Econometrics I

Lecture 12: Model Selection and Machine Learning

Paul T. Scott
NYU Stern

Fall 2021

Overview

- A good (free) text on machine learning: Hastie, Tibshirani, Friedman
- Machine learning is primarily concerned with efficiently estimating models that generate good predictions.
- ML approaches are often referred to as "black boxes". There is typically little-to-no concern for building theories or understanding mechanisms.

Supervised vs unsupervised learning

- Most of what we have seen so far could be described as supervised learning, where there is an explicit distinction between inputs $\mathbf{x}$ and output y. Our goal is to do a good job predicting the output.
- Unsupervised learning simply tries to find patterns in the data $\mathbf{x}$ without specifying which variable we care about predicting. Examples: K-means clustering, mixture models, principal component analysis.

Loss and Risk

- g is a model that maps from inputs to outputs (predictions).
- A **loss** function is a measure of how well a prediction $g(x_i)$ predicts an outcome y_i:

$$L(y_i, g(x_i))$$

- **Risk** is the expected or average loss of a model:

$$\hat{R}(g) = N^{-1} \sum_i L(y_i, g(x_i))$$

- All estimators we've considered so far can be understood as mimizing a notion of risk:
  - OLS takes risk to be squared error
  - MLE takes risk to be negative (log) likelihood

Cross-validation

- **True risk**, expectation of risk: $R(g) = E(L(y, g(x)))$
- **Empirical risk**, risk in sample: $\hat{R}(g) = N^{-1} \sum_i L(y_i, g(x_i))$
- When a model is selected to minimize risk, the model's empirical risk in the sample used for estimation (**training data**) will be a biased estimate of true risk.
- Therefore, if we want to make a realistic assessment of how accurate the model is, we should compute risk in a sample that wasn't used for estimation (**validation data**).

Model Selection with Cross-Validation

- Whatever notion of risk you use and tool you use to minimize it, you could use one sample for training (estimation) and another for validation.
- The risk in the validation sample will be an unbiased estimate of the estimated model's risk (**cross validation**).
- We could use cross validation to do model selection!
- Rather than dividing our data into one training and one validation sample, we could use k-fold cross validation.
- After selecting our preferred model, we might as well go back and estimate it with all of the data.
- We could also think about trying to estimate what a model's out-of-sample fit will be...

Akaike Information Criterion

- Log-likelihood is $\ell\ell\left(\theta|\mathbf{x}\right)$, where θ is a candidate parameter vector, and $\mathbf{x}$ is all the data.
- The AIC is

$$2k - 2\ell\ell\left(\theta^*|\mathbf{x}\right),$$

  where k is the number of parameters. We select the model that yields the lowest value of the AIC, with θ^* selected for each model to maximize the likelihood function.
- This is a formal version of Ockham's Razor: we want to explain the data well, but we want parsimonious models. But how do we arrive at this precise formula?
- Factor of two is there so that it becomes mean squared error for a Gaussian model.

AIC Derivation

- Let model by summarized by parameter vector θ
- Let risk be negative log likelihood:

$$-R\left(\theta\right) = -E\left(\ell\ell\left(x|\theta\right)\right)$$

- Taylor expansion of *true* risk (negative expected log likelihood) around the *true* parameter vector θ^*:

$$\begin{aligned} -R\left(\hat{\theta}\right) &= E\left[\ell\ell\left(x|\theta^*\right)\right] + \left(\hat{\theta}-\theta^*\right) E\left[\nabla\ell\ell\left(x|\theta^*\right)\right] \\ &\quad + \frac{1}{2}\left(\hat{\theta}-\theta^*\right)' E\left[\nabla\nabla\ell\ell\left(\theta^*\right)\right]\left(\hat{\theta}-\theta^*\right) \\ &= E\left[\ell\ell\left(x|\theta^*\right)\right] + \frac{1}{2}\left(\hat{\theta}-\theta^*\right)' E\left[\nabla\nabla\ell\ell\left(x|\theta^*\right)\right]\left(\hat{\theta}-\theta^*\right), \end{aligned}$$

where $E\left[\nabla\ell\ell\left(\theta^*\right)\right] = 0$ is the identification condition.

