C4.5 - pruning decision trees
Quiz 1
Quiz 1

Q: Is a tree with only pure leaves always the best classifier you can have?

A: No.
Quiz 1

Q: Is a tree with only pure leaves always the best classifier you can have?

A: No.

This tree is the best classifier on the training set, but possibly not on new and unseen data. Because of overfitting, the tree may not generalize very well.
Pruning

- **Goal:** Prevent overfitting to noise in the data

- **Two strategies for “pruning” the decision tree:**
  - **Postpruning** - take a fully-grown decision tree and discard unreliable parts
  - **Prepruning** - stop growing a branch when information becomes unreliable
Prepruning

- Based on statistical significance test
  - Stop growing the tree when there is no statistically significant association between any attribute and the class at a particular node
- Most popular test: chi-squared test
- ID3 used chi-squared test in addition to information gain
  - Only statistically significant attributes were allowed to be selected by information gain procedure
Early stopping

- Pre-pruning may stop the growth process prematurely: early stopping

- Classic example: XOR/Parity-problem
  - No individual attribute exhibits any significant association to the class
  - Structure is only visible in fully expanded tree
  - Pre-pruning won’t expand the root node

- But: XOR-type problems rare in practice

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Post-pruning

- First, build full tree
- Then, prune it
  - Fully-grown tree shows all attribute interactions
- Problem: some subtrees might be due to chance effects
- Two pruning operations:
  1. **Subtree replacement**
  2. Subtree raising
Subtree replacement

- Bottom-up
- Consider replacing a tree only after considering all its subtrees
Subtree replacement

- Bottom-up
- Consider replacing a tree only after considering all its subtrees
Subtree replacement

- Bottom-up
- Consider replacing a tree only after considering all its subtrees
Subtree replacement

- Bottom-up
- Consider replacing a tree only after considering all its subtrees
Estimating error rates

- Prune only if it reduces the estimated error
- Error on the training data is NOT a useful estimator
- Use hold-out set for pruning
  - (“reduced-error pruning”)
- C4.5’s method
  - Derive confidence interval from training data
  - Use a heuristic limit, derived from this, for pruning
  - Standard Bernoulli-process-based method
  - Shaky statistical assumptions (based on training data)
Estimating Error Rates
Estimating Error Rates

Q: what is the error rate on the training set?
Estimating Error Rates

Q: what is the error rate on the training set?

A: 0.33 (2 out of 6)
Estimating Error Rates

Q: what is the error rate on the training set?
A: 0.33 (2 out of 6)
Q: what is the error rate on the training set?
A: 0.33 (2 out of 6)

Q: will the error on the test set be bigger, smaller or equal?
Estimating Error Rates

Q: what is the error rate on the training set?
A: 0.33 (2 out of 6)

Q: will the error on the test set be bigger, smaller or equal?
A: bigger
Estimating the error

- Assume making an error is Bernoulli trial with probability $p$
  - $p$ is unknown (true error rate)
- We observe $f$, the success rate $f = S/N$
- For large enough $N$, $f$ follows a Normal distribution
- Mean and variance for $f$: $p$, $p \frac{1-p}{N}$
Estimating the error

- $c\%$ confidence interval $[-z \leq X \leq z]$ for random variable with 0 mean is given by:

\[
\Pr[-z \leq X \leq z] = c
\]

- With a symmetric distribution:

\[
\Pr[-z \leq X \leq z] = 1 - 2 \times \Pr[X \geq z]
\]
z-transforming f

- Transformed value for $f$: 
  \[ \frac{f - p}{\sqrt{p(1-p)/N}} \]  
  (i.e. subtract the mean and divide by the standard deviation)

- Resulting equation: 
  \[ \Pr \left[ -z \leq \frac{f - p}{\sqrt{p(1-p)/N}} \leq z \right] = c \]

- Solving for $p$: 
  \[ p = \left( f + \frac{z^2}{2N} \pm z \sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}} \right) \div \left( 1 + \frac{z^2}{N} \right) \]
C4.5’s method

- Error estimate for subtree is weighted sum of error estimates for all its leaves

- Error estimate for a node (upper bound):
  \[
e = \left( f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}} \right) \div \left( 1 + \frac{z^2}{N} \right)
  \]

- If \( c = 25\% \) then \( z = 0.69 \) (from normal distribution)

<table>
<thead>
<tr>
<th>Pr([X \geq z])</th>
<th>( z )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>2.33</td>
</tr>
<tr>
<td>5%</td>
<td>1.65</td>
</tr>
<tr>
<td>10%</td>
<td>1.28</td>
</tr>
<tr>
<td>20%</td>
<td>0.84</td>
</tr>
<tr>
<td>25%</td>
<td>0.69</td>
</tr>
<tr>
<td>40%</td>
<td>0.25</td>
</tr>
</tbody>
</table>
C4.5’s method

\[ e = \left( f + \frac{z^2}{2N} + z \sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}} \right) / \left( 1 + \frac{z^2}{N} \right) \]
C4.5’s method

\[ e = \left( f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N^2} + \frac{z^2}{4N^2}} \right) / \left( 1 + \frac{z^2}{N} \right) \]

\( f \) is the observed error
C4.5’s method

\[ e = \left( f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}} \right) / \left( 1 + \frac{z^2}{N} \right) \]

\( f \) is the observed error
\( z = 0.69 \)
C4.5’s method

\[ e = \left( f + \frac{z^2}{2N} + z\sqrt{\frac{f^2}{N} - \frac{f^2}{4N^2} + \frac{z^2}{4N^2}} \right) \left/ \left( 1 + \frac{z^2}{N} \right) \right] \]

\( f \) is the observed error
\( z = 0.69 \)
\( e > f \)
\( e = (f + \varepsilon_1)/(1+ \varepsilon_2) \)
C4.5’s method

\[ e = \left( f + \frac{z^2}{2N} + z\sqrt{\frac{f^2}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}} \right) / \left( 1 + \frac{z^2}{N} \right) \]

\( f \) is the observed error

\( z = 0.69 \)

\( e > f \)

\( e = (f + \epsilon_1)/(1+ \epsilon_2) \)

\( N \to \infty, e = f \)
Example

wage increase 1st year
\[ \leq 2.5 \quad > 2.5 \]

working hours per week
\[ \leq 36 \quad > 36 \]

health plan contribution

1 bad 1 good

4 bad 2 good
\[ f=0.33 \quad e=0.47 \]

1 bad 1 good
\[ f=0.5 \quad e=0.72 \]

4 bad 2 good
\[ f=0.33 \quad e=0.47 \]

Universiteit Leiden
Example

Combined using ratios 6:2:6 gives 0.51
Example

Combined using ratios 6:2:6 gives 0.51
Summary

- Decision Trees
  - splits – binary, multi-way
  - split criteria – information gain, gain ratio, ...
  - pruning
- No method is always superior – experiment!