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

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