# *Are ML and Statistics Complementary?*

**Roundtable discussion at the 6th IMS-ISBA meeting on "*Data Science in the next 50 years"***

Max Welling, University of Amsterdam

December 28, 2015

I have been asked to provide some personal remarks on the topic "*Data Science in the next 50 years*", and in particular on the relation between machine learning and statistics. Machine learning, being the much younger discipline of the two is build on the foundations of statistics and has absorbed much of its philosophy and many of its techniques over the years. For instance, In machine learning we almost exclusively follow either the Maximum Likelihood or Bayesian route to estimation and we use expectation maximization (EM) to handle latent variables. Also, the previous "hype" in machine learning (before deep learning) was about nonparametric Bayesian methods, clearly a core domain of statistics. At the same time, there are cultural differences between the two fields: where statistics is more focussed on statistical inference, that is, explaining and testing properties of a population from which we see a random sample, machine learning is more concerned with making predictions, even if the prediction can not be explained very well (a.k.a. "a black-box prediction").

The paradigm shift towards deep learning which we are witnessing today provides a beautiful illustration of the cultural gap between machine learning and statistics. It follows a decade of convergence between the two disciplines when graphical models and nonparametric methods were the tools of choice, and learning/inference methods like expectation maximization (EM) and MCMC ruled the day.

Deep learning's successes can be explained by three factors:

1. Scaling up computation (e.g. by using GPUs).
2. Collecting and processing massive amounts of data.
3. Using models with massive amounts of parameters, even if they are unidentifiable and uninterpretable.

Factors 1 and 2 are second nature to a machine learner because of his or her education in computer science (and not mathematics). Factor 3 is a result of 1 and 2 and a focus on making accurate predictions rather than statistical inference. Increasingly, the paradigm in deep learning seems to be: collect a (massive) dataset, determine the cost function you want to minimize, design a (massive) neural network architecture such that gradients can propagate "end-to-end", and finally apply some version of stochastic gradient descend to minimize the cost until time runs out. Like it or not, the surprising fact is that nothing out there seems to beat this paradigm in terms of prediction.

After a decade of graphical models, the pendulum thus seems to swing once more in a direction away from statistical principles and towards computational ones. There is little hope of interpreting the billions of parameters of a neural architecture. There even seems to be a certain reluctance to attach properly calibrated probabilities to outcomes, in an attempt to quantize uncertainties in predictions.

I predict that the two disciplines will not divorce however. They represent two key aspects of data science that should become integrated in the long run. (And yes, it would help if we don't house the two disciplines in different departments). To the statistician I would say: in today's world it would be silly to ignore the massive amounts of data available for analysis (and indeed many statisticians acknowledge this fact). And to analyse massive data one needs to worry about storage and (distributed) computation. I have often heard researchers with appointments in statistics departments moan about the programming skills of their students. It therefore seems reasonable to include computer science classes in a statistics curriculum. How about complex black-box prediction models that are hard to interpret? It clearly depends on the problem, but it seems that under the assumption that the real world is “infinitely complex”, model complexity should scale with data volume and perhaps this is a price we have to pay.

One place where statistics and computation seem to converge beautifully is when the model is expressed as a simulation. This is in fact the way that most scientific disciplines express their expert knowledge of a problem domain (think e.g. weather forecasting). All variables have clear semantic interpretations and it is the task of the statistician to perform inference over these variables. This task is however highly computationally demanding and requires careful thought on where and how to spend the available computation. This field is called “approximate Bayesian computation” (ABC) in statistics. In machine learning a new paradigm is appearing called “probabilistic programming” that aims to solve the same inference tasks but in addition develops specialized programming languages (e.g. based on graphical models) in which to express these models.

A key question is whether statistics and machine learning will also converge in fields such as deep learning? In other words, will statisticians adopt the computation heavy deep learning modeling paradigm and will the machine learner adopts some of the statistical tools to enrich this field? My prediction is yes, it will, and here is why. While for certain applications mere predictions seem sufficient, there are many for which it is not. Take the example of predicting what ads to show on a webpage. In principle just making accurate predictions will create profit. However, determining which factors *cause* certain outcomes (a form of statistical inference) will create insight and greatly help with designing predictors that are more robust to shifting input domains. Also, having access to *calibrated uncertainty estimates* will help determine whether to base these predictions on covariates (content based filtering) or previous user click behavior (collaborative filtering). Additionally it can help balance exploitation (serving ads that we know the user will like) and exploration (serving ads from which we can learn something about the user).

More generally I would say that we need the tools of statistics (e.g. causal reasoning, calibrated error-bars) when we deploy predictors in the real world, i.e. have them interact with humans and make decisions based on them. For example, a physician would like to understand why an algorithm thinks this patient will develop Alzheimer’s disease and s/he would also like to know what the probability is that this prediction is correct. Also, a self-driving car will need to know when it does not understand the road situation and hand the steering wheel back to the human driver. Thus, for many applications, in order to successfully interact with humans, machines will need to explain their reasoning, including some quantification of confidence, to humans.

Finally, machine learners have the tendency to focus more on the practical, methodological aspects of modeling (although there is a whole subfield of machine learning theory dealing with theoretical questions). As a result, there are a number of “accepted” methods in machine learning that have very limited theoretical foundation, or stated differently, of which the theoretical properties have not been studied in depth. For instance, while a substantial group of researchers use or develop semi-supervised learning methods, it is not all that well known under what conditions they work. It is also here that the more theoretically inclined statistician has much to contribute to machine learning in my opinion.

There are many more intriguing data science questions where statistics and machine learning meet, such as 1) making fair decisions (e.g. decisions that do not depend in any way on race or gender), 2) removing bias from data collected in the wild (i.e. not by following a random sampling protocol), 3) privacy preserving predictions, 4) disentangling correlation from causation, 5) developing sound statistical procedures in high dimensions, and many more.

It is my hope that both disciplines continue to realize that their tools are truly complementary, and more collaborations between the fields will emerge as a result in the coming years.