Fairness and Transparency in Trustworthy Cloud-based Analytics Services

Leandro Balby[1] **Flavio Figueiredo**[2]
On behalf of the members of ATMOSPHERE's Work Package 6

[1]Universidade Federal de Campina Grande [2] Universidade Federal de Minas Gerais

1. Introduction

Machine learning (ML) is nowadays ubiquitous, providing mechanisms for supporting decision making in any data rich scenario. This increasing importance of ML raises societal concerns about the trustworthiness of systems that depend on it. The highest accuracy for large datasets is often achieved by complex models that humans struggle to interpret, creating a trade-off between accuracy and interpretability, both of which affect trust in the system. In this context, the new General Data Protection Regulation (GDPR) demands that organizations take the appropriate measures to protect individuals' data, and use them in a fair and transparent fashion.

The ATMOSPHERE project (atmosphere-eubrazil.eu) considers trustworthiness as depending on many properties such as security, dependability, and privacy assurance, among others. Moreover, data become a first class citizen, as trustworthiness also depends greatly on respecting data subject's rights.

In the project, fairness and transparency emerge as key properties for the trustworthiness, with both terms being related. However, given the impact of ML systems on society, their definitions require care and may change according to the context. We view transparency as a means to capture interpretability of ML models. Complementary, fairness is related to both biases which may exist in datasets used to train ML systems, as well as biases which the ML system may incur on society.

2. Transparency

Interpretability may be defined as the degree to which a human can understand the cause of a decision [Miller 2017]. There is still no consensus on how interpretability should be assessed. In ATMOSPHERE, we adopt the view proposed by Doshi-Velez and Kim [Doshi-Velez and Kim 2017]:

- Application level: The explanations are integrated into the outputs of the final model such that domain experts may evaluate it.
- Human level: This corresponds to a simplified version of the application level. The difference is that here domain experts are not required, thus making experiments cheaper and more feasible.
- Function level: Here humans are not required. This is more efficient when the models used are already well understood by humans.

In addition to assessing the interpretation of ML systems, we intend to promote it by equipping any complex model with human intelligible explanations. This is being done through model-agnostic interpretable algorithms such as LIME [Ribeiro et al. 2016] and SHAP [Lundberg and Lee 2017].

3. Fairness

In order to understand fairness, consider for instance a classifier trained with the task of identifying health risks based on subject features.

- Selective Sampling/Labeling: The classifier above may be trained with data that was either sampled or labeled selectively. For instance, crowd-workers may label data based on preconceived relations between gender and health risks. Also, the dataset may be biased towards certain social-demographic variables. On the deployment phase, identifying under represented or over-represented groups of subjects may mitigate such issue. Columns of a table, or features, may also present fairness biases. Some of such features may be even illegal to train ML algorithms on [Daumé III 2018]. Currently, we are using the Aequitas Toolkit[1] to tackle biases. Using the toolkit, we shall evaluate the representation of subjects/labels, helping end-users identify biases and act accordingly.
- Loss Functions: The loss function of ML algorithms may also be subject to biases and fairness issues. In fact, loss functions prone to less biases are gaining the attention of the scientific community. Recently, novel techniques exist, which can deal with representation issues directly in the training phase [Chierichetti et al. 2017, Berk et al. 2017]. Such techniques may be provided to the end-user.
- Feedback loops: Finally, feedback loops occur when user actions guided by ML algorithms decrease trustworthiness. In our example, medical actions guided by an algorithm (e.g., a correlation between race and a certain risk) may increase biases when such data is used to re-train new models. Even though feedback loops rise due to user actions, it is possible to detect their presence [Ensign et al. 2017].

4. Concluding Remarks

Algorithms that measure transparency and fairness are currently being implemented using Lemonade [Santos et al. 2017], which is a visual platform for distributed computing, aimed to enable machine learning applications. The platform will include a set of fairness and transparency metrics that are available to the user.

We plan to evaluate fairness and transparency measurements on real world datasets from the medical domain. In particular, we are currently evaluating methodologies such as LIME, SHAP and Aequitas, already implemented on Lemonade, in both open[2] and closed domain data.

References

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. (2017). A convex framework for fair regression. In *FatML*.

Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. (2017). Fair clustering through fairlets. In *NIPS*.

Daumé III, H. (2018). *A course in machine learning*. Online at `http://ciml.info`.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *CoRR*, abs/1702.08608.

[1] https://dsapp.uchicago.edu/aequitas/

[2] https://mimic.physionet.org/

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. (2017). Runaway feedback loops in predictive policing. In *FatML*.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *NIPS*.

Miller, T. (2017). Explanation in artificial intelligence: Insights from the social sciences. *CoRR*, abs/1706.07269.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). ”why should i trust you?”: Explaining the predictions of any classifier. In *KDD*.

Santos, W., Carvalho, L. F. M., de P. Avelar, G., Silva, Jr., A., Ponce, L. M., Guedes, D., and Meira, Jr., W. (2017). Lemonade: A scalable and efficient spark-based platform for data analytics. In *CCGrid*.