Analysis: the big picture

2014-08-04 02:34:43

What are we trying to do when we analyze (model) data?

- “learn from the data”; “answer scientific and management questions” (vague!)
- describe (really?)
- understand or explain something (slippery/subjective)

  Breiman (2001)

  > as data becomes more complex, the data models become more cumbersome and are losing the advantage of presenting a simple and clear picture of nature’s mechanism.

Paradigm conflict across fields

- Platonists vs (?) Aristotelians; e.g. constructive empiricists
- Biology/ecology: Strong inference (Platt 1964); Peters (1991) *Critique for Ecology*
- linguistics: Norvig vs Chomsky
- arguments about microfoundations in economics; Big Data in econometrics
- Chris Anderson: “The End of Theory”

Methods

Models are *always* simplifications: otherwise they they don’t help us understand, or predict, reality (Borges)

- constancy
- linearity
- independence
- smoothness
- discrete classes

Classical

- Linear models: mostly model-based, but:

  - least-squares/MVUE interpretation
  - very efficient for Big Data (large-scale linear algebra)
- extended linear models: GLMs, correlations, zero-inflation, etc.
  - more/different parametric assumptions in pursuit of efficiency & interpretability
- hierarchical/mixed models
  - ancestor (ANOVA) mostly used for hypothesis testing
  - relatively efficient way to do grouping
  - works well for large N, small n within clusters
  - computationally challenging
- classical (rank-based) nonparametrics [weak assumptions about conditional distributions]: mostly hypothesis-testing (provide *only* p-values)

Algorithmic

- modern nonparametrics
  - generalized additive models (technically still 'linear models', with attendant advantages)
  - kernel density estimators (*smoothing*)
  - quantile regression
  - great for description, but difficult for decomposing descriptions (interpretability)
  - interactions possible (tensor product splines, multidimensional KDEs) but comp. intensive
- classification and regression trees (plus extensions: random forests/bagging/boosting etc.)
  - mostly ignore interactions
- support vector machines
  - computationally powerful high-dimensional categorization
- penalized/regularized approaches (ridge regression, lasso, ...)
  - mostly description-oriented; confidence intervals etc still difficult

Model building

Many tradeoffs (Levins 1966):

- Realism

- Computational feasibility (especially if resampling)
- Conformity with existing models
- Interpretability
- Flexibility

etc. etc. etc. …

Deciding on a model?

- no free lunch
- bias-variance tradeoff = under/overfitting
- **BE VERY, VERY CAREFUL WHEN USING THE DATA TO DECIDE ON A MODEL**, especially if doing hypothesis testing (*data snooping*)
- in- vs out-of-sample prediction
  - bad in-sample prediction → bad model
  - good in-sample prediction: maybe overfitted?

Model checking and diagnostics

- Graphical tools
- Goodness-of-fit measures (*avoid hypothesis testing!*)
  - Compare to saturated and null model
- Explore residuals
- Posterior predictive sampling
- Assessment of predictive skill:
  - hold-out data
  - cross-validation: this document points to `boot::cv.glm`; `rms::validate.*` (*But* see Wenger and Olden (2012))
- Fit to simulated data
  - Simulated from estimation model (= positive/negative controls)
  - Simulated from a different model (robustness)

References

Breiman, Leo. 2001. "Statistical Modeling: the Two Cultures." *Statistical Science* 16 (3) (August): 199–215. http://www.jstor.org/stable/2676681.

Levins, R. 1966. "The Strategy of Model Building in Population Biology." *American Scientist* 54: 421–431.

Peters, R. H. 1991. *A Critique for Ecology.* Cambridge, UK: Cambridge University Press.

Platt, John R. 1964. “Strong Inference.” *Science,* Series 3 146 (October): 347–353. http://links.jstor.org/sici?sici=0036-8075/%2819641016/%293/%3A146/%3A3642/%3C347/%3ASI/%3E2.0.CO/%3B2-K.

Wenger, Seth J., and Julian D. Olden. 2012. “Assessing Transferability of Ecological Models: an Underappreciated Aspect of Statistical Validation.” *Methods in Ecology and Evolution* 3 (2) (April): 260–267. doi:10.1111/j.2041-210X.2011.00170.x. http://doi.wiley.com/10.1111/j.2041-210X.2011.00170.x.