POINTS OF SIGNIFICANCE

Ensemble methods: bagging and random forests

Many heads are better than one.

Just as we might consult multiple experts about a problem and then combine their advice to come to a consensus decision, repeated statistical analyses on the same data can be combined to form a single result called an ensemble or consensus estimator. This is particularly useful when the outcome of the original analysis is sensitive to small changes in the sample. This month, we will discuss how bootstrap samples[1] can be used to create an ensemble to improve regression (prediction of a quantitative outcome) and classification (prediction of a categorical outcome)[2]. This differs from our previous use of the bootstrap to assess the variability of an analysis.

Bagging is a common ensemble method that uses bootstrap sampling[3]. Random forest is an enhancement of bagging that can improve variable selection. We will start by explaining bagging and then discuss the enhancement leading to random forest.

We'll illustrate bagging by improving a regression tree fit[4] of noisy data sampled from a parabola (**Fig. 1**). Because our sample is relatively small ($n = 30$), our regression tree prediction based on the entire sample is coarse (**Fig. 1a**). We begin bagging by generating bootstrap samples of size n by sampling n observations with replacement from our sample and then calculating a regression tree prediction for each bootstrap sample (**Fig. 1b**). Finally, we combine the individual bootstrap predictions into a consensus estimate, which can be done for regression by averaging the fitted values (**Fig. 1c**).

Our consensus regression fit in **Figure 1c** is smoother than the single fit based on the entire sample and reflects the shape of the parabola more closely. This suggests that if we increase the number of bootstraps we could obtain an even better fit—but how many should we use? Using more samples reduces the variance of the fit, but because many bootstrap samples are similar, at some point more bootstraps will merely increase computation time without improving the estimates.

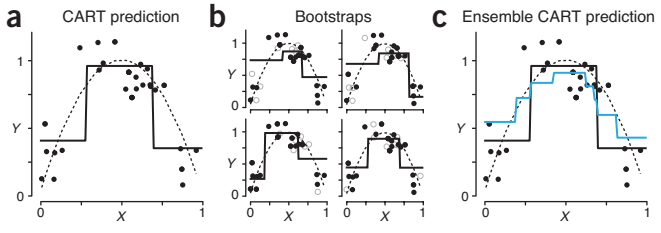

Figure 1 | Bagging applied to regression using a regression tree. (**a**) A sample of size $n = 30$ generated from a parabola (dashed line) with added noise and the associated fit from a regression tree (solid black line). (**b**) Four different bootstrap samples from the sample in **a** and their corresponding regression tree predictions. Solid points are in the bootstrap sample, and some are represented more than once. Hollow points are not in the bootstrap sample and are called out-of-bag (OOB) points. (**c**) Ensemble regression (blue line) formed by averaging bootstrap regressions in **b**. The original regression tree fit from **a** is also shown.

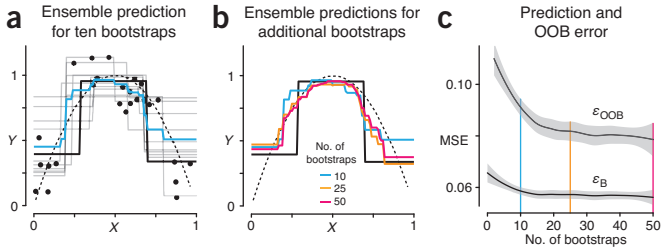

Figure 2 | Ensemble regressions improve in quality, up to a point, as the number of bootstraps is increased. (**a**) The consensus regression (blue line) for ten bootstrap iterations (gray lines) for data in **Figure 1**. (**b**) Ensemble regressions for 10, 25 and 50 bootstrap iterations. (**c**) The bagged and OOB errors (ε_B, ε_{OOB}) as a function of the number of bootstraps. The curve is fit to ten simulations at each bootstrap level using locally weighted smoothing. The gray band is the fit's 95% confidence interval.

