Information Visualisation for Explainable AI

Aidan Slingsby, a.slingsby@city.ac.uk
giCentre, Computer Science
City, University of London

My research is on the design and application of interactive visualisation to help understand complex data and the phenomena that they represent. One application of this that interests me is to help understand **why models give the answers they do**. Recent advances in machine learning are producing impressive results, but may be contributing to making people more accepting of predictions from black box models. Visualisation may have an important role in helping humans judge the output so that they do not accept them blindly. I am interested in the role of visualisation in the field of AI, a field in which it is largely absent, where the use of black-box AI in applications appears to be encouraged. An important characteristic of visualisation is its ability to depict richer information that high-level summaries. I think its ability to convey distributions, uncertainties and variation, make it a suitable technique to help make AI explainable.

One way to address this is to accompany the model output with details of the characteristics that were important for leading to the outputs and perhaps a summary of the characteristics of the historical cases that accompany the output. For example, models that make decisions about whether a loan application should be accepted could be accompanied with the characteristics of the application that were most important in the decision and a summary of the characteristic of the cases from the training data. This would help judge, for example, whether the training cases are adequate for the current case. There are also techniques in which the sensitivity of the output to the inputs. Visualisation can help us compare model variation within and between different parameter ranges, but the degree to which is this possible depends on the nature of the models used.

One piece of work I did (Slingsby et al, 2011) was to design and apply interactive visualisation to "unpick" simple black box models that are taken at face-value. This model was a simple classification statistical model. I used interactive visualisation used to depict the fuzzy classification of this, how it varied amongst the categories, and the geographical distribution of these. I used this to try and better inform the model output use in applications. This was a relatively simple approach, where I was relating model inputs and outputs, quantifying uncertainty and allowing this to be explored by category and geographically using interactive visualisation. This has similarities to the simplest form of explainable AI (Doran et al, 2017): "opaque systems that offer no insight into its algorithmic mechanisms", but also considering the process by which the model gets there to different degrees, gets us to some of the more sophisticated types of explainable AI described by Doran et al (2017).

Where models explicitly model the *process* of the phenomenon under considerations, this can be reasonably straightforward. But most statistical models simply map inputs and outputs, in order that outputs can be predicted for any given set of inputs. Many AI model also do not model the process, but establish relationships of inputs and outputs. Some complex deep neural networks establish internal states that are used internally, but unlikely to relate to concepts that humans understand. But when combined with symbolic representations, they may be proxies for human-understandable processes. This is often observed in computer-vision applications.

My main question is "in which ways can visualisation help explainability of AI?" I'm sure it can help for the simple cases of relating inputs to outputs and showing difference between different ranges of inputs, but I am also interested it its potential for helping understand higher-level relationships which might correspond to internal states of deep learning networks as proxies for human-understandable concepts.

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? *A new conceptualization of perspectives. arXiv preprint arXiv:1710.00794.*

Slingsby, A., Dykes, J. and Wood, J. (2011). Exploring Uncertainty in Geodemographics with Interactive Graphics. *IEEE Transactions on Visualization and Computer Graphics, 17(12), pp. 2545-2554. doi: 10.1109/TVCG.2011.197*