Cross-Validation and Predictor Choice

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Cross-validation-based choice of predictor

- Armed with cross-validation errors for multiple prediction methodologies, the most obvious way to use them to choose among predictors is to simply "**pick the** (cross-validation error) **winner**"

- Other possible methods have been suggested (including a "one standard error" rule) but they seem to lack convincing motivation and clear evidence of superiority to the pick-the-winner rule

- A somewhat subtle but important point is that the "winning cross-validation error" is NOT a reliable indicator of the likely performance of the (whole pick-the-winner) strategy actually strategy employed! To obtain such a thing one needs to cross-validate the whole methodology (picking a typically different "winner" inside each of *K* remainders)

Cross-validating "Pick-the-CV-Winner"

- In order to assess the likely performance of

$$\tilde{f} = \arg\min_{\hat{f}} CV\left(\hat{f}\right)$$

via cross-validation, **inside each remainder $\boldsymbol{T} - \boldsymbol{T}_k$** one must

1. **split into K folds,**
2. **fit on the K remainders,**
3. **predict on the folds and make a cross-validation error,**
4. **pick a winner for the function in 3., say $\tilde{f}^k$, and**
5. **then predict on $\boldsymbol{T}_k$ using $\tilde{f}^k$.**

It is the values $\tilde{f}^{k(i)}\left(\boldsymbol{x}_i\right)$ that are used to make a cross-validation error for $\tilde{f}$ applied to the whole training set.

Cross-validating pick-the-winner performance

- The basic principle at work here (and always) in making valid cross-validation errors is that **whatever one will ultimately do in the entire training set to make a predictor must be redone** (in its entirety!) **in every remainder and applied to the corresponding fold**

- This point holds whether one is choosing between fundamentally different methodologies or "just" picking a tuning parameter for a single basic methodology