

# alessia\_project

Alessia

2023-06-06

## BAGGING PREDICTORS, LEO BRAIMAN

### WHY BAGGING WORKS

#### Numeric Prediction

Instability is important in bootstrap aggregating because it helps improve the performance of the ensemble model by reducing variance and increasing accuracy. Bagging is a technique that combines multiple models (often decision trees) trained on different bootstrap samples of the dataset to make predictions. The key idea behind bagging is to introduce diversity among the base models by using different subsets of the training data. This diversity is achieved through the bootstrap sampling process, where each base model is trained on a randomly selected subset of the original data, allowing for variations in the training sets. If the base models used in bagging are stable, meaning they produce similar predictions when trained on slightly different datasets, the ensemble model's performance may not significantly improve. This is because the models will produce similar predictions and will not provide enough diversity to effectively reduce variance. On the other hand, if the base models are unstable, meaning they are sensitive to small changes in the training data, the ensemble model will benefit from the aggregation of these diverse models. The variations among the models can help to capture different aspects of the data and lead to improved generalization and prediction accuracy. Therefore, instability, or the ability of the base models to exhibit different behaviors with different training data, is important in bagging as it contributes to increased diversity and better performance of the ensemble model.

Let each  $(y, \mathbf{x})$  case in  $L$  be independently drawn from the probability distribution  $P$ . Suppose  $y$  is numerical and  $\Phi(\mathbf{x}, L)$  the predictor. Then the aggregated predictor is the average over  $L$  of  $\Phi(\mathbf{x}, L)$ , i.e.  $\Phi_A(\mathbf{x}) = E_L[\Phi(\mathbf{x}, L)]$ . Take  $\mathbf{x}$  to be a fixed input value and  $y$  an output value. Applying the inequality  $E(Z^2) \geq (E(Z))^2$  then  $E_L[(y - \Phi(\mathbf{x}, L))^2] \geq (y - \Phi_A(\mathbf{x}))^2$ . Integrating both sides over the joint  $y, \mathbf{x}$  output-input distribution, we get that the mean-squared error of  $\Phi_A(\mathbf{x})$  is lower than the mean-squared error averaged over  $L$  of  $\Phi(\mathbf{x}, L)$ . How much lower depends on how unequal the two sides of  $[E_L \delta(\mathbf{x}, L)]^2 \leq E_L \delta^2(\mathbf{x}, L)$  are. The effect of instability is clear. If  $\delta(\mathbf{x}, L)$  does not change too much with replicate  $L$  the two sides will be nearly equal, and aggregation will not help. The more highly variable the  $\delta(\mathbf{x}, L)$  are, the more improvement aggregation may produce. But  $\delta_A$  always improves on  $\delta$ . Now  $\Phi_A$  depends not only on  $\mathbf{x}$  but also the underlying probability distribution  $P$  from which  $L$  is drawn, i.e.  $\Phi_A = \Phi_A(\mathbf{x}, P)$ . But the bagged estimate is not  $\delta_A(\mathbf{x}, P)$  but rather  $\delta_B(X) = \delta_A(\mathbf{x}, P_L)$  where  $P_L$  is the distribution that concentrates  $1/N$  at each point  $(y_n, \mathbf{x}_n) \in L$ . Then if the procedure is unstable,  $\delta_B$  can give improvement through aggregation. On the other side, if the procedure is stable, then  $\delta_B$  will not be as accurate for data drawn from  $P$  as  $\delta_A(\mathbf{x}, P)$ . There is a cross-over point between instability and stability at which  $\delta_B$  stops improving on  $\delta_A(\mathbf{x}, P)$  and does worse.

