Training/Testing and Regularization

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

KAIST

Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

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

Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

CONCEPT OF BIAS AND VARIANCE

Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

Up To This Point…

- Now, you are supposed to have some knowledge in classifications
  - Naïve Bayes
  - Logistic Regression
  - Support Vector Machine
- SVM is still a commonly used machine learning algorithm for classifications

- Functioning is ***kind of*** done
- Efficiency and accuracy now becomes a problem

KAIST
Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

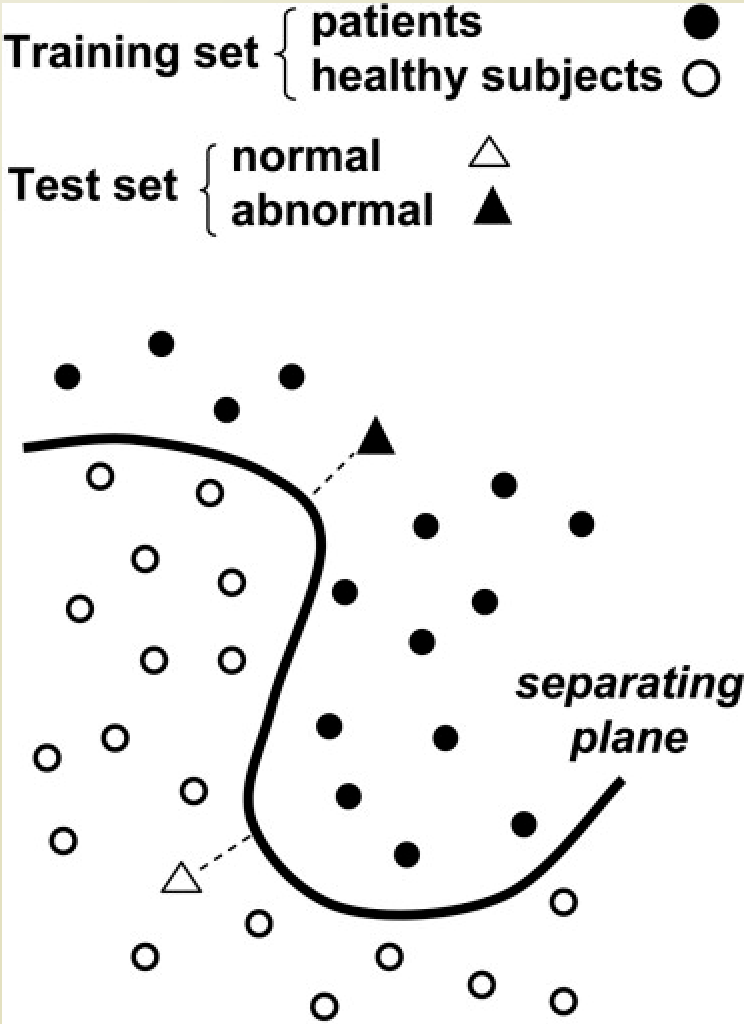
Better Machine Learning Approach?

- Accurate prediction result
  - Ex) with this NB classifier, I can filter spams with 95% accuracy!
- Is this a right claim?
  - The validity of accuracy
    - No clear definition
    - Why not use other performance metrics? Such as Precision/Recall, F-Measure
  - The validity of dataset
    - Spams??
    - How many spams?
    - Where did you gathered?
    - Big variance in the spams?
    - Is the spam mail evolving?
      - From Nigerian prince scheme to something else?

Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

Training and Testing

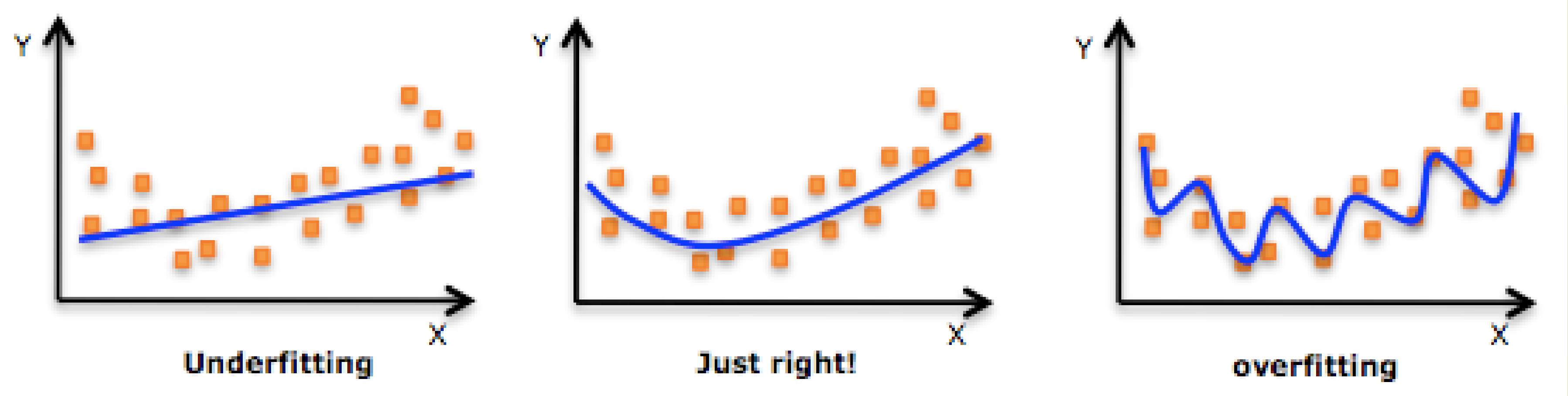
- Training
  - Parameter inference procedure
  - Prior knowledge, past experience
  - There is no guarantee that this will work in the future
    - ML's Achilles gun is the stable/static distribution of learning targets.
  - Why ML does not work in the future?
    - The domain changes, or the current domain does not show enough variance
    - The ML algorithms inherently have problems
- Testing
  - Testing the learned ML algorithms/the inferred parameters
  - New dataset that is unrelated to the training process
  - Imitating the future instances
    - By setting aside a subset of observations



Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

Over-Fitting and Under-Fitting

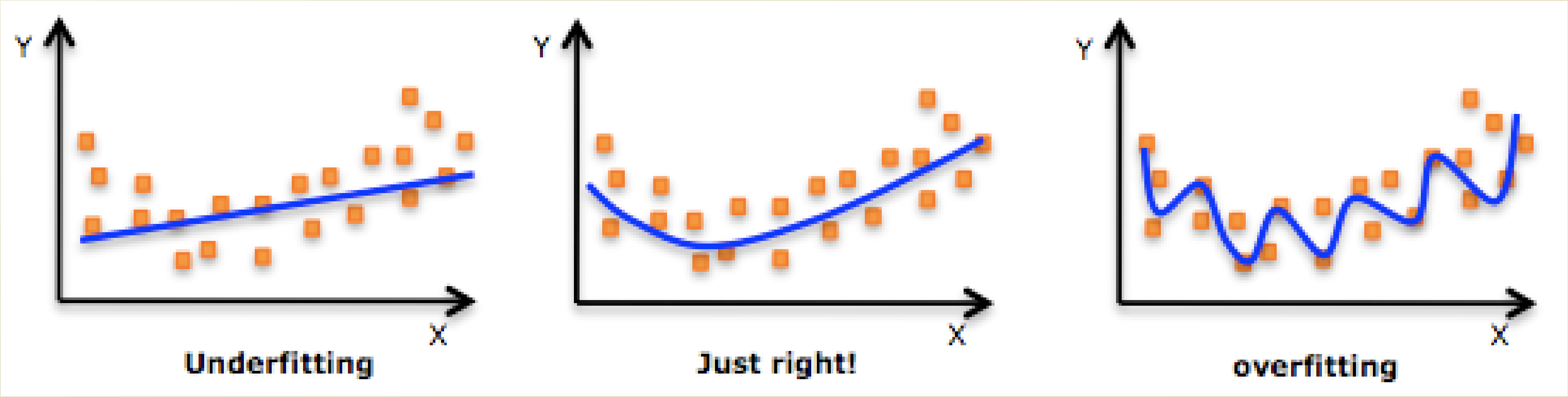
- Imaging this scenario
  - You are given N points to train a ML algorithm
  - You are going to learn a simple polynomial regression function
    - Y=F(x)
    - The degree of F is undetermined. Can be linear or non-linear
- Considering the three Fs in the below, which looks better?



Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

Tuning Model Complexity

- One degree, two degree, and N degree trained functions
  - As the degree increases, the model becomes complex
  - Is complex model better?
- Then, where do we stop in developing a complex model?
  - Is there any measure to calculate the complexity and the generality?
- There is a trade-off between the complexity of a model and the generality of a dataset.



Sources of Error in ML

- Source of error is in two-folds
  - Approximation and generalization
- $E_{out} \leq E_{in} + \Omega$
  - E_{out} is the estimation error, considering a regression case, of a trained ML algorithm
  - E_{in} is the error from approximation by the learning algorithms
  - Ω is the error caused by the variance of the observations
- Here, we define a few more symbols
  - f: the target function to learn
  - g: the learning function of ML
  - $g^{(D)}$: the learned function by using a dataset, D, or an instance of hypothesis
  - D: an available dataset drawn from the real world
  - $\bar{g}$: the average hypothesis of a given infinite number of Ds
    - Formally, $\bar{g}(x) = E_D[g^{(D)}(x)]$

Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

Bias and Variance

- $E_{out} \leq E_{in} + \Omega$
- Error of a single instance of a dataset D
  - $E_{out}(g^{(D)}(x)) = E_X[\left(g^{(D)}(x) - f(x)\right)^2]$
- Then, the expected error of the infinite number of datasets, D
  - $E_D[E_{out}\left(g^{(D)}(x)\right)] = E_D[E_X[\left(g^{(D)}(x) - f(x)\right)^2]] = E_X[E_D[\left(g^{(D)}(x) - f(x)\right)^2]]$
- Let's simplify the inside term, $E_D[\left(g^{(D)}(x) - f(x)\right)^2]$
  - $E_D\left[\left(g^{(D)}(x) - f(x)\right)^2\right] = E_D\left[\left(g^{(D)}(x) - \bar{g}(x) + \bar{g}(x) - f(x)\right)^2\right]$
  - $= E_D\left[(g^{(D)}(x) - \bar{g}(x))^2 + \left(\bar{g}(x) - f(x)\right)^2 + 2(g^{(D)}(x) - \bar{g}(x))\left(\bar{g}(x) - f(x)\right)\right]$
  - $= E_D\left[(g^{(D)}(x) - \bar{g}(x))^2\right] + \left(\bar{g}(x) - f(x)\right)^2 + E_D\left[2(g^{(D)}(x) - \bar{g}(x))\left(\bar{g}(x) - f(x)\right)\right]$
- $E_D\left[2(g^{(D)}(x) - \bar{g}(x))\left(\bar{g}(x) - f(x)\right)\right] = 0$
  - Because of the definition of $\bar{g}(x)$
- Then, eventually the error becomes ….
  - $E_D[E_{out}\left(g^{(D)}(x)\right)] = E_X[E_D\left[\left(g^{(D)}(x) - \bar{g}(x)\right)^2\right] + \left(\bar{g}(x) - f(x)\right)^2]$

Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

Bias and Variance Dilemma

- $E_D[E_{out}\left(g^{(D)}(x)\right)] = E_X[E_D\left[\left(g^{(D)}(x) - \bar{g}(x)\right)^2\right] + \left(\bar{g}(x) - f(x)\right)^2]$
- Let's define
  - Variance(x)=$E_D\left[\left(g^{(D)}(x) - \bar{g}(x)\right)^2\right]$
  - Bias^2(X)=$\left(\bar{g}(x) - f(x)\right)^2$
- Semantically, what do they mean?
  - Variance is an inability to train a model to the average hypothesis because of the dataset limitation
  - Bias is an inability to train an average hypothesis to match the real world
- How to reduce the bias and the variance?
  - Reducing the variance
    - Collecting more data
  - Reducing the bias
    - More complex model
- However, if we reduce the bias, we increase the variance, and vice versa
  - Bias and Variance Dilemma
  - We will see why this is in the next slide by empirical evaluations….

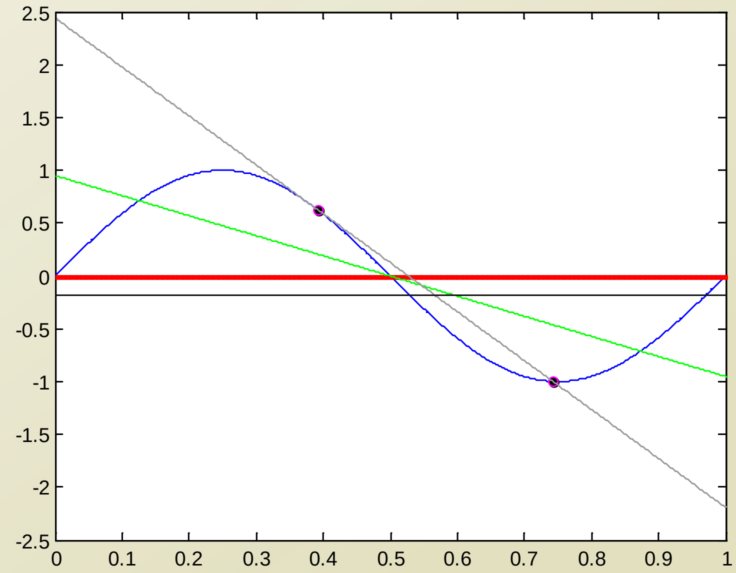
KAIST
Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

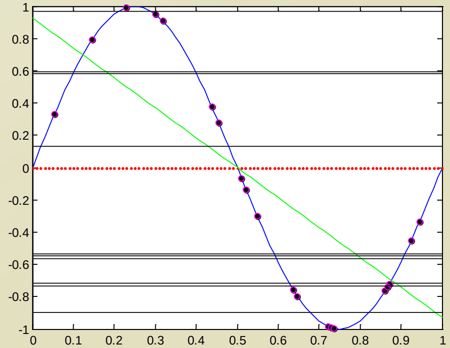
PERFORMANCE MEASUREMENT

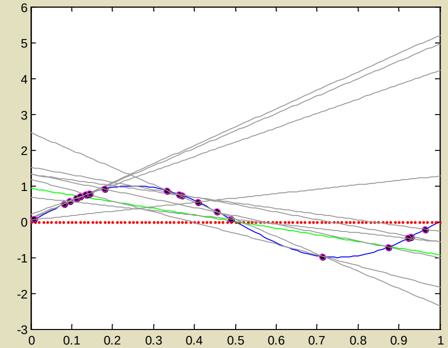
Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

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?

