Model Assessment

Wei Li

Syracuse University

Spring 2021

Effective Degrees of freedom

Some concepts related to errors

Bias-variance decomposition (revisit)

Optimism of the training error rate

Estimates of in-sample prediction error

LOOCV for linear smoothers

Effective Degrees of freedom

Linear smoothers:

The prediction of f at the training data points, which is of the form:

$$\hat{f} = Sy$$

for some matrix depending on data points and possibly on some tuning parameter (bandwidth or penalizaton coefficient) but not y.

Degree of freedom:

The effective number of parameters used by the procedure, it provides a quantitative measure of the estimator complexity.

examples

- For projection-based regression, for instance regression splines, $M = tr(H_P)$ givens the dimension of the projection space, which is also the number of basis funcions, and the number of parameters involved in the fit.
  - For regression splines of degree d with K knots, then $M = d + K + 1$.
  - the df for a regression natural spline is equal to the number of knots
  - For B-splines, it is equal to the number of interior knots plus order of the splines.
- For (cubic) smoothing-splines, $df_\lambda = tr(S_\lambda) = \sum_{i=1}^{n} \frac{1}{1+\lambda d_i}$.
  - Note that $df_\lambda \to n$, $S_\lambda \to I$ as $\lambda \to 0$;
  - $df_\lambda \to 2$, $S_\lambda \to H_{ols}$ (the hat matrix of OLS), as $\lambda \to \infty$.

If $y = f(x) + \epsilon$ where $var(\epsilon) = \sigma^2$.

- For projection linear smoothers: Let $\hat{y}$ denote the linear fit with d-inputs or basis functions (basis-regression), then

  $$\sum_{i=1}^{n} cov(\hat{y}_i, y_i) = d\sigma^2$$

  the covariance *is conditional on the predictors or treating predictors as fixed.*

- For shrinking linear smoother, $\hat{y} = Sy$ for some S that does not depend on y,

  $$\sum_{i=1}^{n} cov(\hat{y}_i, y_i) = tr(S)\sigma^2.$$

One can define the so-called effective d.f. of a fitted model as

$$d.f.(\hat{f}) = \frac{\sum_{i=1}^{n} cov(\hat{y}_i, y_i)}{\sigma^2} = tr(S)$$

examples:

- ridge regression or smoothing splines, $d.f.(\hat{f}) = tr(S_\lambda)$.
- a best subset selection of size k (k fixed), $d.f.(\hat{f})$ would be greater than k
- k-nearest-neighbor average, $\hat{y} = Sy$ where $S_{i,j} = w(x_i, x_j)$ where $w(x_i, x_j) = K_k(x_i, x_j)/\sum_{i=l}^{n} K_k(x_i, x_l)$, and $K_k(x_0, x) = 1(\|x - x_0\| \leq \|x_{(k)} - x_0\|)$, where x_k is the training observation ranked k-th in distance from x_0.
  $tr(S) = \sum_{i=1}^{n} w(x_i, x_i) = n/k$.

Some concepts related to errors

Some concepts related to errors

Suppose we have i.i.d. sample $\boldsymbol{\tau} = (y_i, x_i)_{i=1}^n$, where y are continuous and generated as $y_i = f(x_i) + \epsilon_i$. Let $\hat{f}$ denote the estimate of f, based on the training data $\boldsymbol{\tau}$.

- **Training error**: the average loss over the training samples

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

- **Expected Training error**:

$$E_{\boldsymbol{\tau}}(\overline{err})$$

- **Test error** (generalization error):

$$Err_{\boldsymbol{\tau}} = E[L(Y^*, \hat{f}(X^*))|\boldsymbol{\tau}]$$

  where (Y^*, X^*) is a new draw that are independent of $(y_i, x_i)_{i=1}^n$. It can be estimated by

$$\overline{Terr} = \frac{1}{n'} \sum_{i=1}^{n'} L(y_i^*, \hat{f}(x_i^*))$$

- **Expected prediction error**:

$$\text{EPE} = \text{Err} = E_{(Y^*, X^*, \boldsymbol{\tau})} \left[L\left(Y^*, \hat{f}(X^*) \right) \right] = E_{\boldsymbol{\tau}}[Err_{\boldsymbol{\tau}}]$$

  The first expectation is taken w.r.t. all random quantities.

