Bias-variance trade-off

We can call it the bias-generality tradeoff

Terence Parr
MSDS program
University of San Francisco

UNIVERSITY OF SAN FRANCISCO

Many poor descriptions of this concept on web

(and with highest variance of definitions 🤪)

- For example, Wikipedia starts a paragraph with "*Models with high bias are usually more complex*" and finishes that same paragraph with "…*models with higher bias tend to be relatively simple*…" (the latter bit is correct)
- This blog is pretty good: https://elitedatascience.com/bias-variance-tradeoff
- When you hear "*bias-variance*," think "*bias-generality*"
- It's a trade-off because increasing accuracy (reducing bias) usually means reducing generality (and vice versa)

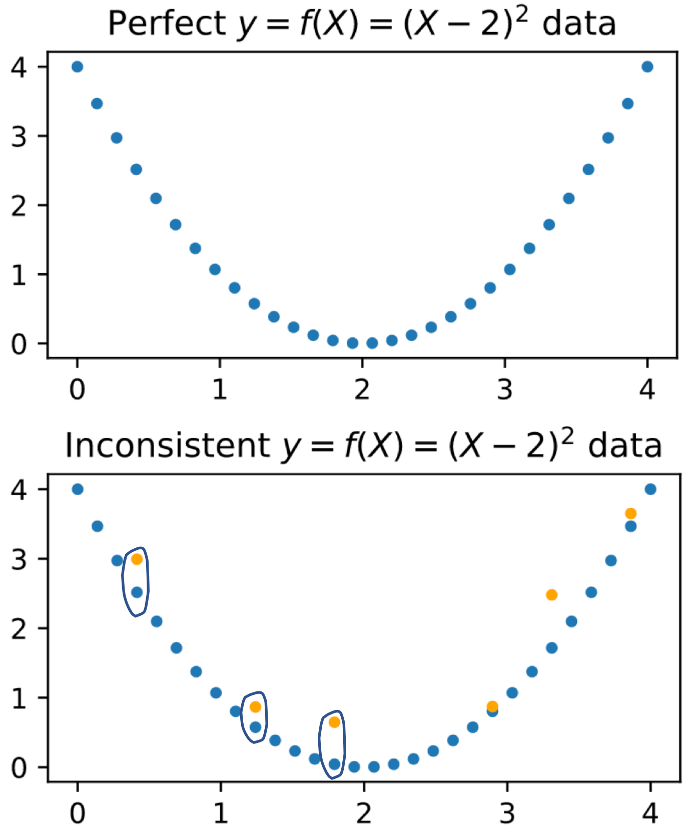
https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff

Sources of prediction error

- We're given (X, y) training data and we fit a model $\hat{f}(X)$
- Test error *Err* = $\left(\hat{f}(x_0) - y_0\right)^2$ from single (x_0, y_0) test case
- There are 3 sources of errors in that *Err* number:
  1. *noisy* data, such as inconsistent X → y data
  2. model *underfitting* or *bias*; too weak or simple; doesn't capture X → y
  3. model *overfitting*; model too specific to training data; not general
- Conceptually: *Err = "noise"* + *"bias"* + *"overfitting"*
- Stats nerds use *Err* = *Irreducible Error* + *Bias^2* + *Variance*
- Why they use "variance" will make sense shortly

1. *Noise* can lead to inconsistent data

- Noise can cause inconsistent training observations, such as:
  [18,1,9] → 91
  [18,1,9] → 99
- No model can predict two different y values for same x vector
- Pick mean or either y value; model will have *Err*>0 no matter what
- This is called the *irreducible error*
- Noise comes from faulty sensors, typos, self-reporting issues, etc…
- Nothing we can do about the irreducible error



