Prof. Petkovic (SFSU) Feedback on questions from ZOOM chat related to on-line class on "AI Explainability" at Frankfurt University, 05-13-20

Prof. Petkovic Feedback is below in ***italics bold,*** prefaced by ***DP,*** after each question.

Petkovic@sfsu.edu

05-14-20

14:26:39 From Ilenna Jones : Q2: Would you adopt a AI system that is blackbox or explainable?

I would compare the explainable alg to human performance to make this decision
DP Good, but in my opinion comparison with human performance is just one component of decision making. Generally, people compare "best possible AI result" (often a black box), with human, and if comparable they assume AI system is a good candidate for adoption. Explainability then comes a bit independently and may vary from area to area: in some low risk areas like misrelating it may not be so important. In some like health or policing or loan approval, where risks and legal liability are hi, explainability is more important. Finally, and important: explainability is often used to QA and audit black box approach, as shown in our case studies. Black box AI can produce seemingly great result but all due to the wrong reasons, only discoverable only by explainability. OR it can increase our trust by showing that most important factors in decision makeing are intuitive..

14:26:56 From iPad Pro NSt : Black Box could be acceptable if I can prove the reliability.
DP Agree (I assume reliability is related to our trust and related to explainability providing positive results)

14:29:31 From Fotis Fitsilis : ... and yes, (the degree of) explainability belongs to the important parameters to be defined
DP Thanks and yes, it is one parameter that has to be checked, whether black box shows good or bad results, kind of an audit necessary to provide user trust

14:42:30 From Ilenna Jones : how would LIME choose the sample words that it would perturb? How does it know what words are best for explanability?
DP It randomly perturbs them from a larger set, then looks for combi9oanton that is best in its local linear classification. Note that they only approximate unknown black box AI using local linear c classifier with only simple explainable features (e.g. words not any complex relationship among words)

14:50:41 From Ilenna Jones : in LRP. what is relevance? Do we have to label what is relevant first for this kind of analysis?

DP I assume you talk about layer-wise relevance propagation? Relevance is derived from ultimate correct outcome or decision at the end/last layer, and propagated back. Explained here
W. Samek et al:"Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models", ITU Journal ICT Discoveries, Special issues No 1, 13 Oct 2017

15:08:17 From Ilenna Jones : In the papers on the data provided for the case study, were the mRNA levels relative levels or absolute levels? It could be that a gene has level 100 active and 10 suppressed, and another gene is 10 active and 1 suppressed. If the threshold for active and suppressed is based on absolute levels, then this could lead to an inaccurate labelling of gene activity
DP I am not biologist but as far as I know they are absolute, details are here
Aevermann B., Novotny M., Bakken T., Miller J., Diehl A., Osumi-Sutherland D., Lasken R., Lein E., Scheuermann R.: "Cell type discovery using single cell transcriptomics: implications for ontological representation", Human Molecular Genetics 27(R1): R40-R47 · March 2018

15:24:11 From Ilenna Jones : Could we collect the same information output in the RFEX table for explaining Neural networks too?
DP It would be hard if you want to relate it to pixels. RFEX would work if one could base it on higher level features which can be arranged in a table and treated relatively independently….Another necessary ingredient for RFEX is some form of ranking of those features

15:29:22 From Ilenna Jones : The 60 samples used for the neighborhood of k-nearest neighbor - is 60 a hyperparameter that needs to be adjusted for each task?
DP As in all KNN approaches, one has to decide on K. We based K on the following: good number of samples to show some meaningful statistics but also to ensure possibly good mix of + and – samples. That is why we set it up to 20% of the number + or smaller class samples, kind of a compromise, allowing enough of samples, especially wrt. the smaller class

15:33:57 From Ilenna Jones : On the topic of outliers like the one on the problematic sample slide, is there a standard way of determining the threshold for excluding or flagging outlier samples? This could be very useful for either finding possible human error, but also be used as another way to check the quality of the trained model outside of just using accuracy (kind of like how a high standard error of accuracy indicates low quality performance of a model)
DP Thresholds are tricky. Always hard to have absolute ones woring for all applications, plus it is data dependent. Our approach is to have conservative threshold first, say 10% of lowest votes, then investigate and revise of necessary. Human judgment is important here but RFEX helps at least filter down the candidates.

15:35:46 From Ilenna Jones : Explainability tools like RFEX could establish explainability standards for ML algorithms. I definitely see a broad application of collecting these statistics for many ML algorithms beyond random forest
DP Yes, as long as features are kind of separate not like pixels and signals, and as long as you have some feature ranking mechanism

15:37:31 From Ilenna Jones : It could be that explainability standards could help answer the question of responsibility - the more explainable something is, the more responsibility is on the user. The less explainable it is, the more responsibility is on the company/developer
DP Maybe…complicated. Auditors can establish explainability but then legal responsibility may vary? Explainabilty seems like necessary but not a sufficient condition for ethical AI….

15:39:57 From Ilenna Jones : That's an interesting ppoint - how would companies make money with explainable AI? If we have policies put in place to protect data, and AI is transparent and explainable, what niche is there for companies to make money? This might help predict the direction of development in our profit-driven system
DP I agree. I was a bit shocked when I read about this issue it and told myself – this is a problem…

Thank you all., Pleas etry RFEX or similar. If any questions read our paper below

- ***https://www.biorxiv.org/content/10.1101/819078v1***

or contact Petkovic@sfsu.edu