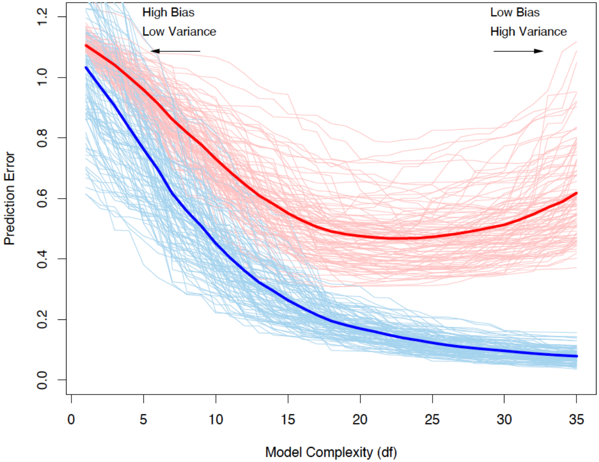
- **Expexcted prediction error at point** x_0:

$$Err(x_0) = \text{EPE}(x_0) = E_{(Y^*, \boldsymbol{\tau})} \left[L\left(Y^*, \hat{f}(X^*) \right) |X^* = x_0 \right]$$

  Then $EPE = \text{Err} = E_{X_0}[Err(X_0)]$.

Bias-variance decomposition (revisit)

Quantitative response



ESL Fig7.1

The training error rate $\overline{err}$ is not a good estimate of test error rate $Err_{\boldsymbol{\tau}}$:

- The training error rate often is quite different from the test error rate.
- Training error can dramatically underestimate test Error.
- Training error decreases with model complexity.
- A model with zero training error overfits the training data; over-fitted models typically generalize poorly.

For regression $Y = f(X) + \epsilon$, it can be shown that the expected (squared) prediction error at x_0 is given by

$$\begin{aligned} EPE(x_0) &= E[(Y^* - \hat{f}(X^*))^2 | X^* = x_0] \\ &= \sigma^2 + E_{\boldsymbol{\tau}}(\hat{f}(x_0) - f(x_0))^2 \\ &= \sigma^2 + bias^2(\hat{f}(x_0)) + var_{\boldsymbol{\tau}}(\hat{f}(x_0)) \end{aligned}$$

where $bias(\hat{f}(x_0)) = E_{\boldsymbol{\tau}}\hat{f}(x_0) - f(x_0)$.

Categorical response

If Y takes values $\{1, \ldots, K\}$, the common loss functions are

$0-1$ loss:
$L(Y, \hat{h}(X)) = I(Y \neq \hat{h}(X)),$ where $\hat{h}(x) = argmax_k \hat{p}_k(x)$

negative log-likelihood (cross-entropy, deviance):
$L(Y, \hat{p}_Y(X)) = -2 \sum_{k=1}^{K} I(Y = k) \log \hat{p}_k(X)$
$= -2 \log \hat{p}_Y(X)$

where $\hat{p}_k(X) = \hat{Pr}(Y = k|X)$ and $\hat{p}_Y(X)$ is the estimate of the probability $Pr(Y|X)$.

Categorical response 0/1 loss

The bias-variance tradeoff behaves differently for $0 - 1$ loss than it does for squared error loss.

But the prediction error ($0 - 1$ loss) is no longer the sum of squared bias and variance, because the squared bias is not suitable for measuring $0 - 1$ loss.

What matters is that $E\hat{f}(x_0)$ and $f(x_0)$ is on the same side of $1/2$ (thus correct classiffication).

For $(X,Y) \in \mathbb{R}^p \times \{0,1\}$, consider the regression function, defined as usual as

$$f(x) = E(Y \mid X = x) = P(Y = 1 \mid X = x)$$

The Bayes classifier is given by

$$h^*(x) = \begin{cases} 0 & \text{if } f(x) \le 1/2 \\ 1 & \text{if } f(x) > 1/2 \end{cases}$$

The plug-in classifier is given by

$$\hat{h}(x) = \begin{cases} 0 & \text{if } \hat{f}(x) \le 1/2 \\ 1 & \text{if } \hat{f}(x) > 1/2 \end{cases}$$

$$\begin{aligned}\mathrm{Err}\left(x_0\right) &= \mathrm{P}\left(Y^* \neq \hat{h}\left(X^*\right) \mid X^* = x_0\right) \\ &= \mathrm{Err}_{\mathrm{B}}\left(x_0\right) + \left|2f\left(x_0\right) - 1\right| \mathrm{P}\left(\hat{h}\left(X^*\right) \neq h^*\left(X^*\right) \mid X^* = x_0\right)\end{aligned}$$

where $\mathrm{Err}_{\mathrm{B}}\left(x_0\right) = \mathrm{P}\left(Y^* \neq h^*\left(X^*\right) \mid X^* = x_0\right)$, the irreducible Bayes error at x_0.

