A COMPARISON BETWEEN TRADITIONAL AND DATA-DRIVEN STATISTICAL PREDICTIVE TECHNIQUES.

[1]Dr. Dhane Neeta K, [2]Dr. Avinash Jagtap, [3]Dr. Jaya Limbore
[1,]Assistant Professor, [2]Associate Professor, [3]Assistent Professor
Department of Statistics
Tuljaram Chaturchand College of Arts, Science and Commerce, Baramati
District-Pune, Maharashtra, India

Abstract: Statistical techniques are used for various purposes, beginning with descriptive and inferential statistics, ending with predictive models that help us to predict possible consequences of our decisions. The paper aims to make a comparison between traditional and data driven statistical predictive techniques. The comparison includes similarities as well as differences between the two. While these two differ from each other technically as well as procedurally, the basic reason for the confusion is a lack of clarity in many textbooks regarding traditional and data-driven statistical predictive techniques .

Keywords: statistical predictive techniques, data-driven predictive models

INTRODUCTION

Prediction attempts to determine the value of a variable as prediction uses sample values on several (predictor) variables and makes a statement about another (response) variable. Prediction attempts to determine unobserved values of the variable of interest also called as the response variable when values of some strongly related variables, known as predictor variables, are available. Traditional statistical predictive techniques first establish a relationship between the predictor variables and the response variables. This relationship is then used for making predictions on the response variable when predictor variables are observed for some individual sampling units in the population.

TRADITIONAL STATISTICAL PREDICTION TECHNIQUES

In traditional statistical prediction techniques the statisticians must understand how the data was collected, statistical properties of the estimator, the underlying distribution of the population they are studying and the kinds of properties they would expect if they did the experiment many times. They need to know precisely what they are doing and come up with parameters that will provide the predictive power. The traditional statistical predictive modeling techniques are usually applied to low dimensional data sets. It is all about sample, population, hypothesis, etc. Traditional statistical predictive techniques are very useful in linear, repeatable, scientific analysis, as science deals with the environment and have very stable relationships. In Linear regression if the boundary between categories is linear and the predictive relationships are linear, the regression can perform just as well. When the problem is less well-behaved, a more flexible model may be necessary. Traditional predictive models are controlled by the assumptions about the underlying probability distributions and optimization techniques based on the specified loss function, like the square error loss function leading to the ordinary least squares (OLS) method. Gaussian distribution is central to most of these methods, justified by the central limit theorem or the laws of large numbers. Homogeneity is essential for validity of these methods given data is also part of veracity. In traditional predictive models the learning spectrum is called analytical learning, (deductive learning), where data is often scarce or it is preferred (or customary) to work with small samples of it. There is also good prior knowledge about the problem and data. Traditional statistical predictive models are conservative in its approaches and techniques and often make tight assumptions about the problem, especially data distributions.

DATA-DRIVEN STATISTICAL PREDICTION TECHNIQUES

Its main focus is the study and design of systems that can "learn from data" and its focus is inductive learning (learning by examples). Data-driven statistical prediction techniques require no prior assumptions about the underlying relationships between the variables. The algorithm processes the data and discovers patterns, using which you can make predictions on the new data set. It is generally applied to high dimensional data sets, the more data you have, the more accurate your prediction is. Data-driven statistical prediction techniques and approach heavily relies on computing power. Cheap computing power and availability of large amounts of data allowed data scientists to train computers to learn by analyzing data. The preferred learning method in data-driven predictive models is inductive learning. At its extreme, in inductive learning the data is plentiful or abundant, and often not much prior knowledge exists or is needed about the problem and data distributions for learning to succeed. Data-driven models often have no theoretical or mathematical model.

POINTS OF COMPARISON BETWEEN TWO

There are different points of comparison between the traditional statistical prediction techniques and data-driven statistical prediction techniques, which clearly indicates the similarities and the differences between the two.

They belong to different fields.

Traditional statistics is a subfield of mathematics which deals with finding relationships between variables to predict an outcome. Data-driven predictive techniques are a part of computer science and artificial intelligence which deals with building systems that can learn from data, instead of explicitly programmed instructions.

They came up in different eras

Traditional statistical predictive techniques have been there for centuries now. However, Data-driven predictive techniques are a very recent development. The steady advances in digitization and cheap computing power enabled data scientists to stop building finished models and instead train computers to do so. The unmanageable volume and complexity of the big data that the world is now swimming in have increased the potential of data-driven predictive models—and the need for it.

Extent of assumptions involved

Traditional statistical predictive techniques work on number of assumptions. For instance a linear regression assumes:

1. Linear relation between independent and dependent variable
2. Homoscedasticity
3. Mean of error at zero for every dependent value
4. Independence of observations
5. Error should be normally distributed for each value of dependent variable

Similarly Logistic regressions come with its own set of assumptions. Even a non linear model has to comply to a continuous segregation boundary. Data-driven predictive techniques do not based on any assumption like distribution of dependent or independent variable, independence of observations etc.

