Benchmarking Random Forest against Naive Bayes

Nghia Nguyen
Brandon King
Anand Subramanian
Objectives

1. How does the performance of Random Forest (RF) compare with that of Naive Bayes (NB)?

   Performance both in metrics of:
   ○ Classifier **Accuracy**: % of test data correctly labeled
   ○ **Run-Time**

2. How does the difference in performance depend upon data set characteristics?
Approach

- Tested performance of RF and NB on 8 small to medium-scale UCI data sets
  - Personal Income
  - Credit Approval
  - Annealing
  - Haberman - Breast cancer surgery 5+ year survival
  - Breast Cancer
  - Letter Image Recognition
  - ISOLET (Isolated Letter Speech Recognition)
  - Ionosphere - Radar returns of ionosphere

Decision Tree Algorithm

1. For data set X, calculate entropy of every feature.
2. Find feature for which information gain is maximum
   ○ For continuous features, sort the data based on the feature and then find the threshold that maximizes the information gain.
3. Make a decision node corresponding to that feature
4. Split set X into subsets using the attribute for which information gain is maximum.
5. Recurse on subsets.

Stopping Condition: Subsets have very small # of samples OR Very small info gain from additional split
Random Forest (RF)  
(aka Breiman's Algorithm)

N: # of training samples, M: # of classifier features

For each tree

1. **Create training set**
   - Randomly select \( n \) samples with replacement from the \( N \) available training samples

2. **Create decision rule** at each node of the tree,
   - Randomly select \( m \) features for the decision at that node.
   - Select feature (from \( m \) available) that maximizes information gain in the training set.

3. Each tree is fully grown and not pruned.
Naive Bayes (NB) Algorithm

- Assume features are independent (Naive)

- Feature Models
  - Categorical: Multinomial
  - Continuous: Gaussian OR Multinomial (binned)

- Perform Laplace ('add-one') smoothing

- Use log probabilities to deal with possible underflow issues
Census Income Prediction

  - 40 Features: age, class, industry code, race, sex, labor union, capital gains, marital status, state of residence, weeks worked per year, etc.
  - 2 Classes: >= 50K, < 50K

- training: 199523 samples
- test: 99762 samples

<table>
<thead>
<tr>
<th></th>
<th>DT</th>
<th>RF</th>
<th>NB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy:</td>
<td>0.927</td>
<td>0.952</td>
<td>0.827</td>
</tr>
<tr>
<td>Run-Time [s]:</td>
<td>31.7</td>
<td>57.7</td>
<td>1.88</td>
</tr>
</tbody>
</table>
ISOLET - Isolated Letter Speech Recognition

• 150 speakers saying each letter of the alphabet twice.
  ○ 617 Features: Continuous Linguistic attributes (sonorants, contours, spectral coefficients)
  ○ 26 Classes: Letters of the alphabet

• training: 6238 instances (120 speakers)
• test: 1559 instances (30 speakers)

<table>
<thead>
<tr>
<th></th>
<th>DT</th>
<th>RF</th>
<th>NB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy:</td>
<td>0.812</td>
<td>0.940</td>
<td>0.767</td>
</tr>
<tr>
<td>Run-Time [s]:</td>
<td>291.5</td>
<td>337.2</td>
<td>5.24</td>
</tr>
</tbody>
</table>
Annealing - Metallurgy

● Prediction of 6 annealing classes
  ○ 38 Features: steel type, carbon content, hardness, strength, thickness, width, etc.
  ○ 6 Classes: Result of annealed steel

● training: 599 samples
● test: 199 samples

<table>
<thead>
<tr>
<th></th>
<th>DT</th>
<th>RF</th>
<th>NB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy:</td>
<td>0.992</td>
<td>0.987</td>
<td>0.742</td>
</tr>
<tr>
<td>Run-Time [s]:</td>
<td>0.044</td>
<td>0.175</td>
<td>0.027</td>
</tr>
</tbody>
</table>
Haberman - Breast Cancer Survival

- Predicting survival rate (after 5 years) of breast cancer patients.
  - 3 Features: Age, Year of operation, # of axillary lymph nodes
  - 2 Classes: Patient survived?

- training: 230 patients
- test: 76 patients

<table>
<thead>
<tr>
<th></th>
<th>DT</th>
<th>RF</th>
<th>NB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.653</td>
<td>0.736</td>
<td>0.732</td>
</tr>
<tr>
<td>Run-Time [s]</td>
<td>0.019</td>
<td>0.007</td>
<td>0.005</td>
</tr>
</tbody>
</table>
Letter Image Recognition

- Predicting letter from primitive image features.
  - 16 Features: Image characteristics such as width, height, # of pixels, etc..
  - 26 Classes: Letters of the alphabet

- **training:** 15000 samples
- **test:** 5000 samples

<table>
<thead>
<tr>
<th></th>
<th>DT</th>
<th>RF</th>
<th>NB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.870</td>
<td>0.959</td>
<td>0.725</td>
</tr>
<tr>
<td>Run-Time [s]:</td>
<td>3.27</td>
<td>9.01</td>
<td>0.37</td>
</tr>
</tbody>
</table>
Conclusion

Random Forests, as compared to Naive Bayes,

1. offers **consistent** and **marked** improvements in accuracy
   ○ Particularly true for multiple class classification tasks

2. requires far **more processing time**
   ○ At least an order of magnitude
   ○ Computational burden increases with **# of features**.