Using the approximation $\hat{f}\left(x_0\right) \sim N\left(\mathrm{E}\hat{f}\left(x_0\right), \mathrm{Var}\left(\hat{f}\left(x_0\right)\right)\right)$, it can be shown that

$$\Pr\left(\hat{h}\left(X^*\right) \neq h^*\left(X^*\right) \mid X^* = x_0\right) \approx \Phi\left(\frac{\operatorname{sign}\left(\frac{1}{2} - f\left(x_0\right)\right)\left(\mathrm{E}\hat{f}\left(x_0\right) - \frac{1}{2}\right)}{\sqrt{\mathrm{Var}\left(\hat{f}\left(x_0\right)\right)}}\right)$$

The term $\operatorname{sign}\left(\frac{1}{2} - f(x_0)\right)\left(\mathrm{E}\hat{f}(x_0) - \frac{1}{2}\right)$ is a kind of **boundary-bias term**, as it depends on the true $f(x_0)$ only through which side of the boundary $\left(\frac{1}{2}\right)$ that it lies.

The bias and variance combine in a multiplicative rather than additive fashion.

Optimism of the training error rate

Optimism of the training error rate

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

$$Err_{\boldsymbol{\tau}} = E[L(Y^*, \hat{f}(X^*))|\boldsymbol{\tau}]$$

The $\overline{err}$ is less than the true error $Err_{\boldsymbol{\tau}}$.

The quantity $Err_{\boldsymbol{\tau}}$ is **extra-sample error**.

Consider the **in-sample error** (conditional on $\boldsymbol{\tau}$):

$$\mathrm{Err}_{\boldsymbol{\tau},\mathrm{in}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{Y^*}\left[L\left(Y_i^*, \hat{f}\left(x_i\right)\right) \mid \boldsymbol{\tau}\right]$$

The Y_i^* notation indicates that we observe n new response values at each of the training points $x_i, i = 1, 2, \ldots, n$. The expectation above is only with respect to the new response Y_i^* at each of the training points $x_i, i = 1, 2, \ldots, n$.

- $\mathrm{Err}_{\boldsymbol{\tau},\mathrm{in}}$ is not often the direct interest.
- the comparison of in-sample error $\mathrm{Err}_{\boldsymbol{\tau},\mathrm{in}}$ is convenient and often leads to efficient model selection.

Optimiam

$$\mathrm{op} \equiv \mathrm{Err}_{\boldsymbol{\tau},\mathrm{in}} - \overline{\mathrm{err}}$$

Average optimism over *the training set response values* (predictors in the training sets held fixed):

$$\omega \equiv \mathrm{E}_{\mathbf{y}}(\mathrm{op})$$

For squared error, $0-1$, and other loss functions, generally that

$$\omega = \frac{2}{n}\sum_{i=1}^{n} \mathrm{Cov}\left(\hat{y}_i, y_i\right)$$

Here $\mathrm{Cov}(\hat{y}_i, y_i)$ the covariance is only taken w.r.t to the response values in the training set.

$$\mathrm{E}_{\mathbf{y}}\left(\mathrm{Err}_{\boldsymbol{\tau},\mathrm{in}}\right) = \mathrm{E}_{\mathbf{y}}(\overline{\mathrm{err}}) + \frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}\left(\hat{y}_i, y_i\right)$$

For the additive error model $Y = f(X) + \varepsilon$, where $\hat{y}$ is fitted by a linear smoother $\hat{y} = Sy$, due to $\sum_{i=1}^{n}\mathrm{Cov}\left(\hat{y}_i, y_i\right) = tr(S)\sigma^2$, above is equivalent to

$$\mathrm{E}_{\mathbf{y}}\left(\mathrm{Err}_{\boldsymbol{\tau},\mathrm{in}}\right) = \mathrm{E}_{\mathbf{y}}(\overline{\mathrm{err}}) + 2\cdot\frac{tr(S)}{n}\sigma^2$$

Above expression gives the fundamental identity base on which C_p, AIC and BIC can be used to select model (and estimate the in-sample error $\mathrm{Err}_{\boldsymbol{\tau},\mathrm{in}}$).

Estimates of in-sample prediction error

Estimates of in-sample prediction error

This class of method estimate in-sample error $Err_{\boldsymbol{\tau},in}$ by adding the estimated optimism to the $\overline{err}$. These methods include C_p, AIC and BIC.

The general form of the in-sample estimates is

$$\widehat{\mathrm{Err}}_{\boldsymbol{\tau},\mathrm{in}} = \overline{\mathrm{err}} + \hat{\omega}$$

where $\hat{\omega}$ is an estimate of the optimism.

