The problem

Part I:

Goal: Given $\vec{x} = (x_1, \dots, x_p)$

$y = \{$ response vector

Find a machine $m(\cdot)$ s.t.

$\hat{y} = m(\vec{x})$

Idea: Use training set $T = (\vec{x}_i, y_i)_{i=1}^{N}$

to train or build the machine

$m(\cdot)$.

Part 2: Bias-Variance Problem

Assumption: $T = (X_i, Y_i)$

$X_i \overset{iid}{\sim} f_X \qquad Y_i | X_i \overset{iid}{\sim} f_{Y|X}$

So we can think of T as a R.V.

Define the <u>test set</u> $\vec{x}_0$

and corresponding estimate $\hat{y}_0 = M(x_0)$

Part <u>III</u>: How good is my machine?

Expected - Prediction - Error $= EPE(\vec{x}_0)$

Assume we are in the setting

$$Y = f(\vec{x}) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

known

So a given x_0, $y_0 = f(x_0) + e_0$

known   stochastic

$$EPE(\vec{x}_0) \equiv \mathbb{E}_{T, y_0}\left[(y_0 - \hat{y})^2\right]$$

$$= \mathbb{E}_{y_0}\left[\mathbb{E}_T\left[(y_0 - \hat{y}_0)^2\right]\right]$$

$$= \mathbb{E}_{Y_0}\Big[\mathbb{E}_T\Big[\big(Y_0 - \mathbb{E}_{Y_0}(Y_0)\big) + \big(\mathbb{E}_{Y_0}(Y_0) - \mathbb{E}_T(\hat{Y}_0)\big)$$

$$+ \big(\mathbb{E}_T(\hat{Y}_0) - \hat{Y}_0\big)\Big]^2\Big]$$

$$= \mathbb{E}_{Y_0}\Big[\big(Y_0 - \mathbb{E}_{Y_0}(Y_0)\big)^2\Big] +$$

$$\big(\mathbb{E}(Y_0) - \mathbb{E}_T(\hat{Y}_0)\big)^2 +$$

$$\mathbb{E}_T\Big[\big(\mathbb{E}_T(\hat{Y}_0) - \hat{Y}_0\big)^2\Big]$$

+ expectations of cross terms vanish

$$= \mathrm{Var}(Y_0) + \big(\mathbb{E}(Y_0) - \mathbb{E}_T(\hat{Y}_0)\big)^2 + V_T(\hat{Y}_0)$$

$$= \sigma^2 + \big(f(x_0) - \mathbb{E}_T(\hat{Y}_0)\big)^2 + \mathrm{Var}_T(\hat{Y}_0)$$

variance of target

Bias^2

variance of estimator

So

Thrm: $EPE(\vec{x}_0) = \mathbb{E}_{T, Y_0}\left[(Y_0 - \hat{Y}_0)^2\right]$

$$= Var(Y) + MSE(\hat{Y}_0)$$

Linear Regression

Assume $f(\vec{x}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j = X^T \vec{\beta}$

Goal: Find a method to come up with best $\hat{\beta}$.

$$X = \begin{bmatrix} \text{---} & x_1^T & \text{---} \\ & \vdots & \\ \text{---} & x_n^T & \text{---} \end{bmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$$

Strategy:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta^T x_i)^2$$

$$= \underset{\beta}{\operatorname{argmin}} (Y - X\beta)^T (Y - X\beta)$$

$$\frac{\partial}{\partial \beta} (Y - X\beta)^T (Y - X\beta) \equiv \nabla_\beta RSS(\beta)$$

$$= \begin{pmatrix} \frac{\partial}{\partial \beta_0} RSS(\beta) \\ \vdots \\ \vdots \\ \partial/\partial\beta_p \, RSS(\beta) \end{pmatrix} \stackrel{set}{=} \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$= -2X^T(Y - X\hat{\beta}) \stackrel{set}{=} 0$$

$$\Rightarrow \quad X^T Y = X^T X \hat{\beta}$$ normal equations

$\Rightarrow$ If X is full rank

$$\boxed{\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y}$$

Properties

$$\hat{Y} = X\hat{\beta} = \underbrace{X (X^T X)^{-1} X^T}_{H} Y = HY$$

orthogonal projection matrix.

onto $Col(X)$

Preview of next lecture:

Ridge Regression

$$\hat{\beta}_R = \underset{\beta}{\arg\min} \; \underbrace{(Y - X\beta)^T (Y - X\beta)}_{\text{fit stat.}} + \underbrace{\lambda \beta^T \beta}_{\text{Penalization}}$$

$$\hat{\beta}_R = (X^T X + \lambda I)^{-1} X^T Y$$

"Shrinkage estimator"