# Machine Learning in Social Epidemiology: Learning From Experience

Catherine Kreatsoulas, S.V. Subramanian

www.elsevier.com/locate/ssmph

PII: S2352-8273(18)30040-5
DOI: https://doi.org/10.1016/j.ssmph.2018.03.007
Reference: SSMPH244

To appear in: *SSM - Population Health*
Accepted date: 22

Cite this article as: Catherine Kreatsoulas and S.V. Subramanian, Machine Learning in Social Epidemiology: Learning From Experience, *SSM - Population Health,* https://doi.org/10.1016/j.ssmph.2018.03.007

# Machine Learning in Social Epidemiology: Learning From Experience

Catherine Kreatsoulas, MSc, PhD[1,2*] and S.V. Subramanian, PhD[1]

[1]Harvard TH Chan School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115-6096
[2]Dalla Lana School of Public Health, University of Toronto, 155 College Street, Toronto, ON M5T 3M7

*Corresponding author. Catherine Kreatsoulas ckreats@hsph.harvard.edu

In the summer of 1955 at Dartmouth University, a small community of progressive-thinking scientists including John McCarthy, who is credited with coining the term "artificial intelligence (AI)", Marvin Minsky, Nathan Rochester and Claude Shannon, submitted a research proposal seeking to explore,

> *"...every aspect of learning or any other feature of intelligence that can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans and improve themselves."* (McCarthy, Minsky, Rochester, & Shannon, 1955)

Now over 60 years later, with many momentous accolades achieved in parallel with exponential advances in computing, applications of machine learning have infiltrated, improved and continue to augment many aspects of our daily lives. Today machine learning is a mainstay in business, finance, manufacturing, retail, science, technology, mobile computing, social media affecting our behaviours as consumers and creators of data, each interaction deepening our digital footprint. Medicine and disciplines related to health have become the new frontier for machine learning and big data. In particular, fields such as social epidemiology seem well suited to tap into the vast amounts of social data (Gruebner et al., 2017) including credit scores and social networks that could potentially shed some new insights to understanding health behaviours and how social determinants of health may operate. While successful examples of mainstream applications of machine learning offer much excitement for adaptation in the social sciences, we are at a critical moment in history where we can learn from successful machine learning applications, limitations and the potential dangers of mal-adapting these techniques.

Machine learning is "a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision-making under uncertainty"(Murphy, 2013). And while methods from machine learning are closely related to the type of statistics traditionally used in social health research, they differ in probabilistic inference and modeling. The paper by Seligman et al. (2018) sought to compare four machine learning algorithms with a traditional regression to determine if 1) machine learning algorithms lead to better predictions and 2) do they enhance our understanding of how social determinants may result in differences in health outcomes? The authors conclude that traditional regression historically used in social health research faired well when compared to several machine learning methods; neural networks faired best due to their robust ability to allow for interactions and nonlinearity among input variables. However, the interpretation of neural networks is complicated, and the authors base their conclusions almost exclusively on the r square value obtained in cross-validation, a process in itself laden with inherent limitations. While the authors successfully compare results between the different methodologies, it is unclear how these methods enhance our understanding of health outcomes, particularly when the fundamental goal of machine learning is to generalize beyond the algorithm training set. Arguably this may not necessarily be the fault with this study per se but rather a consequence of the infancy of these techniques in the social epidemiologic space.

