Decision Trees

( Most slides by Pedro Domingos)

Reading: CIML chapter 1
Note: These slides contain some information not in the readings!

Warm-up and Review

1. Turn to the person next to you and introduce yourself.
2. If your name come first alphabetically, answer the first question for your neighbor. If your name comes second alphabetically, explain the answer to the second question to your neighbor.
3. If you have time, feel free to answer the last two questions as well.

Q1: What is the difference between the data distribution and the training data?
Q2: What is the difference between validation data and test data?

Review: Important Concepts

• Data: labeled instances, e.g. emails marked spam/ham
  – Training set
  – Held out set
  – Test set
• Features: attribute-value pairs which characterize each x
• Experimentation cycle
  – Learn parameters (e.g. model probabilities) on training set
  – Tune hyperparameters on held-out set
  – Compute accuracy of test set
  – Very important: never “peek” at the test set!
• Evaluation
  – Accuracy: fraction of instances predicted correctly
• Overfitting and generalization
  – Want a classifier which does well on test data
  – Overfitting: fitting the training data very closely, but not generalizing well

Outline

• Quick introduction to decision trees
• What you should learn:
  – What decision trees can represent compactly
  – How to learn a decision tree recursively, including choosing attributes
  – What decision trees can learn
  – How decision trees handle overfitting
Exercise

Consider the following dataset:

<table>
<thead>
<tr>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

- Draw a decision tree that classifies all examples correctly.
- How would your tree classify (0, 1, 1)?
Choosing the Best Attribute

One way to choose the best attribute is to perform a 1-step lookahead search and choose the attribute that gives the lowest error rate on the training data.

```
CHOOSE-BEST-ATTRIBUTE(S)
choose j to minimize Jₖ, computed as follows:
Sₖ = all (x,y) ∈ S with xₖ = k;
Sₖ = all (x,y) ∈ S with xₖ = k;
jₖ = the most common value of y in Sₖ;
jₖ = the most common value of y in Sₖ;
Jₖ = number of examples (x,y) ∈ Sₖ with y ≠ jₖ;
Jₖ = number of examples (x,y) ∈ Sₖ with y ≠ jₖ;
return j;
```

Choosing the Best Attribute (3)

Unfortunately, this measure does not always work well, because it does not detect cases where we are making "progress" toward a good tree.

```
20 10

12 8
10 2

12 0
9 8
8 9
0 2
```

(These slides are important and not in CIML.)

Entropy

The entropy of V, denoted H(V) is defined as follows:

\[ H(V) = \sum_{x} -P(V = x) \log P(V = x). \]

This is the average surprise of describing the result of one "trial" of V (one coin toss).

Mutual Information

Now consider two random variables A and B that are not necessarily independent. The mutual information between A and B is the amount of information we learn about B by knowing the value of A (and vice versa - it is symmetric). It is computed as follows:

\[ I(A; B) = H(B) - \sum_{a} P(A = a) H(B = b) \]

In particular, consider the case Y of coin training example and the value of feature x₁ to be random variables. Then the mutual information quantifies how much x₁ tells us about the value of the coin Y.

\[ P(x=1|y=0) = 0.99997 \]

```
20 10

12 8
10 2

12 0
9 8
8 9
0 2
```

A Better Heuristic From Information Theory

Let V be a random variable with the following probability distribution:

\[
\begin{array}{c|c}
P(V = 0) & P(V = 1) \\
\hline
0.2 & 0.8 \\
\end{array}
\]

The surprise, S(V = v) of each value of V is defined to be

\[ S(V = v) = -\log P(V = v). \]

An event with probability 1 gives us zero surprise.

An event with probability 0 gives us infinite surprise.

It turns out that the surprise is equal to the number of bits of information that need to be transmitted to a recipient who knows the probabilities of the results.

This is also called the description length of V = v.

Fractional bits only make sense if they are part of a longer message (e.g., describe a whole sequence of coin tosses).

\[ H(Y) = 0.9183 \]

\[ H(Y|x=1) = 0.7910 \]

\[ H(Y|x=0) = 0.3713 \]

\[ H(Y|x=1) = 0.3994 \]

\[ H(Y|x=0) = 0.3994 \]
Learning Parity with Noise

When learning exclusive-or (2-bit parity), all splits look equally good. If extra random boolean features are included, they also look equally good. Hence, decision tree algorithms cannot distinguish random noisy features from parity features.

\[
\begin{array}{c|cc|c}
  x_1 & x_2 & x_3 & y \\
  \hline
  0 & 0 & 0 & 1 \\
  0 & 0 & 1 & 1 \\
  0 & 1 & 0 & 1 \\
  0 & 1 & 1 & 1 \\
  1 & 0 & 0 & 1 \\
  1 & 0 & 1 & 1 \\
  1 & 1 & 0 & 0 \\
  1 & 1 & 1 & 0 \\
\end{array}
\]

Overfitting in Decision Trees

Consider a noisy training example:
Sunny, Hot, Normal, Strong, PlayTennis=No
What effect on tree?

Overfitting

Consider error of hypothesis \( h \) over

- training data: \( \text{error}_{\text{train}}(h) \)
- entire distribution \( \mathcal{D} \) of data: \( \text{error}_{\mathcal{D}}(h) \)

Hypothesis \( h \in H \) overfits training data if there is an alternative hypothesis \( h' \in H \) such that

\[
\text{error}_{\text{train}}(h) < \text{error}_{\text{train}}(h')
\]

and

\[
\text{error}_{\mathcal{D}}(h) > \text{error}_{\mathcal{D}}(h')
\]

Avoiding Overfitting

How can we avoid overfitting?

- Stop growing when data split not statistically significant
- Grow full tree, then post-prune

How to select “best” tree:

- Measure performance over training data
- Measure performance over separate validation data set
- Add complexity penalty to performance measure

Summary: Decision Trees

- Representation
- Tree growth
- Heuristics
- Overfitting and pruning