Lecture 15:
Overfit & Cross Validation

Big Data and Machine Learning for Applied Economics
Econ 4676

Ignacio Sarmiento-Barbieri

Universidad de los Andes

September 28, 2021

Announcements

- Problem Set 3 deadline change to October 15. You **should not** work on the break.
- Problem Sets 4, 5, and 6 are up.
- Final Project
  - First deadline. October 22.
    - This submission should be a brief statement of what you plan to do.
    - Max 2 pages.
  - Second deadline. Presentations. November 30th, December 2nd, and 3rd.
  - Final document. December 10th at 6 pm

Agenda

1. Recap
2. Overfit
   - Overfit and out of Sample Prediction
3. Resampling Methods
   - Validation Set Approach
   - LOOCV
   - K-fold Cross-Validation
4. Further Readings

Linear Models

- Last class we talked about the linear model

$$y = X\beta + u \tag{1}$$

- We focused on predicting y and the connection to classical econometrics
- In the classical view, the linearity is given and the notion of complexity is somehow easily defined.

General Models

- However, we can think about more general models

$$y = f(X) + u \tag{2}$$

- where f is some fixed but unknown function of X and u is a mean-zero error which is independent of x
- The emphasis here is that this is a statistical model. There is nothing causal in it. In the model, f represents the systematic relationship between X and y and u represents idiosyncratic deviations from this systematic relationship.
- Note that:
  - f is not restricted in any way – it can be a completely arbitrary and complex function.
  - Predicting y involves *learning* f, that is, f is no longer taken as given, as in the classical view.
  - It implies an iterative process where initial choices for f are revised in light of potential improvements in predictive performance.

Supervised Learning

$$y = f(X) + u \tag{3}$$

- The problem of learning y based on features X is known as a supervised learning task.
- y supervises the learning.
- For prediction we don't care about f itself. We can treat it as a black box, and any approximation $\hat{f}$ that yields a good prediction is good enough.

Whatever works, works.

Supervised Learning

- What is 'what works', i.e., what is a good prediction?
- Formally, a supervised learning algorithm takes as an input a loss function and searches for a function $\hat{f}$ within a function class $\mathcal{F}$ that has a low expected prediction loss
- The goal here is to solve something which looks like

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \{E\left(L(y, f(X))\right\} \tag{4}$$

  for some loss function L, and for some set of predictors $\mathcal{F}$.
- Note that here in a function space, so we are solving for a function not a point.

Error Decomposition

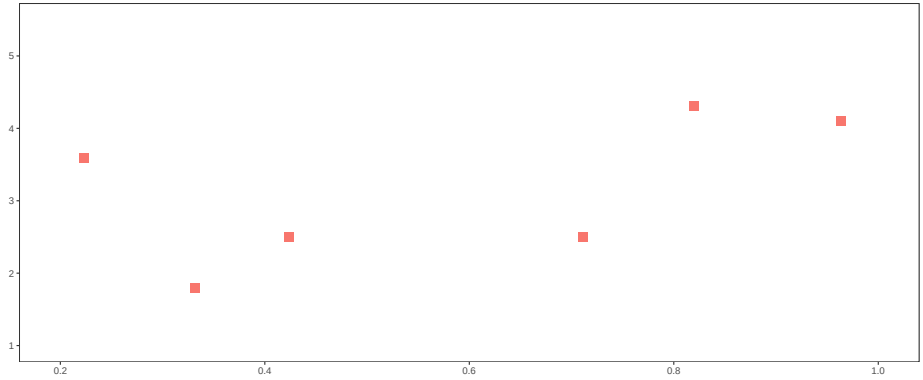
- The most common function is the squared error loss $\rightarrow$ MSE

$$
\begin{aligned}
E\left(L(y,\hat{y})\right) &= E(y-\hat{f})^2 && (5)\\
&= E(f-\hat{f})^2 + E(u^2) && (6)\\
&= \underbrace{MSE(\hat{f})}_{\textit{Reducible error}} + \underbrace{E(u^2)}_{\textit{Irreducible error}} && (7)
\end{aligned}
$$

- There is a lower bound to the accuracy of the best possible prediction for a given f, given by the variance of the idiosyncratic error.

