THE LINK FROM STATISTICS TO DATA SCIENCE

Targeted learning

STAtOR
Mark van der Laan

The foundations of statistical learning from data are based on the following important concepts: data generating probability distribution, statistical model, target estimand, estimator, and sampling distribution of estimator. Firstly, a statistician acknowledges that data can only be interpreted if one understands the experiment that generated the data. The data are viewed as a realization of a random variable with a certain probability distribution, which is often referred to as the data generating distribution.

Secondly, we need to find out as much as possible about this data generating distribution to restrict the set of possible probability distributions that might have generated the data. This latter set is called the

statistical model for the data distribution, which represents our statistical knowledge about the data generating experiment.

Thirdly, the statistician has to specify a so called target parameter mapping from this model to the parameter space (e.g., real line), which specifies for a particular data distribution the feature we aim to learn from the data. That is, one will talk to the scientific collaborator to determine the estimand that represents the best approximation to the scientific question of interest, where the estimand is defined by the target parameter mapping applied to the true data distribution. Determining the statistical model and target estimand thus requires strong interaction with the scientists and data collectors, so that a statistician is naturally part of the scientific team. The statistical estimation problem is now defined.

Fourthly, one has to determine an estimator that maps the data set into an estimate of the target estimand. For example, one might use maximum likelihood estimation to estimate the data probability density and subsequently plug this density estimate in the target parameter mapping to obtain the desired estimate of the target estimand. Finally, we recognize that the estimator is itself a random variable and therefore has a probability distribution, called its sampling distribution. The spread of this sampling distribution represents the uncertainty of the estimator, and an estimator of this sampling distribution can now be used to construct a confidence interval centered at the estimator that will contain with high probability the true estimand.

THE EROSION OF THE NOTION 'STATISTICAL MODEL'

This beautiful foundation of statistical learning appears to have been lost in most of statistical practice. Obviously, the choice of model is crucial since it is supposed to contain the true data distribution. Misspecification of the model, i.e. making assumptions about your data generating distribution that are wrong, will guarantee that the target estimand does not represent the scientific question of interest, and thereby that the resulting confidence interval will most likely not contain the answer to the question of interest.

Typically, we have no knowledge about functional stochastic relations between the different variables we observe. For example, the probability of death will not be a logistic linear function of the baseline measurements on

the subject. Nonetheless, almost all statistical methods are based on regression models such as logistic linear regression, linear regression, Cox-proportional hazard regression, and so on, that precisely assume highly simplistic linear relations between outcomes of interest and covariates.

These models are guaranteed to be wrong. An in-depth philosophical and historical perspective on the erosion of 'truth' in statistics, and its dire consequences for the field and science in general, is provided in (Starmans, 2011).

THE DRAMATIC IMPLICATIONS FOR THE PRACTICE OF STATISTICS: THE WILD-WEST OF STATISTICS

Suppose a scientist consults two different statisticians with a data set, a specification of the question of interest, and description of the data generating experiment.

For example, one might observe on a sample of patients baseline covariates, a binary treatment, and an indicator of heart disease a year after treatment was initiated, as part of an observational study aiming to assess the effect of the treatment on heart disease. Most likely, these two statisticians will only care about knowing the format of the data and will quickly decide that this is a logistic regression problem. Each one will specify a particular form for the logistic linear regression. Of course, there is no reason they would select the same model, so most likely both select quite different linear logistic regression models (may be one includes certain interactions, while the other did not).

They will probably report the coefficients of their logistic model fit with p-values and confidence intervals. Presumably, the focus will be on the coefficient in front of treatment, even though that gets quite complicated when the model would include interactions of treatment and covariates, which is a reason for most practitioners to not include interactions with treatment (and thereby force additional misspecification!). Clearly, for large enough sample size, the answers reported will not only be different but even statistically significantly different: the different choice of model will imply a different 'true' coefficient in front of treatment, so that the two

confidence intervals corresponding with the two statisticians will not overlap each other for large enough sample size. That is, our scientist report random output just depending on what statistician is consulted.

In fact, real practice will typically involve fine tuning the model choice based on interactions with the collaborators till all are satisfied with the reported results and at that point one uses the output in terms of p-values and confidence intervals even though these assumed that the final data and human-adaptively selected model were a priori specified.

As a result, there is no well-defined estimator and thereby a scientifically sound way to determine the uncertainty in the estimator: in particular, one cannot run a bootstrap that reproduces the estimator applied to a random sample from the actual data set, due to these human interventions.