In general, the optimal number of bootstrap samples depends on the problem. Let's look at how we can monitor the quality of our fit to choose the number of bootstraps. Let $\hat{y}$ be our original predictor based on the entire sample and $\hat{y}_B$ be the consensus bagged predictor. We can use these to calculate the mean square prediction error, $\text{MSE} = \Sigma_i(y_i - \hat{y}_i)^2/n$, for both fits, which we'll call ε and ε_B.

It turns out that there is another useful error that we can calculate, the 'out-of-bag' (OOB) error. Because we sample with replacement, in any given bootstrap sample some observations are not selected (hollow points, **Fig. 1b**) while others are represented more than once. The points that are not selected form the OOB sample, which can be used as the validation sample for the fit[5] to assess the regression accuracy for new observations not included in the training data. The OOB error, ε_{OOB}, is calculated analogously to ε_B, except that instead of $\hat{y}_i$ we use $\hat{y}_{OOB,i}$, which is the fit for each y_i averaged from samples in which it is OOB .

To assess the bagging process, we periodically compute ε_{OOB} and continue to create new bootstrap samples until the error stabilizes. Let's perform more bootstraps to the sample in **Figure 1** and see how the errors decrease. The regression tree fit based on the full sample gives an MSE of $\varepsilon = 0.067$. If we run ten bootstraps (**Fig. 2a**), the error drops to $\varepsilon_B = 0.048$ with an OOB error of $\varepsilon_{OOB} = 0.077$. In **Figure 2b** we compare the ensemble regressions for single runs of 10, 25 and 50 bootstraps. There does not appear to be much difference between the fits that use 25 and 50 bootstraps, which we can verify by looking at the profile of ε_B and ε_{OOB} as a function of the number of bootstraps (**Fig. 2c**). We can see that after about 25 bootstraps, both errors remain relatively constant. Using ε_{OOB} gives us a better indication of when to stop, since ε_B appears to stabilize too early (at about 15 bootstraps).

Because bagged regressions are averages, they usually have smaller variance than $\hat{y}$. But because $\hat{y}_{OOB}$ is based on only about 37% of the bootstrap samples (the expected fraction of the sample that is OOB), it is more variable than $\hat{y}_B$. This is reflected in the smaller value of ε_B compared with ε_{OOB} (**Fig. 2c**). However, as an estimator of the true prediction error for new samples, ε_B is too small because it is based on the overfitted training sample. While ε_{OOB} tends to be a bit larger than the true prediction error based on new samples, this conservative bias is usually small. Using ε_{OOB} to assess the fit allows us to use all the data to develop our regression, rather than requiring a hold-out test sample, and hence provides a better fit in general.

Simulations have shown that bagging performs best for algorithms that are highly sensitive to small changes in the data[6]. This sensitivity means that the fitted values $\hat{y}$ will be highly variable from sample to sample without aggregation. When the algorithm is very stable—for example, in linear regression with no influential points—the $\hat{y}_B$ may actually be more variable than $\hat{y}$.

Bagging can easily be applied to classification problems. Instead of using the average regression as a consensus, now a consensus classification is formed by 'voting', where the observation is classified into the class most frequently chosen.

In **Figure 3** we show bagging applied to the two-dimensional classification example we discussed in our previous column[4]. This example uses two predictors (x and y position) and a categorical outcome with four levels. The classification outcome is sensitive to outliers—green outliers in the top left quadrant cause the green class boundary to extend across the full width of the square (**Fig. 3a**). This issue is mitigated when creating bootstrap samples, since the outliers may be left out, which causes the green class boundaries to be more confined to the upper right (**Fig. 3b**). As we increase the number of bootstrap iterations, the boundaries become smoother and less likely to overfit (**Fig. 3c**). As before, we can monitor the bagging and OOB error (**Fig. 3d**) to guide us about the number of bootstrap iterations to perform. The original predictor misclassification rate was 29%, which dropped to 26% at 50 bootstrap iterations with an OOB error of about 40%.

Regression or classification fits generated from different bootstrap samples are correlated because of the observations that have been selected in both samples. The higher the correlation, the more similar the fit from each bootstrap and the smaller the mitigating effect of the consensus in reducing variance. For variable selection problems, strongly predictive variables that are selected in most bootstrap samples induce a strong correlation among the fits, reducing the utility of bagging.