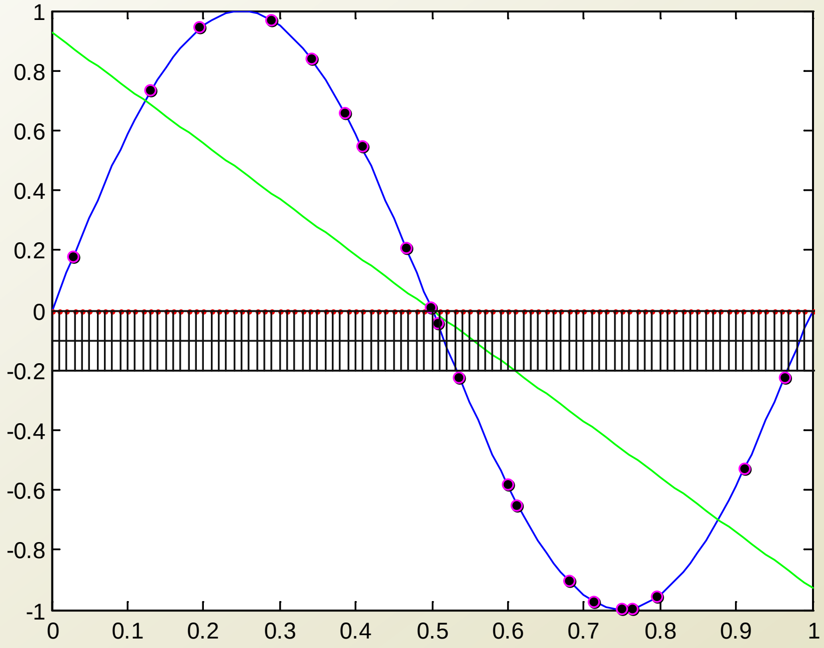


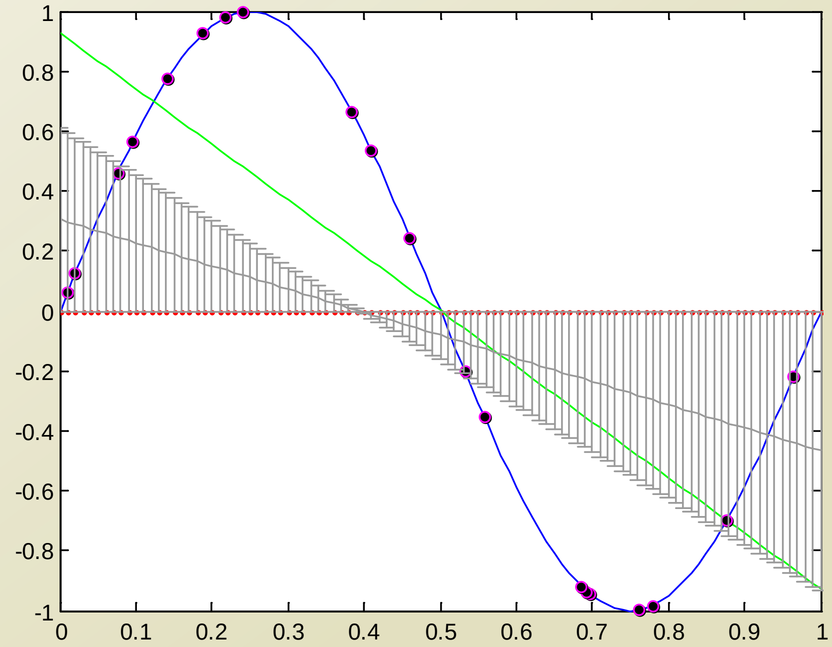




Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

Bias and Variance of Two Hypotheses





Bias = 0.5051
Var. = 0.2410

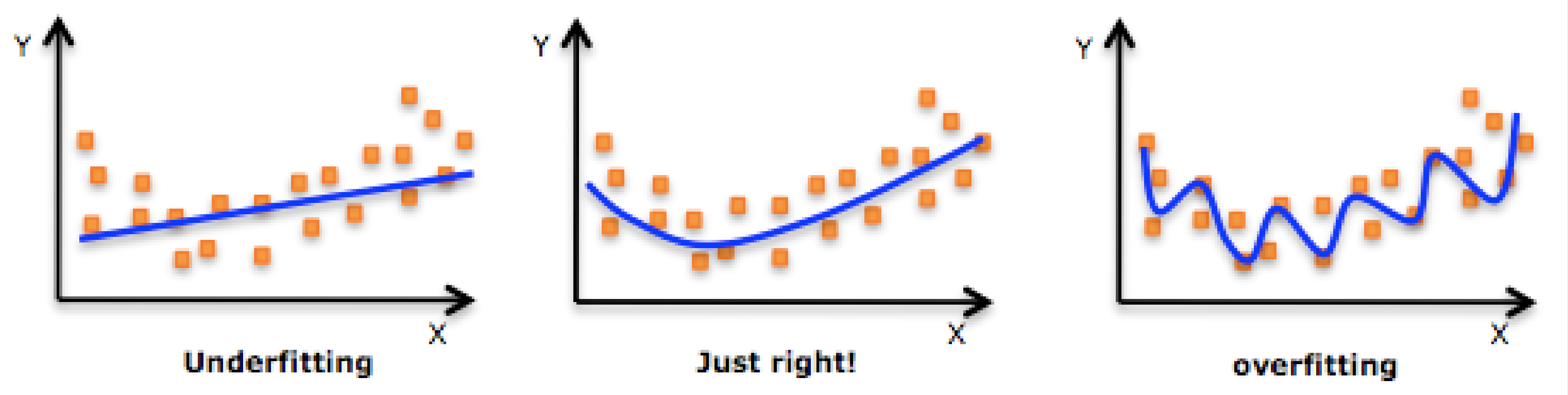
Bias = 0.3092
Var. = 2.0708

- 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

Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

Occam's Razor

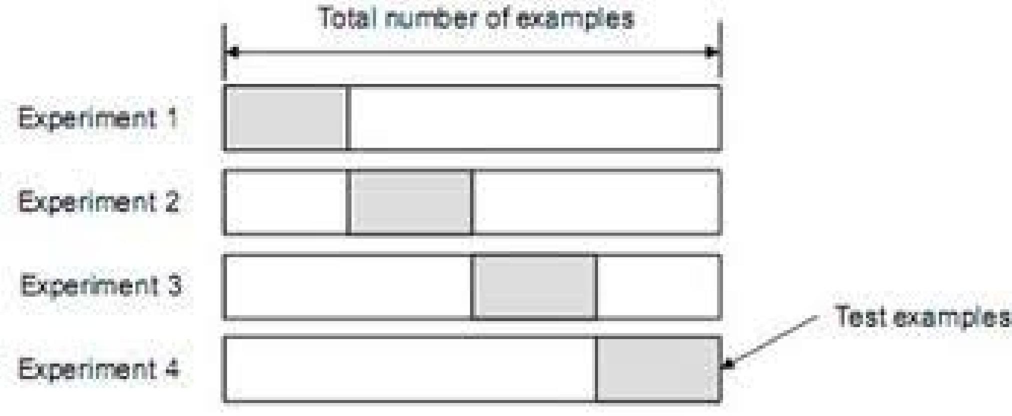
- Occam's Razor
  - Among competing hypotheses, the one which makes the fewest assumption should be selected
- Competing?
  - Relevantly similar error in the prediction
- Fewest assumption
  - Less complex model
- Given the approximately same error, a simple model should be selected
- Reflection of Bias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



Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

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



Copyright © 2010 by Il-Chul Moon, Dept. o

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+TN) / (TP+FP+FN+TN)
  - Precision and Recall
  - F-Measure
  - ROC curve

		Actual Value	
		True	False
Estimated Value	Positive	True Positive ①	False Positive ②
	Negative	False Negative ③	True Negative ④

Copyright © 2010 b

Precision and Recall

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

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

Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

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 = $(1+b^2)$ * (Precision * Recall) / (b^2*Precision + Recall)
  - F_1-Measure= 2 * (Precision * Recall) / (Precision + Recall)
  - $F_{0.5}$ and F_2 are also used.
    - F_2 emphasizes recall
    - $F_{0.5}$ emphasizes precision

