Session 4 - Statistical Modeling

Jongbin Jung

January 10, 2016

Dependencies

- Latest version ($\geq 3.1.2$) of R
  (*free* from https://www.r-project.org/)
- Latest version of Rstudio (also *free* from
  https://www.rstudio.com/)
- Packages

```
install.packages('glmnet')
install.packages('ROCR')
```

Statistical Modeling

- In this session, I'll demonstrate some common statistical modeling techniques used in `R`
- But I *won't*
  - Go into the theory/details of each (or any) model/method; I assume you know all that already
  - Introduce the `carat` package, which is a **great** package, but I'm not dealing with it because:
    1. it's pretty well documented here (http://topepo.github.io/caret/index.html), and
    2. I (very personally) don't think 'teaching people how to use the `carat` package' is equivalent to 'teaching people statistical modeling in `R`'
- I'll focus on regularized linear models with the `glmnet` package, but many of the general concepts work for most models/packages
- At the very least, you'll end up with a set of (hopefully useful) `R` snippets

Data Splitting

Pre-processing

model.matrix()

scale()

Training models

Prediction/Evaluation

Bootstrap