Model Selection and Assessment

Chau-Wai Wong

Electrical & Computer Engineering
North Carolina State University

Contact: *chauwai.wong@ncsu.edu*. Updated: October 5, 2020.

Model Selection Definition

Model Selection: Choose the best model out of a set of candidate models.

Model Assessment: Having chosen a final model, estimating its prediction/generalization error on new data.

Readings: Chapter 7 of Hastie et al.

Model Selection Examples

(1) Time series:

$$\mathcal{S}_1 = \{AR(1), AR(2), AR(3), ...\}$$

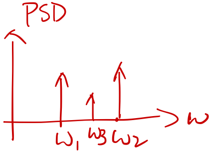
(2) Linear regression:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + e_i, \quad i = 1, \ldots, 50.$$

$$\mathcal{S}_2 = \Big\{\beta_0 \neq 0, \beta_1 \neq 0, \ldots, (\beta_0, \beta_1) \neq \underset{\sim}{0}, (\beta_0, \beta_2) \neq \underset{\sim}{0}, \ldots, (\beta_0, \ldots, \beta_p) \neq \underset{\sim}{0}\Big\}$$

$$\binom{p+1}{1} + \binom{p+1}{2} + \cdots + \binom{p+1}{p+1} = 2^{p+1} - 1$$

Model Selection Examples (cont'd)



(3) Harmonic model:

$$y(n) = \sum_{i=0}^{p} A_i e^{j(\omega_i n + \phi)} + v(n), \quad n = 0, \ldots, 999,$$

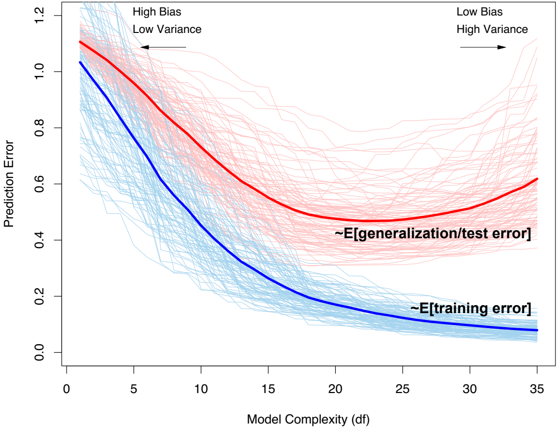
where $v(n) \sim N(0, \sigma_v^2)$, $\phi \sim \text{Uni}(0, 2\pi]$, and (A_i, ω_i) are fixed but unknown parameters.

$$\mathcal{S}_3 = \left\{ A_0 \neq 0, \ldots, (A_0, A_1) \neq \underset{\sim}{0}, \ldots, (A_0, \ldots, A_p) \neq \underset{\sim}{0} \right\}$$

Note that $|\mathcal{S}_2| = |\mathcal{S}_3| = 2^{p+1} - 1$.

Model Selection Criterion: Generalization Performance

A learning method's **generalization performance** is reflected by its prediction capability assessed using **new/test data** drawn from the same population where the data used for training were drawn.



Model Selection in Ideal, Data-Rich Scenario

Split data into two three sets:

Training	Validation	Test

1. Fit K candidate models to the training data.
2. Evaluate the prediction errors using validation data for all models. Select the model with the smallest prediction error. This is called the "validation error."
3. Test the selected model using the test data and evaluate the prediction error. This is called the "test/generalization error."

Model Selection in Ideal, Data-Rich Scenario

Split data into two three sets:

Training	Validation	Test

1. Fit K candidate models to the training data.
2. Evaluate the prediction errors using validation data for all models. Select the model with the smallest prediction error. This is called the "validation error."
3. Test the selected model using the test data and evaluate the prediction error. This is called the "test/generalization error."
4. Question: Why can't validation error be considered as the generalization error? (Hint: Test data mustn't be seen by the model selection process.)

Model Selection in Practical, Data-Limited Scenario

Strategy	**Method**
Sample reuse	Crossvalidation, Bootstrap
Analytically approximate test/generalization step	AIC, BIC, MDL, etc.

