Decompositions of the Prediction Error

Stephen Vardeman
Analytics Iowa LLC
ISU Statistics and IMSE

Quantifying the performance of a predictor

- We suppose that training cases $(\mathbf{x}_i, y_i)$ for $i = 1, 2, \ldots, N$ and test case $(\mathbf{x}, y)$ are iid from distribution P and that predictor $\hat{f}$ is built using the training set $\mathbf{T}$

- Then expected prediction loss (alternatively called the "prediction error," "test error," and "generalization error") suffered using $\hat{f}$ is

$$\mathrm{Err} \equiv \mathrm{E}^{\boldsymbol{T}} \mathrm{E}^{(x,y)} \left[L\left(\hat{f}(x), y \right) \right]$$

- Two decompositions of this help make clear what must be controlled in order to make prediction error small

General decomposition of prediction error

- Use the notation

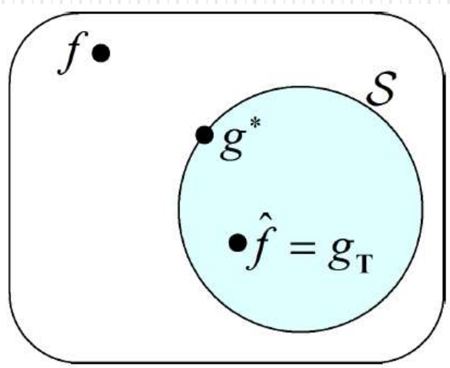
  f for the (theoretically) optimal predictor

  $\mathcal{S}$ for a class of functions g from which a predictor can be chosen

  $\hat{f} = g_{\mathbf{T}}$ for the training-set-dependent element of $\mathcal{S}$ used for prediction

  g^* for a (fixed) element of $\mathcal{S}$ with minimum expected prediction loss

- The situation can then be pictured as below in terms of the optimal, restricted optimal, and fitted predictors



General decomposition of prediction error

- The optimal f is potentially (typically) outside $\mathcal{S}$
- The "closest" (in terms of prediction error) one can get to it inside $\mathcal{S}$ is g^*
- For any fixed training set $\hat{f} = g_{\mathbf{T}}$ can be no better than g^*
- As the training set varies randomly, how much better g^* is than $\hat{f}$ varies randomly

General decomposition of prediction error

- So since here

$$\mathrm{Err} = \mathrm{E}^{\boldsymbol{T}}\mathrm{E}^{(\boldsymbol{x},y)}L\left(\hat{f}\left(\boldsymbol{x}\right),y\right) = \mathrm{E}^{\boldsymbol{T}}\mathrm{E}^{(\boldsymbol{x},y)}L\left(g_{\boldsymbol{T}}\left(\boldsymbol{x}\right),y\right)$$

we have

$$\begin{aligned}\mathrm{Err} = \mathrm{E}^{(\boldsymbol{x},y)}L\left(f\left(\boldsymbol{x}\right),y\right) &+ \left(\mathrm{E}^{(\boldsymbol{x},y)}L\left(g^{*}\left(\boldsymbol{x}\right),y\right) - \mathrm{E}^{(\boldsymbol{x},y)}L\left(f\left(\boldsymbol{x}\right),y\right)\right) \\ &+ \left(\mathrm{E}^{\boldsymbol{T}}\mathrm{E}^{(\boldsymbol{x},y)}L\left(g_{\boldsymbol{T}}\left(\boldsymbol{x}\right),y\right) - \mathrm{E}^{(\boldsymbol{x},y)}L\left(g^{*}\left(\boldsymbol{x}\right),y\right)\right)\end{aligned}$$

- This can be thought of as

$$\begin{aligned}\mathrm{Err} = \text{minimum expected loss possible} &+ \text{modeling penalty} \\ &+ \text{fitting penalty}\end{aligned}$$

SEL decomposition of prediction error

- A more detailed and illuminating decomposition of Err is possible for SEL
- Begin with a measure of prediction performance at input vector $\mathbf{x}$

$$\mathrm{Err}(\boldsymbol{x}) \equiv \mathrm{E}^{\boldsymbol{T}}\mathrm{E}\left[\left(\hat{f}(\boldsymbol{x}) - y\right)^2 |\boldsymbol{x}\right]$$

(noting that $\mathrm{Err} = \mathrm{E}^{\mathbf{x}}\,\mathrm{Err}(\mathbf{x})$)

- Then

$$\begin{aligned}\mathrm{Err}(\boldsymbol{x}) &= \mathrm{E}^{\boldsymbol{T}}\left\{\left(\hat{f}(\boldsymbol{x}) - \mathrm{E}[y|\boldsymbol{x}]\right)^2 + \mathrm{E}\left[(y - \mathrm{E}[y|\boldsymbol{x}])^2 |\boldsymbol{x}\right]\right\} \\ &= \mathrm{E}^{\boldsymbol{T}}\left\{\left(\hat{f}(\boldsymbol{x}) - \mathrm{E}^{\boldsymbol{T}}\hat{f}(\boldsymbol{x})\right)^2 + \left(\mathrm{E}^{\boldsymbol{T}}\hat{f}(\boldsymbol{x}) - \mathrm{E}[y|\boldsymbol{x}]\right)^2\right\} + \mathrm{Var}[y|\boldsymbol{x}] \\ &= \mathrm{Var}^{\boldsymbol{T}}\left(\hat{f}(\boldsymbol{x})\right) + \left(\mathrm{E}^{\boldsymbol{T}}\hat{f}(\boldsymbol{x}) - \mathrm{E}[y|\boldsymbol{x}]\right)^2 + \mathrm{Var}[y|\boldsymbol{x}]\end{aligned}$$

