Training/Testing and Regularization

Il-Chul Moon
Dept. of Industrial and Systems Engineering
KAIST
icmoon@kaist.ac.kr

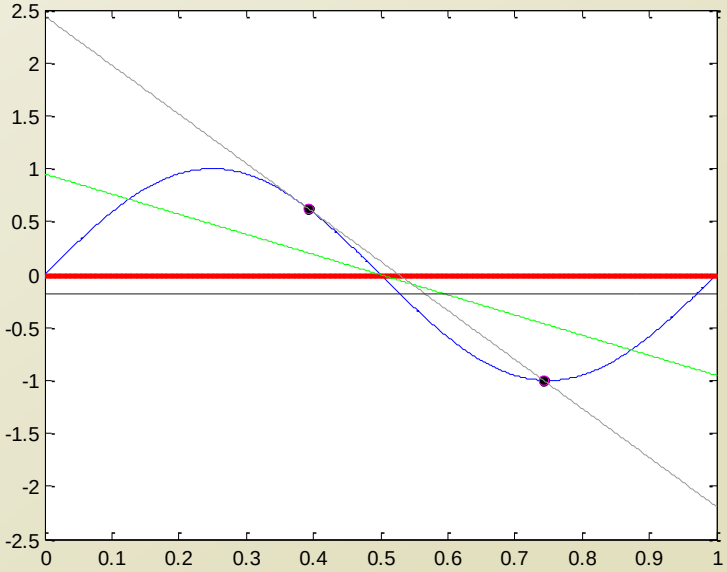
Weekly Objectives

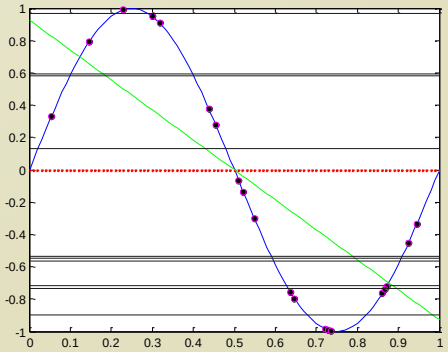
- Understand the concept of bias and variance
  - Know the concept of over-fitting and under-fitting
  - Able to segment two sources, bias and variance, of error
- Understand the bias and variance trade-off
  - Understand the concept of Occam's razor
  - Able to perform cross-validation
  - Know various performance metrics for supervised machine learning
- Understand the concept of regularization
  - Know how to apply regularization to
    - Linear regression
    - Logistic regression
    - Support vector machine

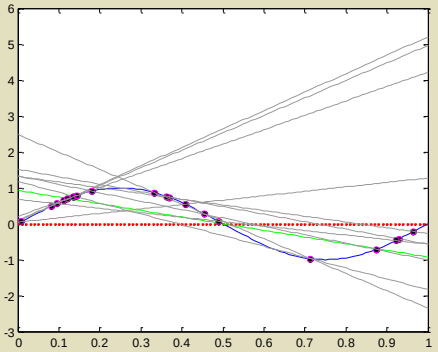
PERFORMANCE MEASUREMENT

Empirical Bias and Variance Trade-off

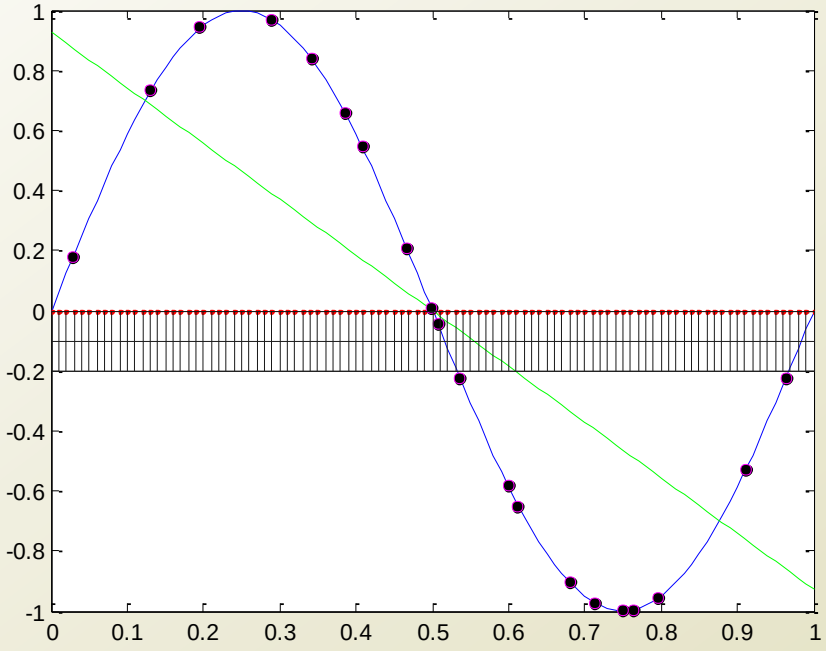
- Consider
  - f(x)=sin(2*pi*x)
  - D={two points|point=(x,sin(2*pi*x)), 0<=x<=1)
  - Two g(x)
    - Zero degree: dark grey line
    - One degree: light grey line
  - Two $\bar{g}(x)$
    - Zero degree: red line
    - One degree: green line
- Which has a greater bias and a greater variance between one degree and zero degree?