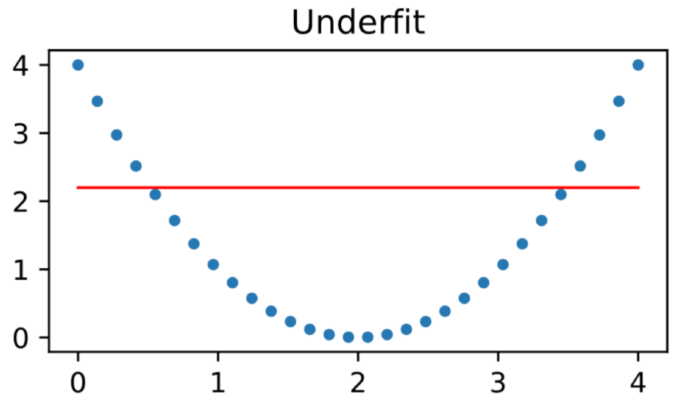
UNIVERSITY OF SAN FRANCISCO

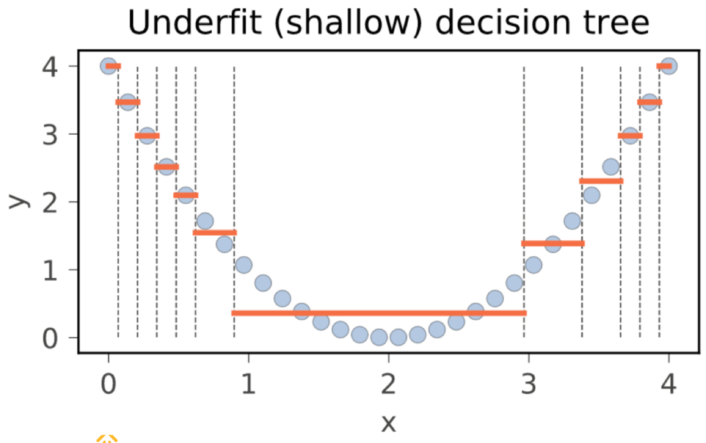
Missing variables looks like noise

- What if inconsistent training observations, such as:
  $[18,1,9] \rightarrow 91$
  $[18,1,9] \rightarrow 99$
  were really just missing a variable we don't have?
  $[18,1,9,10] \rightarrow 91$
  $[18,1,9,7] \rightarrow 99$
- E.g., two apartment observations look identical, say, 2 bedrooms & 1 bath but have very different prices; inconsistency only because we lack "square foot" or "awesome view" vars
- Missing vars are called *exogenous* vars (econ/finance term)

2. Underfitting leads to *biased* models

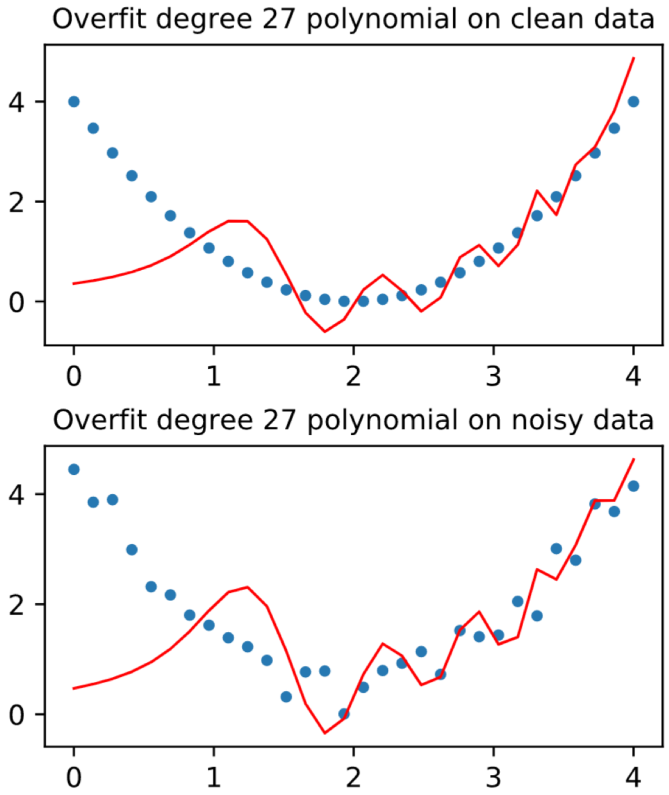
- Now take noise out of picture
- If our model is unable to capture $\mathrm{X} \rightarrow y$ well enough, model is *bias*ed, systematically under- or over-predicting
- Predicting with mean (line) for quadratic is too weak, as is a decision tree that is too shallow to partition x space well
- Increasing complexity of model will typically reduce the bias, increasing accuracy and reducing *Err*





