Bootstrap AGGregatING

Jonas Haslbeck

April 19th, 2016

Statistical Learning Recab

Finite samples $X^n = \{X_1^n, ..., X_p^n\}$ and Y^n.

Assumption: There exists a function f^* such that $y = f^*(x) + \epsilon$

The goal is to find the *best* guess $\hat{f}(x)$ of $f^*(x)$ based on the finite samples X^n and Y^n.

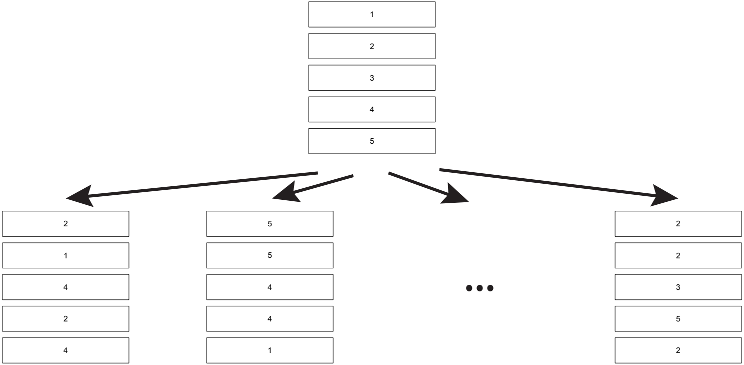
The *best* guess minimizes the risk:

$$E[f^*(x) - \hat{f}(x)]^2 = Bias(\hat{f}(x))^2 + Var(\hat{f}(x))$$

But we observe only the empirical risk:

$$\sum_{i=1}^{n}(y_i - \hat{f}(x))^2$$

How does bagging work?



$\hat{f}_1(x)$ $\hat{f}_2(x)$ $\cdots$ $\hat{f}_B(x)$

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(x)$$

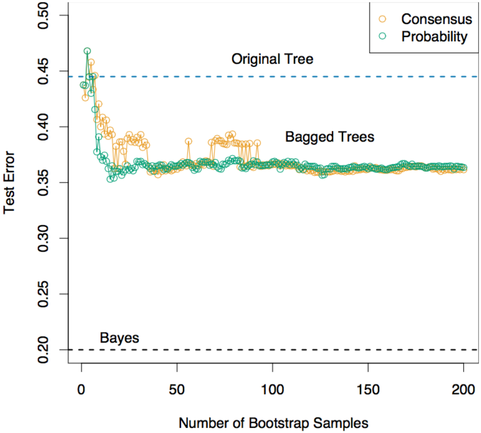
Bootstrap AGGregatING

- $\hat{f}(x)$ estimate on full dataset
- $\hat{f}_{bag}(x)$ bagged estimate
- Claim:

$$E[f^*(x) - \hat{f}(x)]^2 \geq E[f^*(x) - \hat{f}_{bag}(x)]^2$$

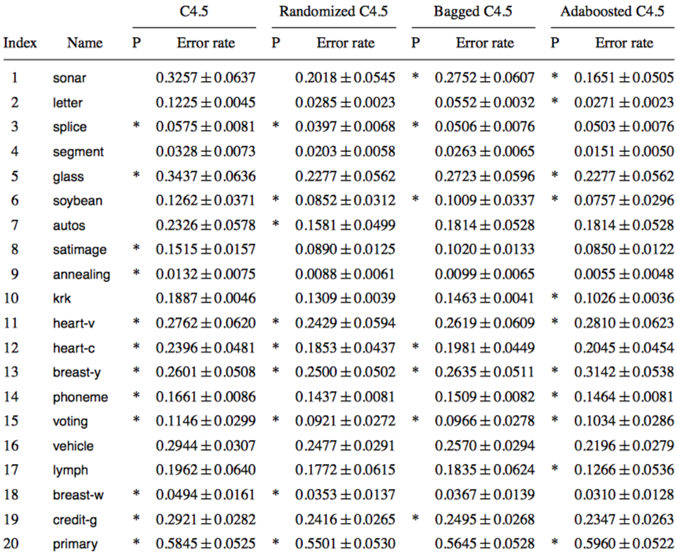
$$Bias(\hat{f}(x))^2 + Var(\hat{f}(x)) \geq Bias(\hat{f}_{bag}(x))^2 + Var(\hat{f}_{bag}(x))$$

