Journal Club: SuperLearner

by Mark van der Laan, Eric Polley and Alan Hubbard (2007)

Alex Hayes

2018-01-23

Motivation

We want to combine multiple models together in a way that achieves minimum prediction error. How do we do it?

Answer: stacking

1. Split the original data X into k folds
2. For each fold:
    - Train each model on the other folds
    - Use this trained model to predict on the current fold
3. Aggregate the predictions on held out folds into new matrix Z.
4. Train a *metalearner* on Z.

To predict on new data:

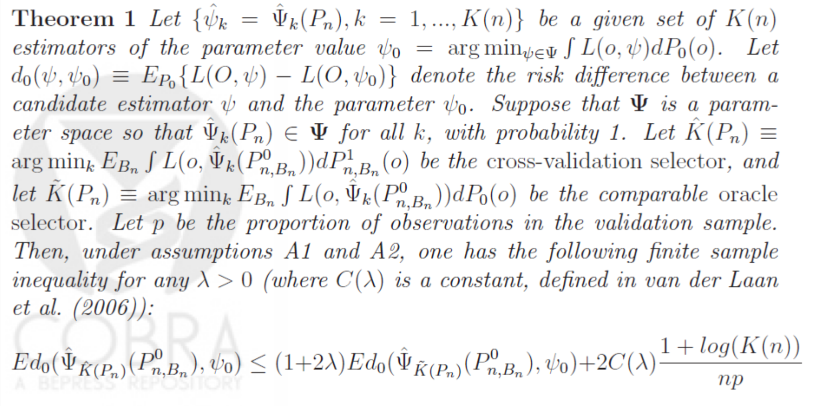
1. Run data through each of the models in the ensemble
2. Use these predictions to create Z'
3. Run Z' through the metalearner

Optimality result: English

> *The super learner performs as well (in terms of expected risk difference) as the oracle selector, up to a typically second order term.*

If one of the candidate models is a correctly specified parametric model, the Super Learner acheives the “almost parametric” rate of convergence $\frac{\log n}{n}$. Otherwise it performs asymptotically as well as the best possible combination of models.

Optimality result: Math



Theorem 1 *Let* $\{\hat{\psi}_k = \hat{\Psi}_k(P_n), k = 1, ..., K(n)\}$ *be a given set of* $K(n)$ *estimators of the parameter value* $\psi_0 = \arg\min_{\psi\in\Psi} \int L(o,\psi)dP_0(o)$. *Let* $d_0(\psi,\psi_0) \equiv E_{P_0}\{L(O,\psi) - L(O,\psi_0)\}$ *denote the risk difference between a candidate estimator* ψ *and the parameter* ψ_0. *Suppose that* $\mathbf{\Psi}$ *is a parameter space so that* $\hat{\Psi}_k(P_n) \in \mathbf{\Psi}$ *for all* k, *with probability 1. Let* $\hat{K}(P_n) \equiv \arg\min_k E_{B_n} \int L(o, \hat{\Psi}_k(P^0_{n,B_n}))dP^1_{n,B_n}(o)$ *be the* cross-validation selector, *and let* $\tilde{K}(P_n) \equiv \arg\min_k E_{B_n} \int L(o, \hat{\Psi}_k(P^0_{n,B_n}))dP_0(o)$ *be the comparable* oracle selector. *Let* p *be the proportion of observations in the validation sample. Then, under assumptions A1 and A2, one has the following finite sample inequality for any* $\lambda > 0$ *(where* $C(\lambda)$ *is a constant, defined in van der Laan et al. (2006)):*

$$Ed_0(\hat{\Psi}_{\hat{K}(P_n)}(P^0_{n,B_n}), \psi_0) \leq (1+2\lambda)Ed_0(\hat{\Psi}_{\tilde{K}(P_n)}(P^0_{n,B_n}), \psi_0) + 2C(\lambda)\frac{1 + log(K(n))}{np}$$

What are assumptions A1 and A2?

A1: The loss function $L(O, \psi) = (Y - \psi(X))^2$ is uniformly bounded

A2: The variance of ψ_0 centered loss function $L(O, \psi) - L(O, \psi_0)$ can be bounded by its expectation uniformly in ψ

Extension + Question

The *Subsemble* algorithm by Erin LeDell is less computationally expensive but achieves the same optimality via partitioning data out to each candidate learner

Why does you have you use V-Fold cross validation instead of, say, the bootstrap?

Paper available at goo.gl/UrxnT7