"Statistical Modeling: The Two Cultures"

By Leo Breiman

Roy Makary

2023/2024

Contact info:
roy.moukari@gmail.com

- We will explore the paper "Statistical Modeling: The Two Cultures" by Leo Breiman.
- I will discuss the concept of two distinct cultures in statistical modeling and their implications.

Leo's Breiman Background

- Leo Breiman, a prominent statistician and machine learning pioneer, wrote the paper.
- He made significant contributions to statistics and machine learning.
- Breiman's work has had a lasting impact on modern statistical thinking.

The two cultures

- Breiman introduced the idea of "The Two Cultures" in statistical modeling.
- These two cultures represent different approaches to modeling and analysis.
- Let's delve into these cultures and their characteristics.

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.

→ Statistics starts with data. Data being “generated” by a black box in which a vector of **input** variables x (⫫) in one side, 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.

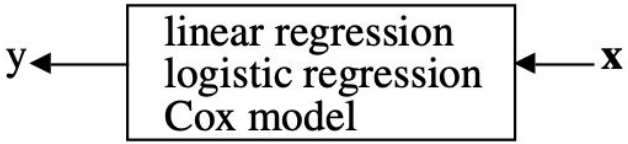
The term "nature" refers to the underlying process or phenomenon that generates the data being analyzed.

1)The data Modeling culture:

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. Model validation: yes–no using goodness-of-fit tests and residual examination.

- Culture 1, data/classical statistical modeling, emphasizes theory and assumptions.
- It relies on mathematical rigor and often employs parametric models.
- Classical statistics is rooted in hypothesis testing and estimation.

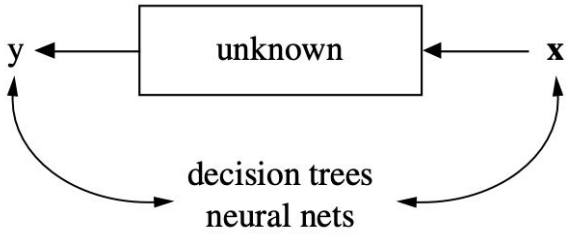
→These methods aim to infer the parameters of these distributions to make predictions or draw conclusions about the underlying process (nature).



2)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(x), an algorithm that operates on x to predict the responses y.

Model of validation: measured by predictive accuracy.



- Culture 2, algorithmic modeling, prioritizes prediction and data-driven methods.
- This culture utilizes machine learning techniques and non-parametric models.
- Algorithms like Random Forests and Gradient Boosting are key tools here.

→ What is the Goal of Leo Breiman in this paper?

He will argue that the focus in the statistical community on data models has:

- Led to irrelevant theory and questionable scientific conclusions,
- Kept statisticians from using more suitable algorithmic models,
- Prevented statisticians from working on exciting new problems.

From Breiman past experience:

→ Data modeling has given the statistics field many successes in analyzing data and getting information about the mechanisms producing the data. But there is also misuse leading to questionable conclusions about the underlying mechanism.

→ In the past fifteen years, the growth in algorithmic modeling applications and methodology has been rapid. It has occurred largely outside statistics in a new community often called machine learning that is mostly young computer scientists.

Back to Breiman's background and Road Map

Leo Breiman became a member of the small second culture (algorithmic modeling culture) after he decided to leave his job of academic probabilist to go freelance consulting. After 13 years of consulting he joined the Berkeley Statistics Department in 1980 and has been there since.

PROJECTS IN CONSULTING

He worked on a diverse set of prediction projects. Here are some examples:

- Predicting next-day ozone levels,
- Toxicity Of Chemicals,
- Predicting the class of a ship from high altitude radar returns,
- Speech Recognition,
- Etc…

→ His Perceptions on Statistical Analysis

As he came back to the university, these were his perceptions on working with data to answer problems:

1. Live with the data before you plunge into modeling,
2. Search for a model that gives good solution, either algorithmic or data,
3. Predictive accuracy on test sets is the criterion for how good the model is,
4. Computer are an indispensable partner.

