Predictive Analytics Lecture 5

Adam Kapelner

Stat 422/722
at The Wharton School of the University of Pennsylvania

February 14 & 15, 2017

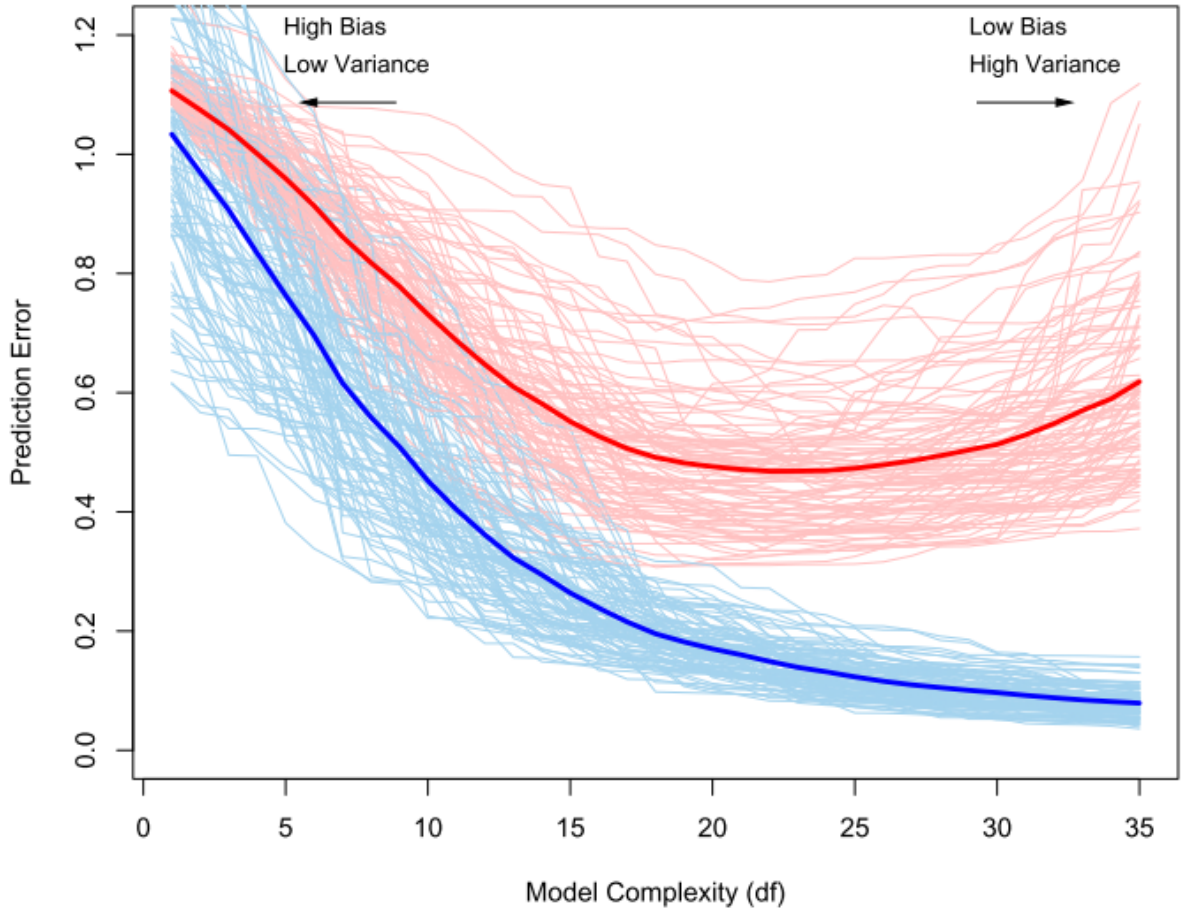
Underfitting & Overfitting

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

Goal of machine learning: fit f as best as possible. When we build models, we do one (or both of the following)

- We fall **short** by underfitting (usually due to too little degrees of freedom and inflexible bases). For example: if the f is a curve and we fit a line, we underfit (recall medicorp sales vs bonus regression).
- We can shoot too **long** by encroaching on and fitting / optimizing to the $\mathcal{E}$. Since $\mathcal{E}$ is independent of $x_1, \ldots, x_p$, this part of $\hat{f}$ is essentially a random fit and it is the opposite of the "data-driven approach".

