Use Stacking to Average Models for Time Series and Panel Data Under Continuous Ranked Probability Score

Yuling Yao

April 12, 2020

1   Method

The CRPS for forecast CDFs with a finite first moment can be written as

$$crps(F, y) = \mathrm{E}_X \left| X - y \right| - \frac{1}{2} \mathrm{E}_{X,X'} \left| X - X' \right|.$$

Given K prediction models and observations y_{tr} (number of observed icu needs) at time $t = 1, \ldots, T$ and region $r = 1, \ldots, R$, we could generate one-day-ahead prediction in each model, and draw K predictive samples $x_{1ktr}, \ldots, x_{Sktr}$ for the k-th model on the day t and region r. The prediction uses all information before t to eliminate over-fitting.

Using these predictive draws, we can compute the crps of the k-th model on the tr-th observation by

$$\widehat{crps}_{ktr} = \widehat{crps}(x_{1ktr}, \ldots, x_{sktr}, y_{tr}) = \frac{1}{S} \sum_{s=1}^{S} \left| x_{sktr} - y_{tr} \right| - \frac{1}{2S^2} \sum_{s,j=1}^{S} \left| x_{sktr} - x_{jktr} \right|.$$

Now the goal is to aggregate these K models. It is easy to show that, when prediction is a mixture with weights $w_1, \ldots, w_s$, the CRPS can be expressed as

$$\widehat{crps}_{agg,tr}(w_1, \ldots, w_K) = \frac{1}{S} \sum_{k=1}^{K} w_k \sum_{s=1}^{S} \left| x_{skt} - y_t \right| - \frac{1}{2S^2} \left(\sum_{k=1}^{K} \sum_{k,k'=1}^{K} w_k w_{k'} \sum_{s,j=1}^{S} \left| x_{skt} - x_{jk't} \right| \right).$$

This expression is quadratic on w. We only need to compute $\sum_{s=1}^{S} \left| x_{skt} - y_{tr} \right|$, $\sum_{s,j=1}^{S} \left| x_{sktr} - x_{jktr} \right|$, and $\sum_{s,j=1}^{S} \left| x_{sktr} - x_{jk'tr} \right|$ for all k, k' pairs once and store them for all weight values in the optimization.

An extra degree of freedom is how we weight the prediction over time and over regions. In the early days we have very few data so all predictions are noisy and we might want to down weight the early phase in the final model evaluation. Mathematically we can pre-fixed introduce a time-varying weight $\lambda_1, \ldots, \lambda_T$, say $\lambda_t = 1.5 - (1 - t/T)^2$, it penalizes early estimates (smaller t). Likewise we take into account the association of regions and individual sample size by a region-specific weight τ_r

Finally we solve a quadratic optimization:

$$\min_{w_1, \ldots, w_K} \sum_{t=1}^{T} \sum_{r=1}^{R} \lambda_t \tau_r \widehat{crps}_{agg,tr}(w),$$

$$s.t. \ 0 \le w_1, \ldots, w_K \le 1, \sum_{k=1}^{K} w_k = 1.$$