Convention: lower vs. upper cases—deterministic vs. random; upper case & bold—deterministic matrix; Tilde below—vector.

Notations: y_i response, $\underset{\sim}{x}_i$ collection of predictors for y_i,
$\mathcal{T} = \left\{(\underset{\sim}{x}_i, y_i), i = 1, \ldots, N\right\}$ deterministic data set,
$\hat{f}_{\mathcal{T}}(\cdot)$ or $\hat{y}_{\mathcal{T}}(\cdot)$ prediction function based on/conditioned on $\mathcal{T}$,
$L(\,\cdot\,,\,\cdot\,)$ loss function, e.g., $L(a,b) = (a-b)^2$ or $L(a,b) = |a-b|$.

Examples when the prediction function is linear:

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}}_{\underset{\sim}{y}} = \underbrace{\begin{bmatrix} \underset{\sim}{x}_1^T \\ \vdots \\ \underset{\sim}{x}_N^T \end{bmatrix}}_{\mathbf{X}} \beta + \underset{\sim}{e},$$

$$\hat{f}_{\mathcal{T}}(\underset{\sim}{x}_0) = \underset{\sim}{x}_0^T \hat{\beta}_{\mathcal{T}} = \underset{\sim}{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \underset{\sim}{y},$$

$$\text{or} = \underset{\sim}{x}_0^T \tilde{\beta}_{\mathcal{T}} = \underset{\sim}{x}_0^T (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \underset{\sim}{y}.$$

Definitions of Test and Training Errors

Generalization/Test error

$$\text{Err}_{\mathcal{T}} = \mathbb{E}\big[L(Y^0, \hat{f}_{\mathcal{T}}(\underset{\sim}{X}^0)|\mathcal{T}\big] \text{ (extra-sample error).}$$

Expected generalization/test error

$$\text{Err} = \mathbb{E}[\text{Err}_{\mathcal{T}}] = \mathbb{E}\Big[\mathbb{E}\big[L(Y^0, \hat{f}_{\mathcal{T}}(\underset{\sim}{X}^0)|\mathcal{T}\big]\Big] = \mathbb{E}\big[L(Y^0, \hat{f}_{\mathcal{T}}(\underset{\sim}{X}^0)\big].$$

Training error

$$\overline{\text{err}} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}_{\mathcal{T}}(\underset{\sim}{x}_i)).$$

Question: How can you modify the definition of training error to define validation error?

Law of total / iterative expectation:

$$\mathbb{E}[X] = \mathbb{E}\Big[\mathbb{E}[X \mid Y]\Big] = \mathbb{E}\Big[\underbrace{\mathbb{E}[X \mid Y=y]}_{g(y)}\,\big|_{y=Y}\Big] = \cdots$$

$$\mathbb{E}[X \mid Y=y] = \int_{x\in\mathbb{R}} x\, f(x \mid y)\, dx$$

$$= \int_{y\in\mathbb{R}} \left(\int_{x\in\mathbb{R}} x\, f(x \mid y)\, dx \right) f(y)\, dy = \int_x x \underbrace{\left(\int_y f(x,y)\, dy \right)}_{f(x)} dx$$

$$= \mathbb{E}[X]$$

$f_{XY}(x,y)$

Joint dist / population

drawn from $f_{XY}(x,y)$ →

$T = \{(\underset{\sim}{x}_i, y_i), i=1, \dots, N\}$

<u>Data</u>, determinstic

↓

$\hat{f}(\cdot) = \hat{f}_T(\cdot)$

Model learned from data T.

equivalent ⇕

$(\underset{\sim}{X}^0, Y^0)$

Random verriables

drawn from $f_{Y|X}(y|x)$

↓

$\{(\underset{\sim}{x}_i, Y_i^0), i=1, \dots, N\}$

$\underset{\sim}{x}_i$, determinstic, Same as in T;

Y_i^0, conditional random on $\underset{\sim}{x}_i$.