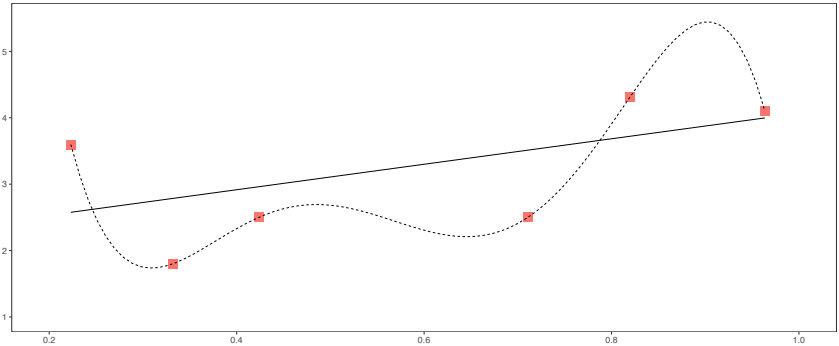
Overfit

- Why don't we estimate the most flexible function that we can?



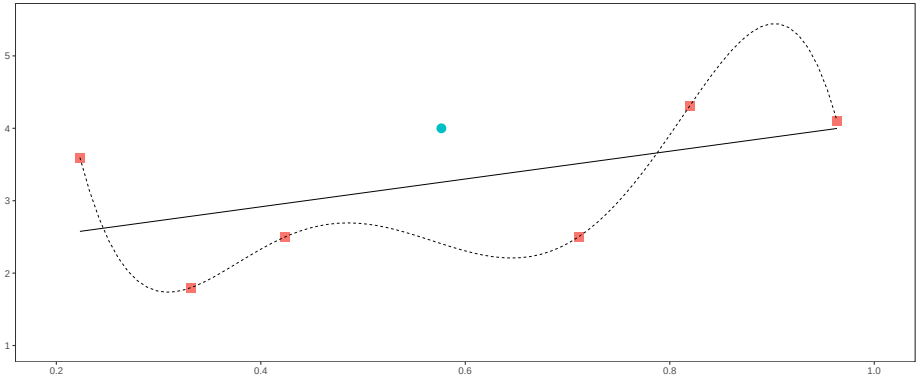
Overfit

- Why don't we estimate the most flexible function that we can?



Overfit

- In practice, it doesn't work $\rightarrow$ leads to a terrible out-of-sample fit.
- A deeper answer says that maximal flexibility leaves all the idiosyncratic noise in the prediction model. A new observation with the same x will have a different idiosyncratic noise, and so the prediction is off.



Overfit and out of Sample Prediction

- ML we care about prediction out of sample
- To avoid overfitting, we need to find find the optimal degree of flexibility.
- Overfitting is an illustration of the bias-variance trade-off

$$
\begin{aligned}
E\left(L(y,\hat{f}\right) &= E(y-\hat{f})^2 && (8)\\
&= E(f-\hat{f})^2 + E(u^2) && (9)\\
&= MSE(\hat{f}) + E(u^2) && (10)\\
&= E(f-\hat{f})^2 + E(\hat{f}-E(\hat{f}))^2 + E(u^2) && (11)\\
&= Bias^2(\hat{f}) + V(\hat{f}) + E(u^2) && (12)
\end{aligned}
$$

- Flexibility reduces bias but increases variance. There is an optimal degree of flexibility

Overfit and out of Sample Prediction

- Choose the right complexity level
- How do we measure the out of sample error? What is 'out of sample'?
- R^2 doesn't work: measures prediction in sample, it's non decreasing in complexity (PS1)

Overfit and out of Sample Prediction

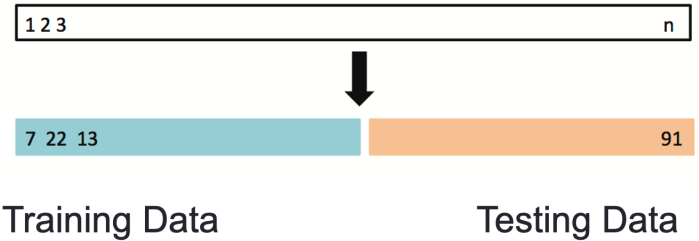
- Choose the right complexity level
- How do we measure the out of sample error? What is 'out of sample'?
- R^2 doesn't work: measures prediction in sample, it's non decreasing in complexity (PS1)
- MSE using resampling methods.

