AGENDA

- Exit survey questions
- Homework
- Project
- Follow up from last time
- Bias Variance Tradeoff
- Cross Validation

BIAS VARIANCE

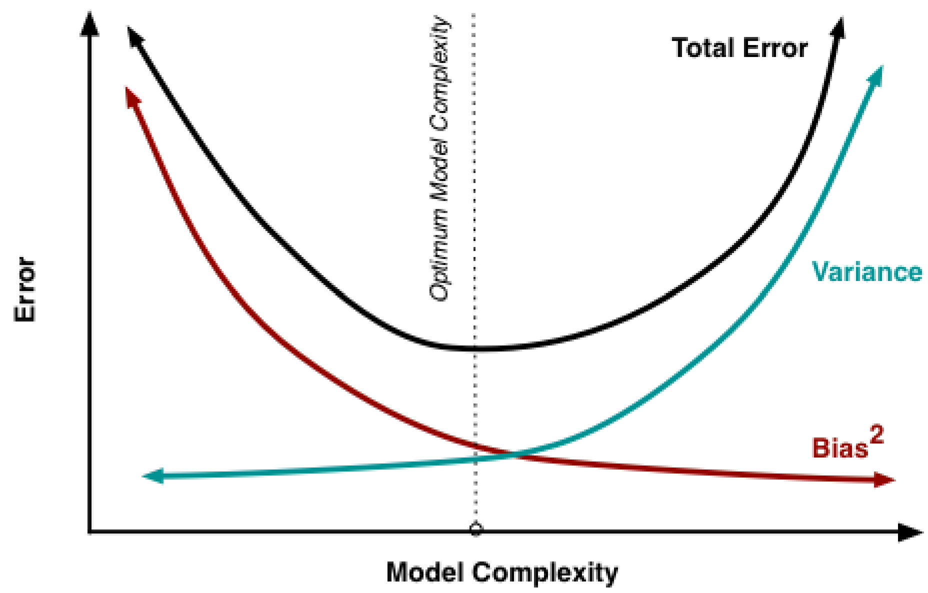
BIAS VARIANCE TRADEOFF

Expected Loss = $(\text{bias})^2$ + variance + noise

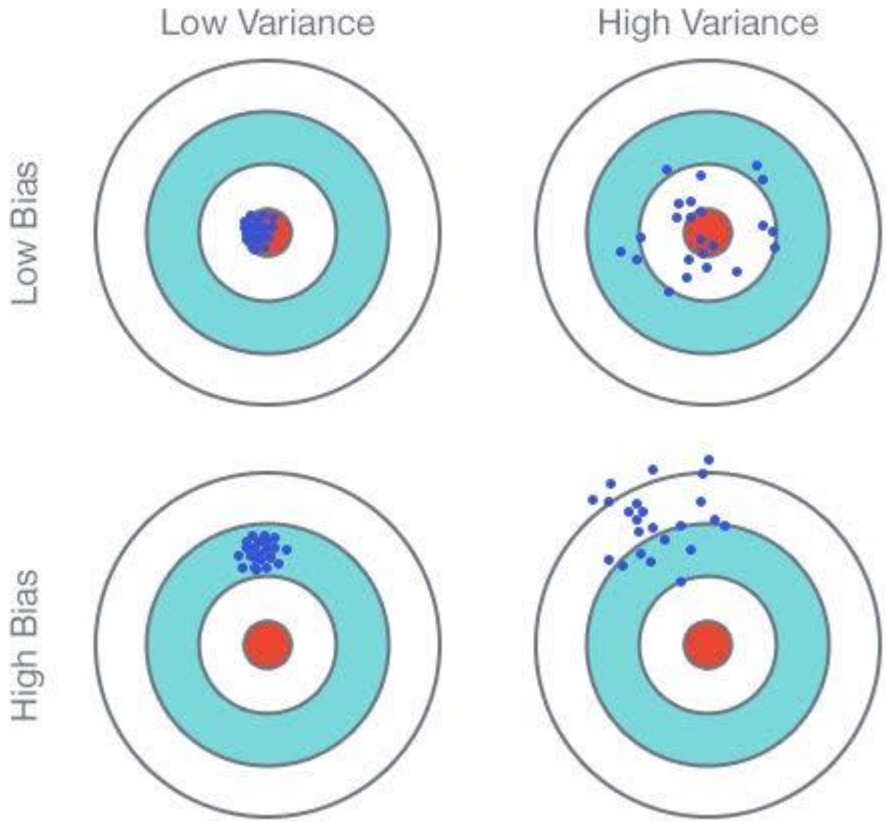
$$E[(y - \hat{f}(x))^2] = Bias[\hat{f}(x)]^2 + Var[\hat{f}(x)] + \sigma^2$$

$$Bias[\hat{f}(x)] = E[\hat{f}(x)] - f(x)$$

$$Var[\hat{f}(x)] = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$$



BIAS VARIANCE TRADEOFF



CROSS VALIDATION

SUPERVISED LEARNING

There are many model options

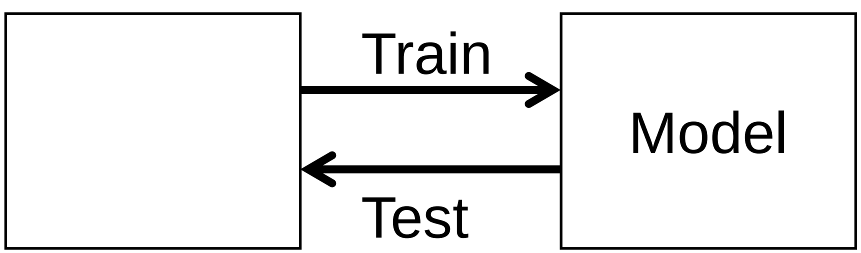
Q: Which one do we choose?

