Bias Variance Tradeoff.

Goal:

$$EPE(f) = E_{XY}\left[(Y - f(X))^2\right]$$

$$EPE(x_0) = E_{\varepsilon}\left[(Y - f(x_0))^2\right]$$

P_{XY} : X is not random

$$Y = f^{true}(X) + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

$$\{x_i, y_i\}_{i=1}^{N} \sim \underbrace{\underline{\underline{P_{XY} \times P_{XY} \times \cdots \times P_{XY}}}}_{N}$$

Some joint dist

$\mathcal{Z}$

$$\underline{\underline{\hat{f}}}$$

$$EPE(x_0) = E_{\varepsilon} E_{\mathcal{Z}} \left[(\underbrace{Y - \hat{f}(x_0)})^2 \right]$$

$Z_1 \quad P_{Z_1}$

$Z_2 \quad P_{Z_2}$

$f(Z_1, Z_2) = Z_1 + Z_2$

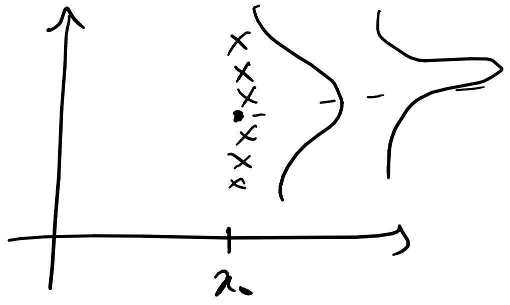
$E_{Z_1} E_{Z_2} [\underline{\underline{f(Z_1, Z_2)}}]$

$$\star \quad EPE(x_0) = \sigma^2 + \text{Bias}^2\left(\hat{f}(x_0)\right) + \text{Var}_Z\left(\hat{f}(x_0)\right)$$

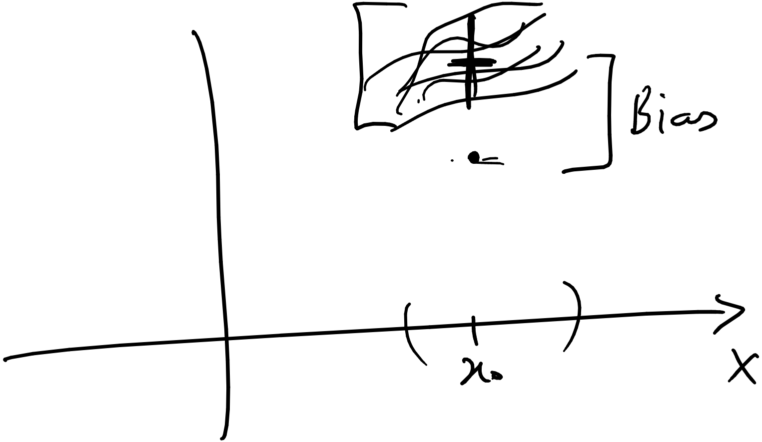
$$\text{Var}_Z\left(\hat{f}(x_0)\right) = E_Z\left[\left(\hat{f}(x_0) - C\right)^2\right]$$

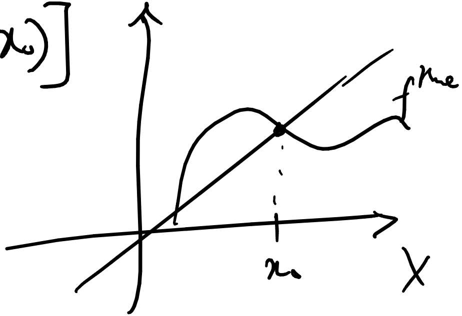
where $C = E_Z\left[\hat{f}(x_0)\right]$

$$\text{Var}(Z) = E\left[\left(Z - E[Z]\right)^2\right]$$



$$\text{Bias}\left(\hat{f}(x_0)\right) = f^{true}(x_0) - E_Z\left[\hat{f}(x_0)\right]$$





Eg: K-nn.

$$\hat{f}(x_0) = \frac{1}{K} \sum_{l=1}^{K} y_l$$

l correspond to the l^{th} closest pt to x_0.

$$y_l = f^{true}(x_l) + \varepsilon_l$$

$$E_Z\left[\hat{f}(x_0)\right] = \frac{1}{K} E_Z\left[y_1 + \cdots + y_K\right]$$