What are resampling methods?

- Tools that involves repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain more information about the fitted model
- Model Assessment: estimate test error rates
- Model Selection: select the appropriate level of model flexibility
- They are computationally expensive! But these days we have powerful computers

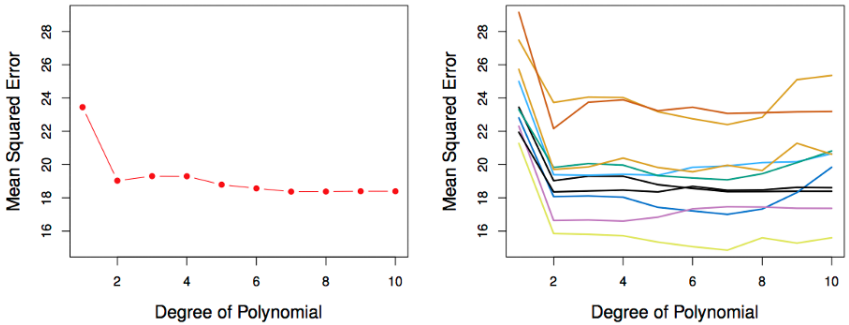
The Validation Set Approach

- Suppose that we would like to find a set of variables that give the lowest test (not training) error rate
- If we have a large data set, we can achieve this goal by randomly splitting the data into training and validation(testing) parts
- We would then use the training part to build each possible model (i.e. the different combinations of variables) and choose the model that gave the lowest error rate when applied to the validation data



The Validation Set Approach

- Model $y = f(x) + u$ where f is a polynomial of degree p^*.
- Left: Validation error rate for a single split
- Right: Validation method repeated 10 times, each time the split is done randomly!
- There is a lot of variability among the MSE's... Not good! We need more stable methods!

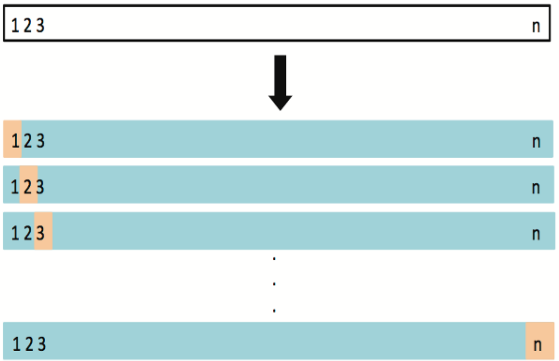


The Validation Set Approach

- Advantages:
  - Simple
  - Easy to implement
  - It may be sensible with really big n
- Disadvantages:
  - The validation MSE can be highly variable, i.e. , is highly dependent on which observations are included in the testing sample.
  - Statistical methods tend to perform worse when trained on fewer observations. They have poor small-sample properties; they perform better with increasing sample size. This leads us to overestimate the test MSE.

Leave-One-Out Cross Validation (LOOCV)

- This method is similar to the Validation Set Approach, but it tries to address the latter's disadvantages



Leave-One-Out Cross Validation (LOOCV)