To limit the impact of such variables, a simple but clever modification of CART bagging is used: a random forest[7]. In this approach, at each node of the tree, a subset m of the p variables in the data is selected at random, and only these m variables are considered for the partition at the node. This random selection of variables reduces the similarity of trees grown from different bootstrap samples—even two trees grown from the same bootstrap sample will likely differ. Once a sufficiently large forest of trees has been grown, the results are bagged in the usual way.

There will be a value of m that optimizes the variance reduction relative to the computational cost. This can be estimated using the OOB error as a function of m. Random forests are quite robust with respect to m, and rules of thumb such as using $m = p/3$ for regression and $m = \sqrt{p}$ for classification are sometimes used[7].

Ensemble methods like bagging and random forest are practical for mitigating both underfitting and overfitting, as we've seen with our regression and classification examples. The use of the OOB sample with each bootstrap is conceptually equivalent to using a test set for out-of-sample assessment but provides a means to use the entire sample to both estimate and assess the fit.

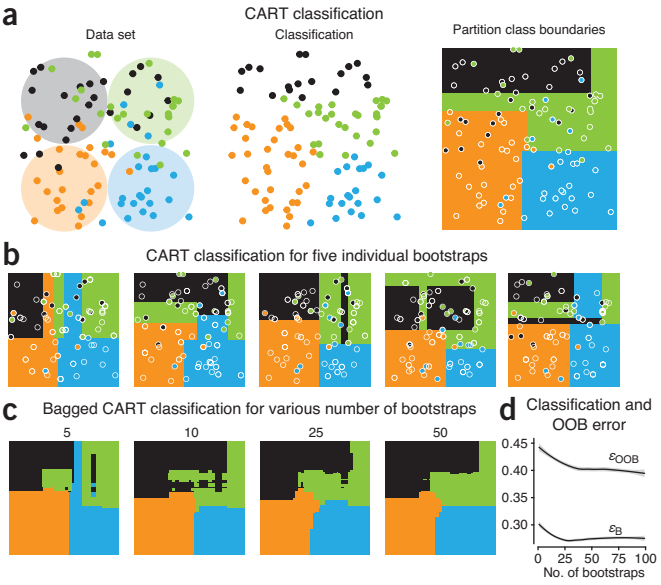

Figure 3 | Application of bagging to classification using a decision tree applied to $n = 100$ two-dimensional data points assigned to one of four color categories. (**a**) The data set is composed of 25 points sampled from the four circles, each with an associated category. Sampling is done from a two-dimensional normal distribution centered on the circle with an s.d. of the circle's radius. Classification is done by a decision tree. The tree's boundaries are indicated by the solid colored regions. (**b**) Classification boundaries based on five different bootstrap samples. The points in the bootstrap are shown as circles, and OOB points are not shown. (**c**) The boundaries of ensemble classification by vote for 5, 10, 25 and 50 bootstrap iterations. (**d**) The bagged and OOB errors (ε_B, ε_{OOB}; MSE) as a function of the number of bootstraps. The error is based on misclassification rate over ten simulations at each bootstrap level. The error curve is presented as in **Figure 2**.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Kulesa, A., Krzywinski, M., Blainey, P. & Altman, N. *Nat. Methods* **12**, 477–478 (2015).
2. Lever, J., Krzywinski, M. & Altman, N. *Nat. Methods* **13**, 541–542 (2016).
3. Breiman, L. *Mach. Learn.* **24**, 123–140 (1996).
4. Krzywinski, M. & Altman, N. *Nat. Methods* **14**, 757–758 (2017).
5. Lever, J., Krzywinski, M. & Altman, N. *Nat. Methods* **13**, 703–704 (2016).
6. Liang, G., Zhu, X. & Zhang, C. in *Proc. 25th AAAI Conference on Artificial Intelligence* (eds. Wang, D. & Reynolds, M.) 1802–1803 (Springer, 2011).
7. Breiman, L. *Mach. Learn.* **45**, 5–32 (2001).

Naomi Altman & Martin Krzywinski

Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre. Naomi Altman is a Professor of Statistics at The Pennsylvania State University.