#### Classification

In classification, a predictor  $\delta(\mathbf{x}, L)$  predicts a class label  $j \in \{1, \dots, J\}$ . Denote  $Q(J|\mathbf{x}) = P(\Phi(\mathbf{x}, L) = j)$ . The interpretation of  $Q(J|\mathbf{x})$  is this: over many independent replicates of the learning set  $L$ ,  $\Phi$  predicts class label  $j$  at input  $\mathbf{x}$  with relative frequency  $Q(J|\mathbf{x})$ . Let  $P(J|\mathbf{x})$  be the probability that input  $\mathbf{x}$  generates class  $j$ . Then the probability that the predictors classifies the generated state at  $\mathbf{x}$  correctly is  $\sum_j Q(J|\mathbf{x})P(J|\mathbf{x})$ . The overall probability of correct classification is  $r = \int [Q(J|\mathbf{x})P(J|\mathbf{x})]P_x(d_x)$  where  $P_x(d_x)$  is the  $\mathbf{x}$  probability

distribution. Note that for any  $Q(j|\mathbf{x})$ ,  $\sum_j Q(j|\mathbf{x})P(j|\mathbf{x}) \leq \max_j P(j|\mathbf{x})$  with equality only if

$$\begin{cases} 1 & \text{if } P(j|\mathbf{x}) = \max_i P(i|\mathbf{x}) \\ 0 & \text{else} \end{cases}$$

The predictor  $\Phi^*(\mathbf{x}) = \operatorname{argmax}_j P(j|\mathbf{x})$  leads to the above expression for  $Q(j|\mathbf{x})$  and gives the highest attainable correct classification rate:  $r^* = \int \max_j P(j|\mathbf{x}) P_X(\mathbf{x}) d\mathbf{x}$ . Call  $\Phi$  order-correct at the input  $\mathbf{x}$  if  $\operatorname{argmax}_j Q(j|\mathbf{x}) = \operatorname{argmax}_j P(j|\mathbf{x})$ . This means that if input  $\mathbf{x}$  results in class  $j$  more often than any other class, then  $\Phi$  also predicts class  $j$  at  $\mathbf{x}$  more often than any other class. The aggregated predictor is:  $\Phi_A(\mathbf{x}) = \operatorname{argmax}_j Q(j|\mathbf{x})$ . For the aggregated predictor the probability of correct classification at  $\mathbf{x}$  is  $\sum_j I(\operatorname{argmax}_i Q(i|\mathbf{x}) = j) P(j|\mathbf{x})$ . If  $\Phi$  is order-correct at  $\mathbf{x}$ , then the previous formula equals  $\max_j P(j|\mathbf{x})$ . Letting  $C$  be the set of all inputs  $\mathbf{x}$  at which  $\Phi$  is order-correct, we get for the correct classification probability of  $\Phi_A$  the expression:

$$r_A = \int_{\mathbf{x} \in C} \max_j P(j|\mathbf{x}) P_X(d\mathbf{x}) + \int_{\mathbf{x} \in C'} [\sum_j I(\Phi_A(\mathbf{x}) = j) P(j|\mathbf{x})] P_X(\mathbf{x})$$

Even if  $\Phi$  is order-correct at  $\mathbf{x}$  its correct classification rate can be far from optimal. But  $\Phi_A$  is optimal. If a predictor is good in the sense that it is order-correct for most inputs  $\mathbf{x}$ , then aggregation can transform it into a nearly optimal predictor. On the other hand, unlike the numerical prediction situation, poor predictors can be transformed into worse ones. The same behavior regarding stability holds. Bagging unstable classifiers usually improves them. Bagging stable classifiers is not a good idea.

### Using the learning set as a test set

In bagging trees, the training set  $L_B$  is generated by sampling from the distribution  $P_L$ . Using  $L_B$  a large tree  $T$  is constructed. The standard CART methodology find the sequence of minimum cost pruned subtrees of  $T$ . The best pruned subtree in this sequence is selected either using cross-validation or a test set. The idea of a test is that it is formed by independently sampling from the same underlying distribution that gave rise to the learning set. In the present context, a test set is formed by independently sampling from  $P_L$ , i.e. we can get a test set by sampling with replacement from the original learning set  $L$ . Infact, consider sampling with replacement a large number of times from  $L$ , each case  $(y_n, \mathbf{x}_n)$  will be selected about the same number of time as any other case. Thus, using a very large test set sampled from  $P_L$  is equivalent to just using  $L$  as a test set.

