Casper Albers
Psychometrie & Statistiek
Rijksuniversiteit Groningen
c.j.albers@rug.nl

Column Casper sees a chance

The statistical approach known as Machine Learning

This issue of the *Nieuw Archief voor Wiskunde* is devoted to Machine Learning. Although this has a very modern sounding name, the majority of techniques in machine learning are borrowed from other fields, most notably statistics, and date back many decades. Yet, there's more to machine learning than simply copying other fields' work.

The field of machine learning has emerged from computing science and, since about thirty years, it has been recognised as a separate branch of science. As such, it is a young and booming field with terms such as '(un)supervised learning', 'Bayesian networks', 'clustering methods' and 'classification trees'. To a large extent, these are new names for old concepts.

Historically, statistics is the scientific field of studying quantitative information under uncertainty. As such it emerged from mathematics, with its roots dating back to renaissance scientists such as Blaise Pascal, Christian Huygens and Jakob Bernoulli. What all statistical methods have in common, whether it concerns the elementary computation of a mean or a complicated statistical test for a high-dimensional data set, is that they aim to extract information from raw data. With the amount of data growing over time, and the emergence of computers in the second half of the previous centuries, statistical methods became more and more computationally intensive. Already in 1977, the International Statistical Association founded the International Association for Statistical Computing. Thus, doing data analysis with computers is not a novelty invented by machine learning.

Same concepts, different words

Stanford statistics professors Robert Tibshiriani, Jerome Friedman and Trevor Hastie — three of the very few that managed to obtain rockstar status among both statisticians and machine learners — offer a course in machine learning. As part of their course material, they provide a table comparing terms in both fields, part of which is reproduced here [5]:

Machine learning	Statistics
Network, graph	Model
Weights	Parameters
Learning	Fitting
Generalization	Test set performance
Supervised learning	Regression, classification
Unsupervised learning	Density estimation, clustering

Photo: James Jowers/George Eastman House Collection

There is some tongue-in-cheek in this table, but they do hit a nerve. Indeed, a basic supervised learning model is nothing more than a linear regression model without a check on the validity of the underlying assumptions.

Statistics

Statistical methods can be roughly subdivided into two disciplines: estimation and inference. Estimation techniques are used to reduce a (relatively) large amount of quantitative information, into a smaller and more comprehensible amount of information. Examples are the mean, standard deviation and visualisations such as a histogram. On the other hand, inferential models form a structured way to use data to infer about two competing hypotheses and, often, decide between two competing actions. An example is the testing of the efficacy of a new drug, by comparing the health of a treatment group to that of the placebo group. Should the difference be sufficiently large, then the statistician would argue that it is so unlikely that such a difference (or even larger) would occur by chance, that it is safe to reject the hypothesis that the drug has no beneficial health effect. Usually such a model relies on a set of assumptions, such as a linear relation between age and health, the same amount of health fluctuations between men and women, normally distributed residuals, et cetera. To build such a model, expert knowledge is required: medical experts can tell you that it is sensible to allow for gender differences when studying health, but that allowing for differences based on the number of letters in your first name is less sensible. Together with the expert, the statistician builds the model.

The goal of statistics is understanding the data-generating process, with the aim to make better decisions. However, the model only ‘works’ if the underlying assumptions are valid. Checking model validity thus is a major part of statistical analyses. Some non-parametric techniques exist that make fewer assumptions (e.g. not assuming normality), but models that make no assumption whatsoever are very rare in statistics.

As George Box said: “All models are wrong, but some models are useful.” The statistician knows that her/his model is a simplification of reality. However, just as the mean and standard information provide less information than the full sample, the statistician is willing to accept this: there is a trade-off between model complexity and model comprehensibility (and, thus, usability). If, for instance, the relation between age and health is not linear (and I’m sure it isn’t), but also not too far from linear, then assuming linearity greatly increases the comprehensibility of the model.

Machine learning

This is where the main difference between statistics and machine learning pops up: machine learning models are fully assumption-free. Furthermore, the machine learner does not care about model validity: the main (and often only) point of interest is prediction accuracy.

Machine learners can make assumption free models by relying on the bootstrap and cross-validation, tools developed by another cross-discipline rockstar, Bradley Efron. A related, common, method is to split the total sample into a training and validation set. The machine learning model is given the training data, e.g. a list of people’s age, gender, health and whether or not they are part of the treatment group. The algorithm then searches this data set for patterns. If, indeed, older people are of lower health, the algorithm will learn this — without the restriction that the relation must be linear. As such, the supervised learning algorithm is nothing more or less than statistical non-parametric, non-linear regression, yet with a different objective: comprehensibility is irrelevant (although some people in machine learning do focus on this). It is perfectly fine if the algorithm is a black box, as long as it learns the relations between the variables as good as possible. The algorithm thus creates some kind of procedure that, given someone's age, gender and drug use information, predicts this person's health as good as possible, without revealing how the exact relation between the variables is.

