What's better than a tree?
Random Forest

- Can often improve performance of decision tree classifiers 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
1. Many trees are created using random subsets of features and bootstrapped data.

2. Each tree votes by predicting target class.

3. Votes are tallied to reach the final prediction.
cf. **Wisdom of the Crowd**

- Statistician Francis Galton observed a 1906 contest to guess an ox’s weight at a country fair. 800 people entered. He noted that their average guess (1,197lb) was very close to the actual weight (1,198lb).

- When getting human annotations training data for machine learning, standard practice is get ≥ 3 annotations and take majority vote.

*cf.* abbreviation (short for *Latin*: confer/conferatur) refer reader to other material to make a comparison
Random Forests Benefits

• Decision trees not the strongest modeling approach
• Random forests make them much stronger
• => more robust than a single decision tree
  – Limit overfitting to given dataset
  – Reduce errors due to training data bias
  – Stable performance if some noise added to training data
Bagging

- Idea can be used on any classifier!
- Improve classification by combining classifications of randomly selected training subsets
- Bagging = **Bootstrap aggregating**
  - An **ensemble** meta-algorithm that can improve stability & accuracy of algorithms for statistical classification and regression
- Helps avoid overfitting
- AKA ensembling
Bagging
Bootstrap Aggregation

Data

Bootstrapped Data

Models

Voting

Outcome

Train model

Voting

Chris Albon

Sampled with replacement

DATA

DATA

DATA

DATA

MODEL

MODEL

MODEL

VOTE

OUTPUT
Choosing training data subsets

• **Classic bagging**: select random subset of training instances **with replacement**

• **Pasting**: select random subset of training instances (i.e., without replacement)

• **Random Subspaces**: use all training instances, but with a random subset of features

• **Random Patches**: random subset of instances and random subset of features

• **Best? YMMV**: depends on problem, training data, algorithm
Examples

• Two examples using Weka
  – UCI Auto mpg prediction dataset
  – UCI Adult income prediction dataset

• RandomForest improves over J48 for the smaller dataset, but not for the larger

• Takeaway: more data is always best
# UCI Auto MGP Dataset (1)

## Auto MPG Data Set

*Download Data Folder Data Set Description*

**Abstract:** Revised from CMU StatLib library, data concerns city-cycle fuel consumption

<table>
<thead>
<tr>
<th>Data Set Characteristics:</th>
<th>Multivariate</th>
<th>Number of Instances:</th>
<th>398</th>
<th>Area:</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute Characteristics:</td>
<td>Categorical, Real</td>
<td>Number of Attributes:</td>
<td>8</td>
<td>Date Donated</td>
<td>1993-07-07</td>
</tr>
<tr>
<td>Associated Tasks:</td>
<td>Regression</td>
<td>Missing Values?</td>
<td>Yes</td>
<td>Number of Web Hits:</td>
<td>430910</td>
</tr>
</tbody>
</table>

**Source:**

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition.

**Data Set Information:**

This dataset is a slightly modified version of the dataset provided in the StatLib library. In line with the use by Ross Quinlan (1993) in predicting the attribute "mpg", 8 of the original instances were removed because they had unknown values for the "mpg" attribute. The original dataset is available in the file "auto-mpg.data-original".

"The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes." (Quinlan, 1993)
UCI Auto MGP Dataset

• Data from 1983
• 398 instances
• Predict auto mpg from seven attributes:
  – Number of cylinders
  – Displacement
  – Horsepower
  – Weight
  – Acceleration
  – Model year
  – Country of origin
Classifier

Choose J48 -C 0.25 -M 2

Test options

- Use training set
- Supplied test set
- Cross-validation
- Percentage split

Classifier output

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances 230 95.8333 %
Incorrectly Classified Instances 10 4.1667 %
Kappa statistic 0.9174
Mean absolute error 0.0453
Root mean squared error 0.1505
Relative absolute error 13.4303 %
Root relative squared error 36.7193 %
Total Number of Instances 240