## A LINEAR REGRESSION ILLUSTRATION

### Forward variable selection

With data of the form  $L = (y_n, \mathbf{x}_n), n = 1 \dots N$  where  $\mathbf{x} = (x_1, \dots, x_M)$  consists of  $M$  predictor variables, a popular prediction method consists of forming predictors  $\delta_1(\mathbf{x}) \dots \delta_M(\mathbf{x})$  where  $\delta_m(\mathbf{x})$  is linear in  $\mathbf{x}$  and depends on only  $m$  of the  $M$   $x$ -variables. Then one of the  $\delta_m(\mathbf{x})$  is chosen as the designated predictor. A common method for constructing the  $\delta_m(\mathbf{x})$ , and one that is used in our simulation, is forward variable entry. If the variables used in  $\delta_k(\mathbf{x})$  are  $(x_{1k}, \dots, x_{mk})$  then for each  $m \notin m_1, \dots, m_K$  form the linear regression of  $y$  on  $(x_{m1}, \dots, x_{mk} + 1, x_m)$ , compute the  $RSS(m)$  and take  $x_{mk+1}$  such that  $m_{k+1}$  minimizes  $RSS(m)$  and  $\delta_{k+1}(\mathbf{x})$  the linear regression based on  $(x_{m1}, \dots, x_{mk+1})$ . What is clear about this procedure is that it is an unstable procedures. The variables are competing for inclusion in the  $\delta_m$  and small changes in the data can cause large changes in the  $\delta_m$ . Forward variable selection is considered an unstable procedure due to the following reasons:

1. Dependency on the order of variable selection: in forward variable selection, the procedure starts with an empty model and adds one variable at a time based on certain selection criteria. However, the initial selection of variables can significantly impact the final result. If the initial set of selected variables differs, the progressive selection of variables can lead to different models. Therefore, the selection of initial variables can cause instability in the results.
2. Sensitivity to correlated variables: forward variable selection may be influenced by correlated variables. When adding a variable to the model, it may appear to provide significant informative value, but it could be strongly correlated with other variables already present in the model. This correlation can

lead to an overestimation of the importance of the added variable and result in selecting a suboptimal model.

3. Impact of small data variations: forward variable selection can be sensitive to small variations in the input data. Even a slight modification of the input data can lead to a completely different variable selection or a different model. This instability can make it challenging to reproduce the results.

To mitigate the potential instability of forward variable selection, techniques such as cross-validation or bootstrap can be used to assess the stability of selected variables and the robustness of the resulting model. These techniques allow for a more accurate estimation of model performance and more reliable identification of the most relevant variables.

### Simulation structure

The simulated data used in this section are drawn from the model  $y = \sum_m \beta_m x_m + \epsilon$  where  $\epsilon$  is  $N(0,1)$ . The number of variables  $M=30$  and the sample size is 60. The subset selection is nearly optimal if there are only a few large non-zero  $\beta_m$ , and its performance is poor if there are many small but non-zero  $\beta_m$ . Subset selection methods can indeed perform well when there are only a few large non-zero coefficients. This is why subset selection can be effective in such cases:

1. Focus on important predictors: subset selection methods aim to identify a subset of predictors that are most relevant for predicting the response variable. When there are only a few predictors with large non-zero coefficients, subset selection can successfully identify and include those important predictors in the model.
2. Improved model interpretability: by selecting a subset of predictors with large non-zero coefficients, subset selection can lead to a more interpretable model. The resulting model will focus on the most influential predictors, providing insights into the relationship between these predictors and the response variable.
3. Reduction in model complexity: by selecting a smaller subset of predictors, subset selection reduces the complexity of the model compared to including all predictors. This reduction in complexity can improve model generalization, making the model more robust and less prone to overfitting.

It's important to note that subset selection methods, like forward selection or backward elimination, are not guaranteed to find the exact optimal model in all cases. They are heuristic approaches that explore different combinations of predictors. The optimality of the selected model depends on factors such as the quality of the selection criteria and the sample size. In practice, it's always recommended to evaluate the performance of the selected model using cross-validation or other validation techniques like bootstrap. This helps to assess the model's predictive accuracy and determine if additional refinement or regularization techniques are needed. In summary, when there are only a few large non-zero coefficients, subset selection methods can be effective in identifying the most important predictors and constructing a simpler, interpretable model. However, it's crucial to validate the model and consider other factors to ensure its performance and reliability.