SEL decomposition of prediction error

- $\mathrm{Var}^{\mathbf{T}}\left(\hat{f}(\mathbf{x})\right)$ is the variance of prediction at $\mathbf{x}$
- $\left(\mathrm{E}^{\mathbf{T}}\hat{f}(\mathbf{x}) - \mathrm{E}[y \mid \mathbf{x}]\right)^2$ is a kind of squared bias of prediction at $\mathbf{x}$
- $\mathrm{Var}\left[y \mid \mathbf{x}\right]$ is an unavoidable variance in outputs at $\mathbf{x}$
- Clearly then

$$\mathrm{Err} = \mathrm{E}^{\mathbf{x}}\mathrm{Var}^{\mathbf{T}}\left(\hat{f}(\mathbf{x})\right) + \mathrm{E}^{\mathbf{x}}\left(\mathrm{E}^{\mathbf{T}}\hat{f}(\mathbf{x}) - \mathrm{E}[y \mid \mathbf{x}]\right)^2 + \mathrm{E}^{\mathbf{x}}\mathrm{Var}\left[y \mid \mathbf{x}\right]$$

and SEL prediction error is a sum of averages (against the marginal of $\mathbf{x}$) of the three quantities at fixed inputs

SEL decomposition of prediction error

- Further analysis of $\mathrm{E}^{\mathbf{x}}\left(\mathrm{E}^{\mathbf{T}}\,\hat{f}(x)-\mathrm{E}[y \mid x]\right)^2$ provides additional insight

- Suppose that $\mathbf{T}$ is used to select a function $g_{\mathbf{T}}$ from some linear subspace $\mathcal{S}$ of the the space of functions h with $\mathrm{E}^{\mathbf{x}}\left(h\left(\mathbf{x}\right)\right)^2 < \infty$ and $\hat{f} = g_{\mathbf{T}}$. Further, let

$$g^{*}\left(\mathbf{x}\right) = \arg\min_{g \in \mathcal{S}} \mathrm{E}^{\mathbf{x}}\left(g\left(\mathbf{x}\right) - \mathrm{E}\left[y|\mathbf{x}\right]\right)^2$$

(the best approximation to the optimal predictor in $\mathcal{S}$ and projection of $\mathrm{E}[y|\mathbf{x}]$ onto $\mathcal{S}$).

SEL decomposition of prediction error

- It's then possible to argue that

$$\mathrm{E}^{\boldsymbol{x}}\left(\mathrm{E}^{\boldsymbol{T}}\hat{f}(\boldsymbol{x})-\mathrm{E}[y|\boldsymbol{x}]\right)^2=\mathrm{E}^{\boldsymbol{x}}\left(\mathrm{E}^{\boldsymbol{T}}\hat{f}(\boldsymbol{x})-g^*(\boldsymbol{x})\right)^2+\mathrm{E}^{\boldsymbol{x}}\left(\mathrm{E}[y|\boldsymbol{x}]-g^*(\boldsymbol{x})\right)^2$$

- The first term on the right is an average (across inputs) squared fitting bias
- The second term is an average (across inputs) squared model bias
- So ultimately for SEL

$$\begin{aligned}\mathrm{Err}&=\mathrm{E}^{\boldsymbol{x}}\mathrm{Var}[y|\boldsymbol{x}]+\mathrm{E}^{\boldsymbol{x}}\left(\mathrm{E}[y|\boldsymbol{x}]-g^*(\boldsymbol{x})\right)^2\\&\quad+\mathrm{E}^{\boldsymbol{x}}\left(\mathrm{E}^{\boldsymbol{T}}\hat{f}(\boldsymbol{x})-g^*(\boldsymbol{x})\right)^2+\mathrm{E}^{\boldsymbol{x}}\mathrm{Var}^{\boldsymbol{T}}\left(\hat{f}(\boldsymbol{x})\right)\end{aligned}$$

SEL decomposition of prediction error

- The SEL decomposition is related to the general one in that

$$\text{minimum expected loss possible} = \text{expected (across } \boldsymbol{x}\text{) response variance} = \mathrm{E}^{\boldsymbol{x}}\mathrm{Var}\left[y|\boldsymbol{x}\right],$$

$$\text{modeling penalty} = \text{expected (across } \boldsymbol{x}\text{) squared model bias} = \mathrm{E}^{\boldsymbol{x}}\left(\mathrm{E}\left[y|\boldsymbol{x}\right] - g^{*}\left(\boldsymbol{x}\right)\right)^{2},$$

$$\text{fitting penalty} = \left(\begin{array}{c}\text{expected (across } \boldsymbol{x}\text{)} \\ \text{squared fitting bias}\end{array}\right) + \left(\begin{array}{c}\text{expected (across } \boldsymbol{x}\text{)} \\ \text{prediction variance}\end{array}\right) = \mathrm{E}^{\boldsymbol{x}}\left(\mathrm{E}^{\boldsymbol{T}}\hat{f}\left(\boldsymbol{x}\right) - g^{*}\left(\boldsymbol{x}\right)\right)^{2} + \mathrm{E}^{\boldsymbol{x}}\mathrm{Var}^{\boldsymbol{T}}\left(\hat{f}\left(\boldsymbol{x}\right)\right)$$

SEL decomposition of prediction error

- The facts that
  - the modeling and fitting penalties have elements of both bias and variance
  - complex predictors tend to have low bias and high variance in comparison to simple ones

lead to the necessity of balancing these elements in predictor development and the so-called **variance-bias trade-off**

- Once more, in qualitative terms, it is the matching of predictor complexity to real information content of a training set that is at issue here