=== Detailed Accuracy By Class ===

<table>
<thead>
<tr>
<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>
</tr>
</thead>
<tbody>
<tr>
<td>0.987</td>
<td>0.025</td>
<td>0.987</td>
<td>0.987</td>
<td>0.987</td>
<td>0.963</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>0.881</td>
<td>0.015</td>
<td>0.925</td>
<td>0.811</td>
<td>0.902</td>
<td>0.883</td>
<td>0.991</td>
<td>0.954</td>
</tr>
<tr>
<td>0.923</td>
<td>0.025</td>
<td>0.878</td>
<td>0.933</td>
<td>0.900</td>
<td>0.880</td>
<td>0.989</td>
<td>0.921</td>
</tr>
</tbody>
</table>

Weighted Avg. 0.958 0.023 0.959 0.957 0.958 0.935 0.995 0.978

=== Confusion Matrix ===

a b c <- classified as
157 1 1 | a = 1
1 37 4 | b = 2
1 2 36 | c = 3
RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Use training set

Correctly Classified Instances 240 100 %
Incorrectly Classified Instances 0 0 %
Kappa statistic 1
Mean absolute error 0.0674
Root mean squared error 0.114
Relative absolute error 19.9659 %
Root relative squared error 27.8064 %
Total Number of Instances 240

Detailed Accuracy By Class

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 2
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 3
Weighted Avg. 1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000

Confusion Matrix

a b c <-- classified as
159 0 0 | a = 1
0 42 0 | b = 2
0 0 39 | c = 3
100% ... Wait, What?

• Results are too good to be true!
  – Something must be wrong
• ML results tend to be asymptotic
  – Asymptotic lines approach a curve but never touch
• Closer you get to F1=1.0, the harder it is to improve
• What did we do wrong?
Results are too good

- Relatively small dataset allows construction of a DT model that does very well
- Using Random Forest still improves on it
- We trained and tested on the same data!
- Very poor methodology since it overfits to this particular training set
- This training dataset has a separate test data set
  - We can also try 10-fold cross validation
Classifier: J48 -C 0.25 -M 2

Test options:
- Supplied test set

Classifier output:

Size of the tree: 49

Time taken to build model: 0.02 seconds

Time taken to test model on supplied test set: 0 seconds

Correctly Classified Instances 112 84.8485%
Incorrectly Classified Instances 20 15.1515%
Kappa statistic 0.7255
Mean absolute error 0.1198
Root mean squared error 0.2915
Relative absolute error 32.9443%
Root relative squared error 66.1432%
Total Number of Instances 132

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>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.987</td>
<td>0.127</td>
<td>0.916</td>
<td>0.987</td>
<td>0.950</td>
<td>0.777</td>
<td>0.967</td>
<td>0.962</td>
</tr>
<tr>
<td>2</td>
<td>0.650</td>
<td>0.063</td>
<td>0.650</td>
<td>0.650</td>
<td>0.650</td>
<td>0.887</td>
<td>0.851</td>
<td>0.660</td>
</tr>
<tr>
<td>3</td>
<td>0.657</td>
<td>0.062</td>
<td>0.793</td>
<td>0.657</td>
<td>0.719</td>
<td>0.887</td>
<td>0.887</td>
<td>0.690</td>
</tr>
</tbody>
</table>

Weighted Avg. 0.848 0.100 0.843 0.843 0.843 0.769 0.928 0.844

Confusion Matrix:

a b c classified as
76 0 1 a = 1
2 13 5 b = 2
5 7 23 c = 3

Status: OK
RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Correctly Classified Instances 115 87.1212%
Incorrectly Classified Instances 17 12.8788%
Kappa statistic 0.7653
Mean absolute error 0.1642
Root mean squared error 0.2605
Relative absolute error 45.1528%
Root relative squared error 59.0951%
Total Number of Instances 132

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.974 0.164 0.893 0.974 0.932 0.831 0.986 0.992 1
0.750 0.036 0.789 0.750 0.769 0.730 0.961 0.838 2
0.714 0.041 0.862 0.714 0.781 0.718 0.965 0.910 3

Weighted Avg. 0.871 0.112 0.869 0.871 0.867 0.785 0.978 0.947

Confusion Matrix:

a b c <- classified as
75 1 1 | a = 1
2 15 3 | b = 2
7 3 25 | c = 3
AUTO MPG Results (2)

• Using an independent test set shows more realistic balanced F1 score of .843
• Using Random Forest raises this to .867
• While the increase is not large, it is probably statistically significant
• F1 scores this high are difficult to increase dramatically
  – Human scores for many tasks are often in this range (i.e., 0.8 – 0.9)
Adult Data Set

Download: Data Folder Data Set Description

Abstract: Predict whether income exceeds $50K/yr based on census data. Also known as "Census Income" dataset.