As quantitative social scientists process and often collate multiple sources of data, there are many alluring features from various techniques in machine learning that offer new methodologic ideas in how to handle and merge structured and unstructured datasets. A distinct advantage of machine learning methods includes the robust handling of large numbers of variables combined in interactive linear and non-linear ways to detect patterns in the data for prediction. While there is a vast array of learning algorithms available, all machine learning algorithms consist of combinations of three key components: 1) representation of the input of data, where a classifier can learn in the hypothesis space, 2) evaluation of the classifiers, and lastly, 3) optimization, a search among classifiers to find the best performing one (Domingos, 2012). In supervised learning, the goal is prediction and includes techniques such as regression and classification or pattern recognition whereas in unsupervised learning, the goal is to find patterns in the data which is sometimes called knowledge discovery (Murphy, 2013). Reinforcement learning, while not as commonly used, is useful for learning how to act or behave when given occasional reward or punishment signals (Murphy, 2013). Table 1 outlines some of the strengths and limitations associated with this comparative study of machine learning methods used to evaluate health outcomes from the Health and Retirement Study dataset. While each type of machine learning offers distinct advantages and disadvantages, in a classic paper entitled, "No free lunch theorems for optimization" (Wolpert & Macready, 1997) the term “no free lunch” has popularly been used to describe that no one type of algorithm is best for every prediction problem. In this classic paper, the authors geometrically demonstrate what it means for an algorithm to be well-suited for an optimization problem and the danger of comparing algorithms by their performance on a small sample of problems (Wolpert & Macready, 1997). In addition to their valuable suggestions, we would like to recommend some additional thoughts when undertaking a machine learning approach to analyzing social data as it relates to health:

### 1. Understand both the underlying mathematical "skeleton" of the optimization theory and how the goals of the analysis should align, *a priori*.

Machine learning techniques have exponentially increased in popularity arguably due to their promise to predict. But it is important to distinguish between prediction and causation; simply put, these are not interchangeable concepts, and the underpinnings of prediction are probabilistic. The work of Judea Pearl (Pearl, 2009) seeks to marry the counterfactual into probabilistic approaches of causation, however, its application to machine learning is still considered to be in its infancy. Equally challenging has been the implementation of causal inference in the social epidemiology space (Kaufman & Cooper, 1999) (Glymour & Rudolph, 2016) particularly the consistency assumption (Rehkopf, Glymour, & Osypuk, 2016). Further, while understanding the baseline assumptions of the research question and how it aligns with the mathematical skeleton of the analysis is imperative in any quantitative analysis, the ability to explain the study results hinges on this. For example, results from regression are relatively simple to explain, whereas machine learning methods such as random forests and neural networks, which are strong in prediction, are complicated to explain and are (literally) black boxes. One must ponder, is probabilistic prediction alone enough and how important is the explanation of the study results? More importantly, there is no substitute for the substantive understanding of the problem with the mechanism, and the corresponding mathematical structure of the analysis, to understand what the results will reveal.

### 2. Understand the data source and composition of the study population; any potential biases may result in overfitting, and can be unintentionally propagated in machine learning algorithms.

Social scientists like data scientists often rely on publically available or longitudinal observational datasets rarely collected for the intended analysis. Within the machine learning community the problem of overfitting, an error in generalization is well-known, yet it is not always immediately apparent. Domingos (Domingos, 2012) in a machine learning overview paper decomposes the problem of overfitting into bias and variance, where bias is the learner's tendency to learn the same thing incorrectly consistently, and variance is the tendency to learn random things irrespective of the real signal. While there is a multitude of techniques to test and combat these challenges, is imperative always to be mindful that machine learning algorithms can only be trained on the data fed. If the goal of the machine learning algorithm is prediction, the algorithm will intrinsically contain an inductive bias. While this in itself is not necessarily a negative bias, if there are any biases in the dataset, they will inherently be propagated. For example, if sex abuse in the population is equally present in men and women, but women are more likely to report it, the algorithm will predict that women are more likely to be sexually abused when in fact this may not be true. Of even greater concern and a notable problem within the machine learning community is that it is virtually impossible to detect or correct for such biases in machine learning algorithms.

### 3. Be aware of limitations in the construction of generalizability and cross-validation techniques of model performance evaluation.

While there are many different versions of cross-validation techniques to evaluate the performance of a machine learning algorithm, almost all contain a training set, a validation, and a test set, split into varying percentages. For example, if 75% of the algorithm is trained on the training set, model selection is then conducted on the validation set and then tested on the test set. If there is an inherent bias in the dataset, such as it was composed of volunteers or a particular gender/ race/ socioeconomic group is underrepresented, validation and test sets will be unable to detect these biases despite using reserved data with acceptable cross-validation metrics. In this scenario, despite metrics suggesting good generalizability, in reality, this remains in question. In fact, (Wolpert & Macready, 1997) demonstrate that the alignment of the underlying probability distribution over the optimization problem determines the performance of the algorithm. There are recent calls to the machine learning community to increase the transparency and publish the code used in machine learning algorithms as the random numbers generated in the training set are highly sensitive and contingent to the data in the initial training (Hutson, 2018). While rarely practiced in machine learning, the best test of validation is to test the algorithms in a completely different dataset altogether to understand the speed-accuracy - complexity trade-offs. After all, one of the hallmarks of science is replicability.

