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

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.

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

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

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

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

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

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

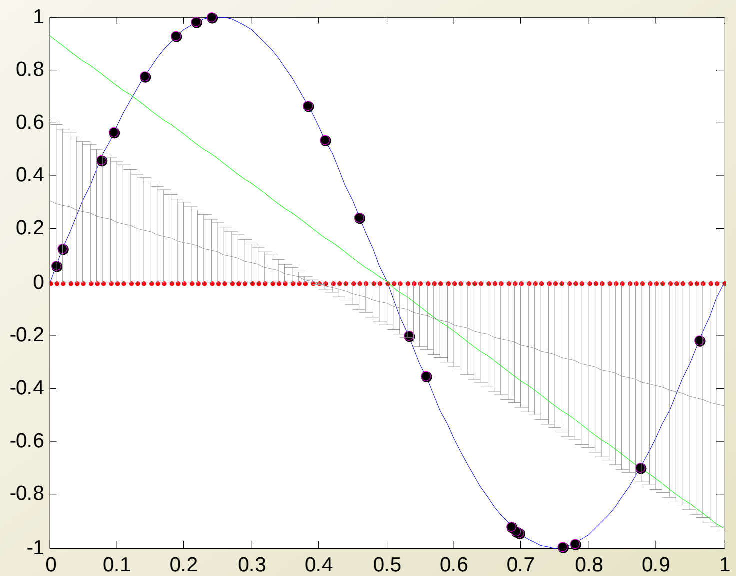
$$-X^T \cdot train + X^T X w + \lambda I w = 0$$

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

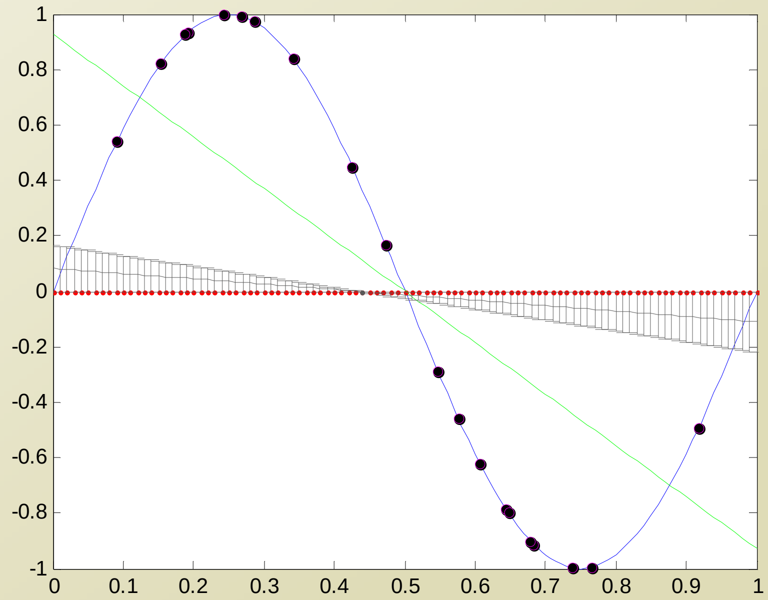
$$(X^T X + \lambda I)w = X^T \cdot train$$

