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. An authoritative reference is Friedman, Hastie, and Tibshirani (2008), written at the entry-year postgraduate level. The ideas in machine learning are general and applicable to economic investigation. Taddy (2018) introduces new technology *artificial intelligence* and the implication of the underlying economic modeling.

The two board classes of machine learning methods are *supervised learning* and *unsupervised learning*. Roughly speaking, the former is about the connection between X and Y, while the latter is only about X. Instances of the former are various regressions; those of the latter are density estimation, principle component analysis, and clustering. These examples are all familiar econometric problems.

From an econometrician's view, supervised 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 $(y_i, x_i)_{i=1}^n$. What would be the best way to predict y_{n+1} from the same experiment if we set 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.

Nonparametric Estimation

Parametric is referred to problems with a finite number of parameters, whereas *nonparametric* is associated with an infinite number of parameters. Nonparametric estimation is nothing new to statisticians. However, some ideas in this old topic is directly related to the underlying principles of machine learning methods.

Consider the density estimation given a sample $(x_1, \ldots, x_n)$. If we know, or assume, that the sample is drawn from a parametric family, for example the normal distribution, then we can use the maximum likelihood estimation to learn the mean and the variance. Nevertheless, when the parametric family is misspecified, the MLE estimation is inconsistent in theory. In practice, what is the correct parametric family is never known. If we do not want to impose a parametric assumption, then in principle we will have to use an infinite number of parameters to fully characterize the density. One well-known nonparametric estimation is the histogram. The shape of the bars of the histogram depends on the partition of the support. If the grid system on the support is too fine, then each bin will have only a few observations. Despite small bias, the estimation will suffer a large variance. On the other hand, if the grid system is too coarse, then each bin will be wide. It causes big bias, though the variance is small because each bin contains many observations. There is an bias-variance tradeoff. This tradeoff the defining feature not only for nonparametric estimation but for all machine learning methods.

Another example of nonparametric estimation is the conditional mean $f(x) = E[y_i|x_i = x]$ given a sample (y_i, x_i). This is what we encountered in the first lecture of graduate econometrics Econ5121A. We solve the minimization problem

$$\min_f E[(y_i - f(x_i))^2]$$

In Econ5121A, we use the linear projection to approximate $f(x)$. But the conditional mean is in general a nonlinear function. If we do not know the underlying parametric estimation of (y_i, x_i), estimating $f(x)$ becomes a non-parametric problem. In practice, the sample size n is always finite. The sample minimization problem is

$$\min_f \sum_{i=1}^{n} (y_i - f(x_i))^2.$$

We still have to restrict the class of functions that we search for the minimizer. If we assume f is a continuous function, one way to estimate it is the kernel method based on density estimation.

An alternative is to use a series expansion to approximate the function. Series expansion generates many additive regressors whose coefficients will be estimated. This is one way to "create" many variables on the right-hand side of a linear regression. For example, any bounded, continuous and differentiate function has a series representation $f(x) = \sum_{k=0}^{\infty} \beta_k \cos(\frac{k}{2}\pi x)$. In finite sample, we choose a finite K, usually much smaller than n, as a cut-off. Asymptotically $K \to \infty$ as $n \to \infty$ so that

$$f_K(x) = \sum_{k=0}^{K} \beta_k \cos\left(\frac{k}{2}\pi x\right) \to f(x)$$

. Similar bias-variance tradeoff appears in this nonparametric regression. If K is too big, f will be too flexible and it can achieve 100% of in-sample R-squared. This is not useful for out-of-sample prediction. Such prediction will have large variance, but small bias. On the other extreme, a very

small K will make $f_K(x)$ too rigid to approximate general nonlinear functions. It causes large bias but small variance.

The fundamental statistical mechanism that governs the performance is the bias-variance tradeoff. Thus we need *regularization* to balance the two components in the mean-squared error. Choosing the bandwidth is one way of regularization, choosing the terms of series expansion is another way of regularization.

A third way of regularization is to specify a sufficiently large K, and then add a penalty term to control the complexity of the additive series. The optimization problem is

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_{k=0}^{K} \beta_k f_k(x_i) \right)^2 + \lambda \sum_{k=0}^{K} \beta_k^2,$$

where λ is the tuning parameter such that $\lambda \to 0$ as $n \to \infty$, and $f_k(x_i) = \cos\left(\frac{k}{2}\pi x_i\right)$. In compact notation, let $y = (y_1, \ldots, y_n)'$ and $X = (X_{ik} = f_k(x_i))$, the above problem can be written as

$$(2n)^{-1}(Y - X\beta)'(Y - X\beta) + \lambda \|\beta\|_2^2$$

This is the *ridge regression* proposed in 1970's. This penalization scheme is very similar to what we will discuss in the next section in variable selection.

The practical question is, given a regularization problem, how to choose the tuning parameter? This is a difficult statistical problem with active research. The main theoretical proposal is either using an *information criterion* (for example, Akaike information criterion or Bayesian information criterion), or *cross validation*.

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

Unstructured Data

- text (Gentzkow, Kelly, and Taddy 2017)
- speech recognition
- photo recognition

Economic Applications

- Lee and Shi (2018, to be released)

- Shi and Huang (2018, to be released)
- Phillips and Shi (2018, to be released)
- Chinco, Clark-Joseph, and Ye (2017)

References

Chinco, Alexander M, Adam D Clark-Joseph, and Mao Ye. 2017. "Sparse Signals in the Cross-Section of Returns." National Bureau of Economic Research.

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.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2008. *The Elements of Statistical Learning*. 2nd ed. Springer.

Gentzkow, Matthew, Bryan T Kelly, and Matt Taddy. 2017. "Text as Data." National Bureau of Economic Research.

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.

Taddy, Matt. 2018. "The Technological Elements of Artificial Intelligence." National Bureau of Economic Research; University of Chicago Press.

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.