Statistical Modeling: The Two Cultures

by Leo Breiman

presented by Jack Morris

https://qdata.github.io/deep2Read/

Roadmap

Data Modeling Culture [Statistics]

Algorithmic Modeling Culture [Machine Learning]

Principles of Statistical Learning

Summary + Leo’s Advice


y
nature
x

Two goals of statistics

Prediction:

To predict responses for future input variables

Information:

To extract some information about what nature is actually doing

The two cultures: Data models

Assume a stochastic model is *actually happening* inside the black box. This means that if we figure out the model, we can figure out what nature is doing!

Popular tools: Linear regression, logistic regression, Cox model

Validation technique: examining residuals, testing model fit, etc.

Estimated population of statisticians (in 2001): 98%

The two cultures: Algorithmic models

We don’t know (or care) what’s happening inside the black box. It’s complex–and fundamentally unknowable. We just want to find some function f(x) that can predict y.

Popular tools: decision trees, neural networks

Validation technique: predictive accuracy

Estimated population of statisticians (in 2001): 2%

Jim Simons: The Ultimate Algorithmic Modelist

“I don’t know why planets orbit the sun. That doesn’t mean I can’t predict them.” –Jim Simons

net worth $15.5 billion

Roadmap

Data Modeling Culture [Statistics]

Algorithmic Modeling Culture [Machine Learning]

Principles of Statistical Learning

Summary + Leo’s Advice

Data Modeling Culture

- As of the time of writing, most of the statistics field was focused on creating **data models**
- Data models give statisticians a job
- they require lots of data analysis to develop hypotheses about how nature is actually functioning... and then model it
- data models **extract information about the underlying mechanism producing the data**

A typical data model

Find a stochastic model of the data-generating process:
y = f(x, parameters, random error)

Data modeling: discerning the model that truly produces the data

- a famous (also infamous) example:

$$y = b_0 + \sum_{1}^{M} b_m x_m + \varepsilon,$$

- y is a function of x with corresponding weights + random error
- *is the rent of some apartments really normally distributed?* 🤔

Data modeling: typical assumptions

- data are generated by a specific stochastic model
- often assumes linearity
- requires lots of data analysis + expert understanding

Data modeling: problems

- Conclusions are made about the model (<u>not</u> about nature)
- Assumptions are often (always?) violated
- Often no real model evaluation– and once the model is released, its predictions are considered gospel
- Focus is on analysis, not prediction
- Data models always fail in areas like image and speech recognition

Roadmap

Data Modeling Culture [Statistics]

Algorithmic Modeling Culture [Machine Learning]

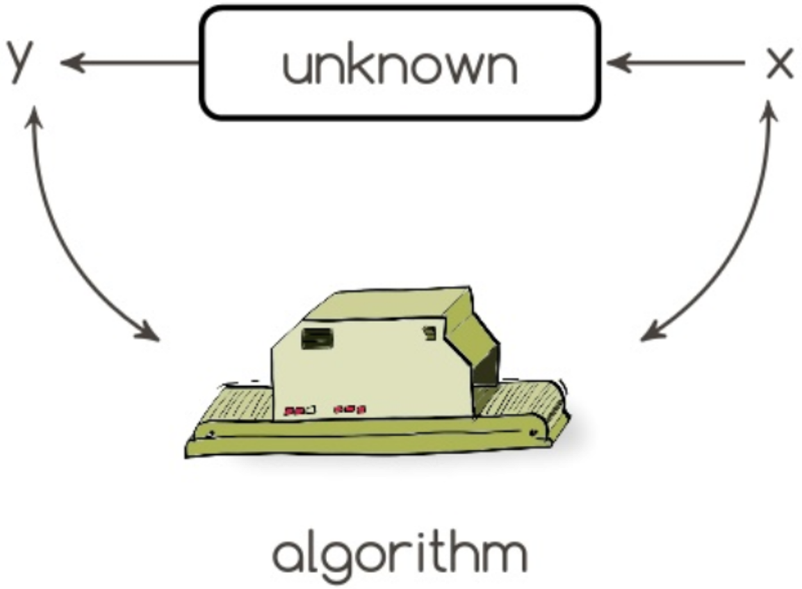
Principles of Statistical Learning

Summary + Leo’s Advice

A singular goal

find a function *f(X)* that minimizes the loss *L(Y, f(X))*.

that's it.



Algorithmic modeling: major differences

- The target is not to find (or understand) the true data-generating mechanism– but to use an algorithm that imitates the mechanism as effectively as possible

- This is machine learning culture

- Summary: data modeling culture tries to find the true data-generating mechanism. Algorithmic modeling culture is comfortable approximating the mechanism as closely as possible.

Algorithmic modeling: major differences

- The target is not to find (or understand) the true data-generating mechanism– but to use an algorithm that imitates the mechanism as effectively as possible

- This is machine learning culture

Algorithmic vs. Data Modeling

- Summary: data modeling culture tries to find the true data-generating mechanism. Algorithmic modeling culture is comfortable approximating the mechanism as closely as possible.