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

Effect of Regularization



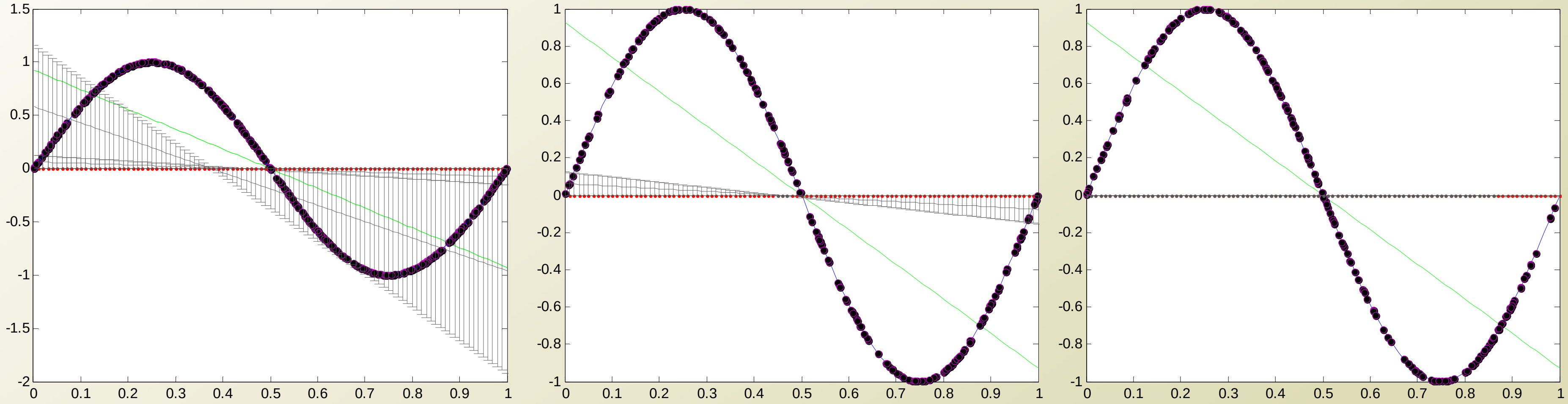
Bias = 0.3092
Var. = 2.0708



Bias = 0.4372
Var. = 0.1167

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

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?

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

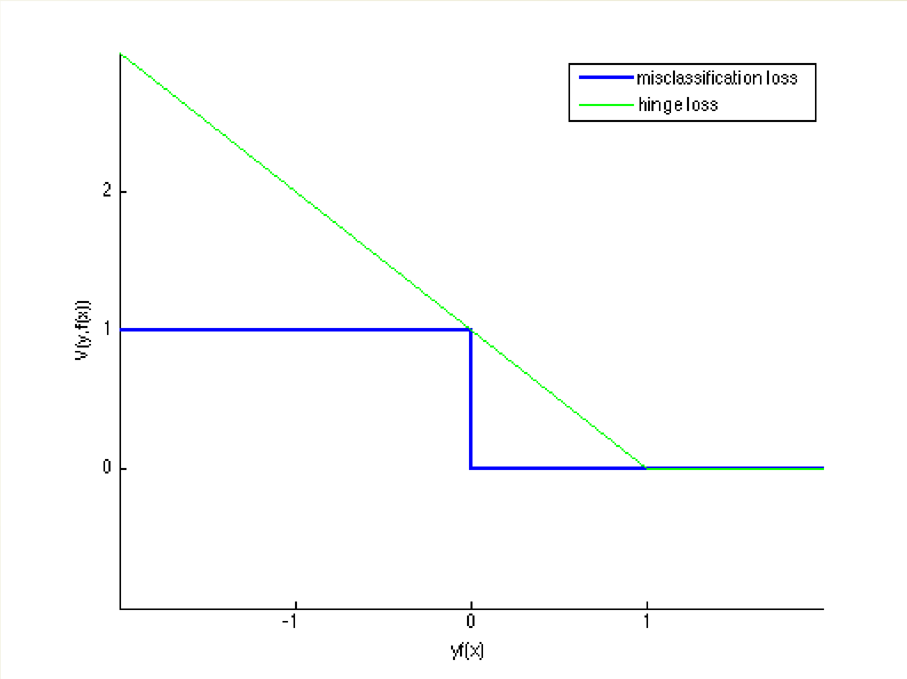
$$\arg\max_{\theta} \sum_{i=1}^{m} \log p(y_i | x_i, \theta) - \alpha R(\theta)$$

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

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

Regularization and SVM

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



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

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

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

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

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

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