Model Assessment and Selection

Matthew S. Shotwell, Ph.D.

Department of Biostatistics
Vanderbilt University School of Medicine
Nashville, TN, USA

March 9, 2020

Model assessment

- for supervised learning, assess model using test error
- different types of test error for different purposes
- different methods to estimate of test error
- today: Mallow's Cp, AIC, BIC

Test error

- test error - average loss using test data
- $\tau = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ - training data
- $\tau_0 = \{(x_{01}, y_{01}), \ldots, (x_{0N_0}, y_{0N_0})\}$ - testing data

$$\mathrm{Err}_\tau = \frac{1}{N_0} \sum_{i=1}^{N_0} L(y_{0i}, \hat{f}_\tau(x_{0i}))$$

- model $\hat{f}_\tau$ depends on training data τ
- synonyms - conditional test error, conditional prediction error
- conditional on training sample τ

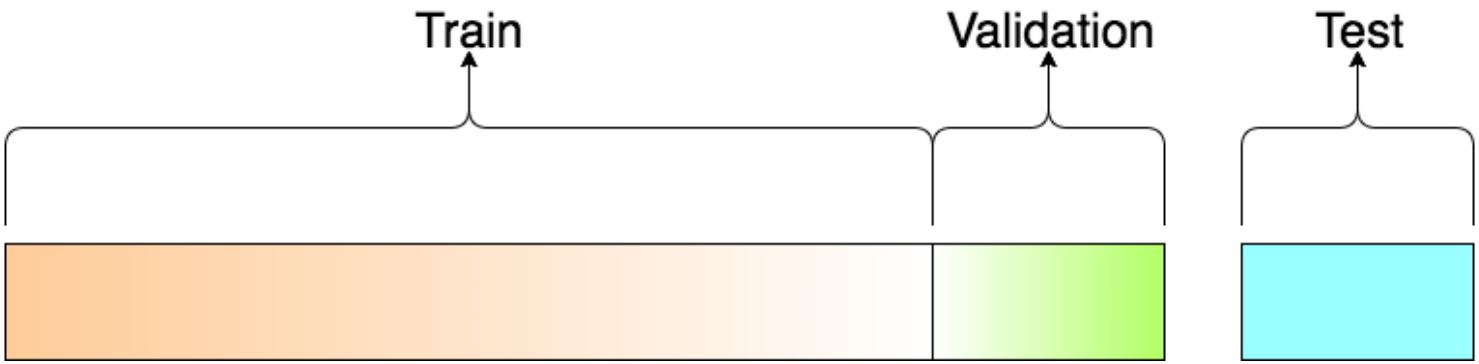
Average test error

- different τ (training data) $\rightarrow$ different $\hat{f}_\tau$
- average test error - test error, averaged over training samples
- $\{\tau_1, \ldots, \tau_B\}$ - multiple training samples
- $\tau_0 = \{(x_{01}, y_{01}), \ldots, (x_{0N_0}, y_{0N_0})\}$ - testing data
- average test error:

$$\begin{aligned} \mathrm{Err} &= \frac{1}{B} \sum_{b=1}^{B} \mathrm{Err}_{\tau_b} \\ &= \frac{1}{B} \sum_{b=1}^{B} \frac{1}{N_0} \sum_{i=1}^{N_0} L(y_{0i}, \hat{f}_{\tau_b}(x_{0i})) \end{aligned}$$