Another painful by-product of this approach to statistics is that the choice of regression model also implies the choice of target estimand, so that also the estimand is selected without regard to what the actual question of interest is. Therefore, this approach minimizes scientific communication between the statistician and the scientific collaborator. In reality, the scientific collaborator might have taught us that he/ she wants to know what the difference in probability of heart disease would have been in a treatment and control arm if one would run a randomized trial. Clearly, these coefficients in the misspecified conditional logistic regression models are not even close in answering this question.

One needs to conclude that statistics has become an art involving human intervention with all its natural biases instead of a science, and it represents an approach that is destined to generate an epidemic of false positives and false claims.

GOING BACK TO OUR FOUNDATIONS OF THE SCIENTIFIC APPROACH

Let's revisit the above example, but now respecting the foundations of statistical learning. Firstly, we would want to know how the data were generated. We might learn that it is fair to assume that the data on each patient were independently generated and that the data can be represented as a repetition of n independent and identical experiments. We would also want to know how the medical doctor made its treatment decision. We

might learn that the medical doctor only takes into account two biomarkers and the age of the patient. Therefore, we might feel comfortable concluding that treatment is conditionally independent of all other baseline measurements, given these three variables. One might conclude that there is no meaningful knowledge about the probability distribution of the covariates and the conditional probability on heart disease as a function of treatment and the covariates. Or may be, we will learn that the outcome is a rare event and that it is perfectly reasonable to assume that the conditional probability on heart disease is always smaller than 0.03.

We can then commit to a statistical model for the probability distribution of the data on a randomly selected patient defined by only assuming the above conditional independence assumption on the conditional distribution of treatment, given the covariates, and (possibly) the bound on the conditional probability of heart disease. Note that this model is not specifying a functional form for the relation between the variables.

Secondly, we would talk to our collaborator, provide causal language using the notion of hypothetical experiments and potential outcomes or nonparametric structural equation models, and determine that our collaborator wants to know the average causal effect defined by the difference in probabilities on heart disease of the two treatment arms in a randomized controlled trial with infinite sample size.

We would then conclude that a best approximation of the answer to this scientific question of interest is the expectation w.r.t. population distribution of covariates of the difference of the conditional probability of heart disease under treatment and control and covariate vector. The latter now represents the estimand of interest, which also defines a mapping from the model to the real line. We can show that this target estimand precisely equals the desired average causal effect if indeed the treatment decision of the doctor was only based on the three variables mentioned and independent noise. We have now defined a realistic statistical model for the data distribution and a target estimand. The statistical estimation problem is defined.

The remaining challenge is now to construct an estimator of the target estimand and estimate its sampling distribution so that we can also construct a confidence interval. Keep in mind though, that after having been so careful in defining the estimation problem, we cannot just return to using

parametric regression models to fit this estimand, but, instead, we need to find a way to learn this estimand whatever the true data distribution might be in our statistical model.

SUPER-LEARNING TO LEARN THE UNKNOWN FUNCTIONAL STOCHASTIC RELATIONS

The only key stochastic relation we will have to learn from the data is the probability of heart disease as a function of treatment and covariates. The expectation over the covariate distribution can be estimated with an empirical average across the patients in our sample.

Since we do not know the functional form of the probability of heart disease as function of treatment and covariates, we do not want to bet on one particular logistic regression estimator. Instead we build an extensive library of candidate logistic regression estimators. This library can include a large number of linear logistic regression models that vary in the choice of main terms and interactions and possibly functional transformations of various baseline covariates of interest. However, we should also include machine learning algorithms that aim to flexibly learn the functional form. A large variety of such machine learning algorithms have been developed in the machine learning and data science community, such as Lasso regression, random forest, polynomial regression, wavelet regression, neural networks and deep neural networks, to name a few (Hastie et al., 2001).

Each of these algorithms depends on the choice of various tuning parameters, so that one might incude many versions of it, or create a tuned version of the algorithm that internally searches the best tuning parameter. Recently, we proposed a new machine learning algorithm named Highly Adaptive Lasso (HAL) (van der Laan, 2015; Benkeser & van der Laan, 2016). We would add this algorithm as well since it has been shown to consistently estimate the true functional relation at a rate faster than $n^{-1/4}$ as a function of sample size n, whenever the true target function has finite variation norm. The latter assumption can be included in our statistical model without any concern for misspecification, since it is unlikely that the true functional relations have infinite variation as a function of the different covariates.

