Questions for CSC2508 Lecture 8 (11/11/2019)

Sean Singh, 996113959

“Why Should I Trust You?” Explaining the Predictions of Any Classifier

1. In section 2, it seems that the emphasis placed on explanations and human made decisions is at odds with the idea of the autonomous systems? Do you feel that there are certain domains where interpretability is not important?
2. In section 5.3, what is the reason for using 25% of the features?
3. In section 6.3, wouldn’t using explanations to engineer features result in a self-fulfilling prophecy of sorts?

On the Robustness of Interpretability Methods

1. In section 1, what is your opinion on the statement that “understanding a complex model with a single point-wise explanation is perhaps too optimistic”?
2. In section 2, how should one interpret the word *reasonable* at the bottom of page 67?
3. Would you agree that there is a trade-off between complexity (and likely performance as well) and interpretability?

Interpreting Deep Learning Models for Entity Resolution: An Experience Report Using LIME

1. In section 1, what are the limitations of using the F-measure described?
2. In section 2, how well would the methods of this paper extend to multiclass classification as opposed to binary classification?
3. In section 5, when they mention using explanation to produce suggestions, what does suggestions mean in this context?