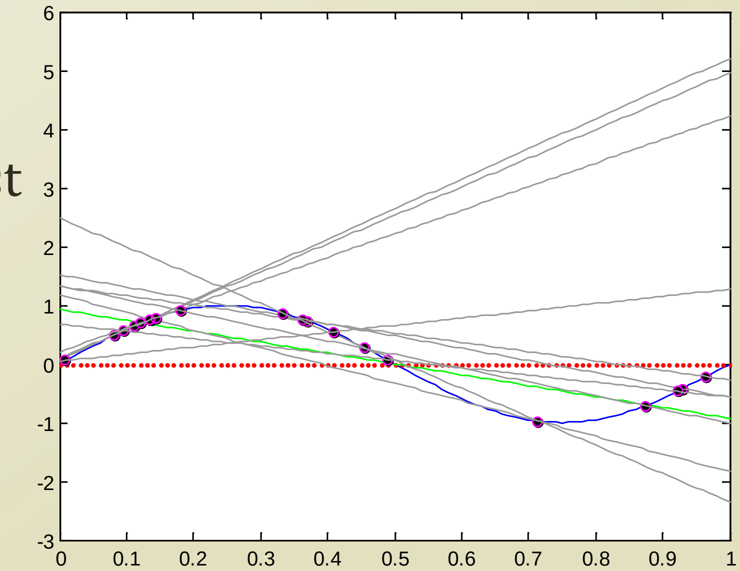
KAIST Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

MODEL REGULARIZATION

Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

Concept of Regularization

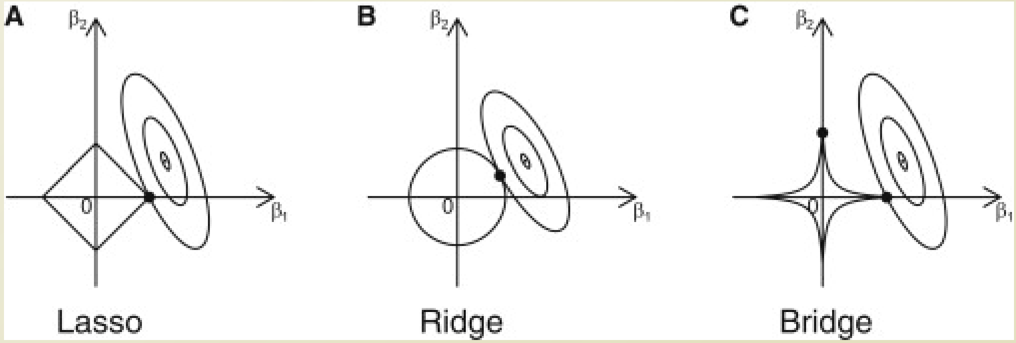
- Disaster in terms of variance
- With regularization
  - We sacrifice the perfect fit
    - Reducing the training accuracy
  - We increase the potential fit in the test
  - Because of the increased model complexity, the bias tends to increase a little bit
  - Eventually, regularization is another constraint for models
    - Existing constraint?
    - Minimizing error from training set
- We add a new term to the MSE



Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

Formal Definition of Regularization

- Regularization is another constraint for the regression
  - The below J(B) is the regularization function to minimize
  - B is the weight of the regression model except the constant term
- There are diverse regularization
  - L1 Regularization == Lasso regularization
    - The first order
  - L2 Regularization == Ridge regularization
    - The second order
  - Depends on the order of the regularization term
    - The order determines the shape of the loss function



$$E(w) = \frac{1}{2}\sum_{n=0}^{N}(train_n - g(x_n, w))^2 + \frac{\lambda}{2}\|w\|^2$$

$$E(w) = \frac{1}{2}\sum_{n=0}^{N}(train_n - g(x_n, w))^2 + \lambda|w|$$

Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

Regularization of Linear Regression

- Let's apply the regularization idea to the linear regression

$$E(w) = \frac{1}{2}\sum_{n=0}^{N}(train_n - g(x_n, w))^2 + \frac{\lambda}{2}\|w\|^2$$

- We can calculate w in the closed form.

$$\frac{d}{dw}E(w) = 0$$

$$\frac{d}{dw}E(w) = \frac{d}{dw}\left(\frac{1}{2}\|train - Xw\|^2 + \frac{\lambda}{2}\|w\|^2\right)$$

$$= \frac{d}{dw}\left(\frac{1}{2}\|train - Xw\|^T\|train - Xw\| + \frac{\lambda}{2}w^T w\right)$$

$$= \frac{d}{dw}\left(\frac{1}{2}(train^T train - 2X^T w \cdot train + X^T X w^T w) + \frac{\lambda}{2}w^T w\right)$$

$$= \frac{d}{dw}\left(train^T train - X^T w \cdot train + \frac{1}{2}X^T X w^T w + \frac{\lambda}{2}w^T w\right)$$

