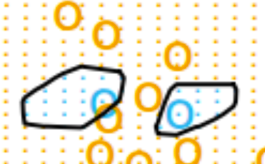
“Why Should I Trust You?” Explaining the Predictions of Any Classifier

1. For an instance x (blue circle), if the decision boundary around it is highly non-linear, how to choose the interpretable model? Do you think the interpretable model can still be local fidelity?
2. Why the submodular pick algorithm can ensure the diversity of picked instance?
3. How to determine the number of instances we are going to pick when applying SP-LIME?

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

1. When we build the *representative text sequence* Tu,v in Mojito, a prefix is added to each token. So, it actually modifies the original attribute values. How does this additional prefix affect the result of LIME?
2. In Figure 3, LIME\_COPY and LIME\_DROP are performed on both hybrid and RNN ER models. Base on the results, which ER model do you think is more trustable?
3. In Figure 2, knowing that the time attribute contributes too much during ER, which is unreasonable. How do you fine-tune the models or training data such that make it more reasonable?

On the Robustness of Interpretability Methods

1. If the pattern of a problem is highly non-linear, which means similarly inputs should map to very different output. Is it reasonable to apply such robustness analysis on the problem?
2. Suppose x0 is the vicinity of x, the small |f(x) – f(x0)| represents the robustness. However, why we maximize L hat in equation (1)?
3. For a machine learning model, is it possible that it’s robust around some decision boundary but not robust around other? In such case, how to measure the overall robustness of the model?