

Intro. to machine learning (CSI 5325)

Lecture 5: decision trees

Greg Hamerly

Some content from Tom Mitchell.

1 Inductive Bias in ID3

2 Overfitting

3 Underfitting

Inductive Bias in ID3

Note H is the power set of instances X

→ Unbiased?

Not really...

- Preference for short trees, and for those with high information gain attributes near the root
- ID3 bias is a *preference* for some hypotheses, rather than a *restriction* of hypothesis space H
- Occam's razor: prefer the shortest hypothesis that fits the data

Occam's Razor

Why prefer short hypotheses?

Arguments in favor:

- Fewer short hypotheses than long ones
- a short hypothesis that fits data **unlikely to be** coincidence
- a long hypothesis that fits data **might be** coincidence

Arguments opposed:

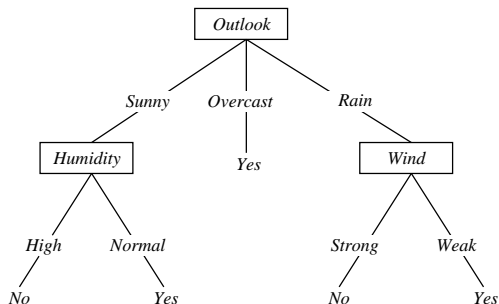
- There are many ways to define small sets of hyps
- e.g., all trees with a prime number of nodes that use attributes beginning with "Z"
- What's so special about small sets based on *size* of hypothesis??

Overfitting in Decision Trees

Consider adding noisy training example #15:

Sunny, Hot, Normal, Strong, PlayTennis = No

What effect does it have on the earlier tree?



Overfitting

Consider error of hypothesis h over

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

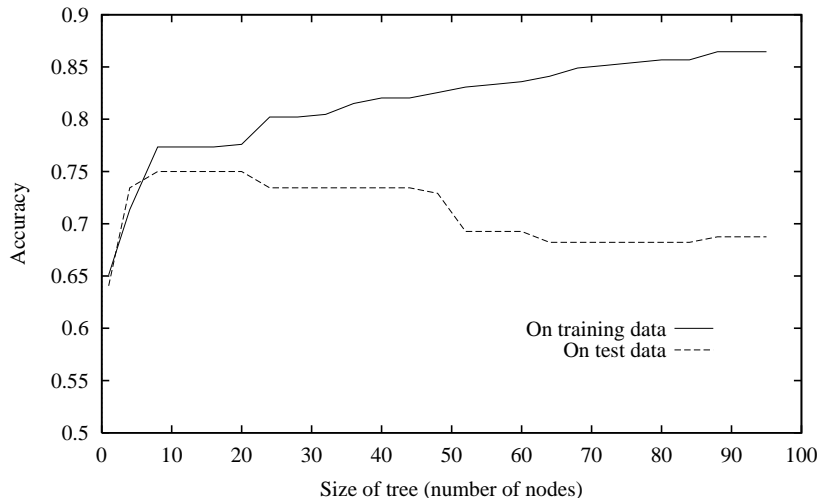
Hypothesis $h \in H$ **overfits** training data if there is an alternative hypothesis $h' \in H$ such that

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

and

$$error_{\mathcal{D}}(h) > error_{\mathcal{D}}(h')$$

Overfitting on a particular task



Data-centric reasons for overfitting

Noise in the data

- Do we want to perfectly model training data with noise?

Too little data

- Spurious correlations could occur.

Avoiding Overfitting

How can we avoid overfitting in decision trees?

- stop growing when data split not statistically significant
 - i.e. information gain due to splitting on an attribute could easily be obtained by some other randomly-chosen split, unrelated to any attribute
- grow full tree, then post-prune

How to select the 'best' tree

- Measure performance over training data
- Measure performance over separate validation data set
- MDL: minimize $size(tree) + size(misclassifications(tree))$

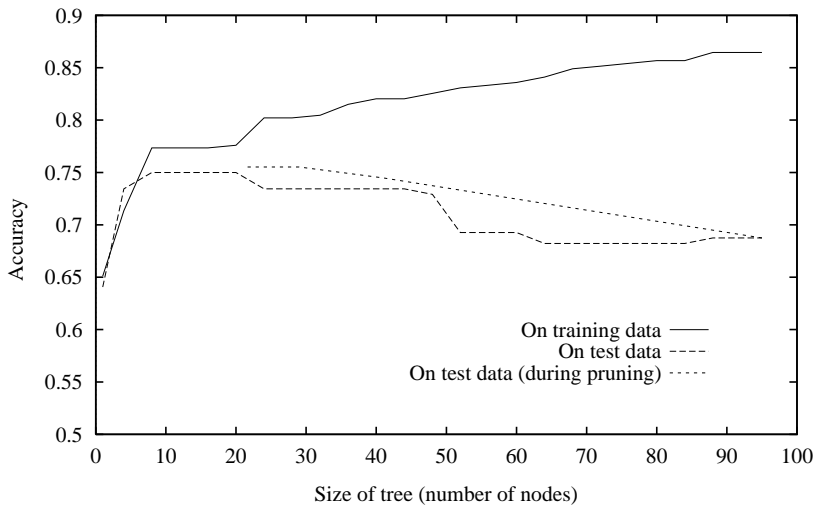
Reduced-Error Pruning

Split data into *training* and *validation* 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 one that most improves *validation* set accuracy
- Produces smallest version of most accurate subtree.
 - What if data is limited?

Effect of Reduced-Error Pruning

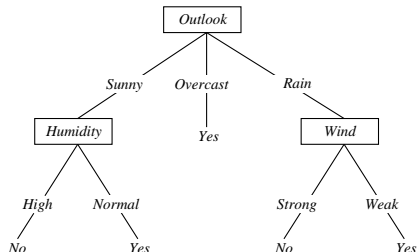


Rule Post-Pruning

- 1 Convert tree to equivalent set of rules
- 2 Prune each rule independently of others
 - Pruning involves removing preconditions of a rule that improve its accuracy
- 3 Sort final rules into desired sequence for use

Perhaps most frequently used method (e.g., C4.5)

Converting A Tree to Rules



IF $(Outlook = Sunny) \wedge (Humidity = High)$
 THEN $PlayTennis = No$

IF $(Outlook = Sunny) \wedge (Humidity = Normal)$
 THEN $PlayTennis = Yes$

...

Overfitting is a general problem

Overfitting is not just a problem for decision trees.

Most machine learning algorithms face overfitting problems. It is more likely to occur when...

- ...the hypothesis space is very large
- ...the hypothesis search is not biased toward simple models
- ...there is little training data
- ...there is a lot of noise in the training data

How can we tell when it happens?

- Radical differences in train/test accuracy
- Overly complex hypothesis/model (e.g. large decision tree)

Underfitting can also occur

Underfitting is the opposite of overfitting

It appears in a different way on our accuracy graphs

- How would you expect an underfit model to do on training data?
- On test data?

Causes for underfitting

- Hypothesis space is too small
- Search strategy is too biased towards simple models
- Attributes do not correlate with the target

Evaluating hypotheses

In any learning algorithm, we want to know 'how well' it is performing.

This is an important and substantial question.

We will talk about it as we go along, and then particularly in chapter 6.