Bias, Variance, Model Complexity

- Quantitative

Loss function: $L(Y, \hat{f}(X)) = \begin{cases} (Y-\hat{f}(X))^2 & \text{squared error} \\ |Y-\hat{f}(X)| & \text{absolute error} \end{cases}$

$\mathrm{Err}_T = E(L(Y, \hat{f}(X)) \mid T)$

$\mathrm{Err} = E[L(Y, \hat{f}(X))] = E[\mathrm{Err}_T]$

Training error: $\overline{err} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$

- Qualitative

Loss function: $L(G, \hat{G}(X)) = I(G \neq \hat{G}(X))$ (0-1 loss)

$$L(G, \hat{p}(X)) = -2\sum_{k=1}^{K} I(G=k)\log \hat{p}_k(X)$$

$$= -2\log p_G(X) \quad (-2 \times \text{log-likelihood})$$

Training error: $\overline{err} = -\frac{2}{N}\sum_{i=1}^{N} \log \hat{p}_{g_i}(x_i)$

- Data-rich situation:

Train	Validation	Test

- Insufficient data: AIC, BIC, MDL, SRM

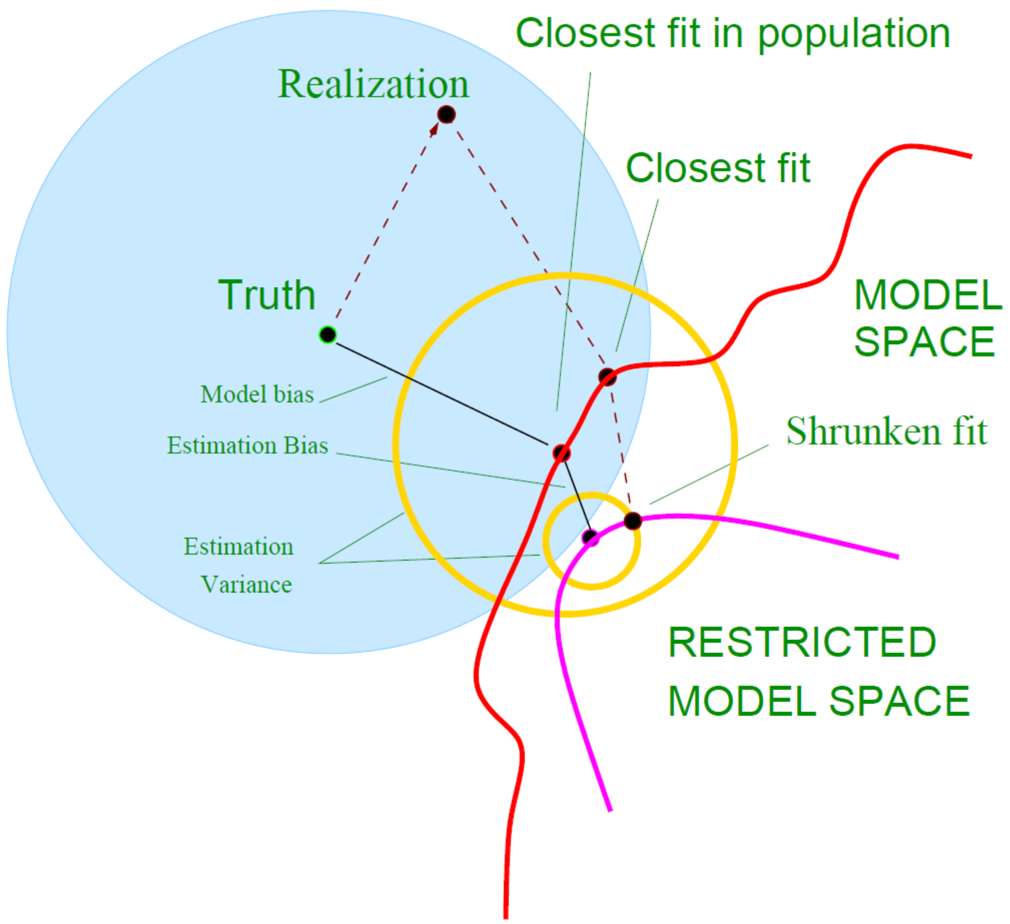
Cross-validation and bootstrap

The Bias-Variance Decomposition

- Assumption: $Y = f(X) + \varepsilon$, $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma_\varepsilon^2$

- Predicted error of a fit $\hat{f}(X)$ at $X = x_0$:

$$
\begin{aligned}
Err(x_0) &= E[(Y - \hat{f}(x_0))^2 \mid X = x_0] \\
&= E[(Y - f(x_0) + f(x_0) - \hat{f}(x_0))^2] \\
&= E[\varepsilon^2] + E[(f(x_0) - \hat{f}(x_0))^2] \\
&= \sigma_\varepsilon^2 + E[f(x_0) - E(\hat{f}(x_0)) + E(\hat{f}(x_0)) - \hat{f}(x_0)]^2 \\
&= \sigma_\varepsilon^2 + [f(x_0) - E(\hat{f}(x_0))]^2 + E(\hat{f}(x_0) - E[\hat{f}(x_0)])^2 \\
&= \sigma_\varepsilon^2 + Bias^2 + Var(\hat{f}(x_0))
\end{aligned}
$$



