PRINCETON UNIVERSITY

Error Reduction from Stacked Regressions

Jason M. Klusowski

Operations Research and Financial Engineering (ORFE)

Joint work with Xin Chen (Princeton) & Yan Shuo Tan (NUS)

Stacked Generalizations and Stacked Regressions

- Data analyst rarely knows a priori what the true model is

Stacked Generalizations and Stacked Regressions

- Data analyst rarely knows a priori what the true model is
- Starts by deriving collection of candidate models $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_M$

Stacked Generalizations and Stacked Regressions

- Data analyst rarely knows a priori what the true model is
- Starts by deriving collection of candidate models $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_M$
- Next step is to select best model based on criterion such as AIC, BIC, or out-of-sample error (e.g., cross-validation)

Stacked Generalizations and Stacked Regressions

- Data analyst rarely knows a priori what the true model is
- Starts by deriving collection of candidate models $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_M$
- Next step is to select best model based on criterion such as AIC, BIC, or out-of-sample error (e.g., cross-validation)
- David Wolpert (1992): Use predictions from these estimators as inputs for another (combined) model

Stacked Generalizations and Stacked Regressions

- Data analyst rarely knows a priori what the true model is
- Starts by deriving collection of candidate models $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_M$
- Next step is to select best model based on criterion such as AIC, BIC, or out-of-sample error (e.g., cross-validation)
- David Wolpert (1992): Use predictions from these estimators as inputs for another (combined) model
- Leo Breiman (1996): Operationalized Wolpert's idea by restricting combined models to have form

$$\sum_{k=1}^{M} \hat{\alpha}_k \hat{\mu}_k$$

Stacked Generalizations and Stacked Regressions

- Data analyst rarely knows a priori what the true model is
- Starts by deriving collection of candidate models $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_M$
- Next step is to select best model based on criterion such as AIC, BIC, or out-of-sample error (e.g., cross-validation)
- David Wolpert (1992): Use predictions from these estimators as inputs for another (combined) model
- Leo Breiman (1996): Operationalized Wolpert's idea by restricting combined models to have form

$$\sum_{k=1}^{M} \hat{\alpha}_k \hat{\mu}_k$$

- Method has found widespread applications (finance, healthcare, commerce, ..., Kaggle competitions)

Stacked Model

- How to learn stacking weights $\hat{\alpha}_k$?

Stacked Model

- How to learn stacking weights $\hat{\alpha}_k$?
- Breiman suggested cross-validation to estimate test error

$$\mathbb{E}_{(x,y)}\left[\left(y - \sum_{k=1}^{M} \alpha_k \hat{\mu}_k(x)\right)^2\right]$$

Stacked Model

- How to learn stacking weights $\hat{\alpha}_k$?
- Breiman suggested cross-validation to estimate test error

$$\mathbb{E}_{(x,y)}\left[\left(y - \sum_{k=1}^{M} \alpha_k \hat{\mu}_k(x)\right)^2\right]$$

- Could regularize with ridge constraint $\sum_{k=1}^{M} \alpha_k^2 = t$, since estimators $\hat{\mu}_k$ are usually highly correlated (trying to estimate same thing)

Stacked Model

- How to learn stacking weights $\hat{\alpha}_k$?
- Breiman suggested cross-validation to estimate test error

$$\mathbb{E}_{(x,y)}\left[\left(y - \sum_{k=1}^{M} \alpha_k \hat{\mu}_k(x)\right)^2\right]$$

- Could regularize with ridge constraint $\sum_{k=1}^{M} \alpha_k^2 = t$, since estimators $\hat{\mu}_k$ are usually highly correlated (trying to estimate same thing)
- Better to make weights non-negative, i.e., $\alpha_k \geqslant 0$

Stacked Model

- How to learn stacking weights $\hat{\alpha}_k$?
- Breiman suggested cross-validation to estimate test error

$$\mathbb{E}_{(x,y)}\left[\left(y - \sum_{k=1}^{M} \alpha_k \hat{\mu}_k(x)\right)^2\right]$$

