During today’s lecture Nathanael briefly mentioned what I think would be a brilliant way to explain the issues with “illuminating the black box” of a complex machine-learning algorithm. He touched on how a baby learns what a chair is, and while I think it was a good example for a learning situation I think the situation should be created where the simple set of rules for classifying is less clear. My example: what are the rules for classifying a sandwich? The class would probably begin saying something simple like “something you can eat covered on both sides by bread”, then the lecturer could counter with “is a burrito a sandwich then?” and this conversation would turn into “well no because that’s only one piece of bread” – following that up with “well is a sub not a sandwich then?” and this back and forth would continue until the class is frustrated. I think this sandwich example demonstrates that we all have an idea of what a sandwich is, based on our upbringing, but it becomes very hard to articulate the rules for your decision making when asked to make it on the spot. The same can be said with a complex machine-learning algorithm (say an SVM with a non-linear kernel): the classifier will make a prediction but determining why (in human-speak) it made the prediction becomes more challenging. This exercise would help derive a parallel between how humans and machines can actually behave similarly sometimes. With more simple problems both the human and the machine can derive a set of rules to follow for classifying. Say you’re trying to classify if this upcoming weekend class has started: the rule could be “is it between 9AM and 4PM on Saturday October 13th?”. Both the human and the machine are capable of coming to this criterion for a conclusion. In the sandwich example however – this simple decision tree is less clear.

Thoughts? Is the way we think more similar to machines than we would like to believe?

Thanks,

-Oliver