<u>Conditional random data</u> created for evaluating generelization/test error.

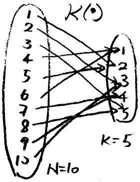
Cross-Validation Motivation & Example

Cross-Validation (CV), sometimes called rotation estimation, or out-of-sample testing.

Data Reuse: Each segment will act as the validation set once, while data in the remaining $K-1$ segments are used to calculate a prediction model.

K-Fold CV, typical choice $K = 5$ or 10. A random partition example when $K = 5$:

Data index:	4, 6	1, 5	2, 10	7, 9	3, 8
Segment index:	1	2	3	4	5
A random partition when $K = 5$	Train	Train	Train	Validation	Train



A training-validation split when the 4th segment is acting as the validation set.

Cross-Validation Error

Cross-Validation error

$$CV(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}^{-\kappa(i)}(\underset{\sim}{x}_i)),$$

where $\kappa : \{1, \ldots, N\} \to \{1, \ldots, K\}$ is a random partition function.

All data points, $(\underset{\sim}{x}_i, y_i), i = 1, \ldots, N$, or all segments, contribute to the CV error.

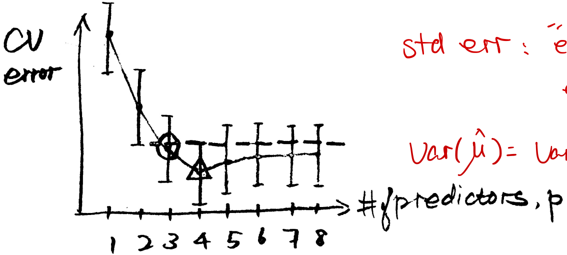
CV error is used to approximate the generalization error.

Note: $CV(\hat{f})$ estimates the expected generalization error, Err, better than the conditional generalization error, $\text{Err}_{\mathcal{T}}$. (See Section 7.12 for more details.)

LOOCV and One SE Rule

Leave-One-Out Cross-Validation (LOOCV): A special case of CV when $K = N$. Approximately unbiased but has large variance as the training datasets are almost the same.

"One standard error rule": Choose the most parsimonious model. Example: CV error for linear regression on polynomials



std err: "estimated std of the estimated value"

$$\mathrm{Var}(\hat{\mu}) = \mathrm{Var}\left(\frac{1}{N}\sum X_i\right) = \frac{\sum \mathrm{Var}(X_i)}{N^2} = \frac{\sigma^2}{N}$$

$$\sqrt{\widehat{\mathrm{Var}}(\hat{\mu})} = \frac{\sigma^2}{N}$$

$\hat{p}_{\text{lowest}} = 4$ and $\hat{p}_{\text{one-std-rule}} = 3$.

Analytic Approximations

Observation: Training error $\overline{\text{err}} < \text{Err}_{\mathcal{T}}$, because the fitted model $\hat{f}_{\mathcal{T}}$ has adapted to data $\mathcal{T}$.

Can we find an correction term and add it to the training error to approximate the generalization error, i.e., $\overline{\text{err}} + \square = \text{Err}_{\mathcal{T}}$?

In-sample prediction error

$$\text{Err}_{\text{in}} = \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\left[L(Y_k^0, \hat{f}_{\mathcal{T}}(\underset{\sim}{x}_k)) | \mathcal{T}\right],$$

which is defined similarly to $\text{Err}_{\mathcal{T}}$ but uses $\{(\underset{\sim}{x}_i, Y_i^0)\}_{i=1}^{N}$ instead of $\{(X_i^0, Y_i^0)\}_{i=1}^{\infty}$.

$\text{Err}_{\text{in}} \approx \text{Err}_{\mathcal{T}}$ if (1) $\underset{\sim}{x}_i$ is uniformly sampled from population, and (2) N is large.