Types of data they deal with

Traditional statistical predictive techniques primarily were concerned with quantitative variables. On the other hand the data-driven predictive models are having the algorithms to run decision trees, support vector machines etc which work well on categorical data.

Size of data they deal with

Data-driven predictive algorithms are wide range tools. These models are capable of learning from trillions of observations one by one. They make prediction and learn simultaneously. The algorithms like Random Forest and Gradient Boosting are also exceptionally fast with big data. Data-driven predictive models do really well with wide (high number of attributes) and deep (high number of observations). However traditional statistical predictive modeling is generally applied for smaller data with less variables.

Approach to model building

Traditional statistical predictive models almost always assumes there is one underlying "data generating model", and good practice requires that the analyst build a model using inputs that have a logical basis for being somehow related to the independent variable. In contrast, data-driven predictive models requires essentially no prior beliefs about the nature of the true underlying relationships, and doesn't even necessary expect that there is just one best model waiting to be discovered.

Formulation

Even when the end goal for both traditional and data-driven predictive models is same, the formulation of two is significantly different.

In a traditional statistical predictive model, we basically try to estimate the function f as

Dependent Variable (Y) = f(Independent Variable) + error function

Whereas data-driven predictive models take away the deterministic function "f" from the above equation and it simply becomes

Input(Y) ----- > Output (X)

It will try to find pockets of X in n dimensions (where n is the number of attributes), where occurrence of Y is significantly different.

Predictive Power and Human Effort

Nature does not assume anything before forcing an event to occur. So the lesser assumptions in a predictive model, higher will be the predictive power. Data-driven predictive models need minimal human effort. It works on iterations where computer tries to find out patterns hidden in data. Because machine does this work on comprehensive data and is independent of all the assumption, predictive power is generally very strong for these models. Traditional statistical predictive techniques are mathematics intensive and based on coefficient estimation. It requires the relation between variables before putting it in.

CONCLUSION

The accuracy of prediction depends not only on the models and methods you used, but also on many other factors, such as how well you understand the causal relationship of the problem and what kind of data you used for your prediction.

The only difference lies in the volume of data involved and human involvement for building a model.

Historically, the traditional statistical techniques were mostly developed in the years where computing power was not an option. As a result, they heavily relies on small samples and heavy assumptions about data and its distributions.

Data-driven predictive models do not based on assumptions about the data and are liberal in approaches and techniques to find a solution, many times using heuristics.

The data-driven predictive models are more appropriate and suitable than the traditional predictive models in situations where data collection is not through a sampling design, but is automated by a process and hence the assumptions required for classical or traditional models cannot be taken for granted.

The traditional models address the questions about models and model parameters rather than questions about the specific data in hand.

Traditional statistical predictive models begin with a model for probability distributions of variables while data-driven predictive models begin with modeling the relationship between predictors and the response variables without any considerations for their probability distributions.

REFERENCES

[1] Ben Krose and Patrick van der Smagt (1996). An Introduction to Neural Networks. The University of Amsterdam.
[2] Raul Rojas (1996). Neural Networks: A Systematic Introduction. Springer.
[3] Abe, N., Pednault, E.,Wang, H., Zadrozny, B., Fan,W., and Apte, C. (2002). Empirical comparison of various reinforcement learning strategies for sequential targeted marketing. In Proceedings of the IEEE International Conference on Data Mining, pages 3–10. IEEE.
[4] John Neter, William Wasserman, and Michael H. Kutner (1989) Applied Linear Regression Models. Richard D. Irwin Inc.
[5] K. Nandhini and S. Saranya (2012). Comparison of Decision Tree and ANN Techniques for Data Classification. International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 5, pp. 323-327.
[6] John O. Rawlings, Sastry G. Pantula, and David A. Dickey (1998). Applied Regression Analysis: A Research Tool. Second Edition. Springer.
[7] Matthew Aryanwu and Sajjan G. Shiva (2009), Comparative Analysis of Serial Decision Tree Classification Algorithms. International Journal of Computer Science and Security. Vol. 3, No. 3, pp. 230-240.
[8] William J. Long, John L. Griffith, Harry P. Selker, and Ralph B. D'Agostino (1993). A Comparison of Logistic Regression to Decision-Tree Induction in a Medical Domain. Computers in Biomedical Research, Vol. 26, pp. 74-97.
[9] https://www.datasciencecentral.com/profiles/blogs/machine-learning-vs-traditional-statistics-different-philosophi-1
[10] https://www.researchgate.net/post/Traditional_predictive_models_in_statistics_vs_fancy_named_computational_methods
[11] https://www.edvancer.in/machine-learning-vs-statistics/
[12] https://www.analyticsvidhya.com/blog/2015/07/difference-machine-learning-statistical-modeling/