Three sets of coefficients are used. In the first set of coefficients there are only three non-zero  $\beta_m$ . In the second set of coefficients there are 15 non-zero  $\beta_m$ . In the third set there are 27 non-zero  $\beta_m$ , all relatively small. For each set of coefficients, the following procedure was replicated 250 times: 1. Data  $L = (y_n, \mathbf{x}_n), n = 1 \dots 60$  was drawn from the model  $y = \sum_m \beta_m x_m + \epsilon$ . 2. Forward entry variables was done using  $L$  to get the predictors  $\delta_1(\mathbf{x}) \dots \delta_M(\mathbf{x})$ . The mean\_squared prediction error of each of these was computed giving  $e_1 \dots e_M$ . 3. Fifty bootstrap replicates  $L_B$  of  $L$  were generated. For each of these, forward step-wise regression was applied to construct predictors  $\delta_1(\mathbf{x}, L_B) \dots \delta_M(\mathbf{x}, L_B)$ . These were averaged over the  $L_b$  to give the bagged sequence  $\delta_1^B(\mathbf{X}) \dots \delta_M^B(\mathbf{X})$ . The prediction errors  $e_1^B \dots e_M^B$  for this sequence were computed.

These computed mean-squared-errors were averaged over the 250 repetitions to give the sequences  $\bar{e}_m^S, \bar{e}_m^B$ . For each set of coefficients, these two sequences are plotted vs  $m$  in Figure1 a,b,c.

### Discussion of simulation results

Looking at Figures1 a,b,c an obvious result is that the most accurate bagged predictor is at least as good as the most accurate subset predictor. In the first set of coefficients (with only three non-zero  $\beta_m$ ), subset

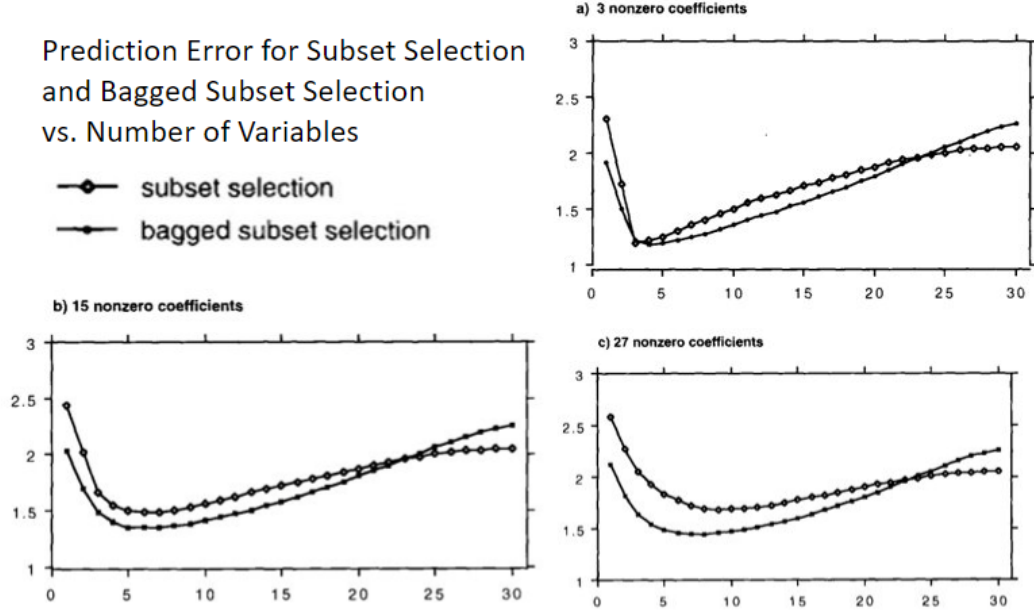


Figure 1: Figure 1. Prediction error for subset selection and bagged subset selection

selection is nearly optimal and there is no improvement. In the second and third set of coefficients there is substantial improvement. *Bagging can improve only if the unbagged is not optimal.* Note that in all three graphs there is a point past which the bagged predictors have larger prediction error than the unbagged. The explanation is this: linear regression using all variables is a fairly stable procedure. The stability decreases as the number of variable used in the predictor decreases. For a stable procedure  $\delta_B = \delta_A(\mathbf{x}, P_L)$  is not as accurate as  $\delta(\mathbf{x}, P)$ . As  $m$  increases, the instability increases and there is a cross-over point at which  $\delta_m^B$  becomes more accurate than  $\delta_m$ .