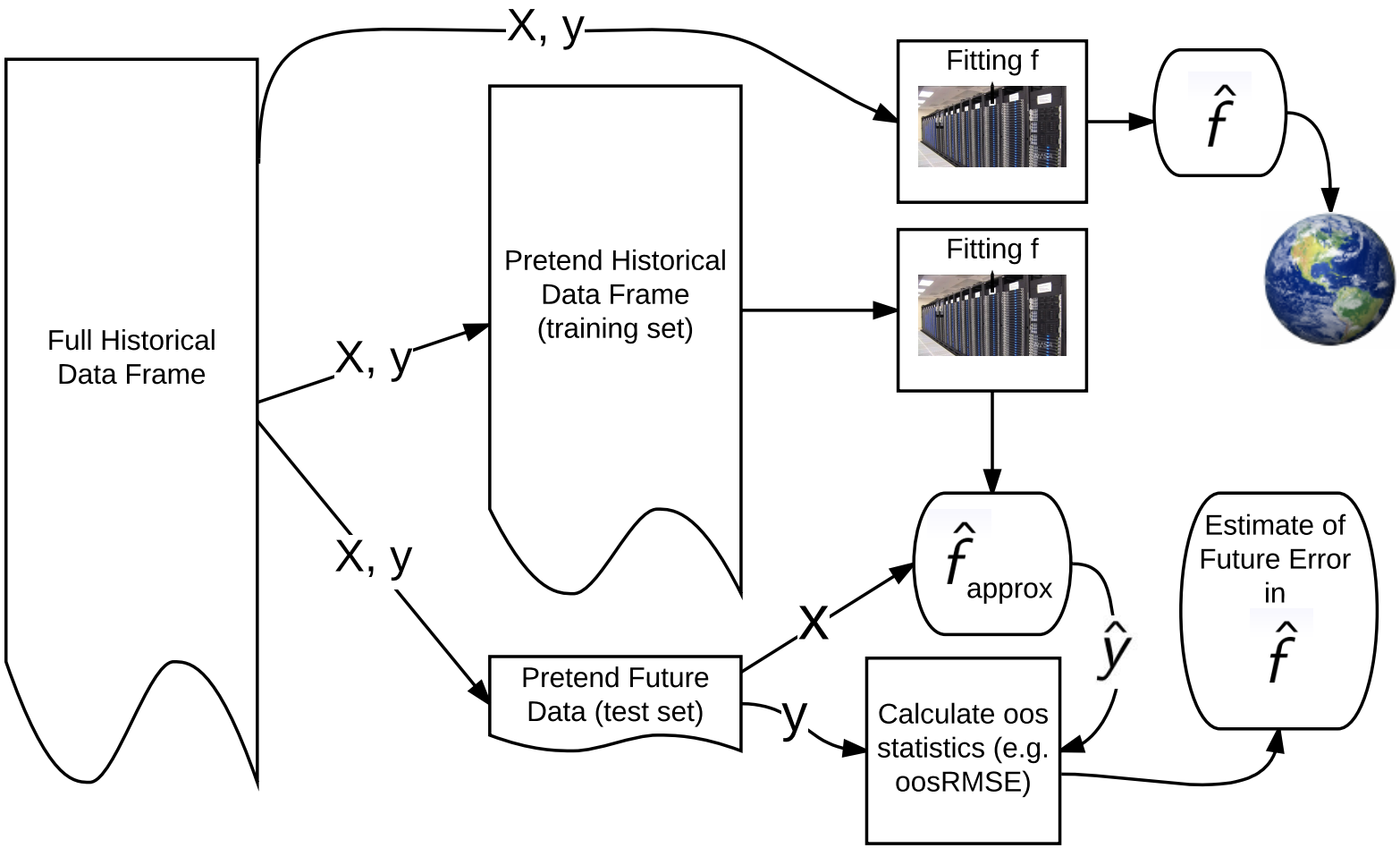
Complexity-Fit Tradeoff



Blue is in-sample fit metric and red is oos fit metric. This is Fig 7.1 from Hastie and Tibsharani (2009).

Assessment: OOS Validation

But knowing where you are on that y-axis would involve knowing the truth. We need to estimate this, so we use oos validation:



Assumptions and Tradeoffs when Splitting

We have a choice to split our dataframe into two pieces. Assuming each data point is independent (the running assumption), you should do this completely randomly. When would this assumption not be true? For example, a time series.

Additionally, we need a non-stationary model relationship. So,

$$Y = f(x_1, \ldots, x_p) + \mathcal{E} \quad \text{and not} \quad Y = f_t(x_1, \ldots, x_p) + \mathcal{E}$$

where f changes with time. In essence non-stationarity is a lack of generalization and when predicting, it is a form of extrapolation.

How large should the test set be? Usual sizes are 10-30%. What's the tradeoff? If the test set is larger, then ...

1. the more accurate the assessment of generalization error would be (less variance) and
2. the less accurate the model will be since it's fitting with less data (more bias)

If the test set is smaller then, vice versa. Note: the in-sample and oos statistics are statistics! Thus, they are random!

