What is bagging?

Bagging, also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset.

In bagging, a random sample of data in a training set is selected with replacement—meaning that the individual data points can be chosen more than once. After several data samples are generated, these weak models are then trained independently, and depending on the type of task—regression or classification, for example—the average or majority of those predictions yield a more accurate estimate.

NOTE: The random forest algorithm is considered an extension of the bagging method, using both bagging and feature randomness to create an uncorrelated forest of decision trees.

*Ensemble learning*

Ensemble learning gives credence to the idea of the “wisdom of crowds,” which suggests that the decision-making of a larger group of people is typically better than that of an individual expert.

Similarly, ensemble learning refers to a group (or ensemble) of base learners, or models, which work collectively to achieve a better final prediction. A single model, also known as a base or weak learner, may not perform well individually due to high variance or high bias. However, when weak learners are aggregated, they can form a strong learner, as their combination reduces bias or variance, yielding better model performance.

Ensemble methods are frequently illustrated using decision trees as this algorithm can be prone to overfitting (high variance and low bias) when it hasn’t been pruned and it can also lend itself to underfitting (low variance and high bias) when it’s very small, like a decision stump, which is a decision tree with one level. Remember, when an algorithm overfits or underfits to its training set, it cannot generalize well to new datasets, so ensemble methods are used to counteract this behavior to allow for generalization of the model to new datasets. While decision trees can exhibit high variance or high bias, it’s worth noting that it is not the only modeling technique that leverages ensemble learning to find the “sweet spot” within the bias-variance tradeoff.

*Bagging vs. boosting*

Bagging and boosting are two main types of ensemble learning methods. As highlighted in this [study](https://www.d.umn.edu/~rmaclin/cs5751/notes/opitz-jair99.pdf) (PDF, 248 KB) (link resides outside IBM), the main difference between these learning methods is the way in which they are trained.

In bagging, weak learners are trained in parallel, but in boosting, they learn sequentially.

This means that a series of models are constructed and with each new model iteration, the weights of the misclassified data in the previous model are increased. This redistribution of weights helps the algorithm identify the parameters that it needs to focus on to improve its performance. AdaBoost, which stands for “adaptative boosting algorithm,” is one of the most popular boosting algorithms as it was one of the first of its kind. Other types of boosting algorithms include XGBoost, GradientBoost, and BrownBoost.

Another difference in which bagging and boosting differ are the scenarios in which they are used. For example,

* bagging on weak learners which exhibit high variance and low bias,
* boosting on low variance and high bias is observed.

How bagging works

In 1996, [Leo Breiman](https://link.springer.com/content/pdf/10.1007/BF00058655.pdf) (PDF, 829 KB) (link resides outside IBM) introduced the bagging algorithm, which has three basic steps:

1. **Bootstrapping:** Bagging leverages a bootstrapping sampling technique to create diverse samples. This resampling method generates different subsets of the training dataset by selecting data points at random and with replacement. This means that each time you select a data point from the training dataset, you are able to select the same instance multiple times. As a result, a value/instance repeated twice (or more) in a sample.
2. **Parallel training:** These bootstrap samples are then trained independently and in parallel with each other using weak or base learners.
3. **Aggregation:**Finally, depending on the task (i.e. regression or classification), an average or a majority of the predictions are taken to compute a more accurate estimate. In the case of regression, an average is taken of all the outputs predicted by the individual classifiers; this is known as soft voting. For classification problems, the class with the highest majority of votes is accepted; this is known as hard voting or majority voting.

Benefits and challenges of bagging

There are a number of key advantages and challenges that the bagging method presents when used for classification or regression problems. The key benefits of bagging include:

* **Ease of implementation:**Python libraries such as scikit-learn (also known as sklearn) make it easy to combine the predictions of base learners or estimators to improve model performance. Their [documentation](https://scikit-learn.org/stable/modules/ensemble.html" \t "_blank) (link resides outside IBM) lays out the available modules that you can leverage in your model optimization.
* **Reduction of variance**: Bagging can reduce the variance within a learning algorithm. This is particularly helpful with high-dimensional data, where missing values can lead to higher variance, making it more prone to overfitting and preventing accurate generalization to new datasets.

The key challenges of bagging include:

* **Loss of interpretability**: It’s difficult to draw very precise business insights through bagging because due to the averaging involved across predictions. While the output is more precise than any individual data point, a more accurate or complete dataset could also yield more precision within a single classification or regression model.
* **Computationally expensive:** Bagging slows down and grows more intensive as the number of iterations increase. Thus, it’s not well-suited for real-time applications. Clustered systems or a large number of processing cores are ideal for quickly creating bagged ensembles on large test sets.
* **Less flexible:**As a technique, bagging works particularly well with algorithms that are less stable. One that are more stable or subject to high amounts of bias do not provide as much benefit as there’s less variation within the dataset of the model. As noted in the [Hands-On Guide to Machine Learning](https://bradleyboehmke.github.io/HOML/" \t "_blank) (link resides outside of IBM), “bagging a linear regression model will effectively just return the original predictions for large enough *b*.”

Applications of Bagging

The bagging technique is used across a large number of industries, providing insights for both real-world value and interesting perspectives, such as in the [GRAMMY Debates with Watson](https://developer.ibm.com/articles/grammy-debates-with-watson-from-the-lab-to-musics-greatest-stage/). Key use cases include:

* **Healthcare**: Bagging has been used to form medical data predictions. For example, [research](https://www.maths.usyd.edu.au/u/pengyi/publication/EnsembleBioinformatics-v6.pdf" \t "_blank) (PDF, 2.8 MB) (link resides outside IBM) shows that ensemble methods have been used for an array of bioinformatics problems, such as gene and/or protein selection to identify a specific trait of interest. More specifically, this [research](https://www.researchgate.net/profile/Nongyao-Nai-Arun/publication/269381309_Ensemble_Learning_Model_for_Diabetes_Classification/links/57512e6708ae1c34b39caac5/Ensemble-Learning-Model-for-Diabetes-Classification.pdf" \t "_blank) (link resides outside IBM) delves into its use to predict the onset of diabetes based on various risk predictors.
* **IT**: Bagging can also improve the precision and accuracy in IT systems, such as ones network intrusion detection systems. Meanwhile, [this research](https://www.sciencedirect.com/science/article/pii/S1877050915007401" \t "_blank) (link resides outside IBM) looks at how bagging can improve the accuracy of network intrusion detection—and reduce the rates of false positives.
* **Environment**: Ensemble methods, such as bagging, have been applied within the field of remote sensing. More specifically, [this research](https://www.mdpi.com/2072-4292/12/10/1683/htm) (link resides outside IBM) shows how it has been used to map the types of wetlands within a coastal landscape.
* **Finance**: Bagging has also been leveraged with deep learning models in the finance industry, automating critical tasks, including fraud detection, credit risk evaluations, and option pricing problems. This [research](https://www.researchgate.net/publication/316510985_Study_of_Data_Mining_Techniques_used_for_Financial_Data_Analysis) (link resides outside IBM) demonstrates how bagging among other machine learning techniques have been leveraged to assess loan default risk. This [study](https://www.sciencedirect.com/science/article/pii/S1877050915007103) (link resides outside IBM) highlights how bagging helps to minimize risk by to prevent credit card fraud within banking and financial institutions.