- Could regularize with ridge constraint $\sum_{k=1}^{M} \alpha_k^2 = t$, since estimators $\hat{\mu}_k$ are usually highly correlated (trying to estimate same thing)
- Better to make weights non-negative, i.e., $\alpha_k \geqslant 0$
- Minimizers $\hat{\alpha}_k$ yield stacked model $\hat{\mu}_{\text{stack}}(x) = \sum_{k=1}^{M} \hat{\alpha}_k \hat{\mu}_k(x)$

What Breiman Found

This resulting predictor $\sum_{k=1}^{M} \hat{\alpha}_k \hat{\mu}_k(x)$ *appears to almost always have lower prediction error than the single prediction* $\hat{\mu}_k$ *having lowest cross-validation error. The word "appears" is used because a general proof is not yet in place. — Leo Breiman (1996)*

What Breiman Found

This resulting predictor $\sum_{k=1}^{M} \hat{\alpha}_k \hat{\mu}_k(x)$ *appears to almost always have lower prediction error than the single prediction* $\hat{\mu}_k$ *having lowest cross-validation error. The word "appears" is used because a general proof is not yet in place. — Leo Breiman (1996)*

Goal of talk is to theoretically confirm this in certain cases

Breiman's Experiments with Nested Regression Trees

- $M = 50$ nested regression trees $\hat{\mu}_k$ from pruning
- Weights $\hat{\alpha}_k$ sum to 0.96

Table 1. Test Set Prediction Errors

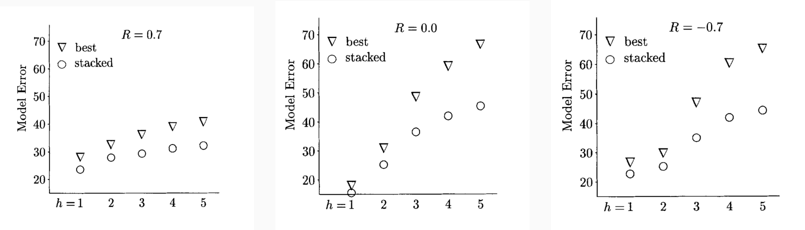
	Data Set			
	Housing		Ozone	
	Best	Stacked	Best	Stacked
Error	20.9	19.0	23.9	21.6

Table 2. Stacking Weights

# Terminal Nodes	Weight
7	.29
10	.13
23	.13
26	.09
29	.12
34	.20

Breiman's Experiments with Subset Regressions

- $M = 40$ linear models $\hat{\mu}_k$ from stepwise deletion
- Weights $\hat{\alpha}_k$ sum to between 0.7 and 0.9



Statistical Model

- Nonparametric regression with fixed design and known variance σ^2:

$$y_i = \mu(x_i) + \sigma\varepsilon_i, \quad i = 1, 2, \ldots, n, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, 1)$$

Statistical Model

- Nonparametric regression with fixed design and known variance σ^2:

$$y_i = \mu(x_i) + \sigma\varepsilon_i, \quad i = 1, 2, \ldots, n, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, 1)$$

- Accuracy of estimator $\hat{\mu}$ measured by in-sample error, i.e.,

$$\|\mu - \hat{\mu}\|^2 = \frac{1}{n}\sum_{i=1}^{n}(\mu(x_i) - \hat{\mu}(x_i))^2$$

Statistical Model

- Nonparametric regression with fixed design and known variance σ^2:

$$y_i = \mu(x_i) + \sigma\varepsilon_i, \quad i = 1, 2, \ldots, n, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, 1)$$

- Accuracy of estimator $\hat{\mu}$ measured by in-sample error, i.e.,

$$\|\mu - \hat{\mu}\|^2 = \frac{1}{n}\sum_{i=1}^{n}(\mu(x_i) - \hat{\mu}(x_i))^2$$

- Training error of $\hat{\mu}$ is

$$\|y - \hat{\mu}\|^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}(x_i))^2$$

Best (Top-performing) Single Model

- Suppose $\hat{\mu}_k$ is least-squares projection of y onto (fixed) space $\mathcal{A}_k$ of dimension d_k