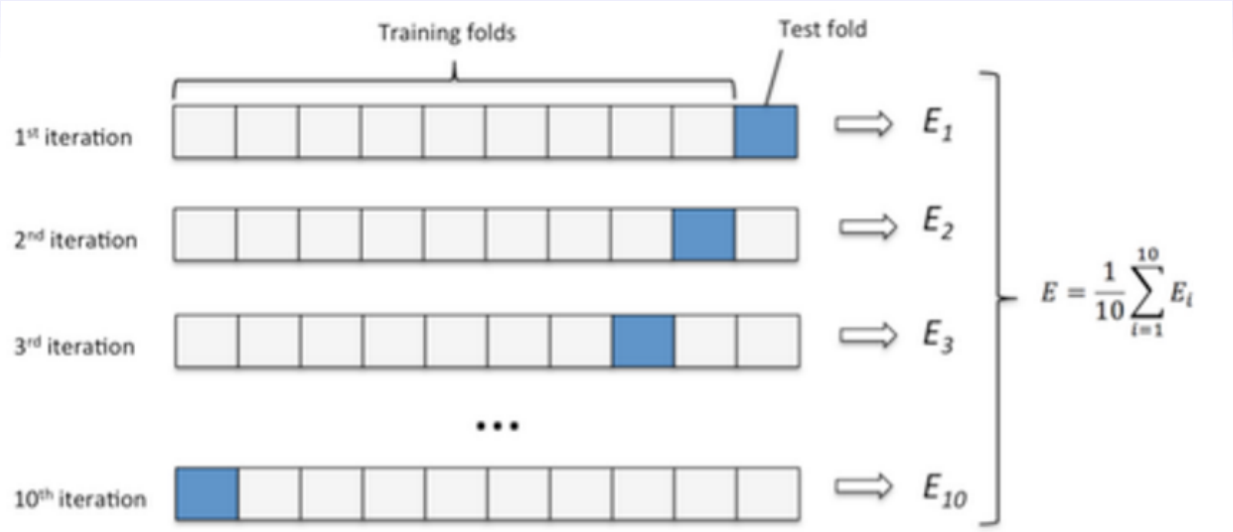
Less Randomness in the OOS Statistics

If we change the observations in the training/test splits, we will get different models and different estimates of future error. Thus, our oosRMSE was really oosRMSE conditional on the idiosyncratic split we happened to get.

We can at the very least... get rid of this idiosyncratic error by ... averaging over all training-test splits. If we have $n = 100$ and the test set is 10%, that means we only have $\binom{100}{10} = 1.73 \times 10^{13}$ split configurations to average over!

We can approximate the averaging over all splits by just taking $100\%/10\% = 10$ random but unique splits called **folds**. Thus, each observation is represented in the test set once. This is known as **K-fold cross validation (CV)** where here $K = 100\%/10\% = 10$ (and this procedure seems to be the industry standard).

10-fold CV



$K = 10$ is arbitrary but it is based on the 10-30% test set proportion. In practice, I've only used 5 or 10 fold CV.

This does not really solve any of our big problems but gives us a little boost in terms of a reduction in standard error of our generalization error estimate. That's OK; we can take all the help we can get if it's costless!!

Note 1: If $K = n$ then we the test set as one sample; this is known as "leave on out CV" (LOOCV) and it not generally recommended — super high variance!
Note 2: bleeding edge of stats — find CI's for generalization error.

Validating Multiple Models

Let's look at three models for the White Wine data. Here the response is wine quality as measured by professional raters and features are 11 features (e.g. acidity, sugar, pH and alcohol content).

- A plain linear model
- B six-degree polynomials for all features
- C six-degree polynomials and all interactions up to 2nd degree (AKA 1st order)
- D six-degree polynomials and all interactions up to 3rd degree (AKA 2nd order)
- E six-degree polynomials and all interactions up to 4th degree (AKA 3rd order)
- F six-degree polynomials and all interactions up to 11th order

[JMP col validation... fit all models with validation ... save prediction formula cols... analyze model... model comparison] Conclusions? Model C looks the best. Note: another popular assessment metric besides oosRMSE is oosAAE which is just average absoute value difference. Strange ... given that linear models optimize for squared error. What did I do that wasn't legal?

A Possible Spin on Validation

Recall the proposal from last class:

1. Split dataframe into training and test.
2. Build model A on training.
3. Predict using the test set.
4. Calculate estimate of future generalization error of model 1.
5. Build a different model B on training.
6. Predict using the test set.
7. Calculate estimate of future generalization error of model 2.
8. ... steps 5-7 for model 3
9. ... steps 5-7 for model 4
10. ...
11. ... steps 5-7 for model M
12. Pick whichever model has better generalization error.

What was wrong with is?

Looking into the Future is Not Legal

The oos validation is only valid if...

you treat the test set as a lockbox. Once you open it up, that's it! And we opened it up M times!

What to do?