Project: Parameter Tuning of Random Forest using Design of Experiments

Lecturer: Alan R. Vazquez  
Affiliation: University of Arkansas

# Task Description

Due to budgetary constraints, the maximum number of tests in the whole experimentation process is 35. That is, you can test a maximum of 35 combinations of the settings of the tuning parameters. If you run all the tests on a single computer, consider the experiment as completely randomized.

In your report, you must provide comprehensive answers to the questions below. In your presentation, you must provide a brief summary of your solution to the final project.

## Part I: Design of the Experiment

#### Question 1. Propose a fractional factorial design for the problem. In addition, propose an experimental design constructed using the optimal design approach.

#### Question 2. Compare the optimal design with the fractional factorial design in practical and statistical terms. For instance, what is the performance of the designs for studying the main effects of the tuning parameters only? Can they estimate all two-parameter interactions? Why or why not? How do they compare in terms of multicollinearity?

#### Question 3. Recommend one experimental design between the two options in Question 1. Motivate your decision.

#### Question 4. Using a commercial software, the TAs and I came up with the experimental design shown in Table 1. How does your recommended design in the previous question compare with this one?

Table 1. Alternative experimental design.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Run** | **ntree** | **replace** | **mtry** | **nodesize** | **maxnodes** | **classwt** | **cutoff** |
| 1 | 100 | 1 | 2 | 11 | 10 | 0.5 | 0.8 |
| 2 | 550 | 0 | 2 | 1 | 10 | 0.5 | 0.2 |
| 3 | 1000 | 1 | 4 | 1 | 10 | 0.5 | 0.2 |
| 4 | 1000 | 1 | 6 | 1 | 1000 | 0.5 | 0.8 |
| 5 | 1000 | 0 | 6 | 1 | 1000 | 0.9 | 0.2 |
| 6 | 100 | 0 | 2 | 1 | 1000 | 0.5 | 0.8 |
| 7 | 1000 | 0 | 2 | 6 | 10 | 0.9 | 0.8 |
| 8 | 100 | 0 | 2 | 11 | 10 | 0.9 | 0.2 |
| 9 | 100 | 1 | 6 | 1 | 10 | 0.7 | 0.8 |
| 10 | 100 | 0 | 6 | 1 | 10 | 0.9 | 0.5 |
| 11 | 100 | 0 | 4 | 11 | 1000 | 0.9 | 0.8 |
| 12 | 1000 | 1 | 2 | 11 | 1000 | 0.5 | 0.5 |
| 13 | 100 | 1 | 6 | 6 | 1000 | 0.5 | 0.2 |
| 14 | 550 | 0 | 4 | 6 | 505 | 0.7 | 0.5 |
| 15 | 100 | 0 | 6 | 11 | 505 | 0.5 | 0.2 |
| 16 | 1000 | 0 | 6 | 11 | 10 | 0.5 | 0.8 |
| 17 | 1000 | 0 | 2 | 11 | 1000 | 0.7 | 0.2 |
| 18 | 1000 | 1 | 6 | 11 | 10 | 0.9 | 0.2 |
| 19 | 100 | 1 | 2 | 1 | 1000 | 0.9 | 0.2 |
| 20 | 1000 | 1 | 2 | 1 | 505 | 0.9 | 0.8 |
| 21 | 550 | 1 | 4 | 6 | 505 | 0.7 | 0.5 |
| 22 | 550 | 1 | 6 | 11 | 1000 | 0.9 | 0.8 |

## Part II: Data Analysis

#### Question 5. Collect data using your recommended design in Question 3.

#### Question 6. Conduct a detailed data analysis. What are the influential tuning parameters? What is the final model that links the tuning parameters to the cross-validation accuracy? Does the final model provide a good fit to the data?