<table>
<thead>
<tr>
<th>Data Set Characteristics:</th>
<th>Multivariate</th>
<th>Number of Instances:</th>
<th>48842</th>
<th>Area:</th>
<th>Social</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute Characteristics:</td>
<td>Categorical, Integer</td>
<td>Number of Attributes:</td>
<td>14</td>
<td>Date Donated</td>
<td>1996-05-01</td>
</tr>
<tr>
<td>Associated Tasks:</td>
<td>Classification</td>
<td>Missing Values?</td>
<td>Yes</td>
<td>Number of Web Hits:</td>
<td>1470139</td>
</tr>
</tbody>
</table>

Source:
Donor:
Ronny Kohavi and Barry Becker
Data Mining and Visualization
Silicon Graphics.
e-mail: ronnyk@live.com for questions.

Data Set Information:
Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AGE>16) & (AGI>100) & (AFNWGT>1)&(HSWK>0))
Prediction task is to determine whether a person makes over 50K a year.

Attribute Information:
Listing of attributes:
UCI Adult Dataset (2)

• Data on adults from 1994 census data
• Large dataset with 48,842 instances
• Predict if person makes over $50K/year
  – Equivalent to ~$90K/year in 2021
• 14 features: age, education, marital status, occupation, race, sex, native country, ...
  – Mixture of numeric (e.g., age) and nominal (e.g., occupation) values
Classifier:
Choose J48 -C 0.25 -M 2

Test options:
- Use training set

Classifier output:
Size of the tree: 911

Time taken to build model: 2.64 seconds

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:

<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>MCR</th>
<th>ROC Area</th>
<th>PRC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;50K</td>
<td>0.631</td>
<td>0.046</td>
<td>0.810</td>
<td>0.631</td>
<td>0.710</td>
<td>0.60</td>
<td>0.907</td>
<td>0.792</td>
</tr>
<tr>
<td>&lt;=50K</td>
<td>0.954</td>
<td>0.369</td>
<td>0.891</td>
<td>0.954</td>
<td>0.921</td>
<td>0.94</td>
<td>0.907</td>
<td>0.960</td>
</tr>
</tbody>
</table>

Weighted Avg.:
- 0.876
- 0.292
- 0.872
- 0.176
- 0.871
- 0.840

Confusion Matrix:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>--- classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>7375</td>
<td>4312</td>
<td>a = &gt;50K</td>
</tr>
<tr>
<td>1727</td>
<td>35428</td>
<td>b = &lt;=50K</td>
</tr>
</tbody>
</table>
Classifier:

```
Choose RandomForest -P 100 -l 100 -num-slots 1 -K 0 -M 1.0 -V 0.001
```
Result

- Significant increase on F1 scores when both trained and evaluated on training set
- This is considered to be poor methodology since it overfits to the particular training set
Create 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
• Large dataset reduced problem of overfitting, so random forest did not help
F = 0.856
F = 0.853
Conclusions

• Bagging can help, especially if amount of training data adequate, but not as large as it should be

• While we explore it using decision trees, it can be applied to any classifier
  – Scikit-learn has a general module for bagging

• In general, using any of several ensemble approaches to classification is often very helpful
Conclusions

• Wait, there’s more...

• A classification problem can change over time
  – E.g.: recognizing a spam message from its content and metadata

• We showed that an ensemble approach can detect a change in the nature of spam
  – Which tells us its time to retrain with new data
Recognizing Concept Drift

• Build ensemble of five models to classify spam comments left on a blog at time T1
• Note the relative level of agreement
• Detect when one of the models starts to diverge from the others with at time T2
  – Time to get new data and retrain
  – Examining disagreements can be enlightening
• Used temporal data spanning several years to prove effectiveness
  – E.g., spam moved from *viagra* to *weight loss*