Conflict of interest
There are no conflicts of interest to report from any of the authors.

## References

Domingos, P. (2012). A few useful things to know about machine learning. *Commun. ACM, 55*(10), 78-87. doi:10.1145/2347736.2347755

Glymour, M. M., & Rudolph, K. E. (2016). Causal inference challenges in social epidemiology: Bias, specificity, and imagination. *Soc Sci Med, 166*, 258-265. doi:10.1016/j.socscimed.2016.07.045

Gruebner, O., Sykora, M., Lowe, S. R., Shankardass, K., Galea, S., & Subramanian, S. V. (2017). Big data opportunities for social behavioral and mental health research. *Soc Sci Med, 189*, 167-169. doi:10.1016/j.socscimed.2017.07.018

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *Science, 359*(6377), 725-726. doi:10.1126/science.359.6377.725

Kaufman, J. S., & Cooper, R. S. (1999). Seeking causal explanations in social epidemiology. *Am J Epidemiol, 150*(2), 113-120.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. doi:citeulike-article-id:7546286

Murphy, K. P. (2013). *Machine learning : a probabilistic perspective*. Cambridge, Mass. [u.a.]: MIT Press.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*: Cambridge University Press.

Rehkopf, D. H., Glymour, M. M., & Osypuk, T. L. (2016). The Consistency Assumption for Causal Inference in Social Epidemiology: When a Rose is Not a Rose. *Curr Epidemiol Rep, 3*(1), 63-71. doi:10.1007/s40471-016-0069-5

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67-82. doi:10.1109/4235.585893

Table 1: Overview of Strengths and Limitations of Machine Learning Techniques in Social Epidemiology

| Technique | Essential Feature | Strengths | Limitations | Generalized prediction |
|---|---|---|---|---|
| Regression | -Attempts to fit a straight hyperplane to data | -Excellent for prediction among linear relationships;<br>-Simple to interpret and understand model because attributes have an additive effect on the model<br>-Can be regularized to deal with overfitting | -Does not handle well non-linear relationships in data<br>-Learning algorithms make a set of assumptions about the data and therefore there is an inductive bias embedded within each algorithm | -Selecting the best model is more challenging than optimizing its parameters once model is fixed<br>-Assumes that any changes in the attributes and output both occur with some regularity and smoothness for generalization |
| | | | | |
| LASSO penalized regression | -Additional variables that do not substantially improve prediction are penalized | -Useful in OLS when many variables are highly correlated (as variance increases in OLS, beta becomes increasingly inaccurate) | -The weighted penalty, lambda, is estimated and tested by a variety of methods each with pros and cons | -Goal is to reduce and select among redundant predictors in generalized linear model to improve prediction |
| | | | | |
| Random forests | -Repeatedly split dataset into random sets of decision trees with if-then rules at branches and interpolation at leaves | -Learning is non-parametric<br>-Variables do not need to be transformed | -Highly prone to overfitting (model can keep branching until the data is memorized) | -Larger forests typically have better prediction (being mindful of |

| | | | | |
|---|---|---|---|---|
| | | -Handles outliers well<br>-Handles missing values well<br>-Ensemble methods that include random forests often perform well | -Black box predictions are difficult to interpret | overfitting and correlated trees) |
| | | | | |
| Neural networks | -Based on neuron/synapse activation structure of human brain using synaptic weights that represent 'hidden layers' between inputs and outputs | -Learning is nonlinear<br>-Handles outliers well<br>-Can learn complex patterns from highly dimensional data<br>-Hidden layers alleviates features engineering<br>-Often best performing algorithm | -Difficult to set up; many parameters require decisions on architecture and hyperparameters of network<br>-Easy to overfit<br>-Often very difficult to interpret<br>-Requires large sample sizes<br>-Computationally very intense to train | -Generalization is difficult without large samples of data |