THE USE OF DATA MODELS

→ Problems in Current Data Modeling

The question of how well the model fits the data is of secondary importance compared to the construction of an ingenious stochastic model.

→ Multiplicity of data models

-One reason for this multiplicity is that goodness-of-fit tests and other methods for checking fit give a yes–no answer.

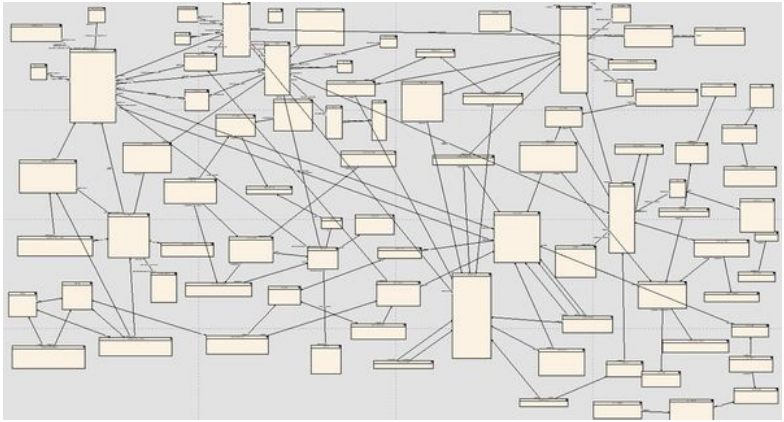
-There Is no way, among the yes–no methods for gauging fit, of determining which is the better model.

→ Predictive Accuracy

The most obvious way to how well the model box emulates nature's box is this: put a case x down nature's box getting an output y. Similarly, put the same case x down the model box getting an output y′. The **closeness** of y and y′ is a measure of how good the emulation is.

Prediction is rarely perfect.There are usually many unmeasured variables whose effect is referred to as "noise."

THE LIMITATIONS OF DATA MODELS



→Usually, simple parametric models imposed on data generated by complex systems, for example, medical data, financial data, result in a loss of accuracy and information as compared to algorithmic models.

→As data becomes more complex,the data models become more cumbersome and are losing the advantage of presenting a simple and clear picture of nature's mechanism.

→To solve a wide range of data problems, a larger set of tools is needed.

ALGORITHMIC MODELING

→Under other names, algorithmic modeling has been used by industrial statisticians for decades.

→The list of statisticians using algorithmic modeling is very short.

→The development of algorithmic methods was taken up by a community outside statistics.

→ **New research community**

In the mid-1980s two powerful new algorithms for fitting data became available: **neural nets** and **decision trees**.

-New research community was born, their goal was predictive accuracy:

-The community consisted of young computer scientists, physicists and engineers plus a few aging statisticians.

-They used the new tools in complex prediction problems where it was obvious that data models couldn't be applicable: image and speech recognition, nonlinear time series predictions etc…

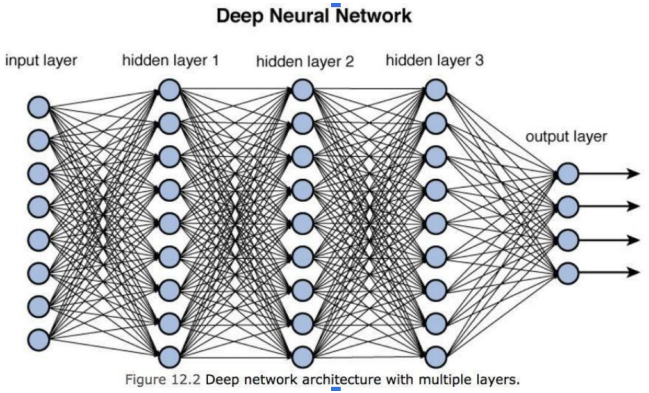


Figure 12.2 Deep network architecture with multiple layers.

