Prediction and/or Understanding

Predictive Modeling & Statistical Learning

Gaston Sanchez

So Far ...

What's coming next

We've talked about Linear Regression—simple and multiple via OLS—from a traditional/classic perspective.

In order to continue introducing more contemporary (modernish) approaches, we need to discuss ideas like:

- modeling purposes
- measuring predictive accuracy
- bias-variance trade-off
- over-fitting
- learning and test sets
- resampling methods (cross-validation and bootstrapping)

Modeling ...
What for?

Statistical Modeling: Two Cultures

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Understanding vs Prediction

Models for Understanding versus Models for Prediction

Gilbert Saporta

Chaire de statistique appliquée & CEDRIC, CNAM
292 rue Saint Martin, Paris, France *saporta@cnam.fr*

Abstract. According to a standard point of view, statistical modelling consists in establishing a parsimonious representation of a random phenomenon, generally based upon the knowledge of an expert of the application field: the aim of a model is to provide a better understanding of data and of the underlying mechanism which have produced it. On the other hand, Data Mining and KDD deal with predictive modelling: models are merely algorithms and the quality of a model is assessed by its performance for predicting new observations. In this communication, we develop some general considerations about both aspects of modelling.

To Explain or to Predict?

Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330

To Explain or to Predict?

Galit Shmueli

Abstract. Statistical modeling is a powerful tool for developing and testing theories by way of causal explanation, prediction, and description. In many disciplines there is near-exclusive use of statistical modeling for causal explanation and the assumption that models with high explanatory power are inherently of high predictive power. Conflation between explanation and prediction is common, yet the distinction must be understood for progressing scientific knowledge. While this distinction has been recognized in the philosophy of science, the statistical literature lacks a thorough discussion of the many differences that arise in the process of modeling for an explanatory versus a predictive goal. The purpose of this article is to clarify the distinction between explanatory and predictive modeling, to discuss its sources, and to reveal the practical implications of the distinction to each step in the modeling process.

Modeling for what?

Goals

Understanding -vs- Prediction

Increasing acknowledge about the Prediction-vs-Understanding spectrum

Paradox 1

Typical conception of a model:

$$Y = f(X; \theta) + \varepsilon$$

Models for understanding

A "good" statistical model is one in which $f()$ is a parsimonious function that helps us explain how Y is related to X;

- we assume there is a "true" function $f()$ that we want to approximate
- we assume there is some stochastic mechanism that generates the data
- a model should be simple with interpretable parameters e.g. rate of change, elasticity, odds-ratio, etc.

Paradox 1

A "good" statistical model does not necessarily give accurate predictions.

For instance, in epidemiology it is more important to find risk factors than having an accurate individual prediction of getting some disease.

Models for Prediction

Prediction?

A "good" statistical model is one that provides accurate predictions: predict new observations with "good" accuracy without necessary providing an explanation about the data generation/variation mechanism.

- no need for a theory of consumer to predict marketing target
- a model may be just simply an algorithm

Paradox 2

We can predict without understanding (i.e. model is a black box).

- The aim is not to approximate the true function $f()$
- The aim is to find $f()$ that performs well by predicting new observations.

Model Performance

Model Performance

How do we define what a "good" model is?

- A model that fits the data well?
  (e.g. minimize resubstitution error)
- A model with optimal parameters?
  (e.g. most likely coefficients)
- A model that adequately predicts new (unseen) observations?
  (e.g. minimize generalization error)

In the Predictive Modeling arena ...

How do we measure prediction accuracy?

What do we mean by "prediction"?

Predictive accuracy

Assessing predictions

Observed y_i -vs- Predicted $\hat{y}_i$

What measure of accuracy can we use?

In the Predictive Modeling arena ...

Used (seen) observations

- Prediction $\hat{y}_i$ for y_i that was used to build the model
- *resubstitution* error: $e_i = y_i - \hat{y}_i$
- "already seen" MSE (less honest measure of accuracy)

Unseen observations

- Prediction $\hat{y}_0$ for y_0 that was NOT used to build the model
- *generalization* error: $e_0 = y_0 - \hat{y}_0$
- "unseen" MSE (more honest measure of accuracy)

In the Predictive Modeling arena ...

- A "good" model is one which gives accurate predictions.
- By *predictions* we mean predictions of new data.
- Therefore we focus on the generalization ability of the model to predict unobserved data
- This involves finding a measure of accuracy for predictions.

Modeling for Understanding

- We typically assume that there is "true" theoretical $f(X)$.
- Our interest is in knowing/estimating $f(X)$ as accurately as possible.
- Attempting recovering $f(X)$ requires that X be known; each and every predictor must be identified.
- Often, real world analysis rarely cooperates with what the theoretical statistical work requires.
- If the goal is to recover $f(X)$, we should preceed with the understanding that our results will typically be biased and inconsistent.

Modeling for Understanding

Keep in mind that no credible data scientist would ever claim that even when all of the necessary predictor are present and prefectly measured, there are one or more statistical learning procedures that will exactly capture the $f(X)$.

The data with which one works will necessarily be an imperfect reflection of the $f(X)$ because of the impact of ε.

References

- **Statistical Modeling: The Two Cultures** by Leo Breiman (2001). *Statistical Science, Vol 16 (3), 199-231*.
- **Models for Understanding versus Models for Prediction** by Gilbert Saporta (2008). COMPSTAT 2008. Physica-Verlag.
- **To Explain or to Predict?** by Galit Shmueli (2010). *Statistical Science, Vol 25 (3), 289-310*.