DOI: 10.20396/revpibic2720192773

XXVII Congresso de Iniciação Científica Unicamp
16 a 18 de outubro de 2019 - Campinas | Brasil

Study and analysis of the Kaggle Competitions to encourage the learning and the exchange of experiences of the Machine Learning community of Campinas

Alexandre T. Okita, Roberto A. Lotufo

Abstract

Kaggle is the main platform of predictive and analytics models competition. The goal of this research project is to study and analyze the main Kaggle competitions, their datasets, and their winner teams solutions, thus, learning the best techniques and practices in the machine learning area. This information, summed up with Kaggle Learn courseware, was used as a subsidy for themes and activities of Kaggle Campinas Meetup meetings coordinated together with the research. The competition analysis produced an updated view of some of the bests techniques for each problem type, and the meetings helped to broaden the machine learning community of Campinas and Unicamp.

Key words:
Machine Learning, dataset, Kaggle.

Introduction

According to the LinkedIn article "LinkedIn's Most Promising Jobs in 2019," Data Science leads the ranking with the highest career advancement score. As there are not many specific courses in Brazil, there is a growing demand for complementary training in the area.

Kaggle is the most popular site in the world with scientific data competitions. It hosts hundreds of public competitions and has thousands of public data sets. Therefore, it is common for users to write Jupyter Notebooks and discussion threads to share their techniques and solutions. The Kaggle community also writes tutorials on some popular topics of machine learning and data mining. Some of these tutorials make up Kaggle's educational line, "Kaggle Learn." Our impressions of "Intro machine learning" course are described in the blog [1]

It is expected that the creation of interest groups could bring new students to the area. The use of Kaggle competitions and discussions is a practical way of collectively studying and updating on new techniques to solve some of the problems of data science.

Figure 1. *First Kaggle Campinas Meetup meeting*

The first meeting of the Kaggle Campinas Meetup (Fig. 1) occurred on 10/3/2018 and had thirty-four members present. On 07/07/2019, the group has 320 members. The meetups unified, besides the academics of the area and market players, people from other areas of computer science. Lessons, lectures and commented solutions to competitions, as well as insights into specific data sets, were produced for the meetings. A student body called "Iris" is being formed at Unicamp to expand the data science community among students.

Results and Discussion

There are four types of data most common in Kaggle 2019 competitions: image data, tabular data, text data, and signal data. From the analysis of the best solutions of eight competitions of 2019 (Table 1), it is possible to link a common problem to a standard solution idea. Tabular data are commonly solved with gradient boosting models, while text data and image data are solved with pre-trained deep learning models, such as Bert and ResNeXt, respectively. The signal data show no visible pattern among the best solutions.

Competition	Data type	Solution
iMaterialist	Image Data	Hybrid Task Cascade with ResNeXt-101-64x4d-FPN
Quick, Drawl Do	Image Data	ResNeXt, CNN, RNN, LightGBM
Instant Gratificati	Tabular Data	Gaussian Mixture Model
Santander Custo	Tabular Data	LightGBM and NN (2.1NN, 1LGBM)
LANL Earthquak	Signal Data	LightGBM, SVR, NN
CareerCon 2019	Signal Data	1D Convolutional NN
Gendered Prono	Text Data	BERT
Jigsaw Unintend	Text Data	BERT

Table 1. *Kaggle competitions analysis*

Conclusions

Kaggle has hundreds of free contests and thousands of public data sets. In addition, it also has a cloud environment that removes the work of beginners from having to install the initial settings to begin studying machine learning, as well as allowing solutions to be easily reproduced in the same environment - which helps the community Kaggle to be more supportive, sharing ideas and working together to resolve competitions. These features show Kaggle as a great learning tool that allows to be updated with the state of the art and is able to assist in the generation of a complementary training in the area of data science.

Acknowledgement

We thank NeuralMind Inteligência Artificial for its valuable comments and discussions.

[1] "Intro to machine learning" Kaggle Learn course: from zero to beginner. Available in: <https://medium.com/@alexandreokita/intro-to-machine-learning-kaggle-learn-course-from-zero-to-beginner-bd65cc362391>