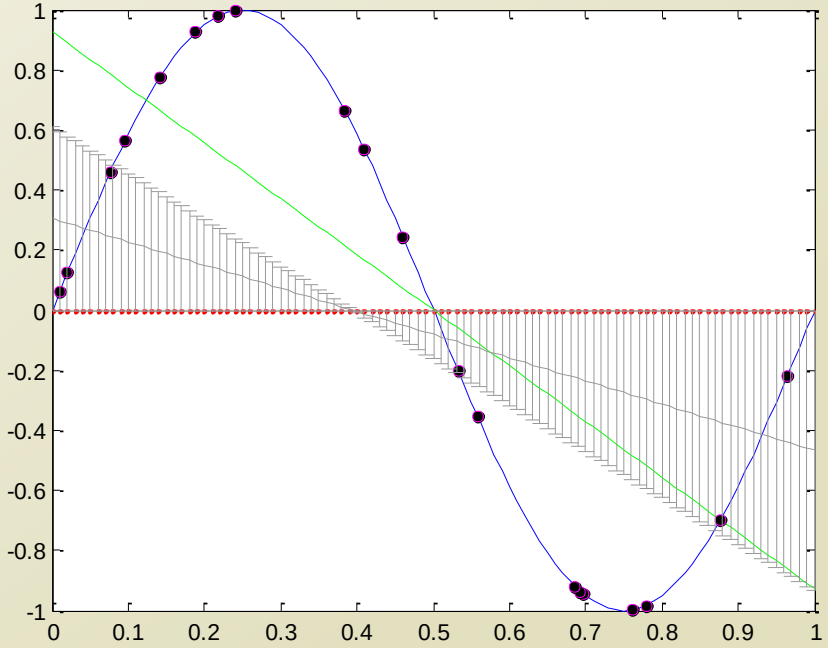




Bias and Variance of Two Hypotheses



Bias = 0.5051
Var. = 0.2410

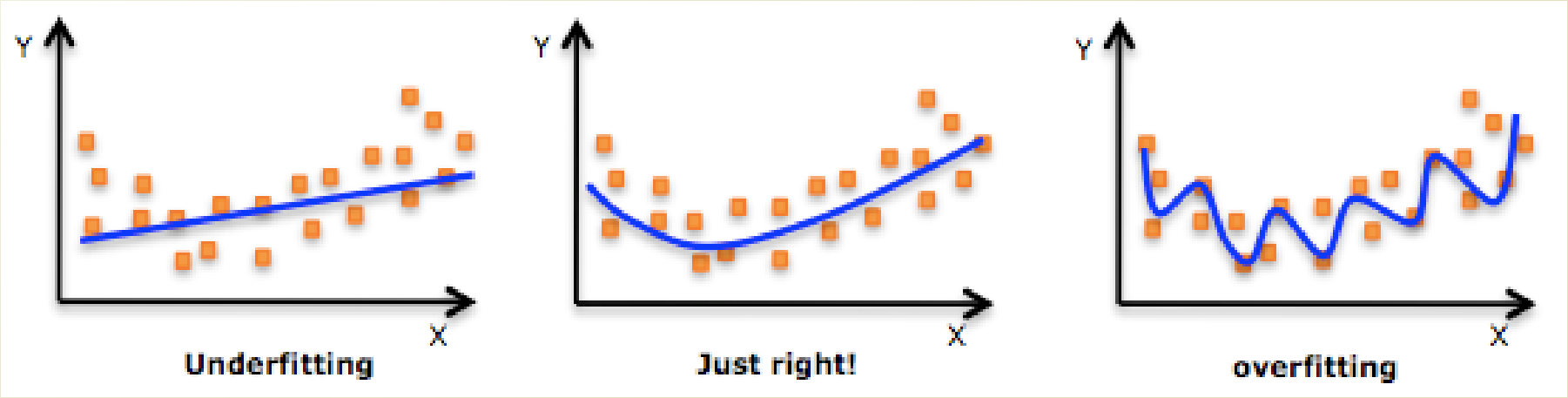


Bias = 0.3092
Var. = 2.0708

- A complex model has a higher variance and a lower bias.
- A simple model has a lower variance and a higher bias.
- Need a balance in the complexity of a ML algorithm

