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 high 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 sometimes vice versa)

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

Sources of prediction error

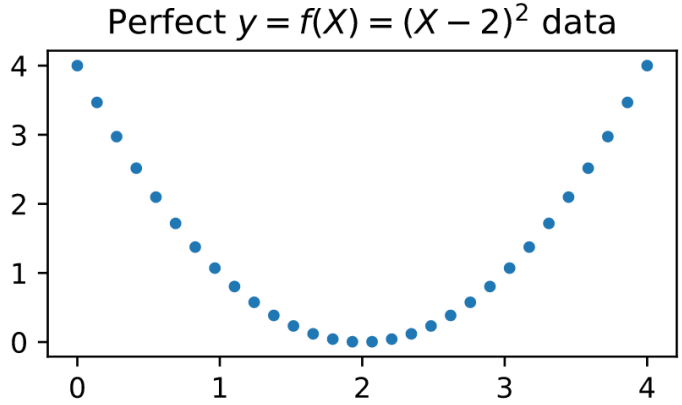
- We're given (X, y) training data and we fit a model $\hat{f}(X)$
- MSE error *Err* = $\left(f(x^{(i)}) - y^{(i)}\right)^2$ from single $(x^{(i)}, y^{(i)})$ test case
- There are 3 sources of errors in that *Err* number:
  1. *noisy* X or y data, such as inconsistent $X \rightarrow y$ data
  2. model *underfitting* or *bias*; too weak or simple; doesn't capture $X \rightarrow y$
  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

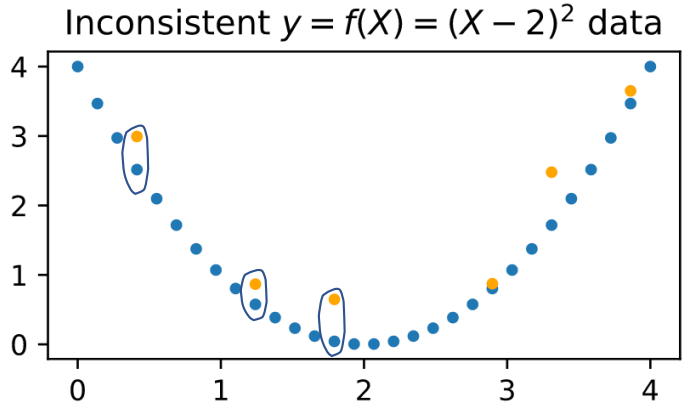
1. *Noise* can lead to inconsistent data

- Noise can cause inconsistent training observations, such as:

$$[18,1,9] \rightarrow 91$$
$$[18,1,9] \rightarrow 99$$

- 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





Missing variables looks like noise

- What if inconsistent training observations, such as:

$$[18,1,9] \rightarrow 91$$
$$[18,1,9] \rightarrow 99$$

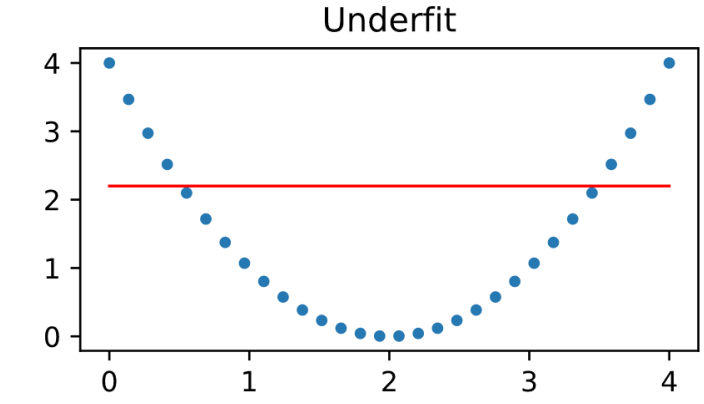
  were really just missing a variable we don't have?

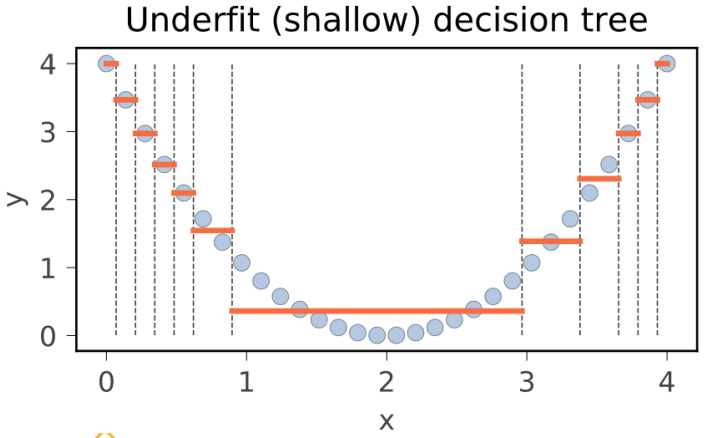
$$[18,1,9,10] \rightarrow 91$$
$$[18,1,9,7] \rightarrow 99$$