The Correction Term: Optimism

Optimism

$$\text{op} \stackrel{\text{def}}{=} \text{Err}_{\text{in}} - \overline{\text{err}}.$$

Expected optimism

$$\omega \stackrel{\text{def}}{=} \mathbb{E}[\text{op} | \{\underset{\sim}{x}_i\}_{i=1}^N].$$

Example: $\omega = \frac{2}{N} \sum_{i=1}^N \text{cov}(\hat{y}_i, y_i)$. The harder we fit, the greater the covariance, and the more op.

Analytic Form of Optimism

$$\mathbb{E}[\mathrm{Err}_{\mathrm{in}}|\{\underset{\sim}{x}_i\}] = \overline{\mathrm{err}} + \frac{2}{N}\sum_{i=1}^{N}\mathrm{cov}(\hat{y}_i, y_i).$$

If $\hat{y}_i$ is from linear model with d predictors, we have

$$\mathbb{E}[\mathrm{Err}_{\mathrm{in}}|\{\underset{\sim}{x}_i\}] = \overline{\mathrm{err}} + 2\cdot\frac{d}{N}\cdot\sigma_e^2.$$

Try to validate the above expression for parameters d, N, and σ_e^2 using a linear regression model as a special case.

Analytic Approximations

Analytic Models: Akaike information criterion (AIC), Bayesian information criterion (BIC), Minimum description length (MDL).

$\star$ One way to estimate the in-sample prediction error $\mathrm{Err}_{\mathrm{in}}$ is to estimate the optimism and then add it to the training error $\overline{\mathrm{err}}$:

$$AIC \text{ or } C_p = \overline{\mathrm{err}} + 2 \cdot \frac{d}{N} \cdot \hat{\sigma}_e^2$$

$$BIC = \frac{N}{\hat{\sigma}_e^2}\left[\overline{\mathrm{err}} + (\log N) \cdot \frac{d}{N} \cdot \hat{\sigma}_e^2\right]$$

Detailed Derivations

Evaluating $\mathbb{E}\left[\text{Err}_{\text{in}} \middle| \{\underset{\sim}{x}_i\}\right]$

$$
\begin{aligned}
\mathbb{E}\left[\text{Err}_{\text{in}} | \{\underset{\sim}{x}_i\}\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left[L(Y_k^0, \hat{f}_{\mathcal{T}}(\underset{\sim}{x}_k)) | \mathcal{T}\right] \middle| \{\underset{\sim}{x}_i\}\right] \\
&= \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left[\mathbb{E}\left[L(Y_k^0, \hat{f}_{\mathcal{T}}(\underset{\sim}{x}_k)) | \{\underset{\sim}{x}_i\}, \{y_i\}\right] \middle| \{\underset{\sim}{x}_i\}\right] \\
&= \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left[L(Y_k^0, \hat{f}_{\mathcal{T}}(\underset{\sim}{x}_k)) | \{\underset{\sim}{x}_i\}\right] \\
&\overset{\text{def}}{=} \frac{1}{N}\sum_{k=1}^{N}\text{Err}(\underset{\sim}{x}_k)
\end{aligned}
$$

The Bias-Variance Decomposition for $\mathrm{Err}(\underset{\sim}{x}_k)$

Let… $\mathrm{Err}(\underset{\sim}{x}_0) = \mathbb{E}\left[L(Y_0^0, \hat{f}_{\mathcal{T}}(\underset{\sim}{x}_0)) \,\middle|\, \{x_i\} \right]$

Note $Y_0^0 = f(\underset{\sim}{x}_0) + e$ ← new error, $\mathbb{E}[e] = 0$, $Var(e) = \sigma_e^2$.