## CONCLUDING REMARKS

### Bagging class probability estimates

Some classification methods estimate probabilities  $p(j|\mathbf{x})$  that an object with prediction vector  $\mathbf{x}$  belongs to class  $j$ . Then the class corresponding to  $\mathbf{x}$  is estimated as  $\arg\max_j p(j|\mathbf{x})$ . For such methods, a natural competitor to bagging by voting is to average the  $p(j|\mathbf{x})$  over all bootstrap replications, getting  $p_B(j|\mathbf{x})$  and then use the estimated class  $\arg\max_j p_B(j|\mathbf{x})$ . In some applications, estimates of class probabilities are required instead of, or along with, the classifications. The evidence so far indicates that bagged estimates are likely to be more accurate than the single estimates. To verify this, it would be necessary to compare both estimates with the true values  $p^*(j|\mathbf{x})$  over the  $\mathbf{x}$  in the test set. For real data the true values are unknown but they can be computed for the simulated waveform data. Using the waveform data, a simulation with learning and test sets both of size 300 and 25 bootstrap replications has been done. In each iteration, the average of  $|p(j|\mathbf{x}) - p^*(j|\mathbf{x})|$  and  $|p_B(j|\mathbf{x}) - p^*(j|\mathbf{x})|$  was computed. This was repeated 50 times and the results averaged. The single tree estimated had an error of .189. The error of the bagged estimates was .124, a decrease of 34%.

### How many bootstrap replicate are enough?

In the paper's experiment, 50 bootstrap replicates was used for classification and 25 for regression. To give some ideas of what the results are as connected with the number of bootstrap replicates we ran the waveform data using 10, 25, 50 and 100 replicates using the same simulation scheme. The results appear

Table 10. Bagged Misclassification Rates (%)

No. Bootstrap Replicates	Misclassification Rate
10	21.8
25	19.4
50	19.3
100	19.3

in Table 10.

The unbagged rate is 29.1, so its clear that we are getting most of the improvement using only 10 bootstrap replicates.

### How big should the bootstrap learning set be?

In the paper’s simulations the bootstrap replicates  $L^B$  were used of the same size as the initial learning set  $L$ . While a bootstrap replicate may have 2,3,... duplicates of a given instance, it also leaves out about .37 of the instances. A reader of the technical report on which the paper is based remarked that this was an appreciable loss of data, and that accuracy might improve if a larger bootstrap set was used. Anyway with a bootstrap learning sets twice the size of  $L$  there was no improvement in accuracy.

### Conclusions

Bootstrap Aggregating can be a valuable approach to improve the performance of a predictive model, especially when the initial model is not very accurate or reliable. Bagging is a technique that involves creating multiple bootstrap samples from the original data and fitting a separate model to each sample. The models are then combined or aggregated to make predictions. Implementing bagging is relatively straightforward, as it requires adding a loop to select bootstrap samples and sending them to the modeling procedure, as well as an aggregation step at the end. However, when using bagging with decision trees as the base model, one may lose the simplicity and interpretability of the individual trees. Decision trees typically have a clear and interpretable structure that allows for easy understanding of the underlying rules. In the bagging process, where multiple trees are combined, the resulting ensemble model may not have a straightforward and interpretable structure. Despite this loss of interpretability, the advantage of using bagging with decision trees is increased accuracy. By combining multiple trees through bagging, the variance is reduced, leading to improved predictive performance.

In summary, bagging is a valuable technique for improving the accuracy of a predictive model, even at the cost of losing the simplicity and interpretability of individual trees and it allows for increased accuracy through the combination of multiple models.