorrect prediction:**
- $P(\text{Training example uncorrupted AND test example corrupted}) = 0.75 \times 0.25$
- $P(\text{Training example corrupted and AND test example uncorrupted}) = 0.25 \times 0.75$
- $P(\text{incorrect prediction}) = 0.375$

**Error = approx 0.375**
E.g: Output corrupted with probability 0.25

The data is noisy. And we have all $2^n$ examples.

Test accuracy for different input sizes

What about the training accuracy?
E.g: Output corrupted with probability 0.25

The data is noisy. And we have all $2^n$ examples.

Test accuracy for different input sizes

What about the training accuracy?

Training accuracy = 100%
Because the learning algorithm *will* find a tree that agrees with the data
E.g: Output corrupted with probability 0.25

The data is noisy. And we have all $2^n$ examples.

Test accuracy for different input sizes

Then, why is the classifier not perfect?
E.g: Output corrupted with probability 0.25

The data is noisy. And we have all $2^n$ examples.

Test accuracy for different input sizes

Then, why is the classifier not perfect?

The classifier **overfits** the training data
Overfitting

• The learning algorithm fits the *noise* in the data
  – Irrelevant attributes or noisy examples influence the choice of the hypothesis

• May lead to poor performance on future examples
Overfitting: One definition

• Data comes from a probability distribution $D(X, Y)$

• We are using a hypothesis space $H$

• Errors:
  - Training error for hypothesis $h \in H$: $\text{error}_{\text{train}}(h)$
  - True error for $h \in H$: $\text{error}_D(h)$

• A hypothesis $h$ **overfits** the training data if there is another hypothesis $h'$ such that
  1. $\text{error}_{\text{train}}(h) < \text{error}_{\text{train}}(h')$
  2. $\text{error}_D(h) > \text{error}_D(h')$
Decision trees will overfit

Graph from Mitchell
Avoiding overfitting with decision trees

• Some approaches:
  1. Fix the depth of the tree
    • *Decision stump* = a decision tree with only one level
    • Typically will not be very good by itself
    • But, short decision trees can make very good features for a second layer of learning
Avoiding overfitting with decision trees

• Some approaches:
  2. Optimize on a *held-out set* (also called *development set* or *validation set*) while growing the trees
    • Split your data into two parts – training set and held-out set
    • Grow your tree on the training split and check the performance on the held-out set after every new node is added
    • If growing the tree hurts validation set performance, stop growing
Avoiding overfitting with decision trees

- Some approaches:
  3. Grow full tree and then prune as a post-processing step in one of several ways:
    1. Use a validation set for pruning from bottom up greedily
    2. Convert the tree into a set of rules (one rule per path from root to leaf) and prune each rule independently
Summary: Decision trees

• A popular machine learning tool
  – Prediction is easy
  – Represents any Boolean functions

• Greedy heuristics for learning
  – ID3 algorithm (using information gain)

• (Can be used for regression too)

• Decision trees are prone to overfitting unless you take care to avoid it