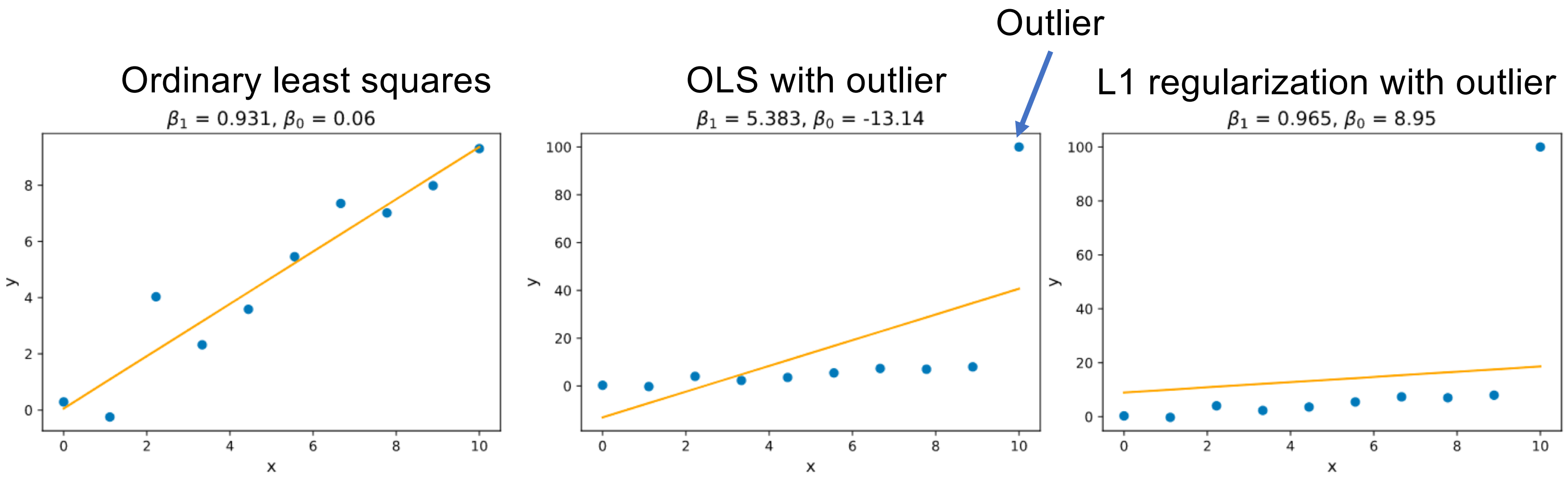
3. Overly-complex models can *overfit*

- Even without noise, models with too much power/flexibility for a training data set can be inaccurate (degree 27 polynomial here)
- We always have noise though, so complex models lead to overfitting not bias
- Model is overfit when it focuses on quirks/details/noise of a specific X, rather than getting the gist of training set X
- Getting the gist means capturing the nature of **distribution** behind X



Recall: even simple models can overfit

- Regularization trades a bit of bias for increased generality
- Below, we see two training sets; center panel has quirk (outlier) that causes OLS to get different model parameters; not general