Instead of betting on one of these algorithms, we carry out a competition by training each of these algorithms on 9/10 of the data, and evaluating its performance in predicting heart disease status of the patient on the 1/10 left-out patients, and apply this competition to a number of random splits of the data in training and validation sample. We evaluate the performance of each algorithm by its average prediction error across the different sample splits, which is called the cross-validated risk of the estimator (e.g. log likelihood risk or squared-error risk). We now select the best algorithm and rerun that algorithm on the complete data set.

This defines a new machine learning algorithm, a so called ensemble algorithm, that combines all these algorithms in the library into a new more powerful and adaptive algorithm. We have named this ensemble algorithm a super-learner, due to its theoretical property that it is asymptoticallly as good as the oracle selector that chooses the best algorithm when given an infinite validation data set, even when the number of candidate algorithm grows polynomial in sample size (van der Laan and Dudoit, 2003; van der Laan et al., 2007; van der Laan and Rose, 2011; van der Vaart et al., 2006). Without much additional effort, one can also compute the cross-validated risk of each weighted convex combination of the candidate algorithms and determine the optimal weighted combination, so that one ends up selecting some weighted combination of the different algorithms in the library.

TARGETED LEARNING TO TUNE THE FIT OF THE DATA DISTRIBUTION TOWARDS THE TARGET ESTIMAND, AND OBTAIN AN APPROXIMATE NORMAL SAMPLING DISTRIBUTION

The super-learning fit of the conditional probability of heart disease is optimized to fit this whole function, and as a result spreads its approximation error across all covariate profile configurations of the population of patients. However, we only care about fitting a specific function/summary measure of this conditional probability function, namely the target estimand. Therefore, the spread of error is not optimized for the target estimand. Another more statistical way of saying this is that the super-learner optimally trades off bias and variance w.r.t. the whole function, while we need to optimally trade off bias and variance of the corresponding estimator of the target estimand (just a real number). The plug-in estimator of the target estimand based on the super-learner will

have variance $1/n$, while its bias will be of the same order as the bias of the superlearner, and thus be larger than $n^{-1/2}$ (e.g. around $n^{-1/4}$). As a consequence, the plug-in estimator based on the super-learner will not even be asymptotically normally distributed.

This is resolved with the so called targeted maximum likelihood estimator (TMLE) which creates a fluctuation of the super-learner fit where the amount of fluctuation is a coefficient we will name ϵ. At zero fluctuation $\epsilon = 0$, this fluctuation returns the super-learner fit, and if we move ϵ away from zero it will fluctuate the super-learner fit in such a way that it maximizes the square change of the target estimand per increase in log-likelihood, at least locally around $\epsilon = 0$. Such a fluctuation strategy is solved by a mathematical optimization problem, and it is referred to as the least favorable parametric model in the efficiency theory literature (Bickel et al., 1997). We now fit ϵ with standard maximum likelihood estimation for this one-dimensional parametric model, resulting in an updated fit of the super-learner. This fitting step is using all the data to purely fit the target estimand and, as a consequence, it removes bias and optimizes the mean squared error for the target estimand. One now plugs this targeted super-learner fit into the estimand representation to obtain the desired TMLE of the target estimand.

This TMLE can now be shown to be behave as an empirical mean of an efficient influence curve transformation of the unit data structure, and is thus approximately normally distributed, and has minimal asymptotic variance (van der Laan, 2015). As a consequence, a variance estimator (e.g., a bootstrap or Wald-type confidence interval) provides an approximate 95% confidence interval for the target estimand.

Above we demonstrated a TMLE for the sake of estimation of the ATE. TMLE is a general method for construction of a two stage estimator of any target estimand for any statistical model, where the first stage uses super-learning to obtain an initial estimator of the data distribution, while the second stage consists of carrying out a TMLE-update step based on a least favorable parametric submodel (van der Laan and Rose, 2011).

DATA SCIENCE AND TARGETED LEARNING

Data science is a flourishing field with growing amount of resources. The machine learning world has played a fundamental role in the definition and growth of data science. Typically, the proposed approaches lack statistical formulation, the recognition of an underlying experiment, the careful definition of a target estimand answering the question of interest and construction of a corresponding theoretically grounded estimator, and above all it lacks assessment of statistical uncertainty. It often appears that statistical thinking is under-appreciated and is not receiving the place it deserves. However, simultaneously, there is a clear and growing recognition that one should not only be concerned with prediction but also assessment of (e.g., causal) effects of interventions on outcomes of interest with proper statistical inference. Due to the massive amounts of data one is often confronted with, standard parametric regression models are not even applicable, so that this community realizes that one will have to use data adaptive estimation and machine learning to make progress.