- synonyms - expected test error, expected prediction error

Conditional test error vs average test error

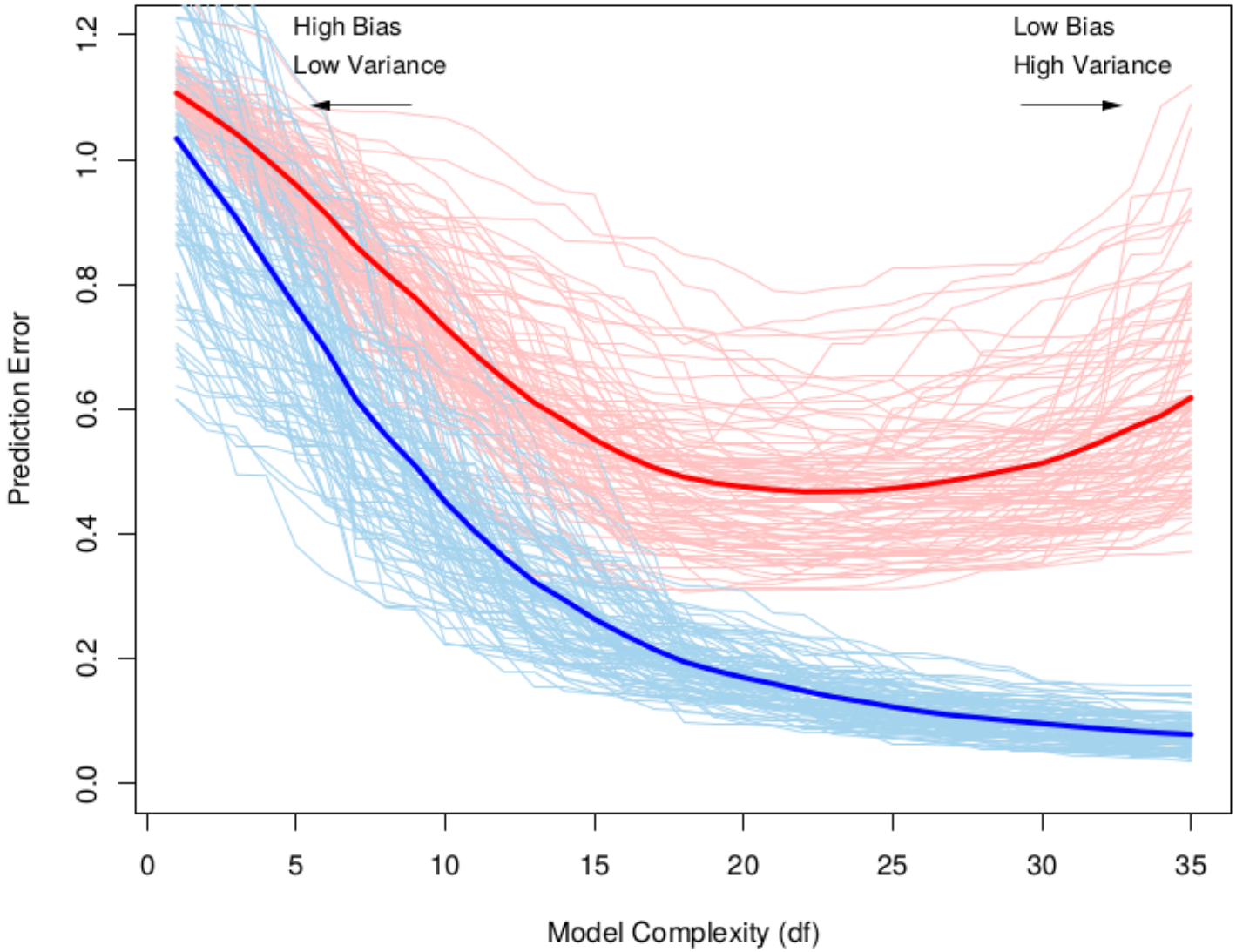


- conditional test error: how will *this* model perform?
- average test error: how does *modeling procedure* perform?
- which to use for model tuning?
- in practice, sometimes used interchangeably

How do we get a good estiamte of average/conditional test error? Next few lectures devoted to this. Today we'll consider methods that start with training error and add some quantity.

- training error too small (optimistic)
- add something to training error to approximate test error

How much to add to training error?



Training error and In-sample error

- training error:

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

- $\tau = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ - training data
- $\tau_y = \{(x_1, y_{01}), \ldots, (x_N, y_{0N})\}$ - testing y at training x
- in-sample error:

$$\text{Err}_{\text{in}} = \frac{1}{N}\sum_{i=1}^{N} L(y_{0i}, \hat{f}_\tau(x_i))]$$

- Err_{in} is easy to work with
- Err_{in} an estimate of Err_τ or Err?
- What can we add to $\overline{\text{err}}$ to estimate Err_{in}

Optimism and expected optimism

- optimism:

$$\mathrm{op} = \mathrm{Err}_{\mathrm{in}} - \overline{\mathrm{err}}$$

- average optimism: $\omega = E_{\tau_y}[\mathrm{op}]$
- E_{τ_y} denotes average conditional on training x_i
- for squared-error loss and 0-1 loss:

$$\omega = \frac{2}{N} \sum_{i=1}^{N} \mathrm{Cov}(\hat{Y}_i, Y_i)$$

- estimate ω, then approximate $\mathrm{Err}_{\mathrm{in}}$ by adding to $\overline{\mathrm{err}}$

