The Intriguing Properties of Model Explanations

Maruan Al-Shedivat       Avinava Dubey       Eric P. Xing

Interpretable ML Symposium
NIPS 2017
**What is Explanation?**

**Explanation** is a *simple model* that approximates the decision boundary of a *complex model*.

---

**Questions:**

1. How do feature selection and feature noise affect explanations?
2. When explanation is a part of the learning and prediction process, how does that affect performance of the model?
3. What insights can we gain by visualizing and inspecting explanations?

---

**Post-hoc (LIME)**
1) Fit the model
2) Fit an explanation

**Joint (CEN)**
Learn a model that generates explanations
Are explanations consistent?

Explanations are as good as the features they use to explain predictions. What is the effect of feature noise on the generated explanations?

Post-hoc explanations may overfit!

Corrupt interpretable features with noise

(more details on the poster)
How using explanations affects performance?

**Explanations** may affect predictive performance of the model. But why and how?

**Turns out:**

- **Abundant data regime**
  CENs perform as well as their vanilla deep network counterparts.

- **Scarce data regime**
  “Learning to explain” regularizes the model and improves performance.

(more details on the poster)
Visualising and inspecting explanations

For more, please come and see our poster!