Decision Trees: Overfitting

These slides were assembled by Byron Boots, with grateful acknowledgement to Eric Eaton and the many others who made their course materials freely available online. Feel free to reuse or adapt these slides for your own academic purposes, provided that you include proper attribution.
Last Time: Which Tree Should We Output

- ID3 performs heuristic search through space of decision trees
- It stops at smallest acceptable tree. Why?

Occam’s razor: prefer the simplest hypothesis that fits the data
Preference bias: Ockham’s Razor

• Principle stated by William of Ockham (1285-1347)
  – “non sunt multiplicanda entia praeter necessitatem”
  – entities are not to be multiplied beyond necessity
  – AKA Occam’s Razor, Law of Economy, or Law of Parsimony

Idea: The simplest consistent explanation is the best

• Therefore, the smallest decision tree that correctly classifies all of the training examples is best
  • Finding the provably smallest decision tree is NP-hard
  • ...So instead of constructing the absolute smallest tree consistent with the training examples, construct one that is pretty small
Overfitting in Decision Trees

• Many kinds of “noise” can occur in the examples:
  – Two examples have same attribute/value pairs, but different classifications
  – Some values of attributes are incorrect because of errors in the data acquisition process or the preprocessing phase
  – The instance was labeled incorrectly (+ instead of -)

• Also, some attributes are irrelevant to the decision-making process
  – e.g., color of a die is irrelevant to its outcome

Based on Slide from M. desJardins & T. Finin
Overfitting in Decision Trees

• Irrelevant attributes can result in *overfitting* the training example data
  – If hypothesis space has many dimensions (large number of attributes), we may find *meaningless regularity* in the data that is irrelevant to the true, important, distinguishing features

• If we have too little training data, even a reasonable hypothesis space will ‘overfit’
Overfitting in Decision Trees

Consider adding a noisy training example to the following tree:

What would be the effect of adding:

<outlook=sunny, temperature=hot, humidity=normal, wind=strong, playTennis=No>?
Overfitting in Decision Trees

Consider error of hypothesis $h$ over

- training data: $error_{train}(h)$
- entire distribution $\mathcal{D}$ of data: $error_{\mathcal{D}}(h)$

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

$$error_{train}(h) < error_{train}(h')$$

and

$$error_{\mathcal{D}}(h) > error_{\mathcal{D}}(h')$$
Overfitting in Decision Trees

![Graph showing accuracy on training and test data as a function of tree size](image)

- **On training data**
- **On test data**

**Legend:**
- Solid line: Accuracy on training data
- Dashed line: Accuracy on test data

**Y-axis:** Accuracy

**X-axis:** Size of tree (number of nodes)
Avoiding Overfitting in Decision Trees

How can we avoid overfitting?
• Stop growing when data split is not statistically significant
• Acquire more training data
• Remove irrelevant attributes (manual process – not always possible)
• 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 (heuristic: simpler is better)
Reduced-Error Pruning

Split training data further into *training* and *validation* sets

Grow tree based on *training set*

Do until further pruning is harmful:
1. Evaluate impact on validation set of pruning each possible node (plus those below it)
2. Greedily remove the node that most improves *validation set* accuracy
Pruning Decision Trees

- Pruning of the decision tree is accomplished by replacing a whole subtree by a leaf node.
- The replacement takes place if a decision rule establishes that the expected error rate in the subtree is greater than in the single leaf.
- For example,

Based on Example from M. desJardins & T. Finin

If we had simply predicted the majority class (negative), we make 2 errors instead of 4.
Effect of Reduced-Error Pruning

On training data it looks great

But that’s not the case for the test (validation) data
Effect of Reduced-Error Pruning

The tree is pruned back to the red line where it gives more accurate results on the test data.

Based on Slide by Pedro Domingos
Summary: Decision Tree Learning

• Widely used in practice

• Strengths include
  – Fast and simple to implement
  – Can convert to rules
  – Handles noisy data

• Weaknesses include
  – Univariate splits/partitioning using only one attribute at a time --- limits types of possible trees
  – Large decision trees may be hard to understand
  – Requires fixed-length feature vectors
  – Non-incremental (i.e., batch method)