Therefore, targeted learning, a pure statistical approach that incorporates the state of the art in machine learning targeting the estimand of interest, while preserving formal statistical inference, brings the statistical foundations and the enormous advances that have been made in our field (e.g., causal inference, empirical process theory, efficiency theory, missing data, censored data, biased sampling, etc) to the forefront in data science. In this manner, statistics deserves its place and can flourish as well as data science as a whole, while moving science forward in the process.

REFERENCES

Benkeser, D, & van der Laan, M.J. (2016). *The highly adaptive lasso estimator. Proceedings of the IEEE Conference on Data Science and Advanced Analytics*. To appear.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y., & Wellner, J. (1997). *Efficient and adaptive estimation for semiparametric models*. Berlin Heidelberg New York: Springer.

Hastie, T.J., Tibshirani, R.J., & Friedman, J.H. (2001). T*he elements of statistical learning: data mining, inference, and prediction*. Berlin Heidelberg New York: Springer.

Starmans. R.J.C.M. (2011). Models, inference and truth: Probabilistic reasoning in the information era. In M.J. van der Laan & S. Rose (Eds.). *Targeted Learning: Causal Inference for Observational and Experimental Studies*, pp. 1 – 20. New York: Springer.

Van der Laan, M.J. (2015). *A generally efficient targeted minimum loss-based estimator.* Technical Report 300, UC Berkeley, 2015. http://biostats.bepress.com/ucbbiostat/ paper343. To appear in International Journal of Biostatistics.

Van der Laan, M.J., & Dudoit, S. (2003). *Unified crossvalidation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples.* Technical Report 130, Division of Biostatistics. Berkeley: University of California.

Van der Laan, M.J., & Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data.* Berlin Heidelberg New York: Springer.

Van der Laan, M.J., Polley, E.C., & Hubbard, A.E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology,* 6(1).

Van der Vaart, A.W., Dudoit, S., & van der Laan, M.J. (2006). Oracle inequalities for multi-fold cross-validation. *Statistics and Decisions*, 24(3), 351 – 371.

ARTICLE INFORMATION

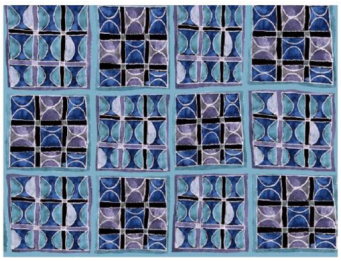
STAtOR 2017 Nr.4 pages 12-16.

AUTHOR INFORMATION

Mark van der Laan, Ph.D., is a Professor of Biostatistics and Statistics at UC Berkeley. His research interests include statistical methods in genomics (i.e., computational biology), survival analysis, censored data, targeted maximum likelihood estimation in semiparametric models, causal inference, data adaptive loss-based super learning, and multiple testing. He is the recipient of the VvS-OR Van Dantzig Award 2005.
E-mail: laan@berkeley.edu

Gepubliceerd op: January 12, 2018

Lees meer



Gedetailleerde en tijdige statistieken over de Nederlandse samenleving

(https://www.vvsor.nl/articles/gedetailleerde-en-tijdige-statistieken-over-de-nederlandse-samenleving/)

COLUMN

Kansrekening, een echt wiskundevak voor havo en vwo

(https://www.vvsor.nl/articles/kansrekening-een-echt-wiskundevak-voor-havo-en-vwo/)

Statistici tegen alcoholmisbruik

(https://www.vvsor.nl/articles/statistici-tegen-alcoholmisbruik/)

EVENT ON DECEMBER 12TH, 2019

Replication Day

(https://www.vvsor.nl/articles/event-sws-section/)

COLUMN

Duisternis en computergestuurde willekeur

(https://www.vvsor.nl/articles/duisternis-en-computergestuurde-willekeur/)

HET ANALYSEREN VAN GEFRAGMENTEERDE DATA

Personal health train

(https://www.vvsor.nl/articles/personal-health-train/)

ALLE ARTIKELEN ()

Vereniging voor Statistiek en Operations Research