On The Selection of Predictive Models in Production

Anonymous Author(s)
Affiliation
Address
email

1 Research Problem

Recent works [5, 3, 11], highlight some of the challenges encountered in predictive serving systems. We are interested in the problem arising in the presence of multiple competing predictive models. This may be the case in large companies in which autonomously developed models about the same phenomenon (a.k.a competing models) are deployed and used during an ad-hoc predictive query. In this context, the training and validation processes used in building competing models may consider different fragments of the modeled phenomenon domain leading to a variation on their predictive accuracy. In this context, predictions used to answer ad-hoc queries must be informed with an approximation of the generalization error [7, 10], as a function of the distance between the query data space and the model building data space. Thus, we investigate the problem of selecting predictive machine learning competing models under this settings.

2 Motivation

The production of an ML model into a framework for user consumption, involves the steps showed in Figure 1:

(a) Summarize the traditional pipeline to train and validate ML models [6, 9], in order to facilitate the implementation of such ML models.

(b) Represent the process for deploying ML models, considering a four-step cycle (monitor, evaluate, compare and rebuild). This is used to analyze the selection and life-span of the ML model(s) for user consumption [2, 13].

(c) Enable prediction serving and decision-making by processing predictive queries formulated by users [3, 8].

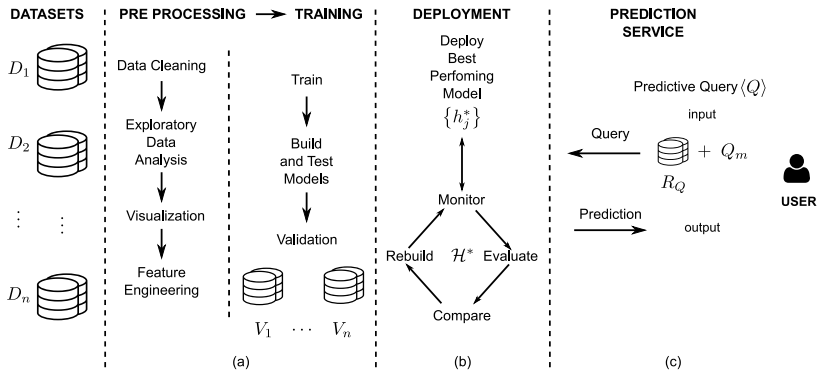

Figure 1: ML Model Lifecycle.

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

Our work focuses on steps (b), (c), and how they interact. Given a set of ML models in a production environment $\mathcal{H}^* = \{h_j^* : j \in J\}$ and a predictive query $Q = \langle R, Q_m \rangle$ we are interested in the following: In (b), the main task consists in monitoring the ML model behavior, in order to identify a possible model decay and the need to rebuild the models; monitoring is based on a thorough analysis of model qualitative characteristics and statistical properties for datasets used to build a model. In (c), given a predictive query and a pool of models available and their respective datasets, we need mechanisms and techniques to choose the adequate model to attend the query.

Regarding the integration of additional data sources: determine how changes in the data distribution affect the models predictive quality. It is known that those distribution changes can occur gradually, seasonally (in certain time intervals) or abruptly [1].

3 Research Questions

- In the case of competing models, the generalization error is not the only metric in deciding which model to use. We will also consider other user-defined metrics, e.g. execution time. Therefore, we need to develop a methodology to evaluate the fitness of an ML model and to choose the best model in a multi-objective scenario.
- In the context of a predictive query, we want to compare the similarity between the query data space and the model building data space. When working with spatio-temporal (ST) data, two main characteristics are relevant: autocorrelation and heterogeneity. The former is related to the correlation between location and timestamps which results in coherence in spatial and smoothness in temporal observations. The latter requires learning different models for varying spatio-temporal regions [4]. We want to study and implement existing techniques to measure the distance between statistical distributions, in order to adapt these techniques to the more complex case of ST data.
- Given the context of the previous research question, we must assume that the data generated by a phenomenon can change its statistical properties over time. Therefore, we want to know when must a model be updated due to a change in the probability distribution of phenomenon features. What are the metrics and thresholds most suitable for updating a model?

4 Technical Contributions

- Provide a theoretical background to understand the different scenarios where new data might induce qualitative variations of an ML production model;
- Formulate strategies to detect changes in the data and categorize their potential impact on the predictive inferences for given regions of the domain.
- Develop a methodology to mitigate the decay in the qualitative properties of ML production models, by either selecting a different model, updating an existing production model or creating a new model based on updated knowledge.

5 Experimental Proposal

We propose to study the behavior of several rainfall forecasting models over the entire Brazilian region [12], including convolutional neural network and AutoRegressive Integrated Moving Average (ARIMA) models. In order to evaluate the predictive quality we plan to synthetically modify statistical properties of existing data, in this way representing the non-stationarity aspect of the phenomenon.

We plan to formulate predictive queries for the average rainfall in sub-regions of Brazil. These will be used to compare the similarity between the query space and the existing domains, and use the most suitable model. For the case where no model is adequate to predict the desired output, we will explore the possibility of updating the models within the specified geographical region.

References

[1] Michèle Basseville and Igor V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[2] Daniel Crankshaw, Peter Bailis, Joseph E. Gonzalez, Haoyuan Li, Zhao Zhang, Michael J. Franklin, Ali Ghodsi, and Michael I. Jordan. The missing piece in complex analytics: Low latency, scalable model management and serving with velox. In *CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings*, 2015.

[3] Daniel Crankshaw, Xin Wang, Guilio Zhou, Michael J. Franklin, Joseph E. Gonzalez, and Ion Stoica. Clipper: A low-latency online prediction serving system. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*, pages 613–627, Boston, MA, 2017. USENIX Association.

[4] Noel Cressie and Christopher K Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2015.

[5] Sindhu Ghanta, Sriram Subramanian, Lior Khermosh, Swaminathan Sundararaman, Harshil Shah, Yakov Goldberg, Drew S. Roselli, and Nisha Talagala. ML health: Fitness tracking for production models. *CoRR*, abs/1902.02808, 2019.

[6] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lotfi A. Zadeh. *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[7] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The elements of statistical learning: data mining, inference, and prediction, 2nd Edition*. Springer series in statistics. Springer, 2009.

[8] Yunseong Lee, Alberto Scolari, Byung-Gon Chun, Marco Domenico Santambrogio, Markus Weimer, and Matteo Interlandi. Pretzel: Opening the black box of machine learning prediction serving systems. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'18, pages 611–626, Berkeley, CA, USA, 2018. USENIX Association.

[9] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[10] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press, 2012.

[11] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. Data lifecycle challenges in production machine learning: A survey. *SIGMOD Rec.*, 47(2):17–28, dec 2018.

[12] Yania Molina Souto, Fábio Porto, Ana Maria de Carvalho Moura, and Eduardo Bezerra. A spatiotemporal ensemble approach to rainfall forecasting. In *2018 International Joint Conference on Neural Networks, IJCNN 2018, Rio de Janeiro, Brazil, July 8-13, 2018*, pages 1–8, 2018.

[13] Wei Wang, Jinyang Gao, Meihui Zhang, Sheng Wang, Gang Chen, Teck Khim Ng, Beng Chin Ooi, Jie Shao, and Moaz Reyad. Rafiki: Machine learning as an analytics service system. *Proc. VLDB Endow.*, 12(2):128–140, October 2018.