Dawei Huang
06/06/2020
Ethics Project Proposal

Project Proposal: Interpretability of Machine Learning

Dear Professor Blase,
Good evening, my name is Dawei Huang. I have already started on my project but I realized that I have given you only a very rough outline of my topic in our discussion and I hope to provide you with more information. I am finished with work from all my classes so I am willing to make fundamental changes to my project or even find a new topic if necessary.

Black Box Method:
Input -> Black Box -> Output
ex. (wind speed, humidity, precipitation) -> black box algorithm -> temperature

For my project, I am interested in developing a new model-agnostic method of interpretability (black box method) based on randomly changing data belonging to a specific variable in the input and then measuring the degree by which the original output will differ. If the variable is heavily weighted, for example, I can expect the original output to deviate significantly from the new output that I will generate by feeding the new input into the black box algorithm. Likewise, variables that are less significant (less weight) should not change the output as much. (Note that I will limit the random range to the min max range of data for that variable after removing possible outliers).

Afterwards I can create a line graph correlating the # of random changes with the amount of deviation from the original output. If my black box algorithm has n variables then I can expect to see n lines for this graph (each representing a variable). Thus the steeper the decline slope the more significant than that variable would have more weight and vice versa.

After implementing this model, I hope to implement this model-agnostic method on a dataset and compare it to other conventional model-agnostic methods (Sharpley values, partial dependence plot, etc) in order to determine its feasibility.

Thank you very much for reading this private post, please let me know if you have any thoughts on what I have been doing for my interpretability in ML topic so far. I am open to implementing any advice you can give me no matter how big or small the required changes will be. This week has been especially busy for me and I apologize for providing you with such a late update. Thank you very much.

Sincerely,
Dawei Huang