A study on Gradient Boosting algorithms

Juliano Garcia de Oliveira

Advisor: Dr. Roberto Hirata Jr.

April,2019

1 Abstract

This project consists in a study of the most popular implementations of *Gradient Boosting Machines* (GBMs), namely the XGBoost and LightGBM library. Gradient Boosting Machines are the state of the art machine learning techniques to deal with structured data. XGBoost and LightGBM implementations are widely used in the industry, and also are part of almost all winning solutions in machine learning competions in Kaggle, according to an industry-wide survey on data science and machine learning (Kaggle [2017]). The performance of GBM models heavily depends on hyperparameter tuning, and the objective of this work is to study the impact of different hyperparameters in the performance of GBM models in different datasets and how different datasets characteristics impact the performance of the models. With this study, it is expected to obtain more insight into the boundaries and capabilities of GBM models and the important aspects that make them so valuable for modern data science solutions in real world applications.

2 Introduction and Motivation

Machine Learning techniques are widely used in the modern industry to leverage useful insights and applications from data. In a range of very different areas, from medical labs to financial institutions, machine learning models are already a fundamental part of the business. There is a rising trend of academic research on the theory and applications of machine learning techniques (Hao [2019]), which is evidence of the growing importance of robust machine learning models for modern scientific and industrial applications. Besides pure predictive accuracy of a model, most machine learning techniques in the industry need to be scalable, flexible and deal with high amounts of data in a distributed environment.

Another important area of development in data science today is about **machine learning interpretability**. The ability to interpret a model and explain the predictions it generates are key for risk minimization and needed the ethics if computational models. As pointed out by Molnar [2019]: "The higher the interpretability of a machine learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made".

The Gradient Boosting Machine algorithm is used for supervised machine learning, and it produces an ensemble of weak learners. The most used implementations of the GBM techniques are LightGBM by Ke et al. [2017], and the XGBoost library by Chen and Guestrin [2016]. Both of these implementations are highly scalable, flexible, and can make use of a highly distributed environment to reduce the training time. They are *gradient boosted trees* implementations, a specific case of GBM where the weak learners are decision trees. Decision trees have very high interpretability, and as such, Gradient Boosted Trees also have good tools and techniques for human interpretable explanations of models. The high predictive performance capability and interpretability are two aspects that make these techniques very interesting to study.

To leverage the full potential of GBMs, hyperparameter tuning is a critical aspect in the machine learning modelling pipeline. Despite the increase in development of automated tuning techniques and the current available hyperparameter methods (e.g. extended grid search, randomized parameter optimization, tree-structured parzen estimator), XGBoost and LightGBM libraries have a very high number of hyperparameters to be optimized, which depending on the amount of data used renders these tuning techniques infeasible. This is why it's important to study more the impact of the hyperparameters in different datasets characteristics, as it can provide useful insights to tackle new and existing machine learning solutions which use GBMs.

3 Objectives

- To study on two implementations of Gradient Boosting Machines algorithms:
  - XGBoost, a "Scalable and Flexible Gradient Boosting" by Chen and Guestrin [2016]
  - LightGBM, a "fast, distributed, high performance gradient boosting framework based on decision tree algorithms", developed by Microsoft Research in Ke et al. [2017].
- Explore hyperparameters effects in predictive accuracy of the GBM models in a range of different datasets with different characteristics: highly skewed features, categorical features with high cardinality, robustness to missing data, outliers, low sample sizes, etc.
- An analytical review about the practical usage of GBM techniques in the industry and machine learning competitions.

4 Work plan

1. Bibliographic research, study of the theory behind Gradient Boosting Machines and Decision Trees
2. Study of XGBoost and LightGBM libraries and hyperparameters
3. Gather and study different datasets from online repositories
4. Implement hyperparameter optimization of XGBoost and LightGBM for different datasets
5. Compare statistical results and measure impact of different combinations hyperparameters and dataset's characteristics
6. Research of ensemble methods and usage of Gradient Boosting techniques in the industry and machine learning competitions
7. Writing the thesis
8. Writing the poster
9. Writing the presentation

Activity	March	April	May	June	July	August	September	October	November
1	X	X	X	X	X				
2		X	X	X	X	X			
3			X	X	X	X	X		
4			X	X	X	X	X	X	
5				X	X	X	X	X	
6					X	X	X	X	
7	X	X	X	X	X	X	X	X	X
8								X	X
9								X	X

5 Methods and Materials

Most of the implementations and comparisons will be made using Python and the main libraries for machine learning and data analysis, such as:

- Scikit-learn
- matplotlib
- seaborn
- numpy

Alongside, using the two GBM libraries mentioned above: XGBoost and LightGBM.

Arxiv and Google Scholar will be the main sources for scientific articles, and Kaggle and the UCI machine learning repository will be the main sources to gather datasets.

References

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 785–794.

Hao, K. (2019, Feb). We analyzed 16,625 papers to figure out where ai is headed next. `https://www.technologyreview.com/s/612768/we-analyzed-16625-papers-to-figure-out-where-ai-is-headed-next/`.

Kaggle (2017). The state of ml and data science 2017. `http://www.kaggle.com/surveys/2017`.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu (2017). Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan,

and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*, pp. 3146–3154. Curran Associates, Inc. http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf.

Molnar, C. (2019). *Interpretable Machine Learning*. Available at https://christophm.github.io/interpretable-ml-book/.