Lecture 8: Out-of-Bag (OOB) Error

Instructor: Prof. Shuai Huang
Industrial and Systems Engineering
University of Washington
The Out-of-Bag (OOB) error

- The out-of-bag (OOB) error in a random forest model provides a computationally convenient approach to evaluate the model without using a testing dataset, neither a cross-validation procedure.
The idea behind the OOB error

• The probability of a data point from the training data is missing from a bootstrapped dataset is 
\[
\left(1 - \frac{1}{N}\right)^N.
\]

• When \(N\) is sufficiently large, we can have 
\[
\lim_{N \to \infty} \left(1 - \frac{1}{N}\right)^N = e^{-1} \approx 0.37.
\]

• Therefore, roughly 37% of the data points from \(S\) are not contained in any bootstrapped dataset \(B_i\).

• And thus, not used for training tree \(i\). These excluded data points are referred as the out-of-bag samples for the bootstrapped dataset \(B_i\) and tree \(i\).
Further develop the line of argument

• As there are 37% of probability that a data point is not used for training a tree, we can infer that, a data point is not used for training about 37% of the trees.

• Therefore, for each data point, in theory, there are 37% of trees trained without this data point. These trees can be used to predict on this data point, which can be considered as testing an unseen data.

• The out-of-bag error estimation can then be calculated by aggregating the out-of-bag testing error of all the data points.

• The out-of-bag error can be calculated after random forests are built, and are significantly less computationally than cross-validation.
A Simple Example

• Suppose that we have a training dataset of 5 instances (IDs as 1,2,3,4,5).

$$\textbf{Table 5.3: The out-of-bag (OOB) errors}$$

<table>
<thead>
<tr>
<th>Bootstrap</th>
<th>Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,1,4,4,5</td>
<td>1</td>
</tr>
<tr>
<td>2,3,3,4,4</td>
<td>2</td>
</tr>
<tr>
<td>1,2,2,5,5</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tree</th>
<th>Training data</th>
<th>1 (C1)</th>
<th>2 (C2)</th>
<th>3 (C2)</th>
<th>4 (C1)</th>
<th>5 (C2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,1,4,4,5</td>
<td>C1</td>
<td>C2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2,3,3,4,4</td>
<td>C1</td>
<td></td>
<td></td>
<td></td>
<td>C2</td>
</tr>
<tr>
<td>3</td>
<td>1,2,2,5,5</td>
<td></td>
<td>C2</td>
<td>C1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• We can see that, as the data instance (ID = 1) is not used in training Tree 2, we can use Tree 2 to predict on this data instance, and we see that it correctly predicts the class as C1.

• Similarly, Tree 1 is used to predict on data instance (ID=2), and the prediction is wrong. Finally, we can see that the overall out-of-bag (OOB) error is 1/6.
R lab

• Download the markdown code from course website
• Conduct the experiments
• Interpret the results
• Repeat the analysis on other datasets