AIC Derivation

- Empirical risk:

$$
\begin{aligned}
-\hat{R}\left(\hat{\theta}\right) &= E\left[N^{-1}\sum_i \ell\ell\left(x_i|\theta^*\right)\right] + \left(\hat{\theta}-\theta^*\right) E\left[N^{-1}\sum_i \nabla\ell\ell\left(x_i|\theta^*\right)\right] \\
&\quad + \frac{1}{2}\left(\hat{\theta}-\theta^*\right)' E\left[N^{-1}\sum_i \nabla\nabla\ell\ell\left(x_i|\theta^*\right)\right]\left(\hat{\theta}-\theta^*\right)
\end{aligned}
$$

- The first term is just the expected likelihood:

$$
E\left[N^{-1}\sum_i \ell\ell\left(x_i|\theta^*\right)\right] = E\left[\ell\ell\left(x|\theta^*\right)\right].
$$

AIC Derivation

- Empirical risk:

$$
\begin{aligned}
-\hat{R}\left(\hat{\theta}\right) &= E\left[N^{-1}\sum_i \ell\ell\left(x_i|\theta^*\right)\right] + \left(\hat{\theta}-\theta^*\right) E\left[N^{-1}\sum_i \nabla\ell\ell\left(x_i|\theta^*\right)\right] \\
&\quad + \frac{1}{2}\left(\hat{\theta}-\theta^*\right)' E\left[N^{-1}\sum_i \nabla\nabla\ell\ell\left(x_i|\theta^*\right)\right]\left(\hat{\theta}-\theta^*\right)
\end{aligned}
$$

- Looking at the second term:

$$
\sum_i \nabla\ell\ell\left(x_i|\theta^*\right) = \sum_i \left(\nabla\ell\ell\left(x_i|\theta^*\right) - \nabla\ell\ell\left(x_i|\hat{\theta}\right)\right) .
$$

Because $\hat{\theta}$ is selected to minimize empirical risk,

$$
\sum_i \left(\nabla\ell\ell\left(x_i|\hat{\theta}\right)\right) = 0
$$

AIC Derivation

- Empirical risk:

$$
\begin{aligned}
-\hat{R}\left(\hat{\theta}\right) &= E\left[N^{-1}\sum_i \ell\ell\left(x_i|\theta^*\right)\right] + \left(\hat{\theta}-\theta^*\right) E\left[N^{-1}\sum_i \nabla\ell\ell\left(x_i|\theta^*\right)\right] \\
&\quad + \frac{1}{2}\left(\hat{\theta}-\theta^*\right)' E\left[N^{-1}\sum_i \nabla\nabla\ell\ell\left(x_i|\theta^*\right)\right]\left(\hat{\theta}-\theta^*\right)
\end{aligned}
$$

- Also,

$$
\sum_i \left(\nabla\ell\ell\left(x_i|\theta^*\right) - \nabla\ell\ell\left(x_i|\hat{\theta}\right)\right) \approx \sum_i \nabla\nabla\ell\ell\left(x_i|\theta^*\right)\left(\theta^*-\hat{\theta}\right)
$$

- We can use $\sum_i \nabla\nabla\ell\ell\left(x_i|\theta^*\right)\left(\theta^*-\hat{\theta}\right)$ to substitute for $\sum_i \nabla\ell\ell\left(x_i|\theta^*\right)$ in the original expression.

AIC Derivation

- Empirical risk:

$$\begin{aligned} -\hat{R}\left(\hat{\theta}\right) &= E\left[N^{-1}\sum_i \ell\ell\left(x_i|\theta^*\right)\right] + \left(\hat{\theta}-\theta^*\right) E\left[N^{-1}\sum_i \nabla\ell\ell\left(x_i|\theta^*\right)\right] \\ &\quad + \frac{1}{2}\left(\hat{\theta}-\theta^*\right)' E\left[N^{-1}\sum_i \nabla\nabla\ell\ell\left(x_i|\theta^*\right)\right]\left(\hat{\theta}-\theta^*\right) \end{aligned}$$

- The second term becomes (approximately)

$$-\left(\hat{\theta}-\theta^*\right)' E\left[N^{-1}\sum_i \nabla\nabla\ell\ell\left(x_i|\theta^*\right)\right]\left(\hat{\theta}-\theta^*\right),$$

  which partially cancels with the third term.

