What is Bagging and Random Forest:

Bagging is a common ensemble method that uses bootstrap sampling[3]. Random forest is an enhancement of bagging that can improve variable selection.

Weaknesses Of Bagging:

- 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 boot- strap samples induce a strong correlation among the fits, reducing the utility of bagging.
- Bagging is computationally expensive.
- Bagging introduces a loss of interpretability of the model. The resultant model can experience multiple biases when proper procedures are not being followed.
- 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 be more variable than $\hat{y}$. But for unstable algorithms it is not a good fit.

Development of Random Forests:

A simple but clever modification of CART bagging is 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.

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.

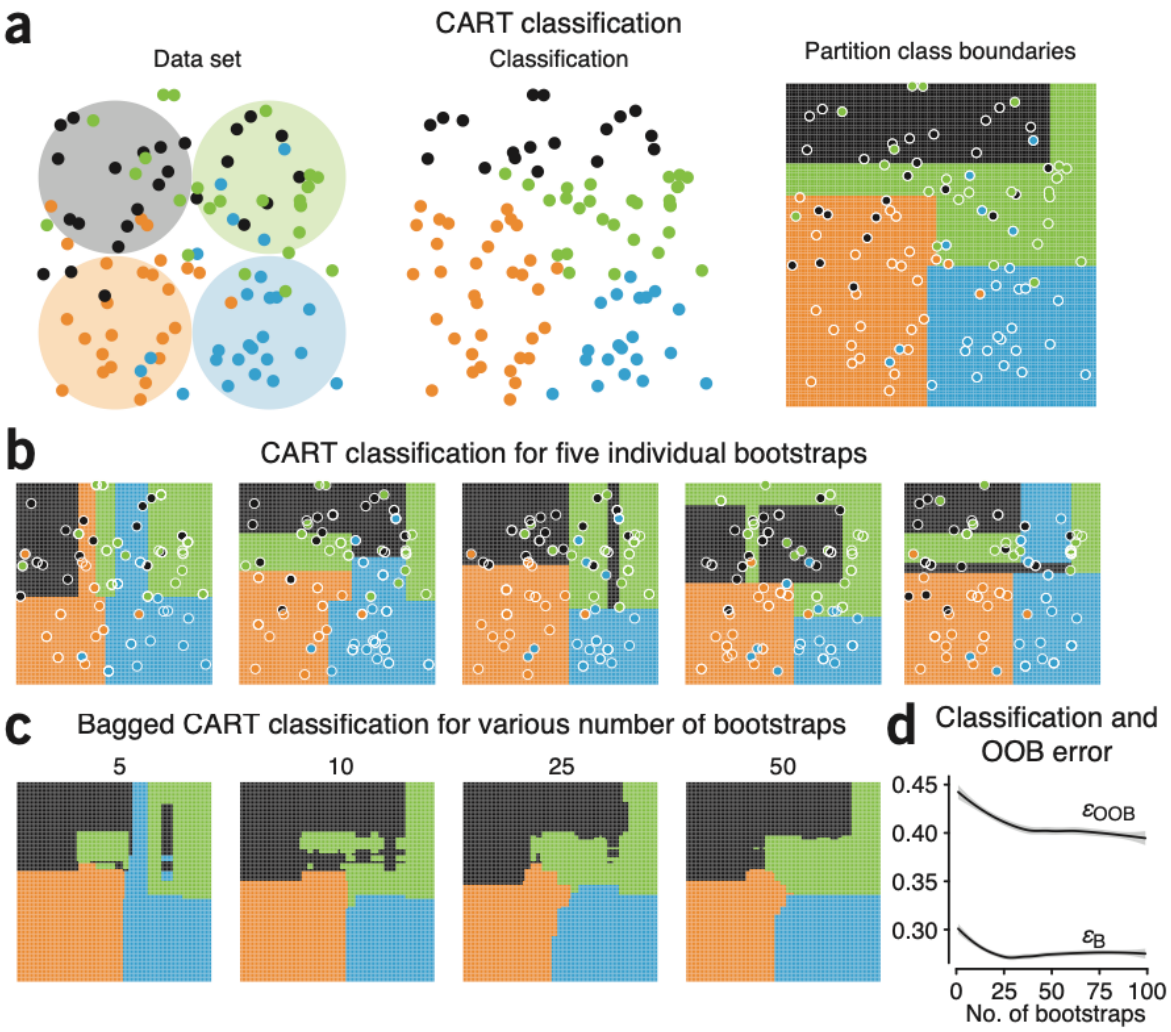

Figure 3 | Application of bagging to classification using a decision tree applied to $n = 100$ two-dimensional data points assigned to one of four color categories. (**a**) The data set is composed of 25 points sampled from the four circles, each with an associated category. Sampling is done from a two-dimensional 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**.

Difference Between Bagging and Random Forests:

Basics

– Both bagging and random forests are ensemble-based algorithms that aim to reduce the complexity of models that overfit the training data. Bootstrap aggregation, also called bagging, is one of the oldest and powerful ensemble methods to prevent overfitting. It is a meta-technique that uses multiple classifiers to improve predictive accuracy. Bagging simply means drawing random samples out of the training sample for replacement to get an ensemble of different models. Random forest is a supervised machine learning algorithm based on ensemble learning and an evolution of Breiman's original bagging algorithm.

Concept

– The concept of bootstrap sampling (bagging) is to train a bunch of unpruned decision trees on different random subsets of the training data, sampling with replacement, to reduce variance of decision trees. The idea is to combine the predictions of several base learners to create a more accurate output. With Random forests, an additional random variation is added into the bagging procedure to create greater diversity amongst the resulting models. The idea behind random forests is to build multiple decision trees and aggregate them to get an accurate result.

Goal

– Both bagged trees and random forests are the most common ensemble learning instruments used to address a variety of machine learning problems. Bootstrap sampling is a meta-algorithm designed to improve the accuracy and stability of machine learning models using ensemble learning and reduce the complexity of overfitting models. The random forest algorithm is very robust against overfitting, and it is good with unbalanced and missing data. It is also the preferred choice of algorithm for building predictive models. The goal is to reduce the variance by averaging multiple deep decision trees, trained on different samples of the data.