- 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. Overly simple leads to *biased* models

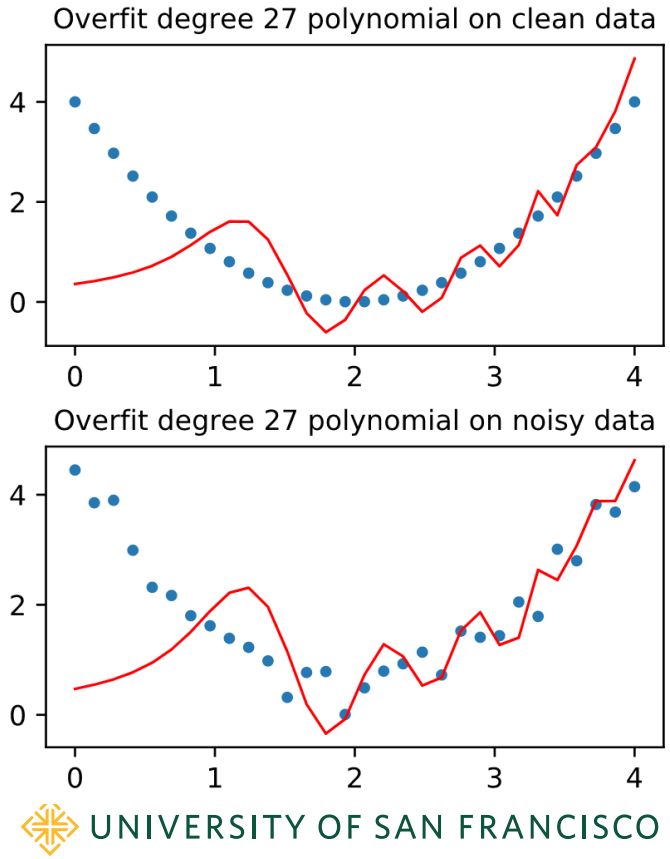
- Now take noise out of the picture
- If our model is unable to capture $X \rightarrow y$ well enough, model is *bias*ed, systematically under- or over-predicting
- Predicting with mean (line) for quadratic is too weak, as is a decision tree that is too shallow to partition x space well
- Increasing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





3. Overly-complex models can *overfit*

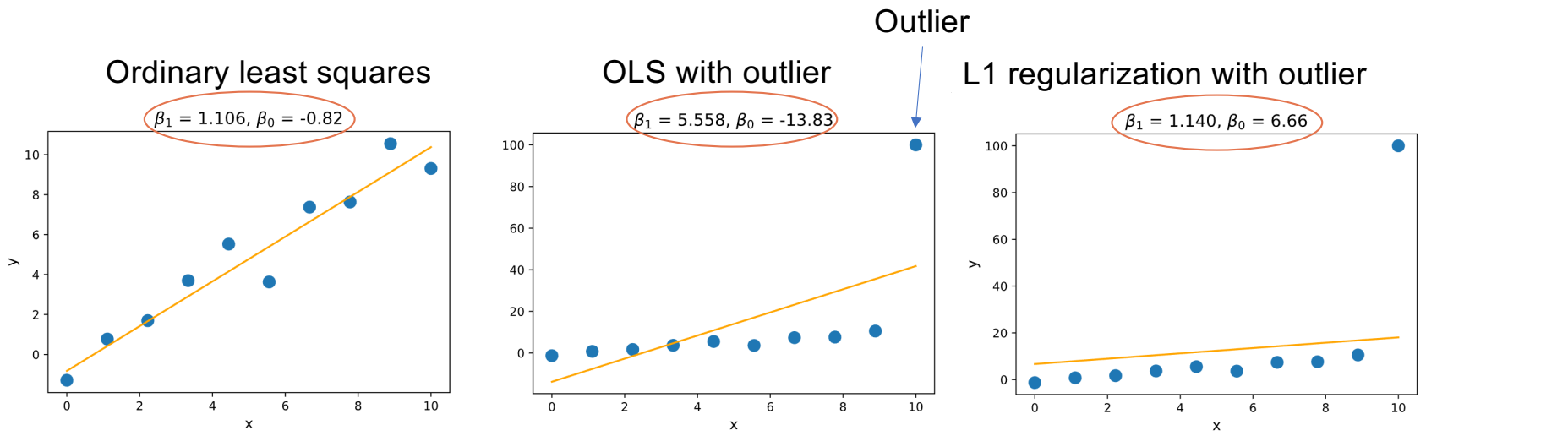
- Even without noise, models with too much power/flexibility for a training data set can be inaccurate (degree 27 polynomial here)
- We always have noise though, so even suitably complex models lead to overfitting
- Model is overfit when it focuses on quirks/details/noise of a specific X, rather than getting the gist of training set X
- Getting the gist means capturing the nature of the underlying **distribution** behind X