Outlier

Ordinary least squares

β_1 = 0.931, β_0 = 0.06

OLS with outlier

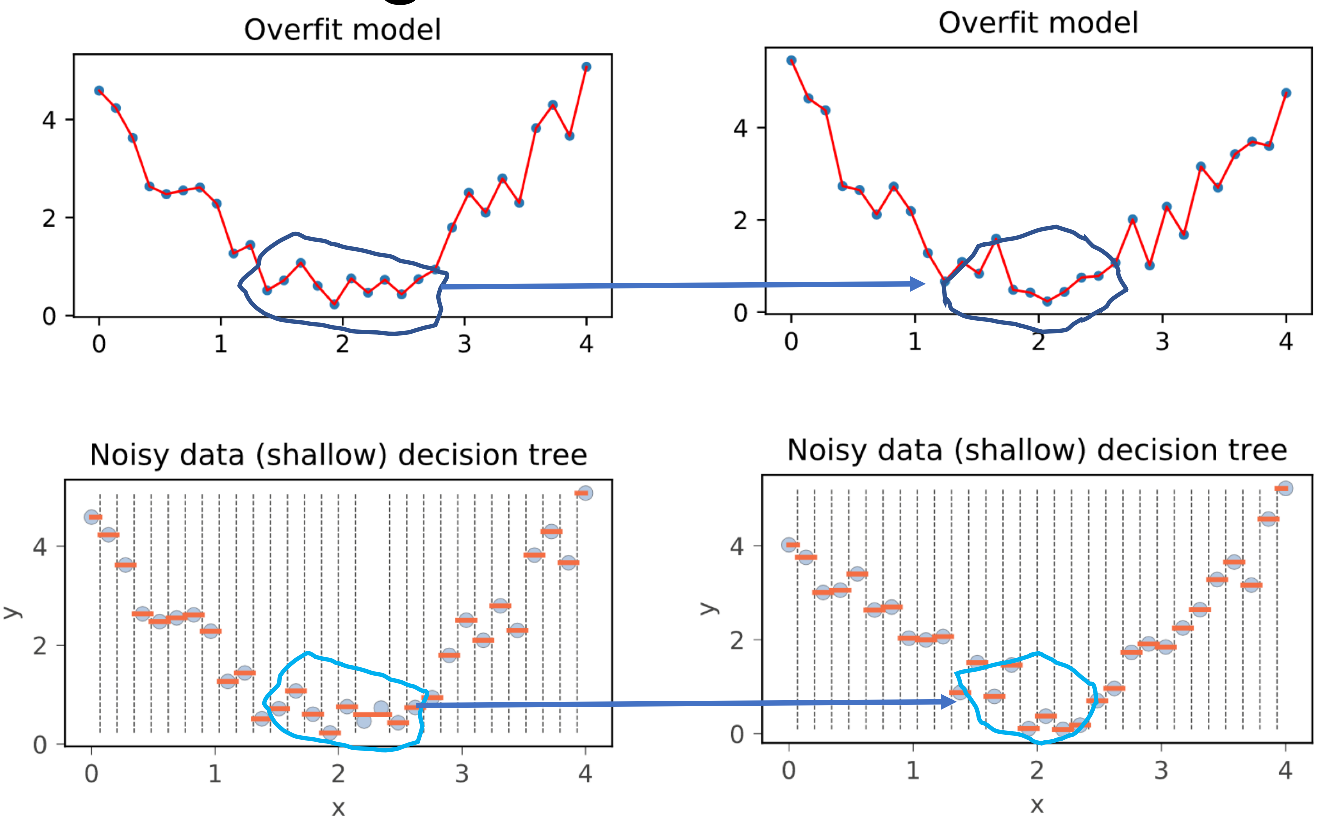
β_1 = 5.383, β_0 = -13.14

L1 regularization with outlier

β_1 = 0.965, β_0 = 8.95

Stats view of overfitting: *variance*

- Small changes in data lead to very different models
- Here are 2 training sets drawn from same distribution
- Same fitting strategy leads to different models
- Variance refers to model parameters not predictions, though predictions also vary



See https://github.com/parrt/msds621/blob/master/notebooks/trees/bias-variance.ipynb