Bagging classification trees



(Elements of Statistical Learning, 2009)

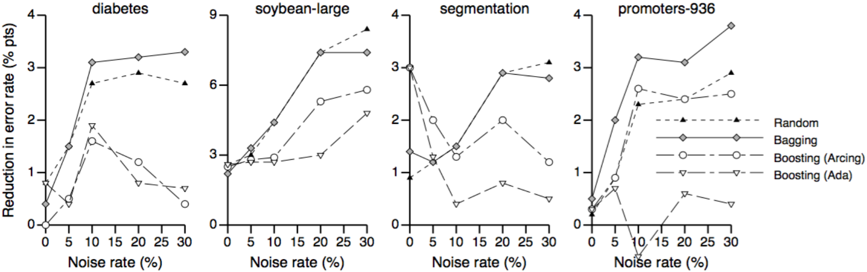
Bagging classification trees (2)



		C4.5		Randomized C4.5		Bagged C4.5		Adaboosted C4.5	
Index	Name	P	Error rate	P	Error rate	P	Error rate	P	Error rate
1	sonar		0.3257 ± 0.0637		0.2018 ± 0.0545	*	0.2752 ± 0.0607	*	0.1651 ± 0.0505
2	letter		0.1225 ± 0.0045		0.0285 ± 0.0023		0.0552 ± 0.0032	*	0.0271 ± 0.0023
3	splice	*	0.0575 ± 0.0081	*	0.0397 ± 0.0068	*	0.0506 ± 0.0076		0.0503 ± 0.0076
4	segment		0.0328 ± 0.0073		0.0203 ± 0.0058		0.0263 ± 0.0065		0.0151 ± 0.0050
5	glass	*	0.3437 ± 0.0636		0.2277 ± 0.0562		0.2723 ± 0.0596	*	0.2277 ± 0.0562
6	soybean		0.1262 ± 0.0371	*	0.0852 ± 0.0312	*	0.1009 ± 0.0337	*	0.0757 ± 0.0296
7	autos		0.2326 ± 0.0578	*	0.1581 ± 0.0499		0.1814 ± 0.0528		0.1814 ± 0.0528
8	satimage	*	0.1515 ± 0.0157		0.0890 ± 0.0125		0.1020 ± 0.0133		0.0850 ± 0.0122
9	annealing	*	0.0132 ± 0.0075		0.0088 ± 0.0061		0.0099 ± 0.0065		0.0055 ± 0.0048
10	krk		0.1887 ± 0.0046		0.1309 ± 0.0039		0.1463 ± 0.0041	*	0.1026 ± 0.0036
11	heart-v	*	0.2762 ± 0.0620	*	0.2429 ± 0.0594		0.2619 ± 0.0609	*	0.2810 ± 0.0623
12	heart-c	*	0.2396 ± 0.0481	*	0.1853 ± 0.0437	*	0.1981 ± 0.0449		0.2045 ± 0.0454
13	breast-y	*	0.2601 ± 0.0508	*	0.2500 ± 0.0502	*	0.2635 ± 0.0511	*	0.3142 ± 0.0538
14	phoneme	*	0.1661 ± 0.0086		0.1437 ± 0.0081		0.1509 ± 0.0082	*	0.1464 ± 0.0081
15	voting	*	0.1146 ± 0.0299	*	0.0921 ± 0.0272	*	0.0966 ± 0.0278	*	0.1034 ± 0.0286
16	vehicle		0.2944 ± 0.0307		0.2477 ± 0.0291		0.2570 ± 0.0294		0.2196 ± 0.0279
17	lymph		0.1962 ± 0.0640		0.1772 ± 0.0615		0.1835 ± 0.0624	*	0.1266 ± 0.0536
18	breast-w	*	0.0494 ± 0.0161	*	0.0353 ± 0.0137		0.0367 ± 0.0139		0.0310 ± 0.0128
19	credit-g	*	0.2921 ± 0.0282		0.2416 ± 0.0265	*	0.2495 ± 0.0268		0.2347 ± 0.0263
20	primary	*	0.5845 ± 0.0525	*	0.5501 ± 0.0530		0.5645 ± 0.0528	*	0.5960 ± 0.0522