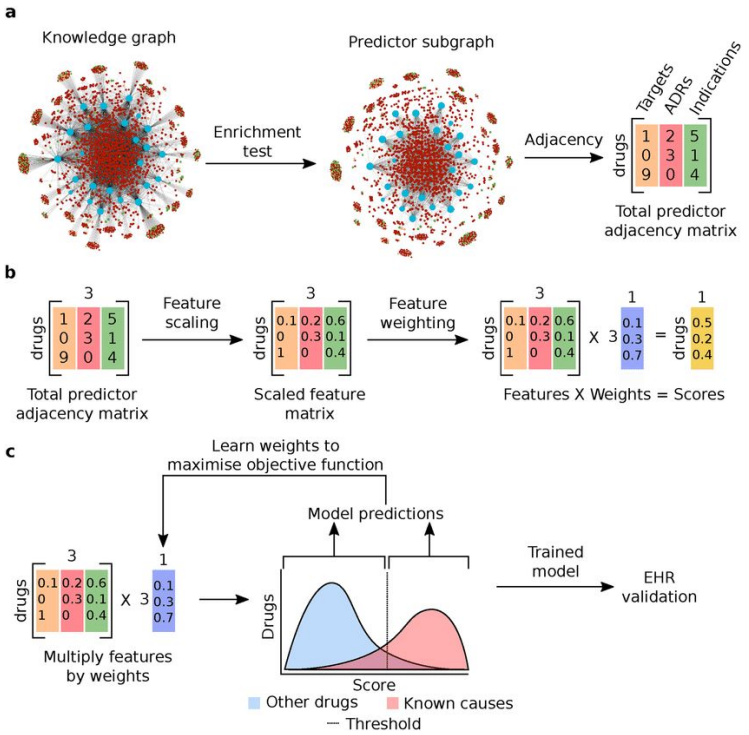
→ **Theory in algorithmic prediction**

-Data models are rarely used in this community. The approach is that nature produces data in a black box whose insides are complex, mysterious, and, at least, partly unknowable.

-What is observed is a set of x's that go in and a subsequent set of y's that come out. The problem is to find an algorithm f(x) such that for future x in a test set, f(x) will be a good predictor of y.

-The theory in this field shifts focus from data models to the properties of algorithms.

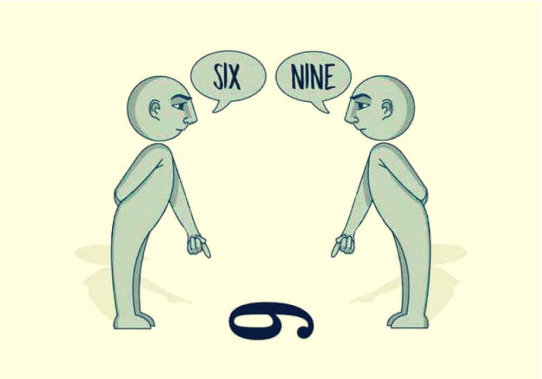
-It characterizes their "strength" as predictors, convergence if they are iterative, and what gives them good predictive accuracy. The **one assumption made in the theory** is that the data is **drawn i.i.d.** from an unknown multivariate distribution.

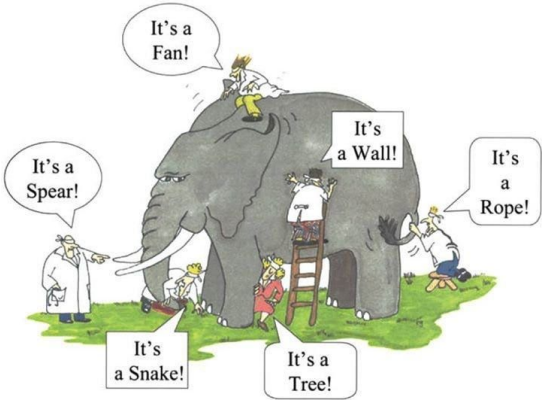


→ **Recent lessons**

-The advances in methodology and increases in predictive accuracy in the mid-1980s that have occurred in the research of machine learning has been phenomenal.
-What has been learned? → Three lessons that seems the most important:

- **Rashomon**: the multiplicity of good models;
- **Occam**: the conflict between simplicity and accuracy;
- **Bellman**: dimensionality, curse or blessing





RASHOMON AND THE MULTIPLICITY OF GOOD MODELS

→Rashomon is a japanese movie in which someone dies and someone is raped and 4 people witnessed the scene from different angles.

