Penalized Fitting and Choosing Complexity

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Penalized training error fitting and complexity

- A way of creating an explicit complexity parameter (and subsequently using it to choose a predictor complexity) involves penalization

Suppose that one can define for every element of the class of functions $\mathcal{S} = \{g\}$ a complexity penalty $J[g] \geq 0$ and for every $\lambda \geq 0$ defines a measure of undesirability for g reflecting both fit to the training data and complexity by

$$\frac{1}{N}\sum_{i=1}^{N} L\left(g\left(\boldsymbol{x}_i\right), y_i\right) + \lambda J[g]$$

If one can optimize this objective over choices of g, call the optimizing function $\hat{f}_\lambda$. The smaller is λ, the more complex will be $\hat{f}_\lambda$.

Examples

- Two simple examples of complexity penalties are below.

For $p = 1$ SEL prediction on $(0, 1)$ with
$\mathcal{S} = \left\{ g | g'' \text{ exists and } \int_0^1 (g''(x))^2 \, dx < \infty \cap \right\}$, a complexity penalty might be

$$J[g] = \int_0^1 (g''(x))^2 \, dx$$

(This penalizes how much g "wiggles.")

For $p = 1$ SEL prediction on $\Re$ and g of the form
$g(x) = \beta_1 x + \beta_2 x^2 + \beta_3 x^3$ (for standardized x) a complexity penalty might be

$$J[g] = \beta_2^2 + \beta_3^2$$

that penalizes how wiggly/non-linear the cubic predictor is.

Pick-the-cv-winner choice of λ

- If λ^{opt} minimizes cross-validation errors and one ultimately employs λ^{opt} and the criterion

$$\frac{1}{N}\sum_{i=1}^{N} L\left(g\left(\boldsymbol{x}_i\right), y_i\right) + \lambda^{\text{opt}} J\left[g\right]$$

  with the whole training set $(\boldsymbol{T})$, the pick-the-winner predictor $\tilde{f} = \hat{f}_{\lambda^{\text{opt}}}$ is produced.