

Machine Learning with R

Joshua Reich

josh@i2pi.com

April 2, 2009

Why R?

How to find out about stuff?

What is Machine Learning?

Show me the money

Learn More

Questions

ML Alternatives

- Matlab
- Weka
- Python
- Stand alone (e.g. Vowpal Wabbit)

*"The best thing about R is that it was developed by statisticians.
The worst thing about R is that it was developed by statisticians."*

–Bo Cowgill, Google (at SF R Meetup)

Why R?

- Working with the CLI - iterative discovery
- Integrated graphics
- Community supported packages (CRAN)
- ODBC Integration
- You already use it

How to find out about stuff?

- ?function
- help.search("search string")
- RSiteSearch("search string")
- <http://rseek.org/>
- names(object) or attributes(object)
- > kmeans
function (x, centers, iter.max = 10, nstart = 1,
algorithm = c("Hartigan-Wong",
"Lloyd", "Forgy", "MacQueen"))
...

What is Machine Learning?

| Statistics | Machine Learning |
|-------------------|------------------|
| Probability Model | Learning Model |
| Observations | Observations |
| Estimation | Training |
| MLE | Optimization |

Semantics v. Pragmatics

- For most statistical models there are either closed form or quick numerical approximations for finding model properties - e.g., confidence intervals. Assuming you believe that your data generating process is accurately captured by your model, then you can make direct statements about unseen events.
- Machine learning is a close cousin to non-parametric techniques and relies on training/testing/validation cycles, bootstrapping and cross-validation to determine measures of reliability.

But invariably, simple models and a lot of data trump more elaborate models based on less data

—Halevy, Norvig & Pereira.

Inductive Bias

In the days when Sussman was a novice, Minsky once came to him as he sat hacking at the PDP-6.

"What are you doing?", asked Minsky.

"I am training a randomly wired neural net to play Tic-tac-toe", Sussman replied.

"Why is the net wired randomly?", asked Minsky.

"I do not want it to have any preconceptions of how to play", Sussman said.

Minsky then shut his eyes.

Inductive Bias

"Why do you close your eyes?" Sussman asked his teacher.

"So that the room will be empty."

What is Machine Learning?

- Regression vs. Classification

$$Y \in \mathbb{R}^P$$

vs.

$$Y \in \{Y_1, Y_2, \dots, Y_N\}$$

- Supervised vs. Unsupervised Learning

$$Y = f(X)$$

vs.

$$X_1, X_2, \dots, X_N$$

General Supervised Learning Framework

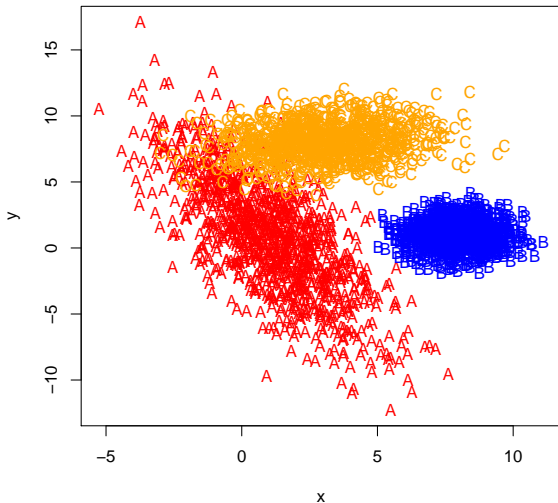
(Ch 7 of ElemStatLearn)

- Training / Validation / Test
- Variance - Bias Decomposition: Overfitting
- Feature Selection / Regularization
- Bootstrapping / Cross-Validation

What we will walk through

- K-Means clustering: `kmeans()`
- K-Nearest Neighbours: `knn()`
- Regression Trees: `rpart()`
- Improving trees with PCA: `princomp()`
- Linear Discriminant Analysis: `lda()`
- Support Vector Machines: `svm()`

The Problem



Machine Learning More

- MachineLearning CRAN view
<http://cran.r-project.org/web/views/MachineLearning.html>
- The caret package is a good one.
- Elements of Statistical Learning
<http://www-stat.stanford.edu/~tibs/ElemStatLearn/>
- Machine Learning (Mitchell)
<http://www.cs.cmu.edu/~tom/mlbook.html>
- Video Lectures
<http://videlectures.net/>

Questions?