Best (Top-performing) Single Model

- Suppose $\hat{\mu}_k$ is least-squares projection of y onto (fixed) space $\mathcal{A}_k$ of dimension d_k
- Define best single model $\hat{\mu}_{\text{best}}$ as $\hat{\mu}_{\hat{k}}$, where

$$\hat{k} \in \underset{k=1,2,\ldots,M}{\arg\min} \; \|y - \hat{\mu}_k\|^2 + \lambda \frac{\sigma^2 d_k}{n}$$

Best (Top-performing) Single Model

- Suppose $\hat{\mu}_k$ is least-squares projection of y onto (fixed) space $\mathcal{A}_k$ of dimension d_k
- Define best single model $\hat{\mu}_{\text{best}}$ as $\hat{\mu}_{\hat{k}}$, where

$$\hat{k} \in \underset{k=1,2,\ldots,M}{\arg\min} \quad \|y - \hat{\mu}_k\|^2 + \lambda \frac{\sigma^2 d_k}{n}$$

- Choosing $\lambda = 2$ corresponds to model selected by AIC, Mallows's C_p

Best (Top-performing) Single Model

- Suppose $\hat{\mu}_k$ is least-squares projection of y onto (fixed) space $\mathcal{A}_k$ of dimension d_k
- Define best single model $\hat{\mu}_{\text{best}}$ as $\hat{\mu}_{\hat{k}}$, where

$$\hat{k} \in \underset{k=1,2,\ldots,M}{\arg\min} \quad \|y - \hat{\mu}_k\|^2 + \lambda \frac{\sigma^2 d_k}{n}$$

- Choosing $\lambda = 2$ corresponds to model selected by AIC, Mallows's C_p
- Choosing $\lambda = \log(n)$ corresponds to model selected by BIC

Best (Top-performing) Single Model

- Suppose $\hat{\mu}_k$ is least-squares projection of y onto (fixed) space $\mathcal{A}_k$ of dimension d_k
- Define best single model $\hat{\mu}_{\text{best}}$ as $\hat{\mu}_{\hat{k}}$, where

$$\hat{k} \in \underset{k=1,2,\ldots,M}{\arg\min} \quad \|y - \hat{\mu}_k\|^2 + \lambda \frac{\sigma^2 d_k}{n}$$

- Choosing $\lambda = 2$ corresponds to model selected by AIC, Mallows's C_p
- Choosing $\lambda = \log(n)$ corresponds to model selected by BIC
- In certain cases, criteria will asymptotically select same model as leave-one-out cross-validation

Best (Top-performing) Single Model

- Suppose $\hat{\mu}_k$ is least-squares projection of y onto (fixed) space $\mathcal{A}_k$ of dimension d_k
- Define best single model $\hat{\mu}_{\text{best}}$ as $\hat{\mu}_{\hat{k}}$, where

$$\hat{k} \in \underset{k=1,2,\ldots,M}{\arg\min} \quad \|y - \hat{\mu}_k\|^2 + \lambda \frac{\sigma^2 d_k}{n}$$

- Choosing $\lambda = 2$ corresponds to model selected by AIC, Mallows's C_p
- Choosing $\lambda = \log(n)$ corresponds to model selected by BIC
- In certain cases, criteria will asymptotically select same model as leave-one-out cross-validation
- Will describe performance of $\hat{\mu}_{\text{stack}}$ relative to $\hat{\mu}_{\text{best}}$

Nested Regressions

- Among various model structures, Breiman focused on stacking

Nested Regressions

- Among various model structures, Breiman focused on stacking
    1. Decision trees resulting from pruning large tree upwards

Nested Regressions

- Among various model structures, Breiman focused on stacking
  1. Decision trees resulting from pruning large tree upwards
  2. Linear regressions resulting from stepwise deletion

Nested Regressions

- Among various model structures, Breiman focused on stacking
    1. Decision trees resulting from pruning large tree upwards
    2. Linear regressions resulting from stepwise deletion
- In both cases, estimators $\hat{\mu}_k$ are least-squares projections of y onto nested subspaces $\mathcal{A}_1 \subset \mathcal{A}_2 \subset \cdots \subset \mathcal{A}_M$