→When they come to testify in court,they all report the same facts, but their stories of what happened are very different.

→What Breiman calls the **Rashomon effect** is that there are often a **multitude of different descriptions** [i.e. equations of f(x)] in a class of functions giving about the same minimum error rate.

Picture 1

$$y = 2.1 + 3.8x_3 - 0.6x_8 + 83.2x_{12} - 2.1x_{17} + 3.2x_{27},$$

Picture 2

$$y = -8.9 + 4.6x_5 + 0.01x_6 + 12.0x_{15} + 17.5x_{21} + 0.2x_{22},$$

Picture 3

$$y = -76.7 + 9.3x_2 + 22.0x_7 - 13.2x_8 + 3.4x_{11} + 7.2x_{28}.$$

This is an example of three equations that are at 1% test error of each other.

Which one is better?

The problem is that each one tells a different story about which variables are important.

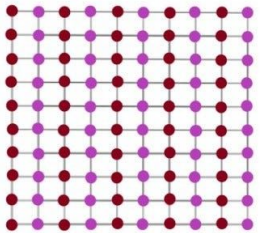
RASHOMON AND THE MULTIPLICITY OF GOOD MODELS

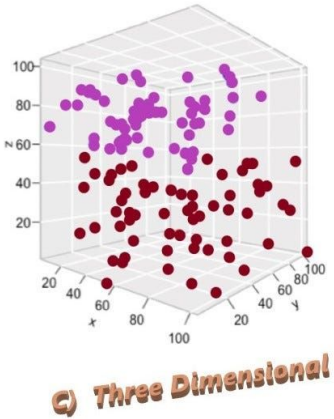
→This effect occurs with decision trees and neural nets. Say by removing a random 2–3%of the data, we can get a tree quite different from the original but with almost the same test set error!

→This effect is closely connected to what he calls **instability**.

→For example in logistic regression and cox model, the common practice of deleting the less important covariates is carried out, then the model becomes unstable.

→The multiplicity problem and its effect on conclusions drawn from models needs serious attention.



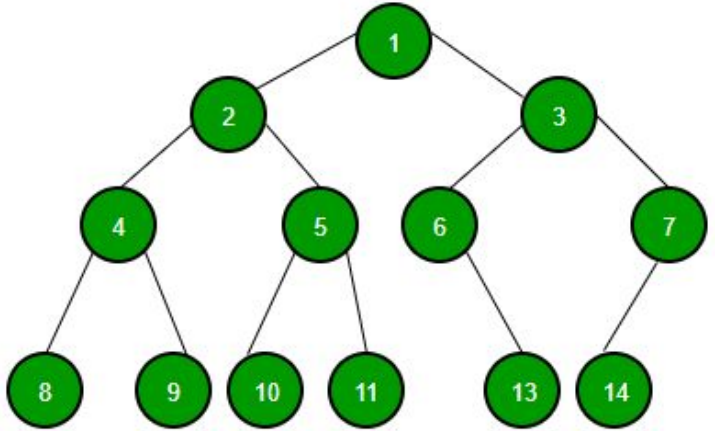


OCCAM AND SIMPLICITY VS ACCURACY

→Occam's razor is a principle says that if you have two competing ideas to explain the same phenomenon, you should prefer the simpler one. Unfortunately, in prediction, simplicity and accuracy are often in conflict!

→For instance, he concluded that trees are very good for **interpretability** (the extent to which a cause and effect can be observed within a system).

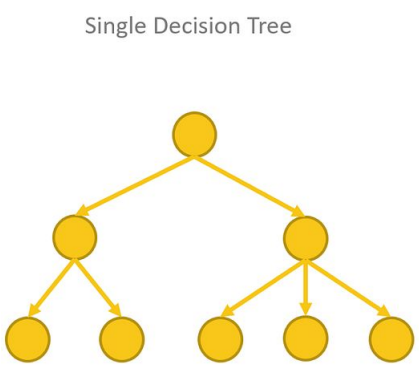
→Trees are A+ for interpretability but B for predictions.

