Truth in Data

David M. Blei
Fall 2009

In COS513, we covered the fundamentals of probabilistic modeling: How to build models, how to fit models to data, and how to infer unknown quantities based on those fitted models. This suite of computational problems is fundamental to many modern machine learning algorithms, with applications in information retrieval, computer vision, computational linguistics, and bioinformatics.

In traditional machine learning tasks, like search and classification, a model is as good as its performance is measured. (A better spam filter will filter more spam; a better object recognizer will recognize more objects.) Many recent applications of probabilistic modeling, however, are towards more exploratory ends. We can build models to identify the hidden community structure of a social network, the hidden thematic structure of a corpus of documents, or the hidden patterns of genes that govern our biology.

Evaluating models for interpretation is tricky, and this is the problem that we will discuss. Many questions arise:

- How and when can we interpret the results of a probabilistic model?
- How should we evaluate our modeling assumptions?
- How should we diagnose where and when they go wrong? When does it matter?
- How should we change our model based on these diagnoses?

Methods for answering these questions are essential to drawing sound conclusions from data.

The seminar is short and our treatment will be incomplete. Loosely, we will be studying four broad issues: (a) model selection (b) model diagnostics (c) model interpretation and (d) causality. We will focus on observational data. (I.e., we won't be discussing experimental design.)

Structure of the course. This is a seminar. We will be closely reading a collection of papers and parts of books, and then discussing them every week. My goal as the coordinator is to foster a lively and interactive atmosphere for discussion. The course will be set up as follows.

1. We will discuss 1-2 readings each week. I will introduce the reading for 10-30 minutes. A different student will lead the discussion.
2. By the Monday before class, please email 1-2 paragraphs of thoughts on the readings to the class mailing list. These will be some of the kindling for our discussion.
3. If you are taking the class for credit, you are responsible for a 5 page report. This can be a novel piece of research or a discussion and synthesis of existing ideas.

Course topics, readings (subject to change)

1. Diagnostics
    - Science and statistics (Box, 1976, 1980)
    - The future of data analysis (Tukey, 1962; Mallows, 2006)
    - Model posteriors (Gelman et al., 1995; Gelman, 2004; Gelman and Hill, 2007)
2. Selection
    - Model selection (Claeskens and Hjort, 2008)
    - Ockham's razor (Jaynes, 2003)
3. Perspectives on causality
    - David Freedman's perspective (Freedman, 2002, 2000)
    - Judea Pearl's perspective (Pearl, 2009)
    - Donald Rubin's perspective (Rubin, 2008)
4. Interpretation
    - Measuring judicial positions (Martin and Quinn, 2002; Quinn et al., 2006)
    - Counterfactuals, and when we can answer them (King and Zeng, 2006)

References

Box, G. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Box, G. (1980). Sampling and Bayes' inference in scientific modeling and robustness. *Journal of the Royal Statistical Society, Series A*, 143(4):383–430.

Claeskens, G. and Hjort, N. (2008). *Model Selection and Model Averaging*. Cambridge University Press.

Freedman, D. (2000). Oasis or mirage. *Chance*.

Freedman, D. (2002). From association to causation: Some remarks on the history of statistics.

Gelman, A. (2004). Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, 13(4):755–779.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995). *Bayesian Data Analysis*. Chapman & Hall, London.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Jaynes, E. T. (2003). *Probability Theory: The Logic Of Science*. Cambridge University Press.

King, G. and Zeng, L. (2006). The dangers of extreme counterfactuals. *Political Analysis*, 14:131–159.

Mallows, C. (2006). Tukey's paper after 40 years. *Technometrics*, 48(3).

Martin, A. and Quinn, K. (2002). Dynamic ideal point estimation via Markov chain Monte Carlo for the U.S. Supreme Court, 1953–1999. *Political Analysis*, 10(2):134–153.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistical Surveys*.

Quinn, K., Park, J., and Martin, A. (2006). Improving judicial ideal point estimates with a more realistic model of opinion content.

Rubin, D. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics*.

Tukey, J. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67.