CS 60050
Machine Learning

Error analysis and validation

Sources of noise

- While learning a target function using a training set
- Two sources of noise
  - Some training points may not come exactly from the target function: stochastic noise
  - The target function may be too complex to capture using the chosen hypothesis set: deterministic noise
- Generalization error: Model tries to fit the noise in the training data, which gets extrapolated to out-of-sample

Ways to handle noise

- Regularization
  - Constraint the model so that the noise cannot be learnt too well
  - Already discussed
- Validation
  - Check performance on data other than training data

Validation

- Divide given data into training set and test set
  - E.g., 80% train and 20% test
  - Better to select randomly
- Learn parameters using training set, check performance on test set, using measures like accuracy, misclassification rate
- Trade-off: more data for training vs. validation

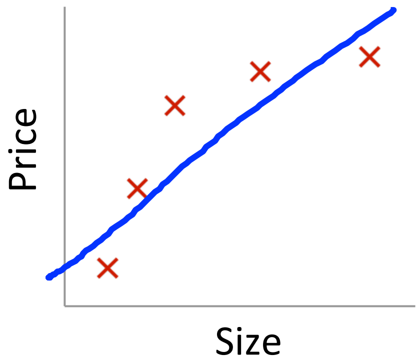
An example: model selection

- Which order polynomial will best fit a given data? Polynomials available: h1, h2, ..., h10
- As if an extra parameter - degree of the polynomial - is to be learned
- Approach
  - Divide into train and test set
  - Train each hypothesis on train set, measure error on test set
  - Select the hypothesis with minimum test set error

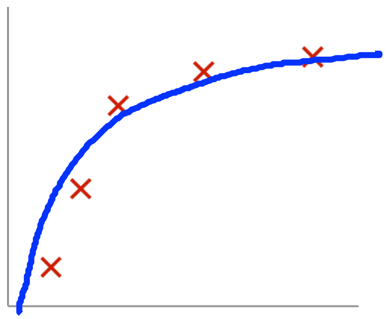
An example: model selection

- Problem with the previous approach
  - The test set error we computed is not a true estimate of generalization error, since our extra parameter is fit to the test set
- Approach 2
  - Divide data into train set (60%), validation set (20%) and test set (20%)
  - Select that hypothesis which gives lowest error on validation set
  - Use test set to estimate generalization error

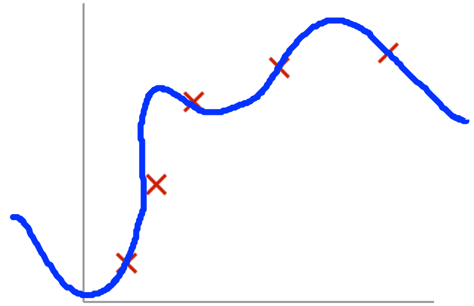
Analysing bias vs. variance



High bias
(underfit)
$d=1$



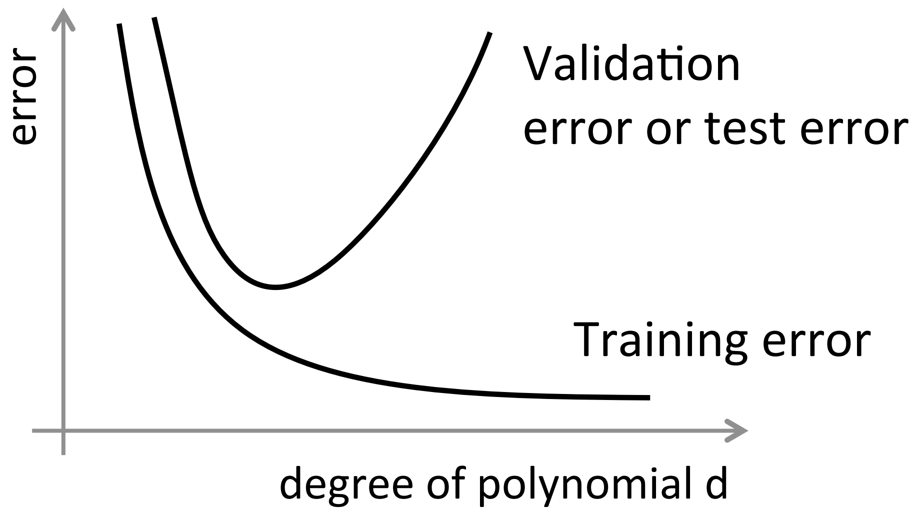
"Just right"
$d=2$



High variance
(overfit)
$d=4$

Analysing bias vs. variance

- Suppose your model is not performing as well as expected. Is it a bias problem or a variance problem?