Nested Regressions

- Among various model structures, Breiman focused on stacking
  1. Decision trees resulting from pruning large tree upwards
  2. Linear regressions resulting from stepwise deletion
- In both cases, estimators $\hat{\mu}_k$ are least-squares projections of y onto nested subspaces $\mathcal{A}_1 \subset \mathcal{A}_2 \subset \cdots \subset \mathcal{A}_M$
- Because models are nested,

$$d_1 < d_2 < \cdots < d_M$$

and

$$\|y - \hat{\mu}_M\|^2 < \cdots < \|y - \hat{\mu}_2\|^2 < \|y - \hat{\mu}_1\|^2$$

Learning Stacking Weights

- Ideally want weights to minimize expected in-sample error

$$\text{Err}(\boldsymbol{\alpha}) = \mathbb{E}\left[\left\|\mu - \sum_{k=1}^{M} \alpha_k \hat{\mu}_k\right\|^2\right],$$

subject to non-negativity constraint $\alpha_k \geqslant 0$

Learning Stacking Weights

- Ideally want weights to minimize expected in-sample error

$$\mathrm{Err}(\boldsymbol{\alpha}) = \mathbb{E}\left[\left\|\mu - \sum_{k=1}^{M} \alpha_k \hat{\mu}_k\right\|^2\right],$$

  subject to non-negativity constraint $\alpha_k \geqslant 0$

- Unbiased estimator of error

$$\mathrm{Err}(\boldsymbol{\alpha}) = \mathbb{E}\left[R(\boldsymbol{\alpha}) + \frac{2\sigma^2}{n}\mathrm{df}(\boldsymbol{\alpha}) - \sigma^2\right]$$

Learning Stacking Weights

- Ideally want weights to minimize expected in-sample error

$$\text{Err}(\boldsymbol{\alpha}) = \mathbb{E}\left[\left\|\mu - \sum_{k=1}^{M} \alpha_k \hat{\mu}_k\right\|^2\right],$$

  subject to non-negativity constraint $\alpha_k \geqslant 0$

- Unbiased estimator of error

$$\text{Err}(\boldsymbol{\alpha}) = \mathbb{E}\left[R(\boldsymbol{\alpha}) + \frac{2\sigma^2}{n}\text{df}(\boldsymbol{\alpha}) - \sigma^2\right]$$

- Training error $R(\boldsymbol{\alpha}) = \left\|y - \sum_{k=1}^{M} \alpha_k \hat{\mu}_k\right\|^2$

Learning Stacking Weights

- Ideally want weights to minimize expected in-sample error

$$\text{Err}(\boldsymbol{\alpha}) = \mathbb{E}\left[\left\|\mu - \sum_{k=1}^{M} \alpha_k \hat{\mu}_k\right\|^2\right],$$

  subject to non-negativity constraint $\alpha_k \geqslant 0$

- Unbiased estimator of error

$$\text{Err}(\boldsymbol{\alpha}) = \mathbb{E}\left[R(\boldsymbol{\alpha}) + \frac{2\sigma^2}{n}\text{df}(\boldsymbol{\alpha}) - \sigma^2\right]$$

- Training error $R(\boldsymbol{\alpha}) = \left\|y - \sum_{k=1}^{M} \alpha_k \hat{\mu}_k\right\|^2$
- Degrees of freedom $\text{df}(\boldsymbol{\alpha}) = \sum_{k=1}^{M} \alpha_k d_k$

First Attempt

- Solve quadratic program with linear constraints:

$$
\begin{aligned}
\text{minimize} \quad & R(\boldsymbol{\alpha}) + \frac{2\sigma^2}{n}\mathrm{df}(\boldsymbol{\alpha}) - \sigma^2 \\
\text{subject to} \quad & \alpha_k \geqslant 0, \quad k = 1, 2, \ldots, M
\end{aligned}
$$

First Attempt

- Solve quadratic program with linear constraints:

$$\begin{array}{ll} \text{minimize} & R(\boldsymbol{\alpha}) + \frac{2\sigma^2}{n}\text{df}(\boldsymbol{\alpha}) - \sigma^2 \\ \text{subject to} & \alpha_k \geqslant 0, \quad k = 1, 2, \ldots, M \end{array}$$

