University of São Paulo
Institute of Mathematics and Statistics
Bachelor of Computer Science

A Study on Gradient Boosting Classifiers

A large-scale experimental analysis of hyperparameter effect on binary classification models

Juliano Garcia de Oliveira

Undergraduate Thesis

mac 499 — Capstone Project

Program: Computer Science
Advisor: Prof. Dr. Roberto Hirata Jr.

São Paulo
November 2019

Too much consistency is as bad for the mind as it is for the body. Consistency is contrary to nature, contrary to life. The only completely consistent people are the dead.
— Aldous Huxley

Resumo

Juliano Garcia de Oliveira. **Um Estudo sobre Classificadores de *Gradient Boosting*: *Uma análise experimental em larga escala do efeito de hiperparâmetros em modelos de classificação binária*.** Monografia (Bacharelado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2019.

O *Gradient Boosting Machines* (GBMs) é um algoritmo supervisionado de aprendizado de maquina que vem obtendo excelente resultados em uma ampla gama de problemas e vencendo diversas competições de aprendizado de máquina. Ao construir um modelo de aprendizado de máquina, a otimização de hiperparâmetros pode se tornar uma tarefa dispendiosa e demorada, dependendo do número e do espaço de busca do procedimento de otimização. Usuários de aprendizado de máquina que não são pesquisadores experientes ou profissionais de ciência de dados podem ter dificuldade para definir quais hiperparâmetros e valores escolher ao iniciar o ajuste do modelo, especialmente com implementações mais recentes de GBMs como a biblioteca XGBoost e LightGBM. Neste trabalho, um experimento em larga escala com 70 conjuntos de dados é realizado usando a plataforma OpenML, medindo a sensibilidade das métricas de avaliação de classificadores binários a alterações em três hiperparâmetros da biblioteca LightGBM. Um arcabouço estatístico sólido é aplicado aos resultados do estudo, analisando o comportamento através de três pontos de vista diferentes: resultados por hiperparâmetros, resultados por características do conjunto de dados e resultados por métrica de desempenho. Os experimentos realizados indicam relações interessantes dos hiperparâmetros nos classificadores de *gradient boosting*, descobrindo quais combinações de hiperparâmetros resultaram em modelos com a maior alteração nas métricas, quais delas são mais sensíveis e quais características dos conjuntos de dados estudados se destacaram. Estes resultados são apresentados aqui para facilitar a criação de modelos de classificação baseados em GBMs em aprendizado de máquina.

Palavras-chave: Importância de Hiperparâmetros. Gradient Boosting. Aprendizado de Máquina Supervisionado. Análise Experimental. Classificação. Seleção de Modelos. Estudo Empírico.

Abstract

Juliano Garcia de Oliveira. **A Study on Gradient Boosting Classifiers:** ***A large-scale experimental analysis of hyperparameter effect on binary classification models***. Undergraduate Thesis (Bachelor). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2019.

Gradient Boosting Machines (GBMs) is a supervised machine learning algorithm that has been achieving state-of-the-art results in a wide range of different problems and winning machine learning competitions. When building any machine learning model, the hyperparameter optimization can become a costly and time-consuming task depending on the number and the hyperparameter space of the tuning procedure. Machine learning users that are not experienced researchers or data science professionals can struggle to define which hyperparameters and values to choose when starting the model tuning, especially with newer GBMs implementations like the XGBoost and LightGBM library. In this work, a large-scale experiment with 70 datasets is conducted using the OpenML platform, measuring the sensitivity of binary classifiers evaluation metrics to changes in three LightGBM hyperparameters. A solid statistical framework is applied to the study results, analyzing the behavior from three different viewpoints: results by hyperparameters, results by characteristics of the dataset and results by performance metric. The carried out experiments indicate insightful relationships of the hyperparameters in gradient boosting classifiers, uncovering which combinations of hyperparameters resulted in models with the highest change in the metrics from the baseline, what metrics are most sensitive and which characteristics of the studied datasets stood out. These results are hereby here presented to facilitate the model building of gradient boosting classifiers for machine learning users.

Keywords: Hyperparameter Importance. Gradient Boosting. Supervised Machine Learning. Experimental Analysis. Classification. Model Selection. Empirical Study.

Lista de Abreviaturas

CFT	Transformada contínua de Fourier (*Continuous Fourier Transform*)
DFT	Transformada discreta de Fourier (*Discrete Fourier Transform*)
EIIP	Potencial de interação elétron-íon (*Electron-Ion Interaction Potentials*)
STFT	Transformada de Fourier de tempo reduzido (*Short-Time Fourier Transform*)
ABNT	Associação Brasileira de Normas Técnicas
URL	Localizador Uniforme de Recursos (*Uniform Resource Locator*)
IME	Instituto de Matemática e Estatística
USP	Universidade de São Paulo

Lista de Símbolos

ω	Frequência angular
ψ	Função de análise *wavelet*
Ψ	Transformada de Fourier de ψ

List of Figures

List of Tables

List of Programs

Contents

Insira o conteúdo dos capítulos do seu trabalho no arquivo “`capitulos.tex`” do diretório “`conteudo`”.

Insira o conteúdo dos apêndices do seu trabalho no arquivo “`apendices.tex`” do diretório “`conteudo`”.

Insira o conteúdo dos anexos do seu trabalho no arquivo “anexos.tex” do diretório “conteudo”.

Index