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

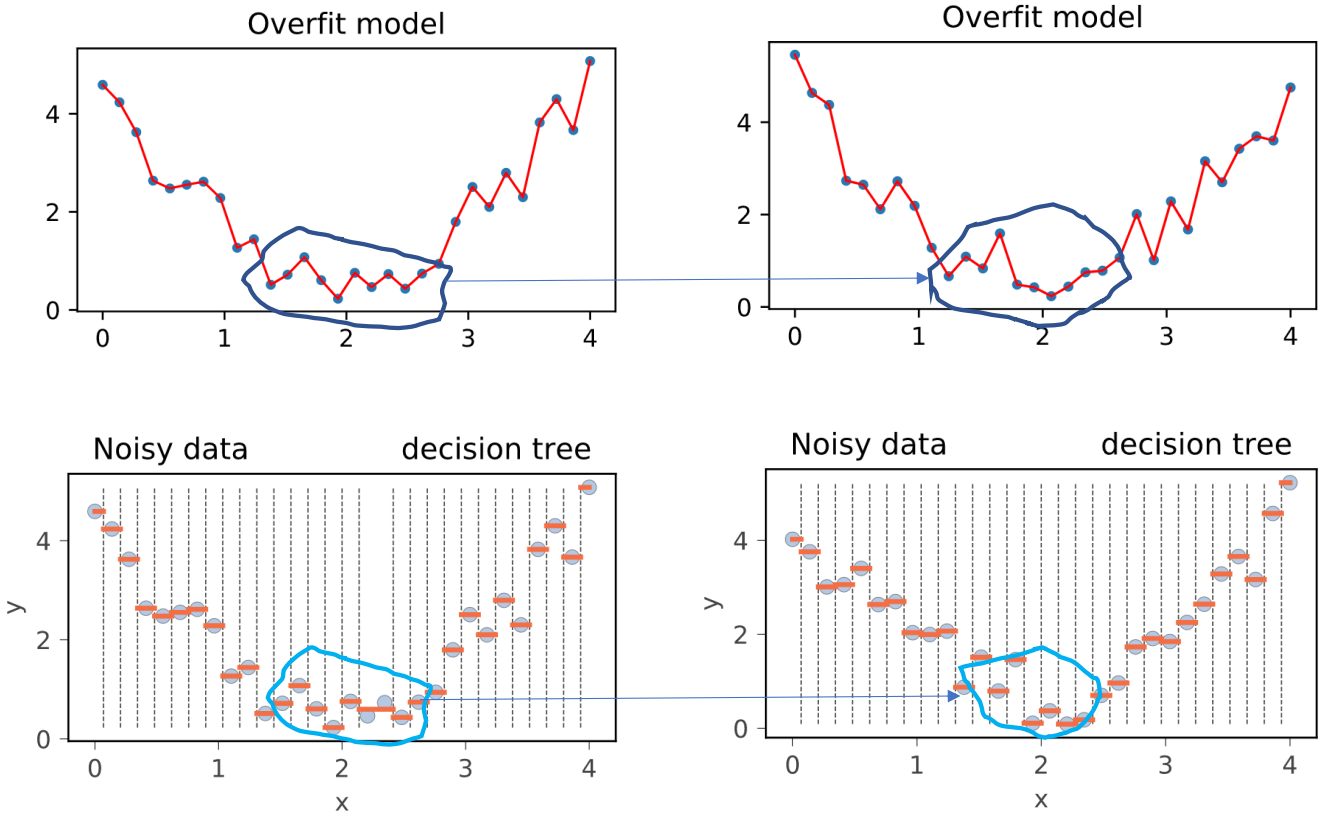
Recall: even simple models can overfit

- 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
- Small change to data set causes dramatic change in β's



Stats view of overfitting: *variance*

- Small changes in training data lead to very different models
- Here are 2 training sets drawn (cols) from same distribution
- Same fitting strategy leads to different models (for interpolation & decision trees)
- Variance refers to model parameters not predictions, though they are related