- Solution $\hat{\alpha}$ satisfies

$$\mathbb{E}\big[\text{Err}(\hat{\boldsymbol{\alpha}})\big] = \mathbb{E}\left[R(\hat{\boldsymbol{\alpha}}) + \frac{2\sigma^2}{n}\text{df}(\hat{\boldsymbol{\alpha}}) - \sigma^2 + \frac{4\sigma^2}{n}\|\hat{\boldsymbol{\alpha}}\|_{\ell_0} + \text{lower order terms}\right]$$

First Attempt

- Solve quadratic program with linear constraints:

$$\begin{aligned} \text{minimize} \quad & R(\boldsymbol{\alpha}) + \frac{2\sigma^2}{n}\text{df}(\boldsymbol{\alpha}) - \sigma^2 \\ \text{subject to} \quad & \alpha_k \geqslant 0, \quad k = 1, 2, \ldots, M \end{aligned}$$

- Solution $\hat{\alpha}$ satisfies

$$\mathbb{E}\big[\text{Err}(\hat{\boldsymbol{\alpha}})\big] = \mathbb{E}\left[R(\hat{\boldsymbol{\alpha}}) + \frac{2\sigma^2}{n}\text{df}(\hat{\boldsymbol{\alpha}}) - \sigma^2 + \frac{4\sigma^2}{n}\|\hat{\boldsymbol{\alpha}}\|_{\ell_0} + \text{lower order terms}\right]$$

- So $R(\hat{\boldsymbol{\alpha}}) + \frac{2\sigma^2}{n}\text{df}(\hat{\boldsymbol{\alpha}}) - \sigma^2$ is no longer unbiased estimator of error for stacked model with adaptive weights

Second Attempt

- Let $\dim(\boldsymbol{\alpha})$ denote dimension of stacked model, $\max_k\{d_k : \alpha_k \neq 0\}$

Second Attempt

- Let $\dim(\boldsymbol{\alpha})$ denote dimension of stacked model, $\max_k\{d_k : \alpha_k \neq 0\}$
- Solve similar (but non-convex) program:

$$\begin{aligned} \text{minimize} \quad & R(\boldsymbol{\alpha}) + \frac{2\sigma^2}{n}\mathrm{df}(\boldsymbol{\alpha}) - \sigma^2 + \frac{\sigma^2}{n}\frac{(\lambda-1)^2}{\lambda}\dim(\boldsymbol{\alpha}) \\ \text{subject to} \quad & \alpha_k \geqslant 0, \quad k = 1, 2, \ldots, M \end{aligned}$$

Second Attempt

- Let $\dim(\boldsymbol{\alpha})$ denote dimension of stacked model, $\max_k\{d_k : \alpha_k \neq 0\}$
- Solve similar (but non-convex) program:

$$\begin{aligned} \text{minimize} \quad & R(\boldsymbol{\alpha}) + \frac{2\sigma^2}{n}\mathrm{df}(\boldsymbol{\alpha}) - \sigma^2 + \frac{\sigma^2}{n}\frac{(\lambda-1)^2}{\lambda}\dim(\boldsymbol{\alpha}) \\ \text{subject to} \quad & \alpha_k \geqslant 0, \quad k = 1, 2, \ldots, M \end{aligned}$$

- Solvable in $O(M)$ time by reducing problem to isotonic regression
  - Same complexity as finding best single model

Second Attempt

- Let $\dim(\boldsymbol{\alpha})$ denote dimension of stacked model, $\max_k\{d_k : \alpha_k \neq 0\}$
- Solve similar (but non-convex) program:

$$\begin{array}{ll} \text{minimize} & R(\boldsymbol{\alpha}) + \frac{2\sigma^2}{n}\mathrm{df}(\boldsymbol{\alpha}) - \sigma^2 + \frac{\sigma^2}{n}\frac{(\lambda-1)^2}{\lambda}\dim(\boldsymbol{\alpha}) \\ \text{subject to} & \alpha_k \geqslant 0, \quad k = 1, 2, \ldots, M \end{array}$$

- Solvable in $O(M)$ time by reducing problem to isotonic regression
  - Same complexity as finding best single model
- Solution satisfies $\sum_{k=1}^{M} \hat{\alpha}_k < 1$, despite no explicit sum constraint