$$= -X^T \cdot train + X^T X w + \lambda w = 0$$

$$-X^T \cdot train + X^T X w + \lambda I w = 0$$

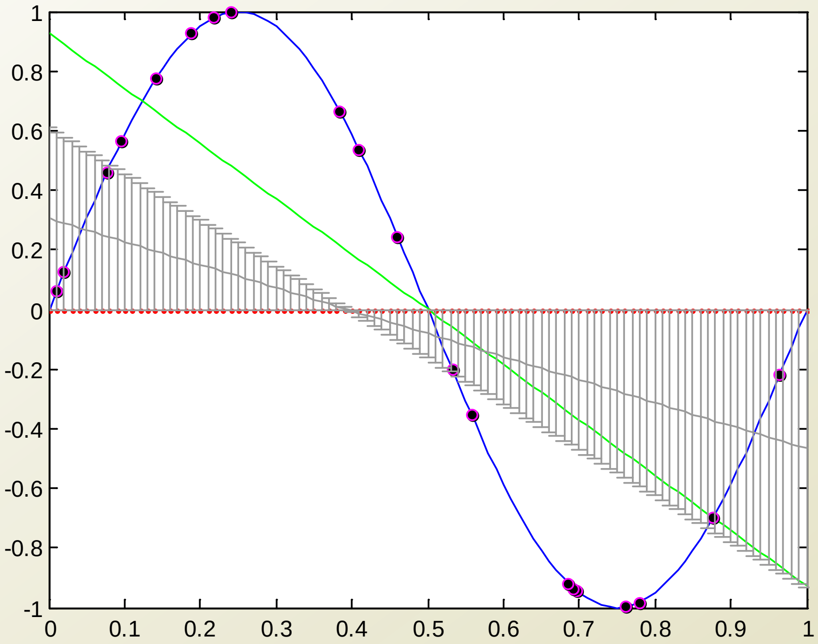
$$-X^T \cdot train + (X^T X + \lambda I)w = 0$$

$$(X^T X + \lambda I)w = X^T \cdot train$$

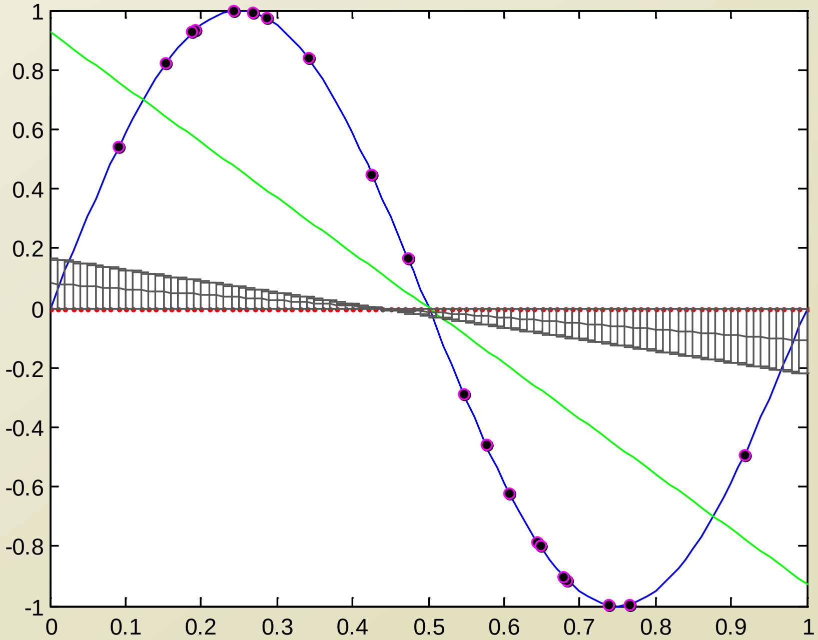
$$w = (X^T X + \lambda I)^{-1} X^T \cdot train$$

Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

Effect of Regularization



Bias = 0.3092
Var. = 2.0708

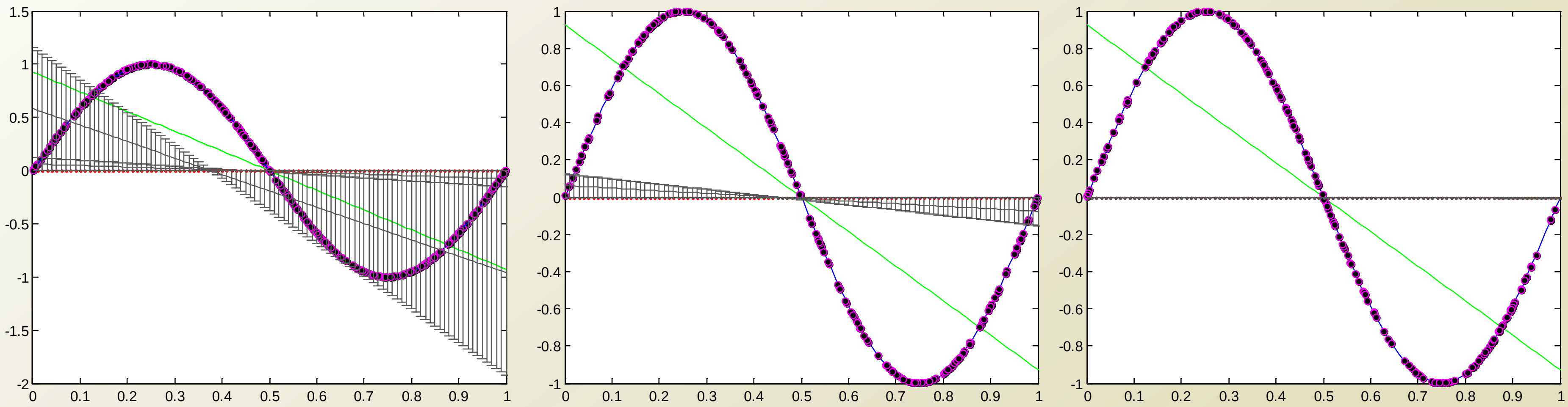


Bias = 0.4372
Var. = 0.1167

- When $\lambda = 1$
  - The bias increases a little bit
  - The variance reduces significantly

Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

Optimizing the Regularization



$\lambda = 0$ $\lambda = 1$ $\lambda = 100$

- We need to optimize λ
  - Too low λ : Too high variance
    - Works like an unregularized model
  - Too high λ : Too low variance
    - Works like a less complex model
    - Converting the first-order model into the constant model
- How to optimize λ?

Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

Regularization of Logistic Regression

- Regularization is applicable to other models
  - Such as logistic regression
- You can search for the closed form and the approximate form of finding θ

$$\arg\max_{\theta} \sum_{i=1}^{m} \log p(y_i \mid \mathbf{x}_i, \theta) - \alpha R(\theta)$$

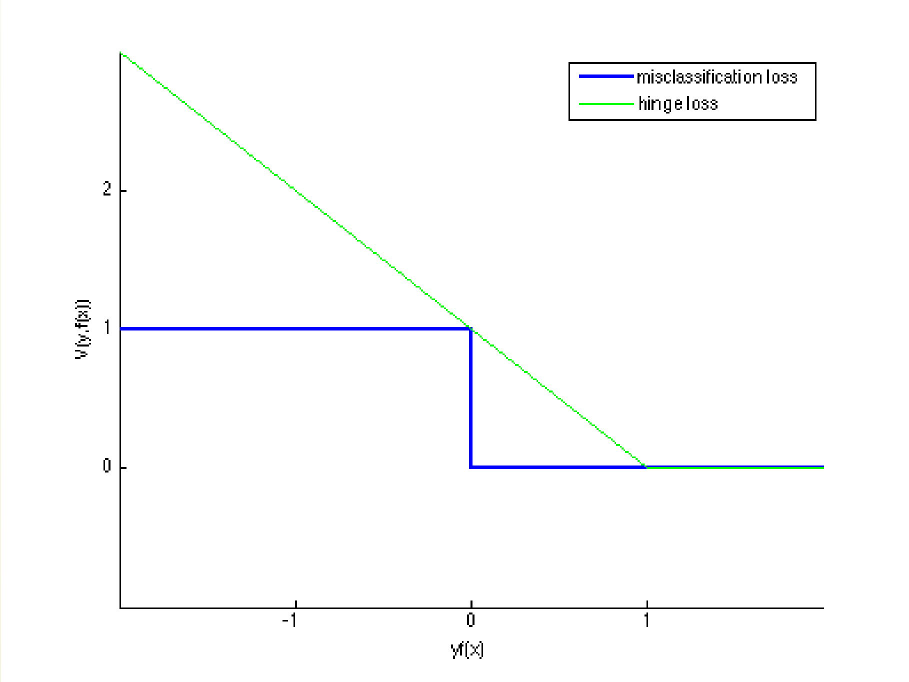
L1: $R(\theta) = \|\theta\|_1 = \sum_{i=1}^{n} |\theta_i|$

L2: $R(\theta) = \|\theta\|_2^2 = \sum_{i=1}^{n} \theta_i^2$

KAIST Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

Regularization and SVM

$$f = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} V(y_i, f(x_i)) + \lambda ||f||_{\mathcal{H}}^2 \right\}$$



$$V(y_i, f(x_i)) = (1 - yf(x))_+$$

$$(s)_+ = max(s, 0)$$

$$f = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - yf(x))_+ + \lambda ||f||_{\mathcal{H}}^2 \right\}$$

$$f = \arg\min_{f \in \mathcal{H}} \left\{ C \sum_{i=1}^{n} (1 - yf(x))_+ + \frac{1}{2} ||f||_{\mathcal{H}}^2 \right\}$$

$$C = \frac{1}{2\lambda n}$$

- Support vector is a special case of regularization with the hinge loss

Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST

Acknowledgement

- This slideset is greatly influenced
  - By Prof. Eric Xing at CMU

Copyright © 2010 by Il-Chul Moon, Dept. of Industrial and Systems Engineering, KAIST