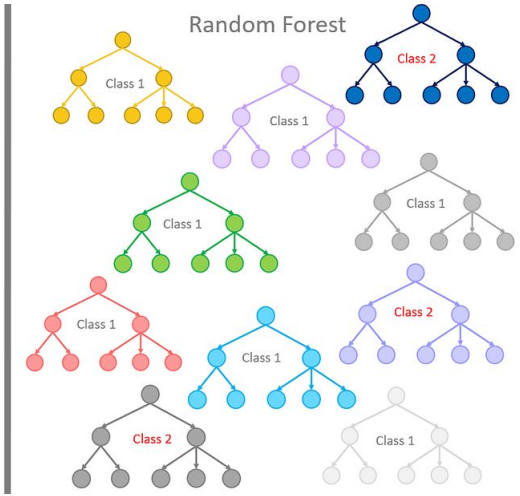


Growing forests for predictions

→In this approach successive decision trees are grown by introducing a random element into their construction.

→The random forest splits the nodes by selecting features randomly. The final prediction will be selected based on the outcome of the obtained trees. The outcome chosen by most decision trees will be the final choice.





Forests compared to trees:

-Some errors are halved.

-Others are reduced by one-third.

TABLE 2
Test set misclassification error (%)

Data set	Forest	Single tree
Breast cancer	2.9	5.9
Ionosphere	5.5	11.2
Diabetes	24.2	25.3
Glass	22.0	30.4
Soybean	5.7	8.6
Letters	3.4	12.4
Satellite	8.6	14.8
Shuttle $\times 10^3$	7.0	62.0
DNA	3.9	6.2
Digit	6.2	17.1

→ **Forests are A+ predictors**

In terms of rank of accuracy, the forest comes 1st in 4 data sets studied in the paper. Forests are therefore the best **classifier**.

→ **The Occam Dilemma**

Forests are A+ predictors but their mechanism for producing a prediction is difficult to understand.

To conclude: accuracy generally requires more complex prediction methods. Simple and interpretable functions do not make the most accurate predictors.

BELLMAN AND THE DIMENSIONALITY CURSE OR BLESSING?

The first step in prediction methodology was to avoid the curse that is when we have too many prediction variables, the recipe was to find a few features (functions of the predictor variables) that "contain most of the information" and then use these features to replace the original variables.

For example, it is common in procedures such as regression, logistic regression, survival models to delete variables to reduce dimensionality.

Digging It Out in Small Pieces

→Reducing dimensionality reduces the amount of information available for prediction. The more predictor variables, the more information.

Let's try going in the opposite direction…

→Instead of reducing dimensionality, increase it by adding many functions of the predictor variables. There may now be thousands of features each potentially containing a small amount of information.

The problem is how to extract and put together these little pieces of information !

Example: The Shape Recognition Forest

→Competition set up on by the NIST for ML to read hand written numerals

→They put together a large set of pixel pictures of handwritten numbers(223,000) written by over 2,000 individuals.

→The competition attracted wide interest, and diverse approaches were tried.

→The Amit–Geman approach defined many thousands of small geometric features in a hierarchical assembly. Shallow trees are grown, such that at each node, 100 features are chosen at random from the appropriate level of the hierarchy; and the optimal split of the node based on the selected features is found.

→Using a 100,000 example training set and a 50,000 test set, the Amit–Geman method gives a test set error of 0.7% close to the limits of human error!!

Conclusion of Leo Breiman

→Breiman argues for a balance between these two approaches, suggesting that both perspectives are valuable, but they serve different purposes.

→Understanding the underlying nature of the data is important

→But so is the ability to make accurate predictions, especially in complex real-world situations.

→Integrating these perspectives can lead to more robust and useful statistical models.


Any
Question