Reduction to Isotonic Regression

- Due to nested structure and non-negative constraints, problem reduces to weighted isotonic regression:

$$\begin{aligned} \text{minimize} \quad & \sum_{k=1}^{M} w_k (z_k - \gamma_k)^2 \\ \text{subject to} \quad & \gamma_1 \leqslant \gamma_2 \leqslant \cdots \leqslant \gamma_M, \end{aligned}$$

where $w_k = \|y - \hat{\mu}_{k-1}\|^2 - \|y - \hat{\mu}_k\|^2$ and $z_k = (d_k - d_{k-1})/w_k$

Reduction to Isotonic Regression

- Due to nested structure and non-negative constraints, problem reduces to weighted isotonic regression:

$$\begin{aligned} &\text{minimize} \quad \sum_{k=1}^{M} w_k (z_k - \gamma_k)^2 \\ &\text{subject to} \quad \gamma_1 \leqslant \gamma_2 \leqslant \cdots \leqslant \gamma_M, \end{aligned}$$

  where $w_k = \|y - \hat{\mu}_{k-1}\|^2 - \|y - \hat{\mu}_k\|^2$ and $z_k = (d_k - d_{k-1})/w_k$

- Closed-form solution:

$$\hat{\gamma}_k = \frac{\sigma^2}{n} \min_{k \leqslant i \leqslant M} \max_{0 \leqslant j < k} \frac{d_i - d_j}{\|y - \hat{\mu}_j\|^2 - \|y - \hat{\mu}_i\|^2}$$

Closed-Form Representations of $\hat{\mu}_{\textbf{stack}}$ and $\hat{\mu}_{\textbf{best}}$

Theorem (Chen, K., & Tan, 2023)

The following representations hold:

$$\hat{\mu}_{best}(x) = \sum_{k=1}^{M} (\hat{\mu}_k(x) - \hat{\mu}_{k-1}(x)) \mathbf{1}(\hat{\gamma}_k < 1/\lambda),$$

$$\hat{\mu}_{stack}(x) = \sum_{k=1}^{M} (\hat{\mu}_k(x) - \hat{\mu}_{k-1}(x))(1 - \hat{\gamma}_k) \mathbf{1}(\hat{\gamma}_k < 1/\lambda)$$

Closed-Form Representations of $\hat{\mu}_{\textbf{stack}}$ and $\hat{\mu}_{\textbf{best}}$

Theorem (Chen, K., & Tan, 2023)

The following representations hold:

$$\hat{\mu}_{best}(x) = \sum_{k=1}^{M} (\hat{\mu}_k(x) - \hat{\mu}_{k-1}(x)) \mathbf{1}(\hat{\gamma}_k < 1/\lambda),$$

$$\hat{\mu}_{stack}(x) = \sum_{k=1}^{M} (\hat{\mu}_k(x) - \hat{\mu}_{k-1}(x))(1 - \hat{\gamma}_k) \mathbf{1}(\hat{\gamma}_k < 1/\lambda)$$

- Best single model performs hard thresholding on predictive differences $\hat{\mu}_k(x) - \hat{\mu}_{k-1}(x)$ across successive sub-models

Closed-Form Representations of $\hat{\mu}_{\text{stack}}$ and $\hat{\mu}_{\text{best}}$

Theorem (Chen, K., & Tan, 2023)

The following representations hold:

$$\hat{\mu}_{best}(x) = \sum_{k=1}^{M} (\hat{\mu}_k(x) - \hat{\mu}_{k-1}(x))\mathbf{1}(\hat{\gamma}_k < 1/\lambda),$$

$$\hat{\mu}_{stack}(x) = \sum_{k=1}^{M} (\hat{\mu}_k(x) - \hat{\mu}_{k-1}(x))(1 - \hat{\gamma}_k)\mathbf{1}(\hat{\gamma}_k < 1/\lambda)$$

- Best single model performs hard thresholding on predictive differences $\hat{\mu}_k(x) - \hat{\mu}_{k-1}(x)$ across successive sub-models
- Stacked model additionally shrinks these predictive differences towards zero by factor $(1 - \hat{\gamma}_k)$

Closed-Form Representations of $\hat{\mu}_{\text{stack}}$ and $\hat{\mu}_{\text{best}}$