Effective degrees-of-freedom

Suppose $\mathrm{var}(Y|X) = \sigma^2$. For some $\hat{Y} = \hat{f}(X)$:

$$\mathrm{df}(\hat{Y}) = \frac{1}{\sigma^2} \sum_{i=1}^{N} \mathrm{cov}(\hat{Y}_i, Y_i)$$

- what happens when $\hat{y} = y$? Or when $\hat{y} = 0$
- $\mathrm{df}(\hat{Y})$ measures where we are on bias-variance spectrum
- larger $\mathrm{df}(\hat{Y})$ means less bias, more variance
- larger $\mathrm{df}(\hat{Y})$ means more flexible model

Effective degrees-of-freedom

- $\mathrm{df}(\hat{Y})$ is sometimes specified (smoothing splines)
- for linear smoothers: $\hat{y} = S_\lambda y$ where $\hat{y}$ is a vector of predictions at training inputs x, and y are the training outputs:

$$\mathrm{df}(\hat{Y}) = \mathrm{trace}(S_\lambda)$$

- works for kernel methods

Optimism and expected optimism

- optimism:

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

- expected optimism: $\omega = E_{\tau_y}[\text{op}]$
- for squared-error loss and 0-1 loss:

$$\omega = \frac{2}{N} \sum_{i=1}^{N} \text{Cov}(\hat{Y}_i, Y_i)$$

- ω proportional to $\text{df}(\hat{Y}) = \frac{1}{\sigma^2} \sum_{i=1}^{N} \text{Cov}(\hat{Y}_i, Y_i)$

Optimism for linear models

- say $Y = X\beta + \epsilon$ where $\mathrm{var}(\epsilon) = \sigma^2$
- d is the number of inputs
- $\omega = \frac{2}{N} \sum_{i=1}^{N} \mathrm{Cov}(\hat{f}(x_i), y_i)$
- $\omega = \frac{2}{N} \mathrm{df}(\beta) \sigma^2$
- $\omega = \frac{2}{N} d \sigma^2$

Estimates of $\mathrm{Err}_{\mathrm{in}}$: Mallow's C_p

- $\widehat{\mathrm{Err}_{\mathrm{in}}} = \overline{\mathrm{err}} + \hat{\omega}$
- for linear models (squared error loss) - $\hat{\omega} = \frac{2}{N} d \hat{\sigma}^2$
- Mallow's $C_p = \overline{\mathrm{err}} + \frac{2}{N} d \hat{\sigma}^2$

Estimates of $\mathrm{Err}_{\mathrm{in}}$: AIC

- consider "entropy loss" $L(Y, \theta) = -2 \log Pr(Y|X, \theta)$
- $\overline{\mathrm{err}} = -\frac{2}{N} \sum_{i=1}^{N} \log Pr(y_i|x_i, \hat{\theta}) = -\frac{2}{N} l(\hat{\theta}|y, x)$
- Akaike showed that $\omega \to \frac{2}{N} d$ asymptotically, for entropy loss, and where d is the number of parameters in θ
- $\mathrm{AIC} = \overline{\mathrm{err}} + \frac{2}{N} d$
- for smoothers, substitute d for effective degrees of freedom

Estimates of $\mathrm{Err}_{\mathrm{in}}$: BIC

- Bayesians select model M by maximizing posterior $Pr(M|Y)$
- Schwarz (father of BIC) showed that:

$$\begin{aligned}\log Pr(M|Y) &\propto \frac{2}{N}\sum_{i=1}^{N} l(\hat{\theta}|Y_i) - \frac{\log N}{N}d \\ &= -\overline{\mathrm{err}} - \frac{\log N}{N}d\end{aligned}$$

- $\mathrm{BIC} = \overline{\mathrm{err}} + \frac{\log N}{N}d$
- maximizing $Pr(M|Y)$ approximately same as minimizing BIC

Mallow's C_p, AIC, BIC

Note that Mallow's C_p, AIC, and BIC:

- approximate (conditional) test error
- training error + estimate of ω (average optimism)
- uses only the training data
- not as good as data splitting (or cross-validation)
- quick and dirty

AIC vs BIC

- $\mathrm{AIC} = \overline{\mathrm{err}} + \frac{2}{N} d$
- $\mathrm{BIC} = \overline{\mathrm{err}} + \frac{\log N}{N} d$
- select model that minimizes AIC or BIC
- which penalizes large models more?
- what to do when d unknown?