Advanced Machine Learning, Lecture 10 (week 11)

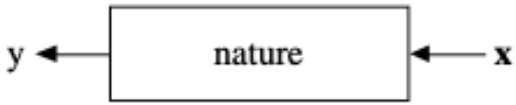
András Kornai

BME 2020 Nov 26

Modeling: the bird's eye view

1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

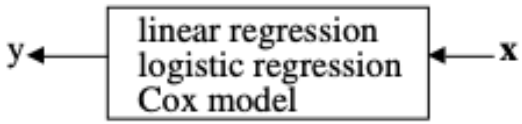
There are two different approaches toward these goals:

The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

response variables = f(predictor variables, random noise, parameters)

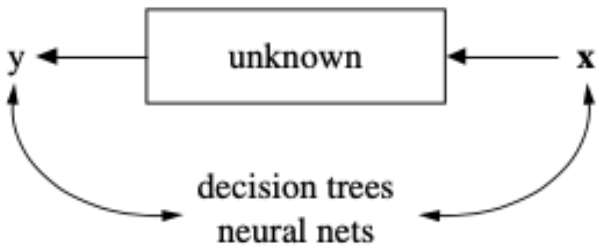
The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on **x** to predict the responses **y**. Their black box looks like this:



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

The "data modeling" and "algorithmic modeling" schools

- Data modelers start with a class of mathematical models: these are highly parametrized, and have only few parameters.
- How few? "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk" (Neumann János)
- They also tend to believe that these models actually reveal something about the internals of the black box
- Main problems: low-hanging fruit all gone, statistical tests become meaningless for millions of datapoints
- Biggest problem (Breiman): fit is no good!

Algorithmic modeling

- The key is prediction accuracy on unseen (future) data
- We don't care if we don't understand the model, black box is good enough
- Many parameters: millions are common, GPT4 has $1.5 \cdot 10^9$ numerical parameters
- Very loosely structured models, e.g. neural nets
- The approach benefits from theorems that show these are universal approximators
- Problems: even low-hanging fruit require very significant CPU reseources
- You may not care if you can't understand the model, but your sponsors will

Decision trees: where the tire meets the road

- Random forests (typically obtained by bagging/boosting) are good, but not interpretable
- Single trees (CART, C5.0) are more interpretable

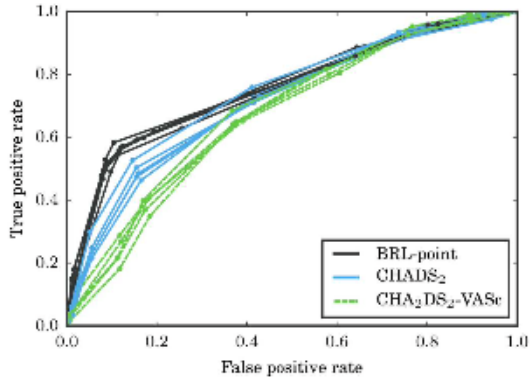


Fig. 4. *ROC curves for stroke prediction on the MDCD database for each of 5 folds of cross-validation, for the BRL point estimate, $CHADS_2$ and CHA_2DS_2-VASc.*

- But even trees may be too general
- Decision lists may be a good compromise

Decision lists

if hemiplegia **and** age $>$ 60 **then** *stroke risk* 58.9% (53.8%–63.8%)
else if cerebrovascular disorder **then** *stroke risk* 47.8% (44.8%–50.7%)
else if transient ischaemic attack **then** *stroke risk* 23.8% (19.5%–28.4%)
else if occlusion and stenosis of carotid artery without infarction **then** *stroke risk* 15.8% (12.2%–19.6%)
else if altered state of consciousness **and** age $>$ 60 **then** *stroke risk* 16.0% (12.2%–20.2%)
else if age $\leq$ 70 **then** *stroke risk* 4.6% (3.9%–5.4%)
else *stroke risk* 8.7% (7.9%–9.6%)

Model comparison

Table 2

Mean, and in parentheses standard deviation, of AUC and training time across 5 folds of cross-validation for stroke prediction. Note that the $CHADS_2$ and CHA_2DS_2-VASc models are fixed, so no training time is reported

	AUC	**Training time (mins)**
BRL-point	0.756 (0.007)	21.48 (6.78)
$CHADS_2$	0.721 (0.014)	no training
CHA_2DS_2-VASc	0.677 (0.007)	no training
CART	0.704 (0.010)	12.62 (0.09)
C5.0	0.704 (0.011)	2.56 (0.27)
ℓ_1 logistic regression	0.767 (0.010)	0.05 (0.00)
SVM	0.753 (0.014)	302.89 (8.28)
Random forests	0.774 (0.013)	698.56 (59.66)
BRL-post	0.775 (0.015)	21.48 (6.78)