CS841: Computational Cognitive Science

Fall 2019 **PROJECT PROPOSAL** Arun Jose

Background

Recent studies (Paxton & Griffiths, 2017) in the Cognitive Science literature have shown that using big and naturally occurring datasets to supplement traditional experimental paradigms is a powerful tool to understand human behavior and cognition. One recent application in this regard was an investigation(Hopman et al, 2018) to predict human word learning accuracy of a second language (L2) from a Duolingo big data set. The paper models ten word level parameters as predictors of L2 word learning accuracy and measures their importance using linear/logistic regression.

In this project, I intend to explore how varying the model used impacts the importance of the predictors of L2 word learning accuracy. Linear/Logistic regression analyses work well in the case of linear decision boundaries, but fail to capture any non-linear relations in the data. Would the results be statistically different on using a non-linear classifier like neural nets or decision trees? In particular, I want to experiment with the state-of-the-art non-linear machine learning models like boosted decision trees (Chen, T. & Guestrin, C., 2016) and measure the importance of the predictors on the same dataset to see how different the results are. In case there is a significant statistical difference between the results obtained and the results of the Hopman et al paper, it could be a clear indication of non-linear patterns in the data, mandating all future experiments to account for this while attempting to model word learning accuracies. Moreover, in a more general sense, I would like to expand this to a few more models and see which predictors are invariant to model changes and which ones vary significantly. This could provide a stronger measure of feature importance.

As additional work (time permitting), I would like to try replicating the same study on a more expansive dataset from Duolingo (Settles & Burr, 2018) and see how the results compare against the older dataset (Settles & Meeder, 2016).

Question

What predictors remain invariant to model changes in L2 word learning accuracy? Is there a non-linear relationship in the data that could prohibit modeling with linear predictors like LR?

Method

The first step in my methodology would be to work on getting the prediction model set up as per the Hopman et al paper. There have been a few modifications to the predictors that the authors have been working on since the paper, and I am planning to set up my system to run the model with these changes on the Settles & Meeder, 2016 dataset. Once this is done, I shall select a state-of-the-art ML model which can capture non-linearities in data and proceed to implement and integrate it with the above set-up. I am planning to choose XG-Boost (Chen, T. & Guestrin, C., 2016) for the same, due to its recent popularity in winning a lot of Kaggle competitions. After this is done, I would capture the importance of each parameter used in the prediction. This can be computed by observing the amount that each attribute split point improves the performance measure (eg. Gini index), weighted by the number of observations the node is responsible for. Using this information, I shall compare the predictor importance values obtained via Logistic Regression and look for similarities/dissimilarities. Finally, I would expand this to a few more models to test for model-invariant predictors. After evaluating the feasibility in doing so, I would also like to run a replication study of the same experiment on the new dataset (Settles & Burr, 2018) and compare how the importance of the various predictors have changed.

References

[1] Hopman, E. W. M., Thompson, B., Austerweil, J. L., & Lupyan, G. (2018). Predictors of L2 word learning accuracy: A big data investigation.

[2] Paxton, A. & Griffiths, T. L. (2017). Finding the traces of behavioral and cognitive processes in big data and naturally occurring datasets. Behavior Research Methods.

[3] Settles, B., & Meeder, B. (2016). A trainable spaced repetition model for language learning. In Proceedings of the 54th Annual Meeting of the ACL.

[4] Settles, Burr (2018). Data for the 2018 Duolingo Shared Task on Second Language Acquisition Modeling

[5] Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System