On "Machine Learning for Sociology"

John A. Bernau

June 18, 2021

Notes and reflections on the article "Machine Learning for Sociology" by Mario Molina and Filiz Garip from the Annual Review of Sociology *(2019).*

Machine learning refers to advanced statistical modeling techniques used primarily by data scientists. In the last ten years, this field has developed different cultures and norms about working with data.

In sociology, you (ostensibly) start with a theory-driven research question about the relationship between an outcome of interest (DV) and one or more predictor variables (IV). You then try to produce an accurate generative model to understand the effect of the IVs on the DV. This usually involves including or excluding certain variables, trying a few interaction effects, etc. Models are typically evaluated according to R^2 (or some variant) and the number of statistically significant predictors. The end result is the model itself, which allows a better understanding of how an outcome came to be. This workflow evolved in a world where data were scarce and computing power was limited.

In data science, you (usually) start with an industry-defined prediction task: *How do we know which videos a user will click on? How do we predict which product customers will buy?* You then try to produce a predictive model to estimate your DV given a set of IVs. This typically involves using as many IVs as are available and letting the statistical model sort out which ones produce accurate predictions. To avoid overfitting, machine learning employs sophisticated models as well as norms about workflow like separating the data into training sets (to fit the model), validation sets (to tune the model), and test sets (to evaluate the model). Models are usually evaluated according to predictive accuracy on the test set. The end results are thus the predictions, which can be used in future decision-making. However, these models are usually too complex to derive meaningful understanding of relationships and can end up producing "black box" predictions. This workflow evolved in a world with abundant data and computing power.

	Sociology	Machine learning
Starts with…	theory-driven RQ	industry-defined prediction task
Creates a…	generative model	predictive model
With the goal of…	understanding relationships (β) between IV and DV	making accurate out-of-sample predictions ($\hat{y}$) of DV
Workflow	use all data to fit model	split into training, validation, test data
Evaluated on…	R^2, significant IVs	prediction accuracy (test set)
Limitations	overfitting, unknown out-of-sample performance	obscure causal pathways, "black box" predictions

Of course, these two emphases are literally two sides of the same coin: a regression equation includes both βs for your IVs and $\hat{y}$ for your predictions. So the distinction is mostly one of emphasis and procedural norms. However, it's important to note that these differences emerged to address the limitations of traditional statistical approaches in a new world of "big data". Let's take a closer look at a few techniques.

First, every statistical model determines parameters using an optimization function. In OLS, this is the sum of squared residuals. See (1) below.

$$SSR = \sum_{i=1}^{n} [y_i - f(x_i)]^2 \tag{1}$$

Penalized regression introduces a "regularizer" that shrinks coefficients towards zero in order to prevent over-fitting. In other words, when using large and complicated datasets with many predictors, interactions, and non-linearities, penalized regression prevents your model from being too complicated. However, it's important to scale variables first (i.e. dividing by standard deviation) so coefficients are evaluated evenly. Three types of penalized regression include lasso, ridge, and elastic net regression.

Lasso regression stands for *Least Absolute Shrinkage and Selection Operator.* It starts with the SSR and adds a regularizer where lambda (λ) is a (researcher-set) penalty introduced for the sum of absolute values of coefficients (β). Because the model is trying to minimize this function, in some cases this pushes coefficients to zero, eliminating unhelpful variables. In the absence of theoretical motivation, this is a useful way to sort through a large number of predictors. See (2) below.

$$SSR = \sum_{i=1}^{n}[y_i - f(x_i)]^2 \; + \; \lambda \sum_{j=1}^{p} |\beta_j| \tag{2}$$

Ridge regression is similar to lasso regression, but adds a regularizer where lambda (λ) is a (researcher-set) penalty introduced for the sum of squared coefficients (β). This will reduce coefficient size, but never all the way to zero. The is useful if the model has mostly informative variables. Thus, ridge regression decreases the complexity of a model but does not reduce the number of variables, just shrinks their effect. See (3) below.

$$SSR = \sum_{i=1}^{n}[y_i - f(x_i)]^2 \; + \; \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3}$$

Elastic net regression is simply a combination of lasso and ridge regression, where lambda (λ) is a (researcher-set) penalty term and alpha (α) (researcher-set) is used to balance both penalties: the sum of absolute values of coefficients (lasso) and the sum of squared coefficients (ridge). Most real-world applications involve testing a model with varying levels of alpha from 0 (ridge regression) to 1 (lasso regression) and seeing which parameters arrive at the best predictions. See (4) below.

$$SSR = \sum_{i=1}^{n}[y_i - f(x_i)]^2 \; + \; \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right) \tag{4}$$

Regression trees are another form of machine learning that model complex relationships between IVs and DVs. These models are best understood visually as a series of branching variables and their interactions. Formally, they "describe a sequence of splits in the input space (X) that predict an output (Y) at the end node" (p31). Essentially, in a linear modeling framework, it takes your data and makes new dummy variables to represent certain threshold cutoff points and interacts them all. The result is a complicated model that is hard to interpret: "Being 50 years or older puts you in a framework where having less than ten years of education means being male improves income." It does a good job of predicting, but it is hard to step back and theorize about "the role of age on income." Like penalized regression it has its own parameters to control how many splits (nodes / leaves) in the tree there are.[1]

Penalized regression and regression trees are both example of *supervised machine learning*: a function that connects known inputs to known outputs. *Unsupervised machine learning*

[1] This is an incredible visual explanation of regression trees from R2D3: https://r2d3.us/visual-intro-to-machine-learning-part-1/.

describes methods to summarize inputs without reference to outputs: principal component analysis, factor analysis, cluster analysis, latent class analysis, sequence analysis, topic modeling, community detection, etc. These techniques attempt to simplify or represent your inputs in a parsimonious and potentially useful way.

The article is a helpful primer as sociology learns to live alongside data science. Duncan Watts has been working in this space for a while and encourages sociologists to focus more on prediction problems: If you claim to know how something works, you need to prove it by making predictions. Philosophers may recognize this as a variant of "Granger causation". Watt's AJS paper[1] explores this and other problems of sociological explanation. In a 2018 talk at Cornell[2], Watts and others (including Filiz Garip) engage in a lively debate about how sociologists might adopt techniques from machine learning.

I am especially interested in the relationship between explanation, causation, and prediction. Karl Pearson and Bertrand Russell famously rejected causation as an old-fashioned metaphysical concept (the former advocating for its replacement: "correlation"). Prediction seems to capture an element of causal relationships, while somehow sidestepping causal understanding. Explanation seems to evoke notions of causation, but divorced from prediction? See Gorski (2018) on 'retroduction'. [3]

Going forward, I hope to translate this theoretical overview into practical examples. How do you run an elastic net regression in R? What are the standard reporting procedures for this type of analysis? Lastly, how will these tools improve the practice of sociology: understanding social things, attributes, relationships, and processes?

---

1 Watts, Duncan. 2015. "Common Sense and Sociological Explanations." *American Journal of Sociology* 120(2): 313–351. https://doi.org/10.1086/678271

2 https://www.youtube.com/watch?v=0O6COXI0cx8

3 Gorski, Philip S. 2018. "After Positivism: Critical Realism and Historical Sociology." Pp. 23–45 in *Political Power and Social Theory*, edited by Timothy Rutzou and George Steinmetz. Emerald Publishing Limited.