Local Interpretable Model-agnostic Explanations

Link to API Reference: LimeTabular

See the backing repository for LIME here.

Summary

Local interpretable model-agnostic explanations (LIME)[1] is a method that fits a surrogate glassbox model around the decision space of any blackbox model's prediction. LIME explicitly tries to model the local neighborhood of any prediction – by focusing on a narrow enough decision surface, even simple linear models can provide good approximations of blackbox model behavior. Users can then inspect the glassbox model to understand how the blackbox model behaves in that region.

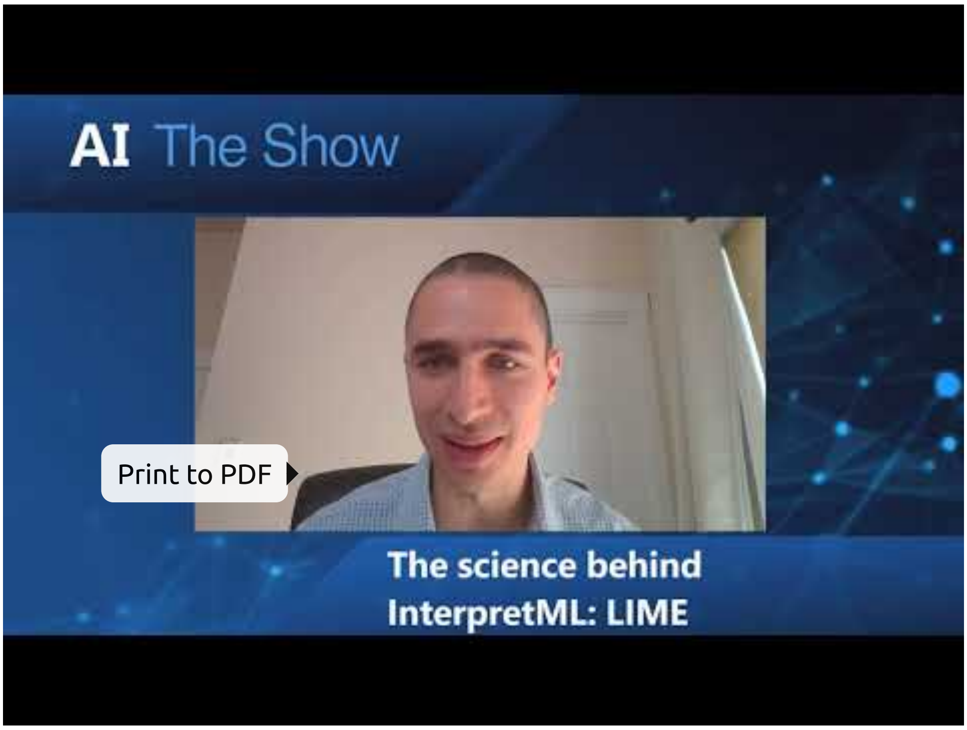
LIME works by perturbing any individual datapoint and generating synthetic data which gets evaluated by the blackbox system, and ultimately used as a training set for the glassbox model. LIME's advantages are that you can interpret an explanation the same way you reason about a linear model, and that it can be used on almost any model. On the otherhand, explanations are occasionally unstable and highly dependent on the perturbation process.

How it Works

Christoph Molnar's "Interpretable Machine Learning" e-book [2] has an excellent overview on LIME that can be found here.

The conceiving paper "Why should i trust you?" Explaining the predictions of any classifier." [1] can be found on arXiv here.

If you find video as a better medium for learning the algorithm, you can find a conceptual overview of the algorithm by the author Marco Tulio Ribeiro below:



Code Example

The following code will train a blackbox pipeline for the breast cancer dataset. Aftewards it will interpret the pipeline and its decisions with LIME. The visualizations provided will be for local explanations.

```
from interpret import set_visualize_provider
from interpret.provider import InlineProvider
set_visualize_provider(InlineProvider())
```

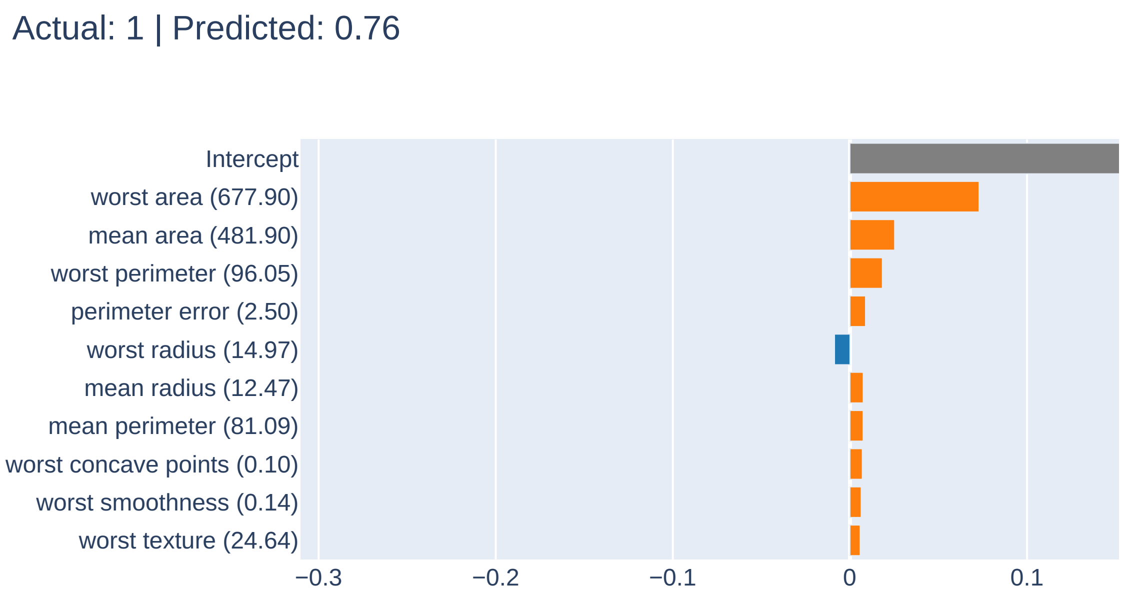
```
import numpy as np
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split

from sklearn.ensemble import RandomForestClassifier
from sklearn.decomposition import PCA
from sklearn.pipeline import Pipeline

from interpret import show
from interpret.blackbox import LimeTabular

seed = 42
np.random.seed(seed)
X, y = load_breast_cancer(return_X_y=True, as_frame=True)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20,

pca = PCA()
rf = RandomForestClassifier(random_state=seed)

blackbox_model = Pipeline([('pca', pca), ('rf', rf)])
```

```
blackbox_model.fit(X_train, y_train)

lime = LimeTabular(blackbox_model, X_train)

show(lime.explain_local(X_test[:5], y_test[:5]), 0)
```

Select Component to Graph

0 : Actual (1) | Predicted (0.76) | Resid (0.24)

LimeTabular_0



Further Resources

- LIME GitHub Repository
- LIME Video for KDD2016

- Conceptual Video for LIME

Bibliography

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ” why should i trust you?” explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 1135–1144. 2016.

[2] Christoph Molnar. Interpretable machine learning. Lulu. com, 2020.