Bias (underfit):
Both training error and validation / test error are high

Variance (overfit):
Low training error
High validation / test error

Will more training data help?

- A learnt model is not performing as well as expected. Will having more training data help?

- Note that there can be substantial cost for getting more training data.

Will more training data help?

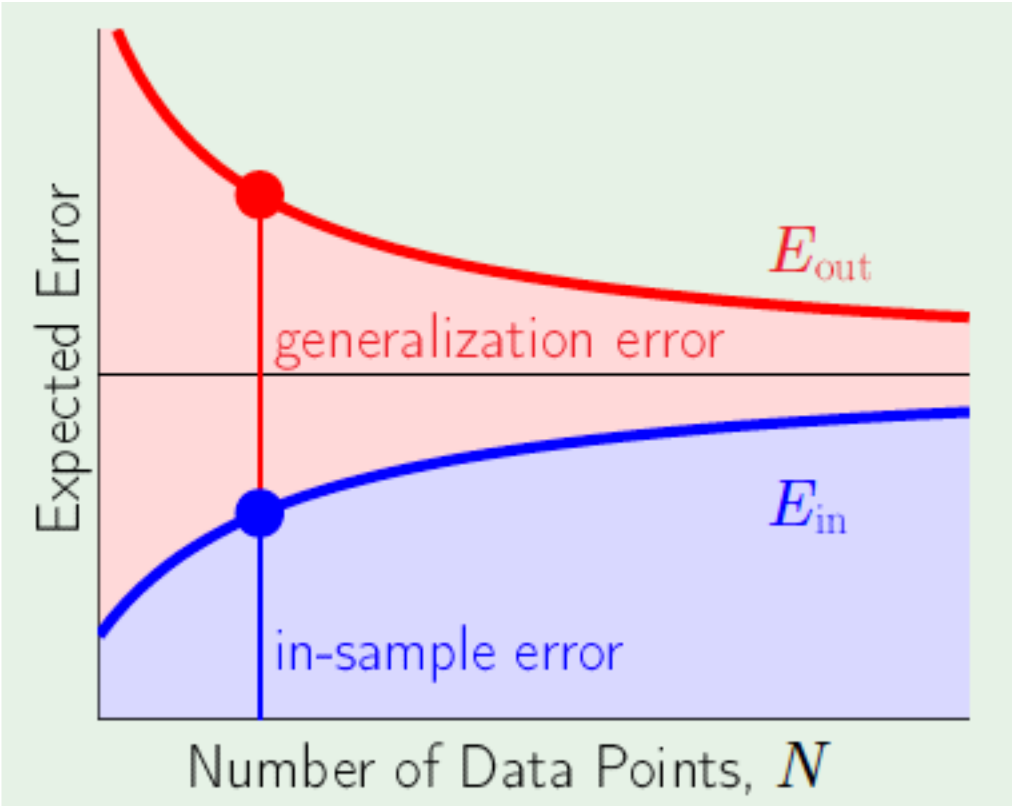
- A learnt model is not performing as well as expected. Will having more training data help?

- Note that there can be substantial cost for getting more training data.

- If model is suffering from high bias, getting more training data will not (by itself) help much.
- If model is suffering from high variance, getting more training data is likely to help

Learning curves

- How do training error (in-sample error) and test or validation error (out-of-sample error) generally vary with number of training points?



Practical approach

- Divide data into training set and validation set
- Start with simple algorithm, train on different amounts of training data, test performance on validation set
- Plot learning curves to decide if more training data, more features likely to help
- Error analysis: Manually examine the examples (in validation set) where algorithm made errors. Any systematic trend in what type of examples it is making errors on?

Skewed classes

- Often the class of interest is a rare class (y=1)
  - Spam emails / social network accounts
  - Cancerous cells
  - Fraud credit card transactions
- Precision: Out of all examples for which model predicted h=1, for what fraction is y=1?
- Recall: Of all examples for which y=1, for what fraction did model correctly predict h=1?

Precision / Recall

		Predicted Label	
		$\hat{y} = 1$	$\hat{y} = -1$
True Label	$y = 1$	True positive	False negative
	$y = -1$	False positive	True negative

Precision: (True positive) / (True positive + False positive)

Recall: (True positive) / (True positive + False negative)

Precision vs. Recall: tradeoff

- Predict y=1 if h > some threshold
- Predict y=1 only if highly confident: high precision, lower recall
- Avoid missing too many cases with y=1: high recall, lower precision

- F-score: harmonic mean of Precision and Recall

$$2\frac{PR}{P+R}$$