Theorem (Chen, K., & Tan, 2023)

The following representations hold:

$$\hat{\mu}_{best}(x) = \sum_{k=1}^{M} (\hat{\mu}_k(x) - \hat{\mu}_{k-1}(x))\mathbf{1}(\hat{\gamma}_k < 1/\lambda),$$

$$\hat{\mu}_{stack}(x) = \sum_{k=1}^{M} (\hat{\mu}_k(x) - \hat{\mu}_{k-1}(x))(1 - \hat{\gamma}_k)\mathbf{1}(\hat{\gamma}_k < 1/\lambda)$$

- Best single model performs hard thresholding on predictive differences $\hat{\mu}_k(x) - \hat{\mu}_{k-1}(x)$ across successive sub-models
- Stacked model additionally shrinks these predictive differences towards zero by factor $(1 - \hat{\gamma}_k)$
- Performs model selection and adaptive shrinkage simultaneously

Main Result

Theorem (Chen, K., & Tan, 2023)

If $d_k \geqslant d_{k-1} + 4$ for all k, then

$$\mathbb{E}\left[\|\mu - \hat{\mu}_{\text{stack}}\|^2\right] < \mathbb{E}\left[\|\mu - \hat{\mu}_{\text{best}}\|^2\right]$$

- Theoretically confirms Breiman's empirical findings

Error Gap Between $\hat{\mu}_{\text{stack}}$ and $\hat{\mu}_{\text{best}}$

- Error gap $\mathbb{E}\big[\|\mu - \hat{\mu}_{\text{best}}\|^2 - \|\mu - \hat{\mu}_{\text{stack}}\|^2\big]$ can be lower bounded by

$$\frac{\sigma^2}{n}\mathbb{E}\left[\min_{1\leqslant k\leqslant M} \frac{(d_k - 4k)^2}{(n/\sigma^2)(\|y\|^2 - \|y - \hat{\mu}_k\|^2)}\right]$$

Error Gap Between $\hat{\mu}_{\mathbf{stack}}$ and $\hat{\mu}_{\mathbf{best}}$

- Error gap $\mathbb{E}\big[\|\mu - \hat{\mu}_{\text{best}}\|^2 - \|\mu - \hat{\mu}_{\text{stack}}\|^2\big]$ can be lower bounded by

$$\frac{\sigma^2}{n}\mathbb{E}\left[\min_{1 \leqslant k \leqslant M} \frac{(d_k - 4k)^2}{(n/\sigma^2)(\|y\|^2 - \|y - \hat{\mu}_k\|^2)}\right]$$

- Similar to improvement from applying James-Stein shrinkage (non-adaptively) to individual model

Error Gap Between $\hat{\mu}_{\text{stack}}$ and $\hat{\mu}_{\text{best}}$

- Error gap $\mathbb{E}\big[\|\mu - \hat{\mu}_{\text{best}}\|^2 - \|\mu - \hat{\mu}_{\text{stack}}\|^2\big]$ can be lower bounded by

$$\frac{\sigma^2}{n}\mathbb{E}\left[\min_{1 \leqslant k \leqslant M} \frac{(d_k - 4k)^2}{(n/\sigma^2)(\|y\|^2 - \|y - \hat{\mu}_k\|^2)}\right]$$

- Similar to improvement from applying James-Stein shrinkage (non-adaptively) to individual model
- As with James-Stein shrinkage, gap tends to be larger when signal-to-noise ratio $\|\mu\|/\sigma$ or sample size are small

Conclusion

In past statistical work, all the focus has been on selecting the "best" single model from a class of models. We may need to shift our thinking to the possibility of forming combinations of models... — Leo Breiman (1996)

Future Work

- Stack non-nested models, such as ridge regressions

Future Work

- Stack non-nested models, such as ridge regressions
- Characterize complexity of stacked model (usually larger than best single model)

Future Work

- Stack non-nested models, such as ridge regressions
- Characterize complexity of stacked model (usually larger than best single model)
- Connect to other ensemble methods like random forests (randomization + model selection)

Thank you!

Chen, K., & Tan, *Error Reduction from Stacked Regressions* (2023)

Available at `klusowski.princeton.edu`