New Book, "The Art of Machine Learning"

and Intro to the qeML Package

Norm Matloff
University of California, Davis

East Bay R Users Group
December 12, 2023

Why Yet Another ML Book?

- Almost all books are either
  - math-heavy or
  - “cookbooks,” step-by-step “recipes,” or
  - both
- ML is an *art*, not a science
  - Note my previous NSP “Art of ” books:
    - *The Art of R Programming*
    - *The Art of Debugging*
  - ML is typically taught in a “What function should I call, and with what arguments?” mode
- My goal is to enable the reader to *use* ML in the real world.
- NO MATH IS USED (just slope of line), but INTUITION is centrally important. What do these methods REALLY do?

Chapter Outline

- Prologue: Regression problems, illustrated with k-NN
- Prologue: Classification problems, illustrated with k-NN
- Bias, Variance, Overfitting
- Dealing with Large Numbers of Features
- Decision Trees
- Tweaking the Tress
- Finding a Good Set of Hyperparamters
- Linear, generalized linear models
- Shrinkage-based models
- Support Vector Machines
- Neural networks
- Image classification
- Time Series and Text

Recurring Sections: the Bias-Variance Tradeoff

- Supremely important—18,400,000 results to my Google query.
- Yet most books just devote one or two *very vague* sentences to it.
- Sections 1.7, all of Chapter 3, 4.3.6, 6.1, 6.3.5, 9.3.2, 11.10, 13.4
- Example: k-Nearest Neighbors, Section 1.7
  - if k is small, not many neighbors, a small "sample"—hence large **variance**
  - if k is large, some neighbors are quite distant, hence a **bias**; e.g. Y = weight, X = height
- Advantages and disadvantages of parametric models, including polynomial regression.

Recurring Sections: Pitfalls

- Sections 1.13, 1.14, 1.15, 1.16, 2.2.1, 2.2.2, 2.2.5, 2.4, 2.7.5, 5.3.1, 11.8, Appendix D
- Example: Random Forests, Setion 5.3.1:
  - NYC taxi data (n=10000 version)
  - potentially 29,315 pickup and dropoff combinations!
  - we aim roughly for $p < \sqrt{n}$ (though note *double descent* etc.)
  - **partykit** package error message, "too many levels"
  - possibly consolidate or even use latitude-longitude embedding

Statistics vs. CS

- Old Breiman "Two Cultures" essay still applies.
- Sampling variation vs. "the data."
- E.g. grid search for hyperparameter tuning includes standard errors.
- Statistics $\iff$ CS Translator, e.g. *prediction* $\iff$ *inference*

The qeML Package

- On CRAN.
- Independent of the book.
- "Quick and Easy" ML
- Uniform, SIMPLE user interface.

  ```
  z ← qeRF(svcensus,'wageinc')
  ```

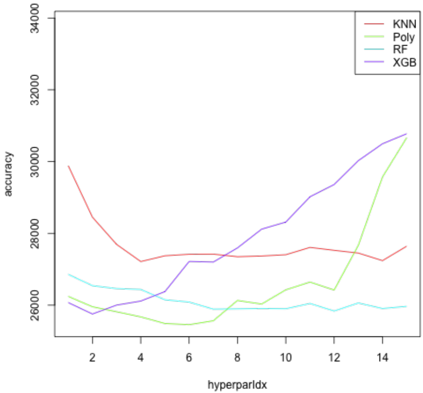
  One simple call, that's all! No clumsy setup needed.
- Various default options.
- "Easy for learners, powerful for advanced users"
- Excellent for teaching:
  - SIMPLE user interface.
  - Many built-in datasets.
  - Includes a number of built-in ML tutorials vignettes, no background needed.
- Various utlities, e.g. for factor manipulation.

Example: Comparison of Various ML Methods

- All **qeML** predictive functions do automatic cross-validation.
- Test accuracy in the **$testAcc** component of the returned object.
- Also **$baseAcc**, accuracy of prediction without X, for comparison.

Example

Predict wage income in 2000 Census dataset, from age, gender, education and tech occupation.



Horizontal axis is (indexed) k, min leaf size etc.
Winner is good ol’ polynomial regression!