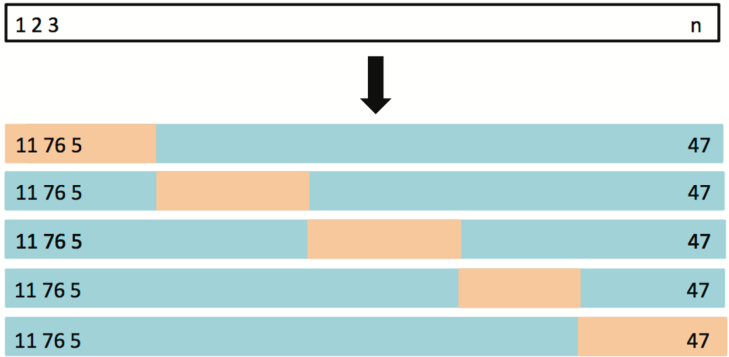
- Use a single observation for validation and (n - 1) observations for training.
- Fit the model leaving-one-out observation $\rightarrow \hat{y}_{-i}$.
- Cycle though all n observations
- The LOOCV estimate for the test MSE is

$$CV(n) = n \sum MSE_{-i} \tag{13}$$

$$= \frac{1}{n} \sum (y_i - \hat{y}_{-i})^2 \tag{14}$$

K-fold Cross-Validation

- LOOCV is computationally intensive, so we can run k-fold Cross Validation instead



K-fold Cross-Validation

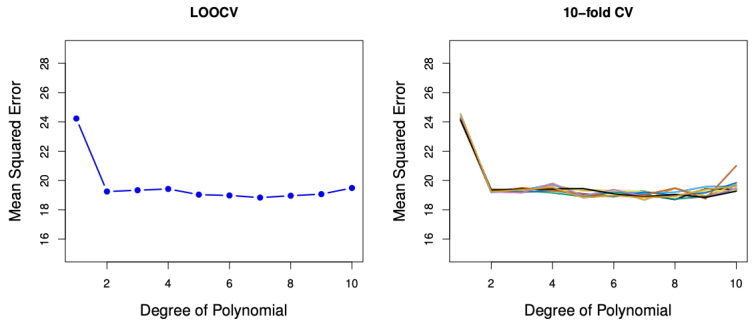
- Split the data into K parts ($n = \sum_{j=1}^{k} n_j$)
- Fit the model leaving out one of the folds $\rightarrow \hat{y}_{-k}$
- Cycle though all k folds
- The CV(k) estimate for the test MSE is

$$CV_{(k)} = \frac{1}{k} \sum_{j=1}^{k} MSE_j \tag{15}$$

$$= \frac{1}{k} \sum_{j=1}^{k} (y_j^k - \hat{y}_{-k}) \tag{16}$$

K-fold Cross-Validation

- Left: LOOCV error curve
- Right: 10-fold CV was run many times, and the figure shows the slightly different CV error rates
- LOOCV is a special case of k-fold, where k = n
- They are both stable, but LOOCV (generally) is more computationally intensive!



Bias- Variance Trade-off for k-fold CV

- Bias:
  - Validation set approach tends to overestimate the test error set (less data, worst fit)
  - LOOCV, adds more data $\rightarrow$ less bias of the test error
  - K-fold an intermediate state
- Variance:
  - LOOCV we average the outputs of n fitted models, each is trained in almost identical set of observations $\rightarrow$ highly (positively) correlated
  - K-fold this correlation is smaller, we are averaging the output of k fitted model that are somewhat less correlated
- Thus, there is a trade-off between what to use
  - We tend to use k-fold CV with (K = 5 and K = 10)
  - It has been empirically shown that they yield test error rate estimates that suffer neither from excessively high bias, nor from very high variance Kohavi (1995)

Spatial K-fold Cross-Validation

- 'First law' of geography states that points close to each other are, generally, more similar than points further away
- Points are not statistically independent because training and test points in conventional CV are often too close to each other
- To alleviate this problem 'spatial partitioning' is used to split the observations into spatially disjointed subsets

Review & Next Steps

- Today:
  - Overfit and out of Sample Prediction
  - Resampling Methods
    - Validation Set Approach
    - LOOCV
    - K-fold Cross-Validation
- Next class: Model selection and Regularization

Further Readings

- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).
- Lovelace, R., Nowosad, J., & Muenchow, J. (2019). Geocomputation with R. CRC Press. (Chapters 2 & 6)