$$\mathrm{Err}(x_0) = \mathbb{E}\left[\left(f(\underset{\sim}{x}_0) + e - \hat{f}_{\mathcal{T}}(\underset{\sim}{x}_0) \right)^2 \,\middle|\, \{x_i\} \right] = \mathbb{E}\left[\left(f(\underset{\sim}{x}_0) - \hat{f}_{\mathcal{T}}(\underset{\sim}{x}_0) \right)^2 \,\middle|\, \{x_i\} \right] + \sigma_e^2$$

$$= \mathbb{E}\left[\left(f(x_0) - \mathbb{E}[\hat{f}_{\mathcal{T}}(x_0) \mid \{x_i\}] + \mathbb{E}[\hat{f}_{\mathcal{T}}(x_0) \mid \{x_i\}] - \hat{f}_{\mathcal{T}}(x_0) \right)^2 \,\middle|\, \{x_i\} \right] + \sigma_e^2$$

$$= \mathbb{E}\left[\left(f(x_0) - \mathbb{E}\hat{f}_{\mathcal{T}}(x_0) \right)^2 \,\middle|\, \{x_i\} \right] + \mathbb{E}\left[\left(\mathbb{E}\hat{f}_{\mathcal{T}}(x_0) - \hat{f}_{\mathcal{T}}(x_0) \right)^2 \,\middle|\, \{x_i\} \right] + \sigma_e^2$$

$$+ 2\mathbb{E}\left[\underbrace{\left(f(x_0) - \mathbb{E}\hat{f}_{\mathcal{T}}(x_0) \right)}_{\text{value}} \left(\mathbb{E}[\hat{f}_{\mathcal{T}}(x_0) \mid \{x_i\}] - \hat{f}_{\mathcal{T}}(x_0) \right) \,\middle|\, \{x_i\} \right]$$

$$= \mathbb{E}\left[\mathrm{bias}^2(\hat{f}_{\mathcal{T}}(x_0)) \,\middle|\, \{x_i\} \right] + Var\left(\hat{f}_{\mathcal{T}}(x_0) \,\middle|\, \{x_i\} \right) + \sigma_e^2$$

$$= \text{bias}^2 + \text{variance} + \text{irreducible error}$$

$$\mathbb{E}\left[\mathrm{Err}_{in} \,\middle|\, \{x_i\} \right] = \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\left[\mathrm{bias}^2(x_k) \right] + \frac{1}{N} \sum_{k=1}^{N} Var\left(\hat{f}_{\mathcal{T}}(x_k) \,\middle|\, \{x_i\} \right) + \sigma_e^2 \qquad (6)$$

Special Case for the Linear Regression Model

Linear model $\underset{\sim}{y} = \mathbf{X}\beta + \underset{\sim}{e}$ using $\hat{\beta}_{LS} = (X^T X)^{-1} X^T y$ as example:

...

$$\hat{y} = \hat{f}_{\mathcal{J}}(x_0) = \underset{\sim}{x}_0^T \hat{\beta}_{LS} = \underset{\sim}{x}_0^T (X^T X)^{-1} X^T y \quad (7)$$

$$Var\left(\hat{f}_{\mathcal{J}}(x_0) \mid \{x_i\}\right) = \underset{\sim}{x}_0^T (X^T X)^{-1} X^T \underbrace{Var(y \mid \{x_i\})}_{\sigma_e^2 \mathbb{I}} X (X^T X)^{-1} \underset{\sim}{x}_0$$

$$= \left[\underset{\sim}{x}_0^T (X^T X)^{-1} \underset{\sim}{x}_0\right] \sigma_e^2 \quad (8)$$

2^{nd} term for Eq.(6)

$$= \frac{\sigma_e^2}{N} \sum_{k=1}^{N} \underset{\sim}{x}_k^T (X^T X)^{-1} \underset{\sim}{x}_k = \frac{\sigma_e^2}{N} \sum_{k=1}^{N} tr\left\{\underset{\sim}{x}_k^T (X^T X)^{-1} \underset{\sim}{x}_k\right\} \quad (9)$$

$$= \frac{\sigma_e^2}{N} tr\left\{\sum_{k=1}^{N} \underset{\sim}{x}_k \underset{\sim}{x}_k^T (X^T X)^{-1}\right\} = \frac{\sigma_e^2}{N} tr\left\{\mathbb{I}_{p \times p}\right\} = \frac{p}{N} \sigma_e^2 .$$