$$= \frac{1}{K} \sum_{l=1}^{K} f^{true}(x_l)$$

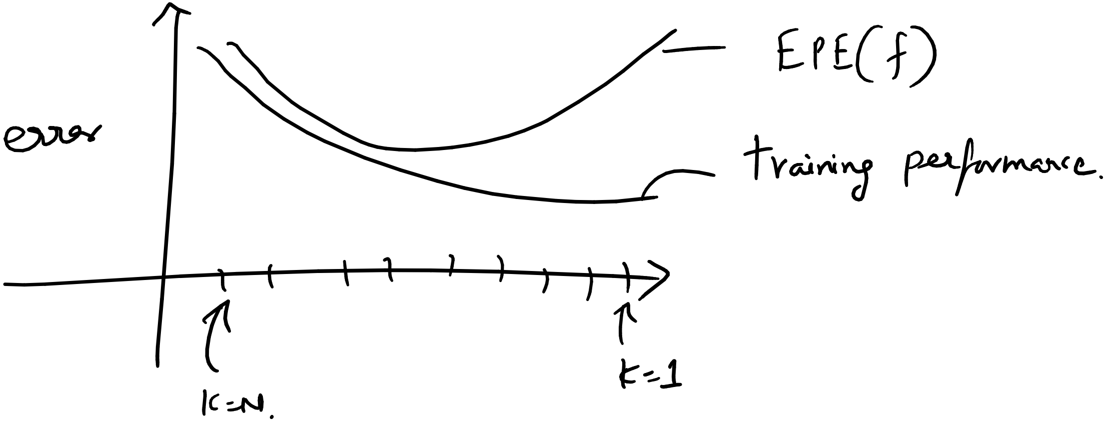
$$EPE(x_0) = \sigma^2 + \underline{\underline{\left(f^{\text{true}}(x_0) - \frac{1}{K}\sum_{\ell=1}^{K} f^{\text{true}}(x_\ell)\right)^2}} + \underline{\frac{\sigma^2}{K}}$$

$$Var_{\tau}(\;) = E_{\tau}\left[\left(\frac{1}{K}\sum_{\ell=1}^{K}\left(f^{\text{true}}(x_\ell) + \varepsilon_\ell\right) - \frac{1}{K}\sum_{\ell=1}^{K} f^{\text{true}}(x_\ell)\right)^2\right]$$

$$= E_{\tau}\left[\left(\frac{1}{K}\sum_{\ell=1}^{K}\varepsilon_\ell\right)^2\right] = \frac{1}{K^2}\left[\sigma^2 + \sigma^2 + \cdots \sigma^2\right]$$

$$= \frac{\sigma^2}{K}$$

$$E[\varepsilon_\ell \varepsilon_j] = 0$$


EPE(f)
training performance.
error
k=N.
k=1

Linear Regression $\rightarrow$ $\underset{p\times 1}{\hat{\beta}} = (X^T X)^{-1} X^T Y$

P_{XY}:

X not random:

$Y = X^T\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$

Fact 1: $\hat{\beta} \sim N(\beta, (X^T X)^{-1}\sigma^2)$

$E[\hat{\beta}] = \beta$ ☆

$Var(\hat{\beta}) = (X^T X)^{-1}\sigma^2$

Null : $\beta_j = 0$

$$z_j = \frac{\hat{\beta}_j}{\sigma\sqrt{V_j}} \sim \mathcal{N}(0, 1)$$

$$\hat{\beta}_{3\times1} \sim N\left(0, \begin{bmatrix} V_1 & x & x \\ x & V_2 & x \\ x & x & V_3 \end{bmatrix}_{3\times3} \sigma^2\right)$$

$$P_{\text{Null}}\left(Z \geq |z_j|\right) < 2\alpha$$

① σ not known

② $X_j \to \beta_j$ Nuance for Categorical variables

$1d: \quad Y = X^T \beta_{p \times 1} \qquad X \quad p \text{ dim}$

$kd: \quad Y_{k \times 1} = \underbrace{\beta^T}_{k \times p} X_{p \times 1}$

$||$

$\tilde{Y}_{1 \times k} = \tilde{X}_{1 \times p} \beta_{p \times k}$

$$\hat{\beta}_{p \times k} = \underbrace{(X^T X)}_{p \times p}^{-1} \underbrace{X^T}_{p \times N} Y_{N \times k}$$

Biasing :

1. Subset selection
2. Ridge regression
3. LASSO

$P \longrightarrow K$

$2^P - 1$

$Y = 1X_1 + \varepsilon$

$Y = 8X_{11} - 7X_{12}$

100 −55

$\{x_i, y_i\}_{i=1}^{N} \longrightarrow$ $K=1$

$\vdots$

$k=k$ ( ) ( ) ( ) ( ) ( )

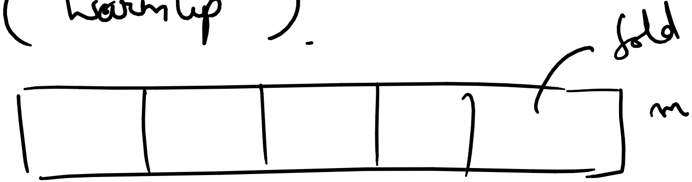
$k=p$

B. "Cross Validation" $EPE(f) = E_{P_{XY}}\left[\left(f(X)-Y\right)^2\right]$

A. Simple training-validation Split

B. CV for k-nn ("warmup").

$\{x_i, y_i\}_{i=1}^{N}$



m observation

$N = 5m.$

for each choice.

for each fold.

hold it out and get $\underbrace{\frac{1}{m}\sum_{i=1}^{m}\left(\hat{f}(x_i) - y_i\right)^2}_{\text{score}_1}$

$\frac{1}{5}\left(\text{score}_1 + \cdots + \text{score}_5\right)$

CV for subset selection.

for each subset size K.

for each subset

get_Score(subset)

min(scores for subsets of size k)

K=4 with features (2, 5, 7, 9) $X_{10\times 1}$

$\hat{\beta}$ using all data.