(Dietterich, 2000)

Bagging Neural Networks



(Maclin & Opitz, 1997)

Why does Bagging work?

Recall:

- $\hat{f}(x)$ estimate on full dataset
- $\hat{f}_{bag}(x)$ bagged estimate
- Claim:

$$E[f^*(x) - \hat{f}(x)]^2 \geq E[f^*(x) - \hat{f}_{bag}(x)]^2$$

$$Bias(\hat{f}(x))^2 + Var(\hat{f}(x)) \geq Bias(\hat{f}_{bag}(x))^2 + Var(\hat{f}_{bag}(x))$$

Approximating average over independent training samples

- $X_1, ..., X_N$ independent
- $Var(X_1) = ... = Var(X_N) = \sigma^2$
- $\hat{f}(x) = N^{-1} \sum_{i=1}^{N} X_i$
- Then $Var(\hat{f}(x) = \frac{\sigma^2}{N}$

Averaging over observations reduces variance!
If we had many training samples we could calculate:

$$\hat{f}_{average}(x) = B^{-1} \sum_{b=1}^{B} \hat{f}^{b}(x)$$

Now the argument is

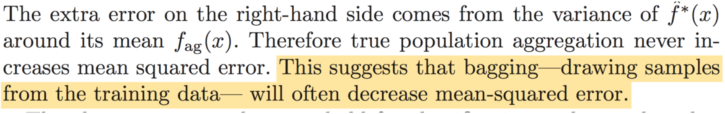
$$\hat{f}_{bag}(x) \approx \hat{f}_{average}(x)$$

(Introduction to Statistical Learning, 2015)

Variance reduction argument 2

- Ideal aggregate estimator $f_{bag} = E_{\mathcal{P}} \hat{f}^{b}(x)$
- $\mathcal{P}$ Population

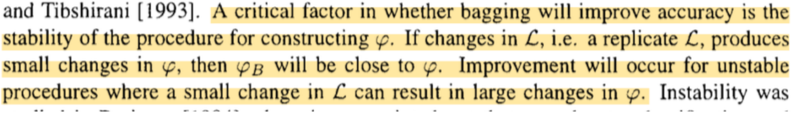
$$
\begin{aligned}
E_{\mathcal{P}}[Y - \hat{f}^{b}(x)]^2 &= E_{\mathcal{P}}[Y - f_{bag}(x) + f_{bag}(x) - \hat{f}^{b}(x)]^2 \\
&= E_{\mathcal{P}}[Y - f_{bag}(x)]^2 + E_{\mathcal{P}}[f_{bag}(x) - \hat{f}^{b}(x)]^2 \\
&\geq E_{\mathcal{P}}[Y - f_{bag}(x)]^2
\end{aligned}
$$



The extra error on the right-hand side comes from the variance of $\hat{f}^*(x)$ around its mean $f_{\text{ag}}(x)$. Therefore true population aggregation never increases mean squared error. This suggests that bagging—drawing samples from the training data— will often decrease mean-squared error.