Let’s choose the model that gives us the best performance

Q: How do we measure performance? How well does it work?

Can we use our dataset for an error estimate?

How would this work? Issues?

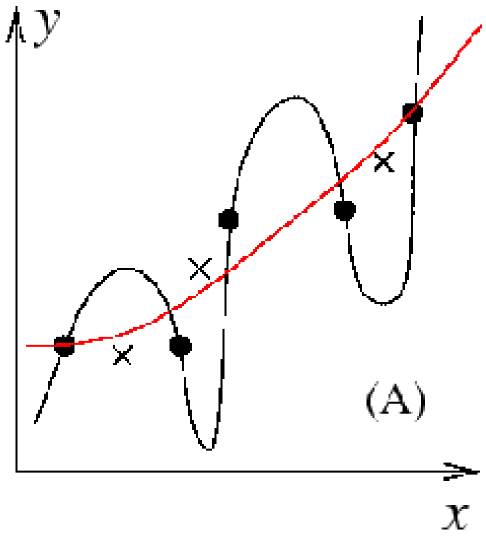


Training Error

Dataset

Q: Are there any issues with training error?
Q: How small can we make our training error?

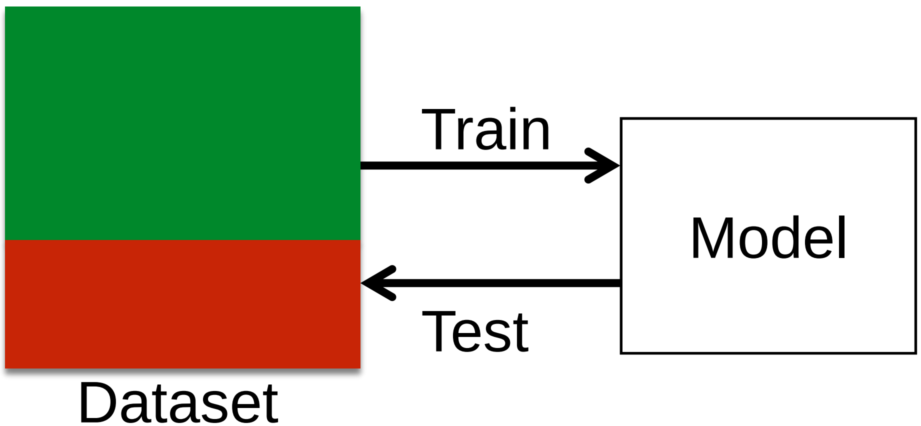
A: We can make the training error go to zero. We just need to memorize.



This is called over fitting

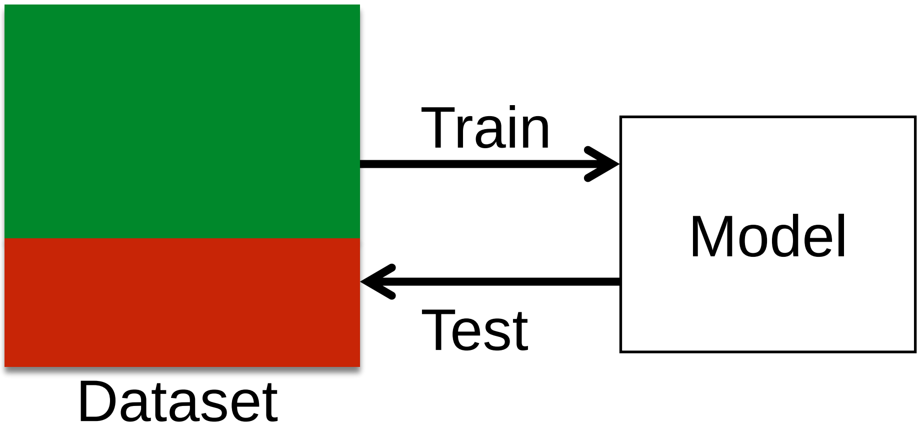
SUPERVISED LEARNING

Want performance on new observations. Data that we haven't seen



Training Error: Error on the training set
Test Error: Error on the test set

SUPERVISED LEARNING



Training Error: Error on the training set
Test Error: Error on the test set

Problem:

1. Error depends on the particular test points which can be highly variable
2. We miss out on some of the data because only a subset is used to train

CROSS VALIDATION

K-Fold Cross Validation

1. Split data set into k subset
2. Use each fold as a validation set once while the union of all others are the training set
3. Combine the generalization error for each fold and combine the results