Zhou Fang

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

Q1. In section 3.3, given a perturbed sample z', how can we recover the sample in the original representation z?

Q2. In section 3.4, it is stated that "we can estimate the faithfulness of the explanation of Z". What is the definition of faithfulness and how can we estimate it if the underlying model is highly non-linear in the locality of the prediction?

Q3. In submodular pick, what is the intuitive explanation for instances that represent completely different features? Does that imply the nonlinearity in the underlying model?

On the Robustness of Interpretability Methods

Q1. Does local stability imply global robustness?

Q2. In section 2, why is the continuous notion of local stability not suitable for models with discrete inputs or those where adversarial perturbations are overly restrictive?

Q3. In section 4, what is the possible explanation for the experiment result that model-agnostic perturbation-based methods are more prone to instability than their gradient-based counterparts?

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

Q1. What will happen to the weights learned by the surrogate when labels C(Ti) do not span uniformly c and other classes?

Q2. Does LIME_COPY provide more advantages compared with LIME_DROP? Or LIME_COPY just complements LIME_DROP?

Q3. In section 3, what is the benefit of processing the input word embedding sequence by both forward RNN and backwards RNN?