Bagging algorithm

Quick note

Abdelwahid Benslimane

wahid.benslimane@gmail.com

One of the main drawbacks of decision trees is that they are unstable, which means that small changes in the data can produce very different trees. The need to respond to this type of problem has led to certain approaches such as bagging (bootstrap aggregating). The underlying idea is to use randomness to produce different data and therefore different models whose results will be averaged, thus reducing the variance.

Let x_i be an example from the learning base described by p attributes $A_1, A_2, \ldots, A_p$, $x_i = \{a_1^i, a_2^i, \ldots, a_p^i\}$.

Let y_i be a realization of the target variable that can take continuous (regression) or discrete (classification) values.

Let us consider that $G(x)$ is a prediction model learned on a sample of data $z = \{(x_i, y_i)\}_{i=1}^n$.

Bagging algorithm:

- we randomly draw from the training base B samples with replacement $z_i, i = 1, \ldots, B$ (each sample having n examples) - called bootstrap samples;
- for each sample i we compute the model $G_i(x)$;
- in the case of regression: we aggregate by the mean $G(x) = \frac{1}{B}\sum_{i=1}^{B} G_i(x)$;
- in the case of classification: we aggregate by the vote $G(x) = majority\ vote\ among\ (G_1(x), \ldots, G_B(x))$.

The diagram below illustrates the procedure very well.

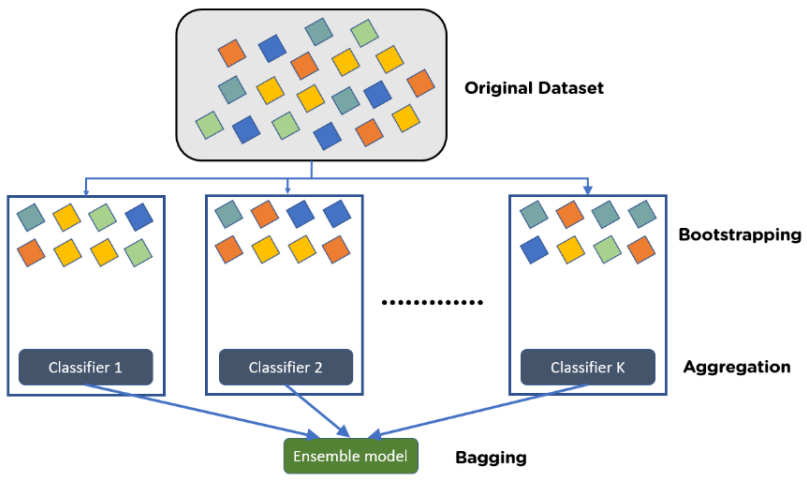

Image source: https://www.simplilearn.com/tutorials/machine-learning-tutorial/bagging-in-machine-learning

In the case of regression, it is the averaging of the estimates that reduces the variance. Indeed, if we have N independent and identically distributed variables of variance σ^2, then the variance of the mean of these variables is $\frac{\sigma^2}{N}$. In the case of classification, the majority vote ensures greater stability of the result. But are the estimators really independent here? This point will be discussed lower in this note.

Choosing the number B of estimators:

The performance criterion for determining the number B of estimators is the OOB (Out Of Bag) error: for each example x_k we construct a predictor (of the random forest type) by combining only the trees G_i in which the example did not participate in the training (so $x_k \notin z_i$ <=> x_k is not part of the bootstrap sample of G_i).

We then aggregate (e.g., by the mean) the OOB errors obtained for each example.

We choose the number B for which the error stabilizes/does not go down anymore. Adding more trees would increase the computation time without improving the results. The OOB error allows to test the efficiency of the resulting model during its construction.

Defects of bagging:

The G_i estimators are in fact not independent. Indeed, they are computed on samples that overlap strongly (sampling with replacement) and therefore they are correlated.

If we have identically distributed but not independent random variables $X_1, X_2, \ldots, X_B$, of variance σ^2 and correlation $\rho = Corr(X_i, X_j), \forall i \neq j$, then $Y = \frac{1}{B}(X_1 + X_2 + \cdots + X_B)$ is of variance:

$Var(Y) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$.

When B is large the 2^{nd} term is negligible but the 1^{st} is not. The improvement proposed by random forests is to lower the correlation between G_i using an additional randomization step.