The four eras of data

Statistical Learning

From Jeff Leek 2016/12/16 simply statistics post

https://simplystatistics.org/2016/12/16/the-four-eras-of-data/

The four eras of data

The three eras of statistics from Brad Efron's book:

- The age of huge census-level data sets were brought to bear on simple but important questions: Are there more male than female births? Is the rate of insanity rising?
- The classical period of Pearson, Fisher, Neyman, Hotelling, and their successors, intellectual giants who developed a theory of optimal inference capable of wringing every drop of information out of a scientific experiment. The questions dealt with still tended to be simple — *Is treatment A better than treatment B?* — but the new methods were suited to the kinds of small data sets individual scientists might collect.
- The era of scientific mass production, in which new technologies typified by the *microarray* allow a single team of scientists to produce *high-dimensional data*. But now the flood of data is accompanied by a deluge of questions, perhaps thousands of estimates or hypothesis tests that the statistician is charged with answering together.

Jeff Leek breakdown goes like this:

1. **The era of not much data** Prior to about 1995, usually we could collect a few measurements at a time. The whole point of statistics was to try to optimaly squeeze information out of a small number of samples - so you see methods like maximum likelihood and minimum variance unbiased estimators being developed.
2. **The era of lots of measurements on a few samples** This one hit hard in biology with the development of the microarray and the ability to measure thousands of genes simultaneously. This is the same statistical problem as in the previous era but with a lot more noise added. Here you see the development of methods for **multiple testing** and **regularized regression** to separate signals from piles of noise.
3. **The era of a few measurements on lots of samples** This era is overlapping to some extent with the previous one. Large scale collections of data from EMRs and Medicare are examples where you have a huge number of people (samples) but a relatively modest number of variables measured. Here there is a big focus on statistical methods for knowing how to model different parts of the data with hierarchical models and separating signals of varying strength with model calibration.
4. **The era of all the data on everything** This is an era that currently we as civilians don't get to participate in. But Facebook, Google, Amazon, the NSA and other organizations have thousands or millions of measurements on hundreds of millions of people. Other than just sheer computing I'm speculating that a lot of the problem is in segmentation (like in era 3) coupled with avoiding crazy overfitting (like in era 2).

References

- Efron and Hastie (2016) Computer-Age Statistical Inference: Algorithms, Evidence, and Data Science, Cambridge University Press