(Elements of Statistical Learning, 2009)

Original paper: Reducing Variance in unstable predictors



and Tibshirani [1993]. A critical factor in whether bagging will improve accuracy is the stability of the procedure for constructing φ. If changes in $\mathcal{L}$, i.e. a replicate $\mathcal{L}$, produces small changes in φ, then φ_B will be close to φ. Improvement will occur for unstable procedures where a small change in $\mathcal{L}$ can result in large changes in φ. Instability was

Same theoretical argument that is based on the assumption that resampling is an approximation to independent samples.

(Breiman,1996 'Bagging Predictors')

A bayesian explanation?

Claim: Bagging reduces a classification learner's error rate because it changes the learner's model space and/or prior distribution to one that better fits the domain.

Empirical test: Find the simplest single decision tree that makes same predictions as the bagged ensemble. If the complexity of the single decision tree is larger than the one of bagged base decision tree, this is evidence, that bagging increases the model space.

Result: Complexity is indeed larger in this single true. But: not very sound.

And: Rao & Tibshirani (1997): bootstrap distribution is an approximation of a posterior distribution based on a symmetric Dirichlet non-informative prior.

(Domingos, 1997)

Reducing variance in non-linear components

Taylor expansion of estimator

$$\hat{f}(x) \approx \hat{f}(x) + \frac{\hat{f}'(x)}{1!}(x-\mu) + \frac{\hat{f}''(x)}{2!}(x-\mu)^2 + ...$$

Claim: the bagged estimator $\hat{f}_{bag}(x)$ is an approximation of:

$$\bar{f}(x) = \hat{f}(x) + E_{\mathcal{P}}\left[\frac{\hat{f}'(x)}{1!}(x-\mu) + \frac{\hat{f}''(x)}{2!}(x-\mu)^2 + ...\right]$$

Empirical evidence!

But: Limited to smooth multivariate functions (e.g. no trees)

(Friedman & Hall, 1999)

Smoothing of regression/classification surfaces

Simplest Example: $\hat{f}(x) = \mathbb{I}_{\{\hat{d} \leq x\}}, x \in \mathbb{R}$

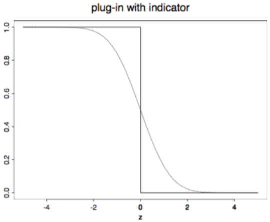


Figure 2.1: Indicator predictor from (2.1) at $x = x_n(0) = d^0$ as in (2.2). Function $g(z) = \mathbf{1}_{[z \leq 0]}$ [solid line] and $g_B(z)$ [dotted line] defining the asymptotics of the predictor and its bagged version [see Proposition 2.1].

For non-differentiable and discontinuous functions (e.g. trees).

(Buehlmann & Yu, 2000)

Negative effect of Bagging on U-Statistics

U-Statistics:

$$U = N^{-1} \sum_{i}^{N} A(X_i) + N^{-1} \sum_{i,j}^{N} B(X_i, X_j)$$

Examples: Variance, Skewness, ...

Main results:

- The influence of bagging on variance depends on specific U-Statistic
- Bagging always increases squared bias

(Buja & Stuetzle, 2000)

Bagging equalizes influence

Different level of analysis: influence of data points on estimator.

Two steps:

1. Bagging equalizes influence
2. Equalized influence explains MSE-reduction due to bagging

(Grandvalet, 2004)

Bagging equalizes influence?

Example: Point estimation

- n = 20 draws from mixture $p(x) = 1 - P\ N(0,1) + P\ N(0,10)$
- compute median or bagged (B = 100) median
- $P = \{0, .04, .2, .7, 1\}$

