What’s better than a tree?
Random Forest

• Can often improve performance of a decision tree classifier using a set of decision trees (a forest)
• Each tree trained on a random subset of training data
• Classify a data instance using all trees
• Combine answers to make classification
  – E.g., vote for most common class
Bagging

• Idea can be used on any classifier!

• Bagging = *Bootstrap aggregating*
  
  An *ensemble* meta-algorithm that can improve stability & accuracy of algorithms for statistical classification and regression

• Helps avoid overfitting
Choosing subsets of training data

- Classic bagging: select random set of training instances *with replacement*
- Pasting: select random subset of training instances
- Random Subspaces: use all training instances, but with a random subset of features
- Random Patches: random subset of instances and random subset of features
- What’s best? YMMV: depends on problem, training data, algorithm
Classifier
Choose J48 -C 0.25 -M 2

Test options
- Use training set
- Supplied test set
- Cross-validation: Folds 10
- Percentage split: % 66

Classifier output
Size of the tree: 911
Time taken to build model: 2.64 seconds

Evaluation on training set:
Time taken to test model on training data: 0.16 seconds

Summary:
- Correctly Classified Instances: 42803 (87.6356%)
- Incorrectly Classified Instances: 6039 (12.3644%)
- Kappa statistic: 0.6325
- Mean absolute error: 0.1861
- Root mean squared error: 0.3048
- Relative absolute error: 51.1076%
- Root relative squared error: 71.4388%
- Total Number of Instances: 48842

Detailed Accuracy By Class:
- TP Rate: 0.631, FP Rate: 0.046, Precision: 0.810, Recall: 0.631, F-Measure: 0.710
- TP Rate: 0.954, FP Rate: 0.369, Precision: 0.891, Recall: 0.954, F-Measure: 0.921
- Weighted Avg.: 0.876, 0.369, 0.891, 0.954, 0.921

Confusion Matrix:
- a b | classified as
- 7375 4312 | a = >50K
- 1727 35428 | b = <=50K
RandomForest

Evaluation on training set:

Time taken to build model: 15.17 seconds

Time taken to test model on training data: 6.52 seconds

Summary:

Correctly Classified Instances 48774 99.8608 %
Incorrectly Classified Instances 68 0.1392 %

Kappa statistic 0.9962
Mean absolute error 0.0737
Root mean squared error 0.1263
Relative absolute error 20.2565 %
Root relative squared error 29.0022 %
Total Number of Instances 48842

Detailed Accuracy By Class:

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>PRC Area</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;50K</td>
<td>0.995</td>
<td>0.000</td>
<td>1.000</td>
<td>0.995</td>
<td>0.997</td>
<td>0.96</td>
<td>1.000</td>
<td>1.000</td>
<td>&gt;50K</td>
</tr>
<tr>
<td>&lt;=50K</td>
<td>1.000</td>
<td>0.005</td>
<td>0.998</td>
<td>1.000</td>
<td>1.000</td>
<td>0.997</td>
<td>1.000</td>
<td>1.000</td>
<td>&lt;=50K</td>
</tr>
</tbody>
</table>

Weighted Avg.:

|       | 0.999   | 0.004   | 0.999     | 0.999  | 0.999     | 0.996| 1.000    | 1.000    |       |

Confusion Matrix:

```
a  b  | <--- classified as
11624 63 | a = >50K
5 37150 | b = <=50K
```
Created a train and test collection

- Train has ~95% of data, test 5%
- Trained models for J48 and random forest using train dataset
- Tested on test data set
- Results were that random forest was (at best) about the same as J48
Correctly Classified Instances 2155  86.2%
Incorrectly Classified Instances 345  13.8%

Kappa statistic 0.5988
Mean absolute error 0.1591
Root mean squared error 0.3196
Relative absolute error 52.5531%
Root relative squared error 74.1954%
Total Number of Instances 2500

=== Detailed Accuracy By Class ===

<table>
<thead>
<tr>
<th></th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>PRC Area</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;50K</td>
<td>0.611</td>
<td>0.052</td>
<td>0.780</td>
<td>0.611</td>
<td>0.686</td>
<td>0.606</td>
<td>0.895</td>
<td>0.759</td>
<td>&gt;50K</td>
</tr>
<tr>
<td>&lt;=50K</td>
<td>0.944</td>
<td>0.360</td>
<td>0.882</td>
<td>0.944</td>
<td>0.912</td>
<td>0.606</td>
<td>0.895</td>
<td>0.953</td>
<td>&lt;=50K</td>
</tr>
</tbody>
</table>

Weighted Avg. 0.862  0.309  0.857  0.862  0.856

=== Confusion Matrix ===

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>376</td>
<td>239</td>
<td>615</td>
</tr>
<tr>
<td>b</td>
<td>106</td>
<td>1779</td>
<td>1885</td>
</tr>
</tbody>
</table>

Total 2500

Time taken to build model: 1.86 seconds
Time taken to test model on supplied test set: 0 seconds
Classifier output

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

=== Re-evaluation on test set ===

User supplied test set
Relation: adult
Instances: unknown (yet). Reading incrementally
Attributes: 15

=== Summary ===

Correctly Classified Instances 2146 85.84 %
Incorrectly Classified Instances 354 14.16 %
Kappa statistic 0.59
Mean absolute error 0.195
Root mean squared error 0.3272
Total Number of Instances 2500

=== Detailed Accuracy By Class ===

<table>
<thead>
<tr>
<th></th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>PRC Area</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;50K</td>
<td>0.610</td>
<td>0.060</td>
<td>0.767</td>
<td>0.610</td>
<td>0.679</td>
<td>0.596</td>
<td>0.893</td>
<td>0.765</td>
<td>&gt;50K</td>
</tr>
<tr>
<td>=50K</td>
<td>0.940</td>
<td>0.390</td>
<td>0.881</td>
<td>0.940</td>
<td>0.909</td>
<td>0.596</td>
<td>0.893</td>
<td>0.959</td>
<td>=50K</td>
</tr>
</tbody>
</table>

Weighted Avg. 0.858 0.309 0.853 0.858 0.853 0.596 0.893 0.911

=== Confusion Matrix ===

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>375</td>
<td>240</td>
</tr>
<tr>
<td>b</td>
<td>114</td>
<td>1771</td>
</tr>
</tbody>
</table>

a = >50K
b = <=50K