Exchanges between Statistics and Machine Learning

Ben Klemens

28/29 November 2017

[The U.S. Treasury takes no position on the issues raised in this presentation.]

Intro and Outline

- A broad discussion of what models have in common
- Examples of comparative statics, validation, and prediction across several types of model

Intro and Outline

- A broad discussion of what models have in common
- Examples of comparative statics, validation, and prediction across several types of model
- What not to expect
  - Q: What is the right model to use?
  - A: That's not a well-formed question

Intro and Outline

- A broad discussion of what models have in common
- Examples of comparative statics, validation, and prediction across several types of model
- What not to expect
  - Q: What is the right model to use?
  - A: That's not a well-formed question
  - Q: What is the best way to evaluate a model?
  - A: That's not a well-formed question

Part I: What is a model?

Problem statement

- "The word 'model' in statistical literature usually refers to an equation to which one tries to fit data via regression analysis." [Complex Systems Modelling Group, *Modelling in Healthcare*, p 49]

Problem statement

One afternoon, I tallied the models in the last 50 papers from the WB working paper series and the U.S. Census Center for Economic Studies w.p. series.

	WB	CES
Papers with regressions only	25	33
Papers including any other model	7	4
Papers w/no model fitting	18	13

WB: excluding no-model papers, 78% regression

CES: excluding no-model papers, 89% regression

Probability vs Statistics

- Probability: Theorems. If the data has some property, as $N \to \infty$, something holds. (Law)
- Statistics: A summary of how the model designer sees the world. (Custom)

A statistical model links parameters and data via likelihoods

- estimation: data $\rightarrow$ parameters
- RNG: parameters + arbitrary sequence $\rightarrow$ data
- predict/conditional expected value: parameters + some data $\rightarrow$ other data
- log likelihood, probability, entropy: parameters + data $\rightarrow$ a measure

Remittances

- Most sent out
  - United States: 263,225 million
  - Saudi Arabia: 87,412
  - United Arab Emirates: 62,242
  - Canada: 43,962
  - United Kingdom: 41,681
  - $\Sigma = 1,263,791 = 174$ countries
- Most received in
  - China: -116,573 million
  - India: -114,090
  - Philippines: -61,261
  - Mexico: -52,026
  - Pakistan: -38,761
  - $\Sigma = -1,263,791 = 197$ countries

[Run `correlations()` in the demo script about here.]

Model I: guessing

- Belize

Model I: guessing

- Belize — net out
- Ecuador

Model I: guessing

- Belize — net out
- Ecuador — net in
- Luxembourg

Model I: guessing

- Belize — net out
- Ecuador — net in
- Luxembourg — net out
- Malta

Model I: guessing

- Belize — net out
- Ecuador — net in
- Luxembourg — net out
- Malta — net in
- Iceland

Model I: guessing

- Belize — net out
- Ecuador — net in
- Luxembourg — net out
- Malta — net in
- Iceland — net in
- Congo, both Republic and Dem. Republic

Model I: guessing

- Belize — net out
- Ecuador — net in
- Luxembourg — net out
- Malta — net in
- Iceland — net in
- Congo, both Republic and Dem. Republic — net out

A list of models (1/2)

- generalized linear regression
  - may include nonlinear terms
  - logit, probit, et cetera
  - also includes systems of equations
  - an incomplete model–see below
- The Normal Distribution (params are μ, σ)
  - Also, χ^2, *t*, Zipf, Lognormal, Poisson

A list of models (1/2)

- generalized linear regression
  - may include nonlinear terms
  - logit, probit, et cetera
  - also includes systems of equations
  - an incomplete model–see below
- The Normal Distribution (params are μ, σ)
  - Also, χ^2, *t*, Zipf, Lognormal, Poisson
- 'non parameteric models': a *lot* of parameters
  - A histogram is a model
  - Number of parameters may be a parameter

A list of models (2/2)

- Decision trees (parameters=cutpoints)
- Bayesian networks (parameters=cross of free submodel params)
  - Build a narrative piece by piece
- Support vector machines [categorization] (params=dividing line parms)
- neural networks (params=network activation params)

Understanding the parameters (comparative statics)

- ceteris paribus:
  - linear regression: β.
  - trees: find the relevant cutpoint; follow it
  - neural network: just try it

Understanding the parameters (comparative statics)

- ceteris paribus:
  - linear regression: β.
  - trees: find the relevant cutpoint; follow it
  - neural network: just try it
- mutis mutandis
  - needs an underlying model for the data
  - linear regression: ¿¿¿???

[`loop_over_models()`]

Part II: validation

Parameter-based

- The parameters have some proven distribution $\rightarrow$ use that.
- Assumptions don't quite fit?
  - Find a theorem deriving the correct distribution.

Parameter-based

- The parameters have some proven distribution $\rightarrow$ use that.
- Assumptions don't quite fit?
  - Find a theorem deriving the correct distribution.
  - Or, just use the Normal distribution anyway.

- Uses the model's likelihood function to evaluate the same model.
- Potentially difficult for non-parametric models past histograms.

Data-based

- How far does the model's implications about data diverge from the data?
- How accurate are its predictions?
- These are always available.

Replication

- The Bootstrap principle: draws from your sample $\approx$ draws from the population.
  - Given this, you can use it to estimate errors on the mean of nearly all parameters.

[`loop_over_models(want_boot=1)`]

An aside: entropy

- Has more real-world validity than most (law, not custom).
- Information loss in actual data $\rightarrow$ fake data from model
  - Kullback-Leibler divergence
  - Can be difficult: models truly falsified by the data have infinite divergence
- Adustment for unknown parameters $\rightarrow$ AIC.
  - Analogy: with unknown μ, sample estimate of $\sigma \neq$ estimate with known μ.

Train & Test

- AKA Cross-validation
- The norm in ML, but usable for any model

Train & Test

- AKA Cross-validation
- The norm in ML, but usable for any model
- We'll summarize via ROC (receiver operating characteristic)

[loop_over_models(want_tt=1)]

PS: What about Belize and Iceland?

	Data	Logit	SVM	Centroid
United States	1	1.00 (1)	1.00 (1)	0.67 (1)
China	0	0.50 (0)	0.40 (0)	0.33 (0)
Ecuador	0	0.29 (0)	0.25 (0)	0.33 (0)
Malta	0	0.30 (0)	0.27 (0)	0.33 (0)
Iceland	0	0.38 (0)	0.36 (0)	0.40 (0)
Belize	1	0.38 (1)	0.41 (1)	0.44 (1)
Luxembourg	1	0.47 (1)	0.45 (1)	0.48 (1)
Congo, Rep.	1	0.50 (1)	0.52 (1)	0.54 (1)

Mark (1) for $> .405$ and (0) for $< .405$
[Output from `make_guesses()`]

Conclusion slide

- Almost everything you can do with a regression, you can do with any model
  - The one exception is parameter-based testing, for a large subset of models
  - Use the wealth of data-space tools
- Almost every tool commonly used with other models, you can use with a regression

Discuss further, ask hard questions, get the code

- ben.klemens@treasury.gov
- ben@klemens.org

- github.com/b-k/ml_for_econometricians