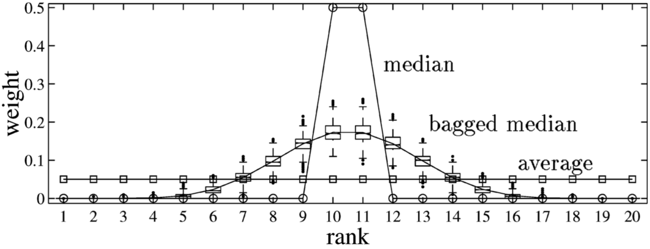
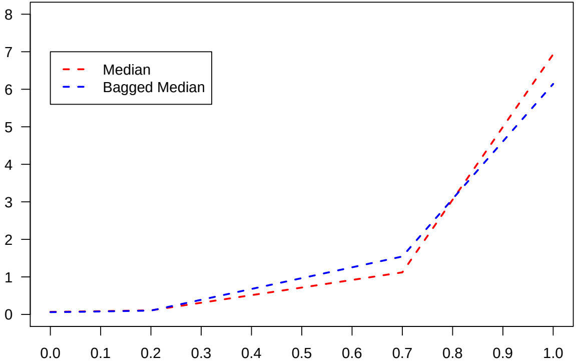


Figure 2. Boxplot of the weight given to the examples x_i versus the rank of x_i, for original and bagged mean estimates.

From equalized influence to prediction error

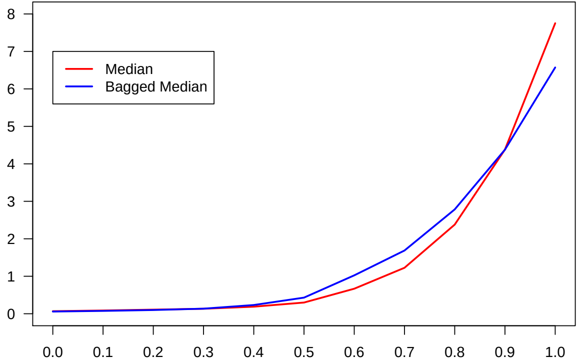
Explaining bagging performance as a function of P:



Data from Grandvalet 2004

From equalized influence to prediction error

Explaining bagging performance as a function of P:



Replicated Simulation

From equalized influence to prediction error

Example 2: Subset selection

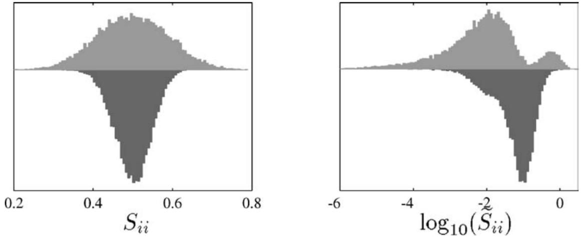
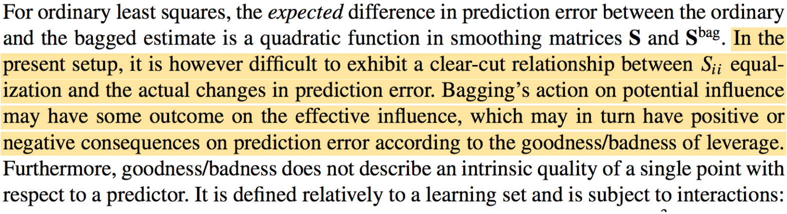


Figure 4. Left: histograms of S_{ii} (up, light grey) and S_{ii}^{bag} (down, dark grey) for ordinary least squares (all variables); right: histograms of $\log_{10}(\tilde{S}_{ii})$ (up, light grey) and $\log_{10}(\tilde{S}_{ii}^{\text{bag}})$ (down, dark grey) for subset selection (one variable).

From equalized influence to prediction error?