Optimism of the Training Error Rate

- $Err_{in} = \frac{1}{N}\sum_{i=1}^{N} E_{Y^0}(Y_i^0 - \hat{f}(x_i))^2$

  $\overline{err} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2$

- $op := Err_{in} - \overline{err}$

  $w := E_y(op) = \frac{2}{N}\sum_{i=1}^{N} Cov(\hat{y}_i, y_i)$

$\hat{y}_i$ is affected by y_i

Y_i^0 is not affected by y_i

proof:

$$w = E_y\left(\frac{1}{N}\sum_{i=1}^{N} E_{Y^0}[L(Y_i^0, \hat{f}(x_i) \mid T)] - \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))\right)$$

$$= E_y\left(\frac{1}{N}\sum_{i=1}^{N} E_{Y^0}[(Y_i^0 - \hat{y}_i)^2] - \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\Big(E_y E_{Y^0}[(Y_i^0)^2] + E_y E_{Y^0}[\hat{y}_i^2] - 2E_y E_{Y^0}[Y_i^0 \hat{y}_i]$$

$$- E_y[y_i^2] - E_y[\hat{y}_i^2] + 2E_y[y_i \hat{y}_i]\Big)$$

$E_y E_{Y^0}[(Y_i^0)^2] = E_y[y_i^2]$

$$= \frac{2}{N}\sum_{i=1}^{N}\left(E(y_i \hat{y}_i) - E(y_i)E(\hat{y}_i)\right)$$

$E_y E_{Y^0}[Y_i^0 \hat{y}_i] = E(y_i)E(\hat{y}_i)$

$$= \frac{2}{N}\sum_{i=1}^{N} Cov(y_i, \hat{y}_i)$$

- $\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) = d \cdot \sigma_\varepsilon^2$ (d inputs)

Estimates of In-sample Prediction Error

- C_p- Statistics $C_p = \overline{err} + 2 \cdot \frac{d}{N} \hat{\sigma}_\varepsilon^2$

- AIC (Akaike information criterion)

$$-2 \cdot E(\log Pr_{\hat{\theta}}(Y)) \approx -\frac{2}{N} \cdot E(\log \text{likelihood}) + 2 \cdot \frac{d}{N}$$

$$\Rightarrow AIC = -\frac{2}{N} \sum_{i=1}^{N} \log P_{\hat{\theta}}(y_i) + 2 \cdot \frac{d}{N}$$

The Effective Number of Parameters

- Linear fitting method: $\hat{y} = Sy$

- Define $df(S) = trace(S)$

- $\sum_{i=1}^{N} Cov(y_i, \hat{y}_i) = tr[Cov(y, \hat{y})] = tr[Cov(y, Sy)]$

$$= tr(\sigma_\varepsilon^2 \cdot S) = \sigma_\varepsilon^2 tr(S)$$

- General definition: $df(\hat{y}) = \frac{\sum_{i=1}^{N} Cov(\hat{y}_i, y_i)}{\sigma_\varepsilon^2}$

- With Penalty: $df(\alpha) = \sum_{m=1}^{M} \frac{\theta_m}{\theta_m + \alpha}$ $\quad$ θ_m is the m-th eigenvalue of Hessian Matrix

The Bayesian Approach and BIC

- $\text{BIC} = -2\log\text{-likelihood} + (\log N)\cdot d$

$$= \frac{N}{\sigma_\varepsilon^2}\left[\overline{\text{err}} + (\log N)\cdot\frac{d}{N}\cdot\sigma_\varepsilon^2\right]$$

AIC has $K=2$

BIC has $K=\log N$

- A set of candidate model $\{M_m\}_{m=1}^{M}$

  Known distribution of parameter θ_m $\Pr(\theta_m|M_m)$

- $P(M_m|Z) \propto \Pr(M_m)\cdot\Pr(Z|M_m)$

$$\propto \Pr(M_m)\cdot\int \Pr(Z|\theta_m, M_m)\cdot\Pr(\theta_m|M_m)\,d\theta_m$$

- Bayes factor: $\text{BF}(Z) = \dfrac{\Pr(Z|M_m)}{\Pr(Z|M_l)} = \dfrac{\Pr(M_m)}{\Pr(M_l)}\cdot\dfrac{\Pr(Z|M_m)}{\Pr(Z|M_l)}$

- With Laplace approximation:

$$\log\Pr(Z|M_m) = \log\Pr(Z|\hat{\theta}_m, M_m) - \frac{d_m}{2}\cdot\log N + O(1)$$

$$\text{BIC} = -2\log\Pr(Z|M_m) = -2\log\Pr(Z|\hat{\theta}_m, M_m) + d_m\cdot\log N$$

- Posterior probability of model M_m:

$$P(M_m|Z) = \frac{e^{-\frac{1}{2}\text{BIC}_m}}{\sum_{l=1}^{M} e^{-\frac{1}{2}\text{BIC}_l}} = \frac{P(Z|M_m)P(M_m)}{\sum_l P(Z|M_l)P(M_l)}$$