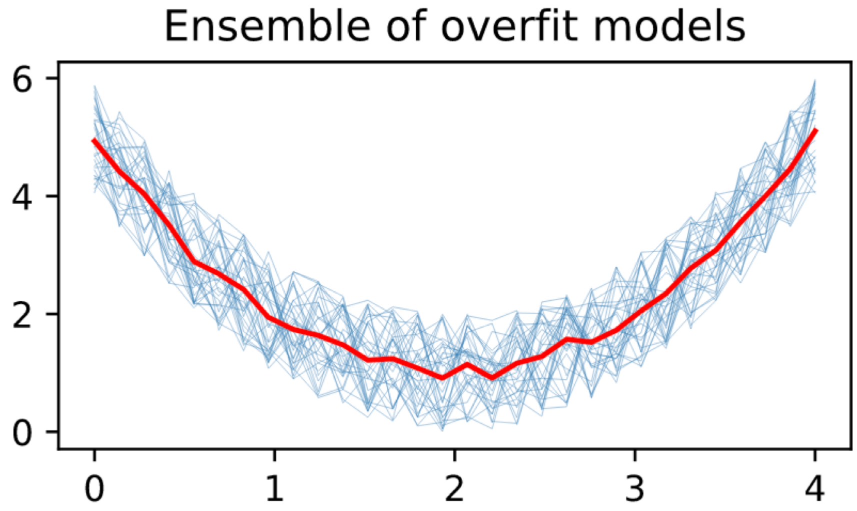
High model variance = overfitting

- The term “variance” is confusing because it refers to the variation of models (and hence prediction errors) trained on multiple data sets but we normally only have one training set. So it’s weird unless you’re really into boostrapping …
- Variance / overfitting leads to poor generality as model doesn’t capture underlying X distribution, which is necessary to make predictions for previously unseen x vectors; model will predict a noisy value as that’s how it was trained
- **Analogy**: multiple unbiased graders grading papers still have their own independent opinions that can vary, bouncing around the “true” score. But, on average the graders get accurate score if each grades each paper

What to do about high variance, overfitting?

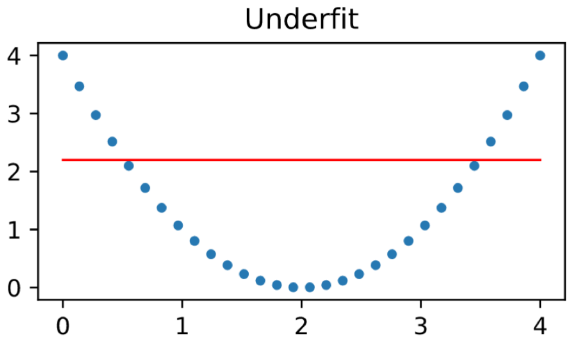
1. Simplify/regularize/restrict model
2. Average results from many overfit models (an *ensemble*)

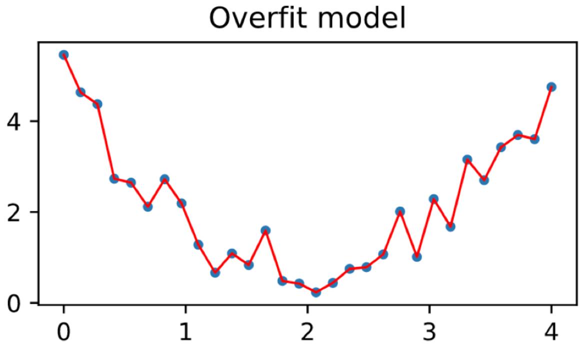
- Since X’s are from same distribution and independent (i.i.d.) here, average of many models should be accurate & with low variance
- Graph shows 35 models fit to noisy data from same distribution, averaged
- Random forests ensemble many overfit decision trees and use a trick to make the trees sort of independent (more in future lecture)

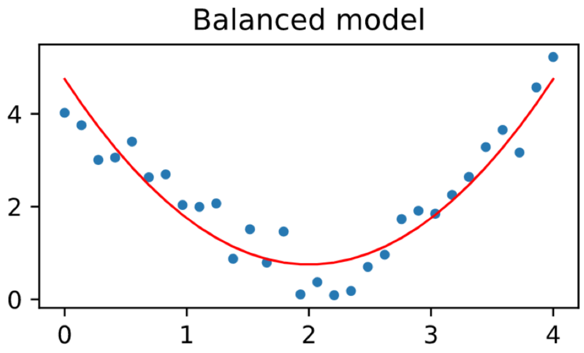


The trade-off

- Must increase the complexity of the model to get more accuracy
- But, increased complexity means more ability to chase quirks of data, making the model overly-specific to the training set
- E.g., decision trees are sensitive to small data changes; change in root split node propagates to all splits below root
- Let the validation error be your guide to appropriate complexity!