Example 2: Subset selection



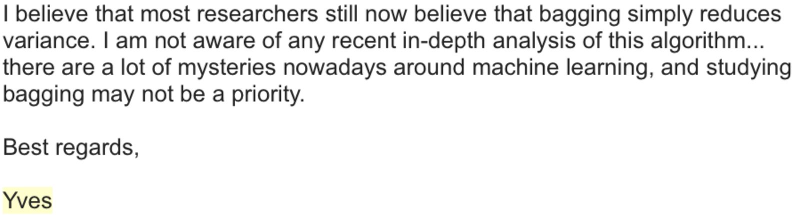
For ordinary least squares, the *expected* difference in prediction error between the ordinary and the bagged estimate is a quadratic function in smoothing matrices $\mathbf{S}$ and $\mathbf{S}^{\text{bag}}$. In the present setup, it is however difficult to exhibit a clear-cut relationship between S_{ii} equalization and the actual changes in prediction error. Bagging's action on potential influence may have some outcome on the effective influence, which may in turn have positive or negative consequences on prediction error according to the goodness/badness of leverage. Furthermore, goodness/badness does not describe an intrinsic quality of a single point with respect to a predictor. It is defined relatively to a learning set and is subject to interactions:

(Grandvalet, 2004)

Summary so far

Paper	Base Estimator	Results
Breiman, 1996	-	Heuristic: B reduces variance of unstable predictor
Domingos, 1997	-	B defines informative prior
Rao & Tibshirani, 1997	-	B doesn't define informative prior
Friedman & Hall, 2000	Smooth	B reduces variance in non-linear components
Buehlmann & Yu, 2000	CART	B performs asymptotic smoothing
Buja & Stueckle, 2000	U-Statistics	B always increases bias (sometimes also cariance)
Grandvalet, 2004	Subset Selection	B equalizes influence

Further developments?



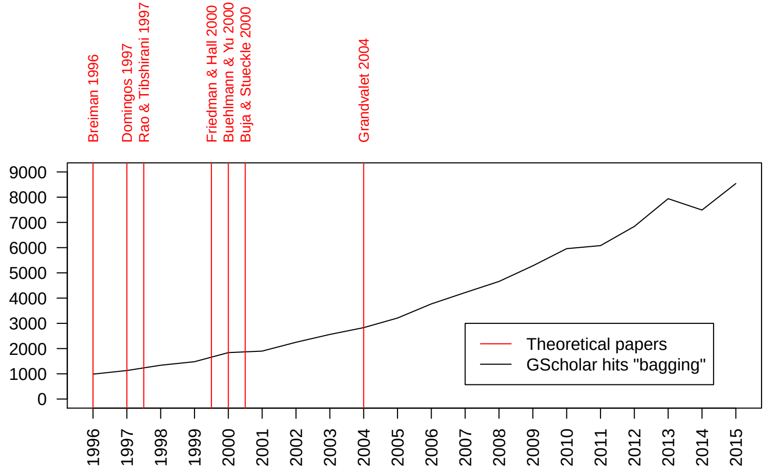
I believe that most researchers still now believe that bagging simply reduces variance. I am not aware of any recent in-depth analysis of this algorithm... there are a lot of mysteries nowadays around machine learning, and studying bagging may not be a priority.

Best regards,

Yves

(Yves Grandvalet)

Bagging: Theory and Practice



Observations & Discussion

Observation: 'Why does it work?' difficult question because ...

- performance of bagging depends on the base predictor *and* the data
- non-smooth estimators are hard to study

Discussion: 'Mysteries in Machine Learning' - does it matter that we know so little?

References

- Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer, 27(2), 83-85.
- Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine learning, 40(2), 139-157.
- Maclin, R., & Opitz, D. (1997). An empirical evaluation of bagging and boosting. AAAI/IAAI, 1997, 546-551.
- Grandvalet, Y. (2004). Bagging equalizes influence. Machine Learning, 55(3), 251-270.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). New York: springer.
- Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.
- Rao, J. S., & Tibshirani, R. (1997). The out-of-bootstrap method for model averaging and selection. University of Toronto.
- Friedman, J. H., & Hall, P. (2007). On bagging and nonlinear estimation. Journal of statistical planning and inference, 137(3), 669-683.

- Grandvalet, Y. (2004). Bagging equalizes influence. Machine Learning, 55(3), 251-270.

Contact

jonashaslbeck@gmail.com

http://jmbh.github.io