- And once you relax your goal– and aspire solely for minimal prediction error– you open a door to a whole host of new algorithms...

Examples of Algorithmic Models

- Boosting
- Support Vector Machines
- Neural networks
- Random forests
- Hidden markov models
- Bayesian networks
- ... many other things

Random forests vs neural networks

- “**Random forests are A+ predictors**” – in a comparison of 18 different classifiers (neural networks, CART, linear regression, nearest, neighbor, etc), random forests placed **1 out of 18** over four datasets
- Fifth dataset: 16x16 pixel grayscale depictions of handwritten numerals
  - **a neural net** ...got **5.1% error** (vs 6.2% for random forest)
- remember this was 2001

Roadmap

Data Modeling Culture [Statistics]

Algorithmic Modeling Culture [Machine Learning]

Principles of Statistical Learning

Summary + Leo’s Advice

Three most important lessons from algorithmic modeling

1. **Rashomon**: there are many equally good models

2. **Occam:** there is a conflict between simplicity and accuracy

3. **Bellman:** dimensionality is a blessing and a curse

[1] Rashomon Effect

- Rashomon is a Japanese movie where four witnesses see a crime from very different angles (but equal accuracy?)
- Models can–and do–have totally different interpretations (@"Attention is not Explanation")
- Algorithmic modelers <u>exploit</u> the Rashomon effect by aggregating the predictions of many models
    - this is what Random Forests do

[2] Occam's Razor

- "The simplest solution is best" (or something like that) –Occam
    - ("Everything should be made as simple as possible, and no simpler" –Einstein)

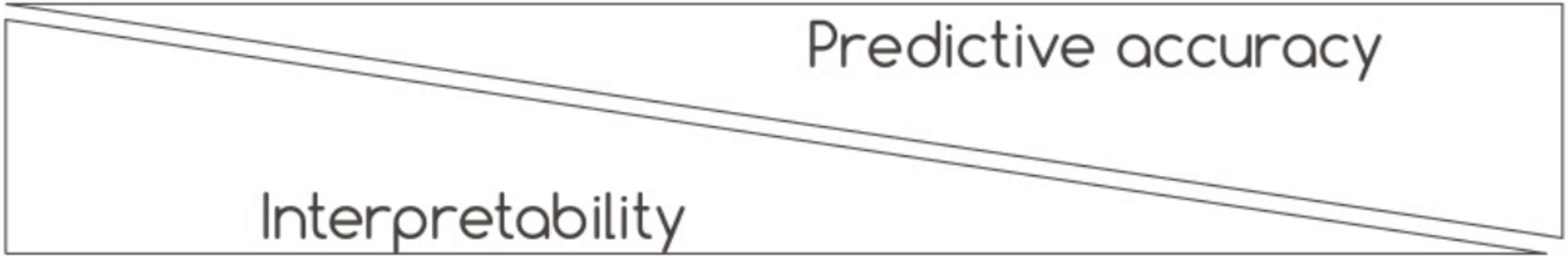
- **There is a natural tradeoff between predictive accuracy and interpretability**

Prediction vs. Interpetation

- Models that are good at prediction are (often) more complex
  - models that are easy to interpret are simple, and therefore, worse predictors
- Decision Trees are super intuitive, but can't model complex processes
- Random Forests have excellent prediction accuracy, but are basically impossible to interpret

Prediction vs. Interpretation

Random Forests
Neural Networks



Decision Trees
Logistic Regression

[3] Bellman and the Curse of Dimensionality

- The higher the dimensionality of the data (# covariates), the more difficult it is to separate signal from noise
- *Common practice in data modeling*: variable selection (done by experts or data analysts) and dimensionality reduction (PCA)
- *Common practice in algorithmic modeling*: engineering extra features (more covariates!) to increase predictive accuracy

Roadmap

Data Modeling Culture [Statistics]

Algorithmic Modeling Culture [Machine Learning]

Principles of Statistical Learning

Summary + Leo’s Advice

Five pieces of advice for statistical analysis

1. Focus on finding a good solution to the problem. That’s what you’re paid for.
2. Live with the data before you plunge into modeling. (!)
3. Search for a model that gives a good solution, be it algorithmic or data.
4. Predictive accuracy on test sets is *the* criterion for how good your model is (at prediction).
5. Computers are an indispensable partner.

Information from a black box

- “A model does not have to be simple to provide reliable information about the relationship between x and y… **the goal is not interpretability, but accurate information.**”

Conclusion

- [1] Higher predictive accuracy --> more reliable information about the underlying data mechanism
  - weaker predictive accuracy --> questionable conclusions
- [2] Algorithmic models can give better predictive accuracy than data models *and* provide better information about the underlying mechanism