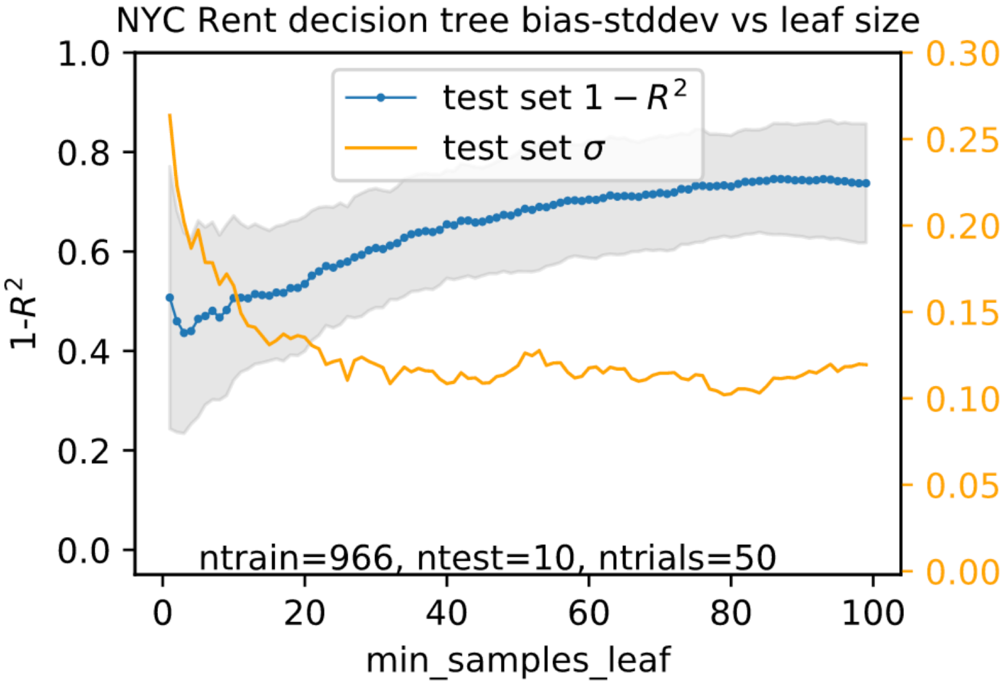






Regressor hold-out accuracy vs generality

- Consider decision tree trained on NYC rent data split into chunks to simulate multiple i.i.d. training sets; hold-out and a training set of 10 obs.
- As we increase leaf size, what is the effect on hold-out prediction error?
- How stable is the hold-out prediction error?