C_p

For regression model $y = f(x) + \epsilon$, using squared-loss, let d be the number of parameters, $\hat{f}$ be a linear fit (maybe using basis functions).

Then

$$C_p = \overline{\text{err}} + 2 \cdot \frac{d}{n}\hat{\sigma}^2 = \frac{RSS_d}{n} + 2 \cdot \frac{d}{n}\hat{\sigma}^2$$

where $RSS_d = \sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$ and $\hat{\sigma}^2$ is an estimate of the noise parameter from a full model or a low-bias model.

Then choose the model d that gives the smallest C_p.

AIC

AIC is derived using the (negative) log-likelihood loss.

$$-2 \cdot \mathrm{E}\left[\log \Pr_{\hat{\theta}}(Y)\right] \approx -\frac{2}{n} \cdot \mathrm{E}[\log \mathrm{lik}] + 2 \cdot \frac{d}{n}$$

Here $\Pr_\theta(Y)$ is a family of densities for Y, $\hat{\theta}$ is the maximum-likelihood estimate of θ, and "loglik" is the maximized log-likelihood:

$$\log \mathrm{lik} = \sum_{i=1}^{n} \log \Pr_{\hat{\theta}}\left(y_i\right)$$

Thus,

$$\mathrm{AIC} = -\frac{2}{n} \cdot \log \mathrm{lik} + 2 \cdot \frac{d}{n}$$

As long as σ^2 assumed known or estimated using a full model,

$$AIC = \frac{RSS_d}{n\sigma^2} + \frac{2d}{n}$$

AIC and C_p are equivalent (up to some factor).

BIC

The Bayesian information criterion (BIC), like AIC, is applicable in settings where the fitting is carried out by maximization of a log-likelihood. The generic form of BIC is

$$\mathrm{BIC} = -2 \cdot \log \mathrm{lik} + (\log n) \cdot d$$

where $loglik$ is the maximized log-likelihood function.

For a Gaussian regression model where σ^2 is known or estiamted using low-bias model,

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

So $BIC \approx AIC$ or C_P (up to the multiplicant $1/\sigma^2$ and with 2 replaed by $\log n$).

Above formula are valid when d are fixed (without being learned from data). Suppose our linear smoother of interest depends on a tuning parameter α(e.g., h for kernel smoothing, λ for smoothing splines, or λ for Mercer kernels), and express this as $\hat{f}_\alpha = S_\alpha y$. Then we could choose the tuning parameter α to minimize the estimated test error, as in

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \frac{1}{n} \|y - S_\alpha y\|_2^2 + \frac{2\sigma^2}{n} \operatorname{tr}(S_\alpha)$$

This is just like the C_p criterion, or AIC, in ordinary linear regression (we could also replace the factor of 2 above with $\log n$ to obtain something like BIC).

LOOCV for linear smoothers

LOOCV for linear smoothers

CV estimates are unbiased for Err (expected prediction error):

$$E(CV(\hat{f})) \approx Err.$$

Leave-one-out CV

$$CV(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}^{-(i)}(x_i))^2$$

where $\hat{f}^{-(i)}$ is fitted using all training data except the i-th observation.

For linear smoothers, the LOOCV has a simple form if $\hat{f} = Sy$ where S does not depend on y,

$$\mathrm{CV}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}^{-(i)}(x_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}(x_i)}{1 - S_{ii}}\right)^2$$

A computationally simpler approximation to LOOCV.

$$\mathrm{GCV}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}(x_i)}{1 - \mathrm{tr}(S)/n}\right)^2 = (1-\nu/n)^{-2}\overline{err}$$

where ν is the effective degrees of freedom and $\overline{err}$ is the training error. This can be of computational advantage in some cases where $\mathrm{tr}(S)$ is easier to compute that individual elements S_{ii}.

Using the approximation $\frac{1}{(1-x)^2} \approx 1 + 2x$ we can write the above as

$$\begin{aligned} \mathrm{GCV}(\hat{f}) &\approx \frac{1}{n} \sum_{i=1}^{N} \left(y_i - \hat{f}\left(x_i\right)\right)^2 \left(1 + \frac{2\,\mathrm{tr}(S)}{n}\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}\left(x_i\right)\right)^2 + \frac{2}{n^2}\,\mathrm{tr}(S) \sum_{i=1}^{n} \left(y_i - \hat{f}\left(x_i\right)\right)^2 \\ &= \overline{err} + \frac{2tr(S)}{n}\overline{err} \end{aligned}$$