Proseminar
Advanced topics in machine learning

Bagging, Boosting, and Ensemble Learning

Author

Florian Zager

Matrikelnummer: *2489018*

Date:

Karlsruhe, December 26, 2023

Abstract — Zusammenfassung

Mandatory. Short summary of the report.

Contents

1 Introduction

Mandatory. Questions like: What is the topic of this work, what's the broader context (topic of the proseminar), why is it relevant?

2 Ensemble Learning

Ensemble learning is an advanced machine learning approach that combines the strengths of multiple smaller learning algorithms to improve predictive performance. The concept behind ensemble learning is analogous to the "wisdom of crowds". Which describes, that a crowd, on average, makes collectively better decisions, than any single member of it.

Just as a diverse group of people can provide a more accurate collective decision than an individual, in ensemble learning, a combination of learning algorithms often predicts more accurately than an individual learning algorithm. This approach is based on the principle that a diverse set of learning algorithms can capture different patterns or trends in the data, leading to more robust and accurate predictions.

To be more precise, ensemble methods use multiple smaller learning algorithms, which specialize in small aspects of the problem. However, by combining these algorithms, the ensemble often achieves better predictive performance than the used algorithm could achieve alone because they complement each others strengths and weaknesses.

So the goal of ensemble learning is to achieve a better predictive performance. Nevertheless, it comes at the cost of increased computational resources for training as well as prediction and storage.

Overall, there are many different ensemble methods, such as bagging and boosting, which we will go into more detail in this report. However, there are many more like stacking and blending.

2.1 Bagging

Bootstrap Aggregating, commonly known as bagging, is an ensemble learning method developed by Breiman (1996). The models for the ensemble get trained individually by using the bootstrapping technique. Bootstrapping involves creating random subsets of the original training dataset. The subsets are created by drawing random data point with replacement and have the same size as the original training dataset. This means that data points can be chosen more than once and that some data points might not be in the subset. The models can be trained in parallel, because of the individual training.

Figure 1 shows how bootstrapping might work on an imaginary dataset. In subset 1 each class is equally distributed. Subset 2 has a focus on the red and orange bubles. The subset N has a strong focus on blue, however it doesn't even have a single orange buble. This will probably mean that the model trained on subset N will be very good at predicting blue, but not very accurate with orange.

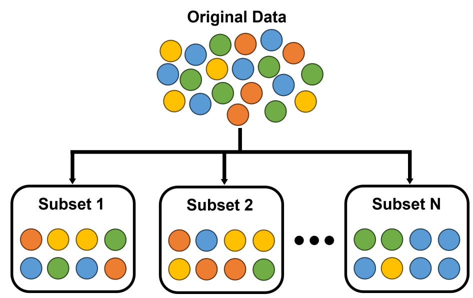

Figure 1: Creating subsets with bootstrapping.

In the bagging ensemble each model makes its prediction independently. Because of that the predictions of the individual models can be ran in parallel, similar to the training. Once every model in the ensemble has made their prediction, they get aggregated to form a final ensemble prediction. The method of aggregation depends on the problem that is being solved by the bagging ensemble. For classification problems a common method is majority voting. Each model votes for a particular class. The class that receives the most votes is chosen for the final ensemble prediciton. For regression problems the predictions of the individual models are typically averaged.

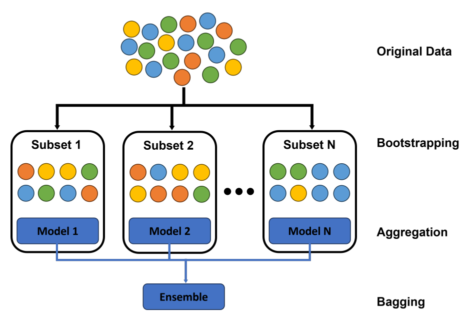

Figure 2: Bagging prediction.

Bagging can be particularly effective when using base learning algorithms that have high variance or tend to overfit quickly. By bootstrapping and averaging the predictions the variance gets reduced and potential overfitting is avoided. Same goes for unstable learning algorithms, that produce significantly different results on small changes in the data. Thats why bagging is often used with decision trees as they tend to be unstable and to have high variance. Additionally, bagging can be very helpful when dealing with noisy, imbalanced datasets or datasets with missing data. The bootstrapping helps to create diverse datasets with each class being adequately represented and also averaging out the noise in combination with aggregation.

All in all, bagging can help to reduce variance, prevent overfitting and to built a more resilient, robust and generalized model.

2.2 Random Forest

2.3 Boosting

2.4 Gradient Boosting

2.5 Extreme Gradient Boosting

3 Examples

3.1 Example 1

3.2 Example 2

4 Summary and conclusion

Mandatory. Short summary of the most important aspects of the report. If possible: What are open challenges?

- Bagging vs. Boosting - whats the difference?

References

Breiman, L. (1996, 01. Aug). Bagging predictors. *Machine Learning*, *24* (2), 123-140. Zugriff auf `https://doi.org/10.1007/BF00058655` doi: 10.1007/BF00058655