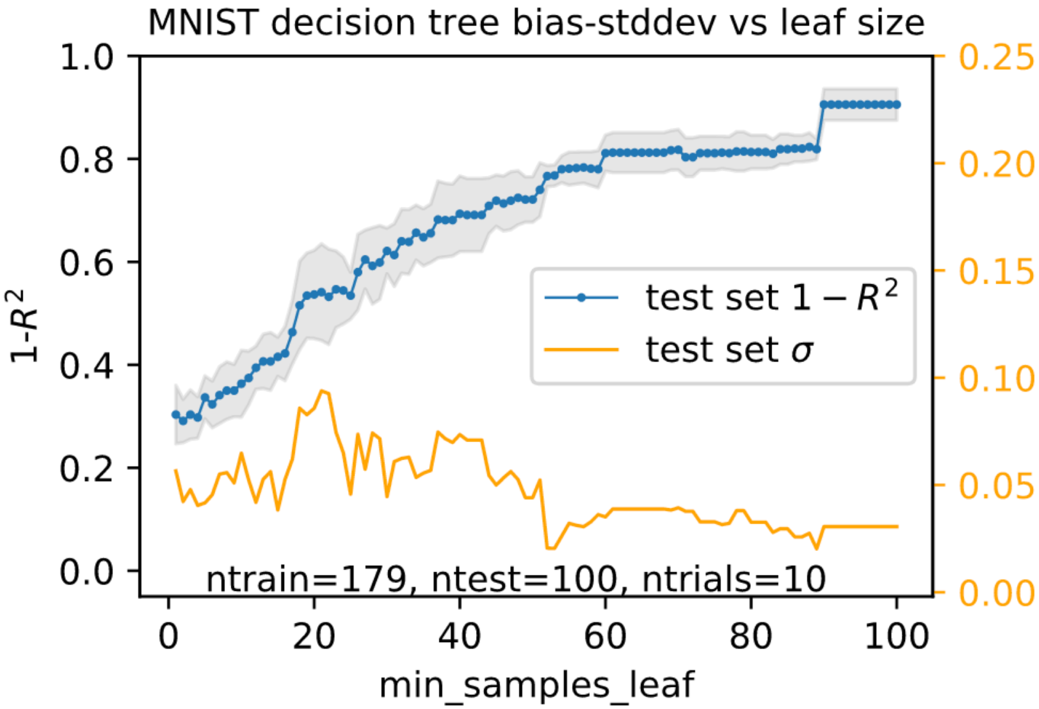


UNIVERSITY OF SAN FRANCISCO

Classifier hold-out accuracy vs generality

(500 records, 20 trials selecting and holding out 5% test set)

- Classifier bias increases as we restrict tree complexity, but variance of prediction error drops
- Trading bias for generality

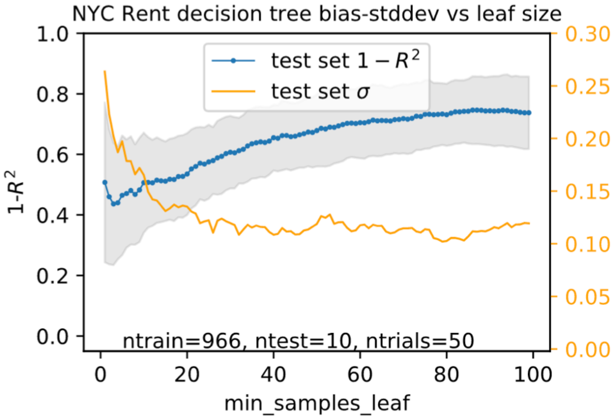


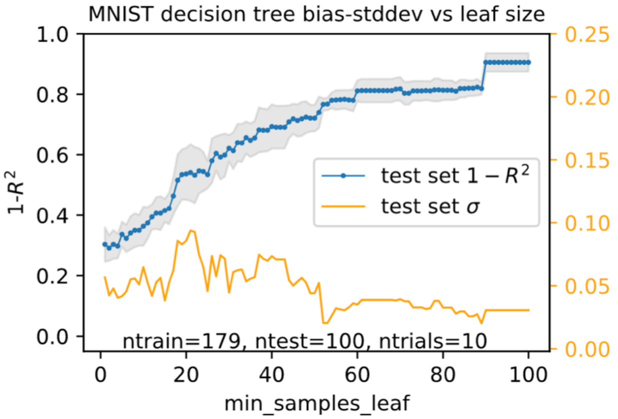
Sidenote: there's mounting evidence that deep neural networks can overfit (getting 0 training error) and still get decent generality. See: https://simons.berkeley.edu/talks/tbd-51

UNIVERSITY OF SAN FRANCISCO

Experimental details / subtleties

- Beware CLT! If test set is big, it can hide lots of variation: test error is **average** of individual prediction errors. Avg reduces var. and then we avg the avg error across trials
- With single training set, simulate multiple i.i.d. X by splitting into nonoverlapping chunks, reserving one chunk for fixed test region
  - Don't pick random subsets; that's boostrapping (i.d.) and not same as using multiple i.i.d. training sets
- We're using test error as proxy for model var.
- Why doesn't variance of test error go to zero?
  - Because chunks and test are not exactly i.d.
  - Some chunks & test set will be similar, others not


NYC Rent decision tree bias-stddev vs leaf size


MNIST decision tree bias-stddev vs leaf size

See https://github.com/parrt/msds621/blob/master/notebooks/trees/bias-variance.ipynb