Predictive Modeling vs Generative Modeling – some reflections

Discussion Forum on Big Data

10th Cross Strait Conference in Statistics and Probability

W.K. Li
HKU
Chengdu, August 11, 2016
Leo Breiman, an early advocate of Data Science, described that there are two cultural outlooks. (See Donoho, 50 years of Data Science, 2015)

1. Prediction → to predict the responses in the future
2. Inference → to infer how nature is associated with the response

In Big data culture (1) is the dominating culture.
• George Santayana: “Those who fail to learn from history are doomed to repeat it.”

• In time series analysis it is easy to find examples that two independent time series match each other closely. For example, 2 independent random walks may appear to be correlated. (Granger and Newbold, 1973).

• “The larger the data set, the more stories; some can be useful or insightful, but many are neither, and even some of the best statisticians can be blind to the difference”

  ➢ Stigler S.M. (The Seven Pillars of Statistical Wisdom, Harvard U.P., 2016)
• Stigler gave the following example:
  - Jevons (1882) associated sunspot activity with commercial crisis.
  - Yule (as quoted by Stigler, 2016) in 1926, “Why do we sometimes get nonsense-correlations between time series?”

• Today we know an AR(2) process can fit the Sunspot series very well.

• The same kind of fallacy still exists today.
4.6 Jevons’s chart showing the relationship of sunspot activity and commercial crises, from 1882. (Jevons 1882)

(Stigler, Harvard U.P., 2016)
Sunspot and prices of commodities – a mere coincidence!
• A recent example is the failure of the Google Flu. Google (2009) claimed that “... we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day.”

• However Google Flu completely missed the 2009 swine flu pandemic and Lazer et al. (2014, Science) show that Google Flu was wrong 100 out of 108 weeks since August 2011. (See Salzburg, www.forbes.com, 2014).
• A new flu check has recently been proposed by Sam Kou and his team at Harvard. The improvement is made by a better understanding of people’s search behavior (Yang et al., PNAS, 2015).

• Challenge:
  ➢ How to strike a balance between the two cultures: (1) predictive modeling and (2) generative modeling?