AIC Derivation

- Empirical risk:

$$\begin{aligned} -\hat{R}\left(\hat{\theta}\right) &= E\left[N^{-1}\sum_i \ell\ell\left(x_i|\theta^*\right)\right] + \left(\hat{\theta}-\theta^*\right) E\left[N^{-1}\sum_i \nabla\ell\ell\left(x_i|\theta^*\right)\right] \\ &\quad + \frac{1}{2}\left(\hat{\theta}-\theta^*\right)' E\left[N^{-1}\sum_i \nabla\nabla\ell\ell\left(x_i|\theta^*\right)\right]\left(\hat{\theta}-\theta^*\right) \end{aligned}$$

- We can rewrite this as:

$$\begin{aligned} -\hat{R}\left(\hat{\theta}\right) &= E\left[\ell\ell\left(x|\theta^*\right)\right] \\ &\quad - \frac{1}{2}\left(\hat{\theta}-\theta^*\right)' E\left[N^{-1}\sum_i \nabla\nabla\ell\ell\left(x_i|\theta^*\right)\right]\left(\hat{\theta}-\theta^*\right) \end{aligned}$$

AIC Derivation

- $-E\left[\nabla\nabla\ell\ell\left(x|\theta^*\right)\right]$ equals the Fisher information matrix, typically written $I\left(\theta^*\right)$.
- Our final expression for expected empirical risk:

$$-\hat{R}\left(\hat{\theta}\right) \approx E\left[\ell\ell\left(x|\theta^*\right)\right] + \frac{1}{2}\left(\hat{\theta}-\theta^*\right)' E\left[I\left(\theta^*\right)\right]\left(\hat{\theta}-\theta^*\right)$$

AIC Derivation

- True risk:

$$-R\left(\hat{\theta}\right) \approx E\left[\ell\ell\left(x|\theta^*\right)\right] - \frac{1}{2}\left(\hat{\theta}-\theta^*\right)' E\left[I\left(\theta^*\right)\right]\left(\hat{\theta}-\theta^*\right)$$

- Expected empirical risk:

$$-\hat{R}\left(\hat{\theta}\right) \approx E\left[\ell\ell\left(x|\theta^*\right)\right] + \frac{1}{2}\left(\hat{\theta}-\theta^*\right)' E\left[I\left(\theta^*\right)\right]\left(\hat{\theta}-\theta^*\right)$$

- The expected difference between true and empirical risk:

$$N\left(R\left(\hat{\theta}\right) - \hat{R}\left(\hat{\theta}\right)\right) \approx N\left(\hat{\theta}-\theta^*\right)' E\left[I\left(\theta^*\right)\right]\left(\hat{\theta}-\theta^*\right)$$

From the asymptotic distribution of the MLE estimator, this should have a Chi-squared distribution with degrees of freedom equal to the number of parameters (k). This has expectation k.

AIC Derivation

- The expected difference between true and empirical risk:

$$N\left(R\left(\hat{\theta}\right)-\hat{R}\left(\hat{\theta}\right)\right)\approx N\left(\hat{\theta}-\theta^{*}\right)^{\prime}E\left[I\left(\theta^{*}\right)\right]\left(\hat{\theta}-\theta^{*}\right)$$

  From the asymptotic distribution of the MLE estimator, this should have a Chi-squared distribution with degrees of freedom equal to the number of parameters (k). This has expectation k.
- Thus, the empirical risk underestimates the true risk by k, the number of parameters.
- The negative log likelihood (risk) plus number parameters is therefore an unbiased estimate of the true risk.
- This suggests using $-\sum_i \ell\ell\left(\hat{\theta}|\mathbf{x}_i\right)+k$ to select models – by selecting the model with the lowest value, you're selecting the model with the lowest expected true risk. Equivalently, use

$$AIC\equiv 2k-2\sum_i \ell\ell\left(\hat{\theta}|\mathbf{x}_i\right)$$

AIC Summary

- There's a good reason to penalize models with lots of parameters; doing so actually leads to selection of models that make better predictions.
- In other words, we want to avoid over-fitting.

AIC: implementation difficulty

- Suppose you have a LOT of x variables that you might include in the model. You could, in principle, use the AIC to consider a model with every possible subset of the variables, but that would be computationally cumbersome.
- Some machine learning tools effectively automate this process.

LASSO

- The LASSO estimator solves

$$\min_{\beta} \left\{ \sum_i \left(y_i - \beta' \mathbf{x}_i \right)^2 \right\} \qquad s.t. \qquad \sum_j |\beta_k| \leq K_{max}$$

- This minimization problem can typically be solved fairly quickly, and LASSO-based predictions tend to avoid overfitting, much like models selected with the AIC. Unsurprisingly, it has become quite popular recently.
- Selecting the K_{max} parameter is important. The AIC is one way to do this.

Decision Trees

- Again, suppose you have a LOT of x variables
- Choose one of your x variables. Consider splitting the data in two groups using that variable within each group, set your prediction of y to minimize risk within each group. Search for the cutoff that minimizes overall risk.
- Within each of the sub-samples created by your first step, choose another x variable and repeat the procedure. If you repeat this k times, you have a k-level decision tree.
- How to choose which variables to look at for each step in your decision tree? **Random forests** make these selections randomly.