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

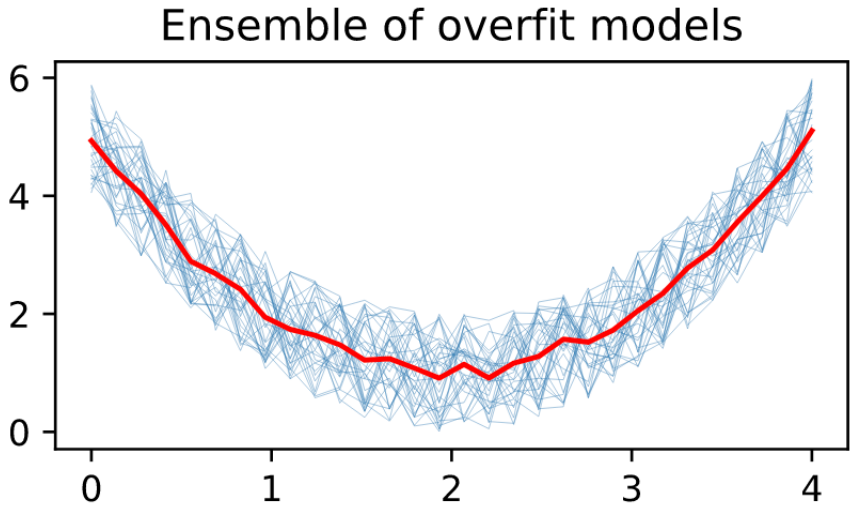
High model variance = overfitting

- The term “variance” is confusing because it refers to the variation of models (and hence prediction errors) trained on multiple, similar data sets, but we normally only have one training set. So it’s weird to think about, unless you’re really into boostrapping …
- High variance implies poor generality because very similar training sets yield very different models; so the model isn't capturing the underlying distribution of X from which the training sets are derived
- That's the same as saying that an overfit model will give very different predictions for test records compared to training records (in same region of feature space)
- Since similar records should give similar predictions, such a model would be inaccurate on unseen test records…the definition of poor generality

What to do about high variance (overfitting)?

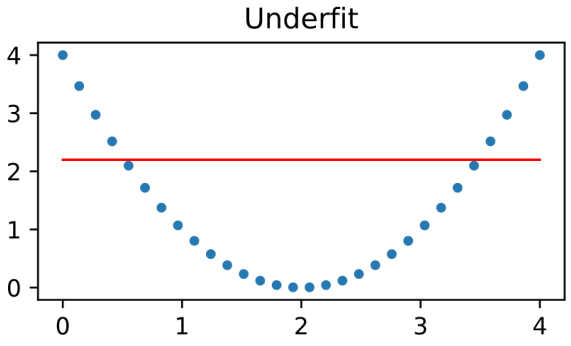
1. Get more data (not always possible)
2. Simplify/regularize/restrict model
3. Average results from many overfit models (an *ensemble*)

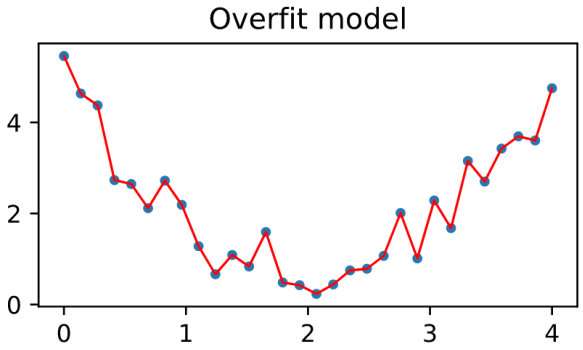
- Sample multiple X's from same distribution and independently (i.i.d.) then average of many models should be accurate & with low variance
- Graph shows 35 models fit to noisy data from same distribution, averaged
- A random forest ensembles many overfit decision trees and uses a trick to make the trees sort of independent (more in RF lecture)

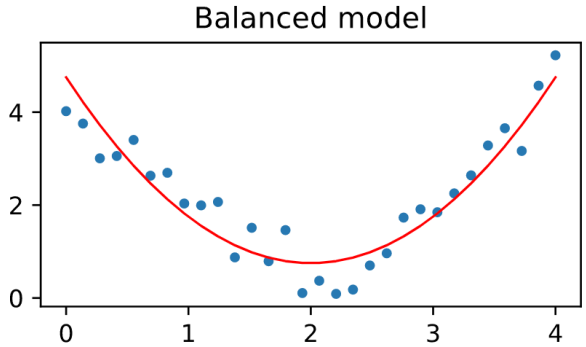


The trade-off

- 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!







Detecting bias and variance

- In the end, practitioners talk about underfitting and overfitting not bias and variance (colloquially biased just means less accurate)
- How do we know that the simple mean model to the right is underfit?
- How do we know that the other model is overfit?
- We need to measure how accurately our model fits the data but which data? Training **and** non-training data
- Models inaccurate on training data are biased
- Models inaccurate on non-training data lack generality
- We'll do a lecture on properly assessing models, followed by a lecture on how to quantify (later)