This algorithm is subsequently applied to the validation set, and the predictions are compared with the actual health scores. A value like the mean squared prediction error provided information on the model performance. As the predictions are made without a model, the machine learner doesn't have to worry about model validity. This advantage comes at the cost that the machine learner also doesn't learn about the data generating process.

Different goals

Thus, many of the techniques used in statistics and machine learning are more or less equivalent. Yet, the purpose of their use, and the chosen strategy, is fundamentally different. Whereas statistics is concerned with understanding processes and providing tools for informed decision making, (most fields in) machine learning is concerned with predicting future observations as good as possible.

In some situations, learning about the data generating mechanisms actually is the core scientific question, in which case statistical models are essential. In other cases, e.g. predicting where traffic queues will occur or whether some stock price will go up or down, a good prediction might be more valuable than a good explanation. Which of the two mindsets is most useful thus depends on the nature of the question you want answered. This is also argued by Leo Breiman [1], in a thought-provoking paper with commentaries from the likes of Sir David Cox and Bradley Efron.

Fawcett and Hardin [3] give a nice metaphor describing the relation between both fields: "[Statistics and machine learning] are like two pairs of old men sitting in a park playing two different board games. Both games use the same type of board and the same set of pieces, but each plays by different rules and has a different goal because the games are fundamentally different. Each pair looks at the other's board with bemusement and thinks they're not very good at the game."

Marketing

According to O'Connor [2], the reason that machine learning invited all these new terms for existing concepts, and has become hugely popular in doing so, is that statistics has a marketing problem: "Machine learning sounds like it's young, vibrant, interesting to learn, and growing; statistics does not." He cites Friedman [4] who states: "One can catalog a long history of Statistics (as a field) ignoring useful methodology developed in other data related fields. Here are some of them [...] that had seminal beginnings in statistics but for the most part were subsequently ignored in our field: pattern recognition, neural networks, machine learning, graphical models/Bayesian networks, chemometrics, data visualization. [...] One view says that our field should concentrate on that small part of information science that we do best, namely probabilistic inference based on mathematics. If this view is adopted, we should become resigned to the fact that the role of Statistics as a player in the 'information revolution' will steadily diminish over time."

Friedman's message is clear: that other fields seemingly took over large parts of data analysis, is statistics' own fault. The field should have embraced computational methodology as fundamental statistical tool. Friedman wrote his piece two decades ago and, fortunately, computational methodology is playing a more relevant role in statistics, most notably in applied statistics, than at the turn of the century. The change has come to late to claim all computational data science to be part of statistics — that ship has sailed. Friedman also paraphrases Efron by stating: "Those who ignore statistics are condemned to reinvent it." At least for the near future, the fields of statistics and machine learning will have to co-exist and co-operate together.

Conclusion

To summarise, I have outlined that many machine learning-ideas have been 'borrowed' from statistics. Some statisticians experience frustration or jealousy from this. Although somewhat understandable, it's like that kid in school that got popular by stealing your joke, I do not share this emotion.

In my experience, teaching statistical thinking, i.e. reasoning under uncertainty, requires different strategies for different audiences. An undergraduate student in mathematics can be taught that the ordinary least squares estimator is 'the optimal' estimator in regression by walking her/him through the elegant proof of the Gauss–Markov theorem. Students in social sciences, however, prefer heuristic explanations based on accessible simulations. I can very well imagine that students with a solid background in computing, will benefit from an explanation in terms of algorithms and other common concepts.

In the end, what matters is that as many students learn how to perform a scientifically validated analysis of quantitative data and, as such. What label they assign to their approach, whether it is statistics, machine learning, or some umbrella term as data science, is virtually irrelevant. If machine learning helps to get more people to do solid data analysis, I'm on board. In this age of fake news and fake information, all scientifically based methods for turning data into comprehensible information should be welcomed with open arms.

References

1 Leo Breiman, Statistical modeling: The two cultures (with discussion), *Statistical Science* 16(3) (2001), 199–231, projecteuclid.org/euclid.ss/1009213726.

2 Brendan O'Connor, Statistics vs. machine learning: Fight! (2008), brenocon.com/blog/2008/12/statistics-vs-machine-learning-fight.

3 Tom Fawcett and Drew Hardin, Machine learning vs. statistics (2017), svds.com/machine-learning-vs-statistics.

4 Jerome Friedman, Data mining and statistics: What's the connection? *Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics* (1997).

5 Robert Tibshiriani, Glossary, statweb.stanford.edu/~tibs/stat315a/glossary.pdf.