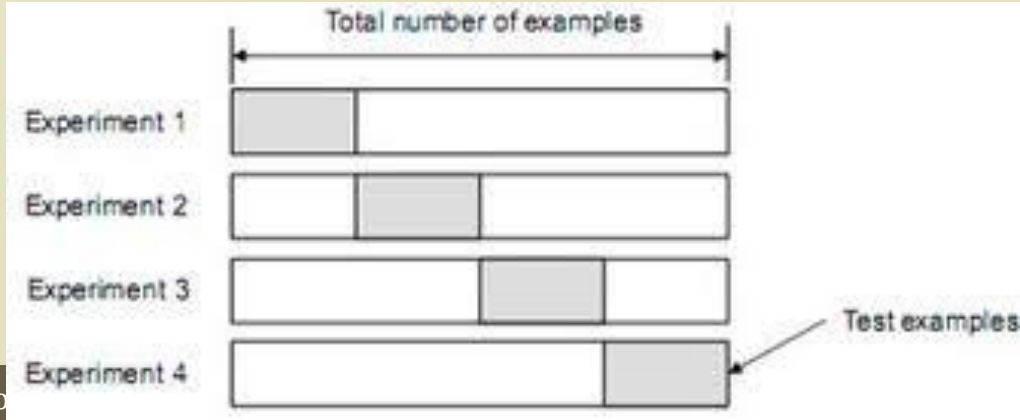
Occam's Razor

- Occam's Razor
  - Among competing hypotheses, the one which makes the fewest assumption should be selected
- Competing?
  - Relevantly similar error in the prediction
- Fewest assumption
  - Less complex model
- Given the approximately same error, a simple model should be selected
- Reflection of Bias and Variance tradeoff!
  - By the way, is it possible to calculate the bias and the variance in the real world setting?



Cross Validation

- We don't have the infinite number of samples observed from the target function
- We have to mimic the infinite number of sampling
  - Where is the number of sampling used in the bias and the variance tradeoff?
    - $\bar{g}$: the average hypothesis of a given infinite number of Ds
    - Formally, $\bar{g}(x) = E_D[g^{(D)}(x)]$
- We need to have many datasets from a fixed number of datasets
- N-fold cross validation
  - We divide a given set of instances into N exclusive subsets.
  - We use (N-1) subsets for training
  - We use 1 subset for testing
- Special case: LOOCV
  - Leave One Out Cross Validation
  - Extreme case of N-fold cross validation



Performance Measure of ML

- Is it possible to calculate the bias and the variance?
  - We don't know the target function, f(X)!
  - We can't compute the average hypothesis, $\bar{g}(x)$!
- Therefore, we can't use the bias and the variance as the performance measures.
- Then, what measures to use?
  - Accuracy= (TP+FN) / (TP+FP+FN+TN)
  - Precision and Recall
  - F-Measure
  - ROC curve

		Actual Value	
		True	False
Estimated Value	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Precision and Recall

- Consider the two cases
  - Building a classifier
    - Spam filter
    - CRM

		Actual Value	
		True	False
Estimated Value	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

- Goals are slightly different
  - Spam filter: classifying spam
    - Safety is first. You don't want to throw out valid emails estimated as spams
    - Reducing the FP is the priority
  - CRM: classifying VIP customer
    - Reaching out is first. You don't want to miss any VIP customers as ordinary ones
    - Reducing the FN is the priority
- **Precision** = TP / (TP+FP)
- **Recall** = TP / (TP+FN)
- Then, which metrics to use in each case?

F-Measure

- Precision and recall are popular metrics, but it has problems in the applications
  - The most safest spam filter == always say 'no spam'
  - The most reaching-out customer filter == always say 'VIP'
- We need a measure that balances the precision and the recall performance
- F-Measure is the derived metric from the precision and the recall
  - F_b-Measure = $(1+b^2) * (\text{Precision} * \text{Recall}) / (b^2 * \text{Precision} + \text{Recall})$
  - F_1-Measure= $2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$
  - $F_{0.5}$ and F_2 are also used.
    - F_2 emphasizes recall
    - $F_{0.5}$ emphasizes precision