Machine Learning

Zhentao Shi

March 7, 2018

Machine Learning

Machine learning has quickly grown into a big field, with applications from scientific research to daily life. The ideas in machine learning are general and adaptable to economic investigation.

From an econometrician's view, machine learning is a set of data fitting procedures that focus on out-of-sample prediction. The simplest illustration is in the regression context. We repeat a scientific experiment for n times, and we harvest a dataset $(y_i, x_i)_{i=1}^n$ in hand. What would be the best way to predict y_{n+1} from the same experiment if we set a x_{n+1}?

Machine learning is a paradigm shift against conventional statistics. When a statistician propose a new estimator, the standard practice is to pursue three desirable properties one after another. We first establish its consistency, which is seen as the bottom line. Given consistency, we want to show its asymptotic distribution. Ideally, the asymptotic distribution is normal. Asymptotic normality is desirable as it holds for many regular estimators and the inferential procedure is familiar to applied researchers. Furthermore, for an asymptotically normal estimator, we want to show efficiency, an optimality property. An efficient estimator achieves the smallest asymptotic variance in a class of asymptotically normal estimators.

Machine learning deviates from such routines. Machine learning researchers dismiss all the three points. First, they argue efficiency is not crucial because the dataset itself is big enough so that the variance is usually small. Second, in many situations statistical inference is not the goal, so inferential procedure is not of interest. For example, the recommendation system on Amazon or Taobao has a machine learning algorithm behind it. There we care about the prediction accuracy, not the causal link why a consumer interested in one good is likely to purchase another good. Third, the world is so complex that we have little idea about how the data is generated. We do not have to assume a data generating process (DGP). If there is no DGP, we lose the standing ground to talk about consistency. Where would my estimator converge to if there is no "true parameter"? With these arguments, the paradigm of conventional statistics is smashed. In the context of econometrics, such argument completely rejects the structural modeling tradition (the Cowles approach).

The above argument have merit, but it is also misleading. In this lecture, we set the ongoing philosophical debate aside. We economists are practical and reasonable souls. We study the most popular machine learning methods that have found growing popularity in economics.

Variable Selection and Prediction

In modern scientific analysis, the number of covariates x_i can be enormous. In DNA microarray analysis, we look for association between a symptom and genes. Theory in biology indicates that only a small handful of genes are involved, but it does not pinpoint which ones are the culprits. Variable selection is useful to identify the relevant genes, and then we can think about how to edit the genes to prevent certain diseases and better people's life.

Many explanatory variables are abundant in economic analysis. For example, a questionnaire from the UK Living Costs and Food Survey, a survey widely used for analysis of demand theory and family consumption, consists of thousand of questions.

Conventionally, applied economists do not appreciate the problem of variable selection, even though they always select variables implicitly. They rely on their prior knowledge to choose variables from a large number of potential candidates. Recently years economists wake up from the long lasting negligence. Stock and Watson (2012) are concerning about forecasting 143 US macroeconomic indicators. They conduct a horse race of several variable selection methods.

The most well-known variable selection method in a regression is the least-absolute-shrinkage-and-selection-operator (Lasso) (Tibshirani 1996). Upon the usual OLS criterion function, Lasso penalizes the L_1 norm of the coefficients. The criterion function of Lasso is written as

$$(2n)^{-1}(Y - X\beta)'(Y - X\beta) + \lambda\|\beta\|_1$$

where $\lambda \geq 0$ is a tuning parameter. In a wide range of values of λ, Lasso can shrink some coefficients exactly to 0, which suggests that these variables are likely to be irrelevant in the regression. However, later research (Zou 2006) finds that Lasso cannot consistently distinguish the relevant variables from the irrelevant ones.

Another successful variable selection is smoothly-clipped-absolute-deviation (SCAD) (Fan and Li 2001). Their criterion function is

$$(2n)^{-1}(Y - X\beta)'(Y - X\beta) + \sum_{j=1}^{d} \rho_\lambda(|\beta_j|)$$

where

$$\rho'_\lambda(\theta) = \lambda\left\{1\{\theta \leq \lambda\} + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} \cdot 1\{\theta > \lambda\}\right\}$$

for some $a > 2$ and $\theta > 0$. This is a non-convex function, and Fan and Li (2001) establish the so-called *oracle property*. An estimator boasting the oracle property can achieve variable selection consistency and (pointwise) asymptotic normality simultaneously.

The follow-up *adaptive Lasso* (Zou 2006) also enjoys the oracle property. Adaptive Lasso is a two step scheme: 1. First run a Lasso or ridge regression and save the estimator $\hat{\beta}^{(1)}$. 2. Solve

$$(2n)^{-1}(Y - X\beta)'(Y - X\beta) + \lambda\sum_{j=1}^{d} w_j|\beta_j|$$

where $w_j = 1\Big/\left|\hat{\beta}_j^{(1)}\right|^a$ and $a \geq 1$ is a constant. (Common choice is $a = 1$ or 2).

In R, `glmnet` or `LARS` implements Lasso, and `ncvreg` carries out SCAD. Adaptive Lasso can be done by set the weight via the argument `penalty.factor` in `glmnet`.

More methods are available if prediction of the response variables is the sole purpose of the regression. An intuitive one is called *stagewise forward selection*. We start from an empty model. Given many candidate x_j, in each round we add the regressor that can produce the biggest R^2. This method is similar to the idea of L_2 componentwise boosting, which does not adjust the coefficients fitted earlier.

To be discussed in the future

- The idea of boosting and its association to reweighting.
- regression tree, bagging (Killian and Inoue), average of subsampling

References

Fan, Jianqing, and Runze Li. 2001. "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties." *Journal of the American Statistical Association* 96 (456). Taylor & Francis: 1348–60.

Stock, James H, and Mark W Watson. 2012. "Generalized Shrinkage Methods for Forecasting Using Many Predictors." *Journal of Business & Economic Statistics* 30 (4). Taylor & Francis Group: 481–93.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*. JSTOR, 267–88.

Zou, Hui. 2006. "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101 (476). Taylor & Francis: 1418–29.