Explainable Machine Learning

Yu Huang

Principal Software Engineer

Argonne National Laboratory, APS

Explainable Machine Learning in healthcare

Content

- What is explainable machine learning?
- LIME algorithm: explaining individual predictions
- Example: How to use LIME to explain individual predictions?
- Reason Code: Model prediction explanations

What is explainable machine learning (1-1)

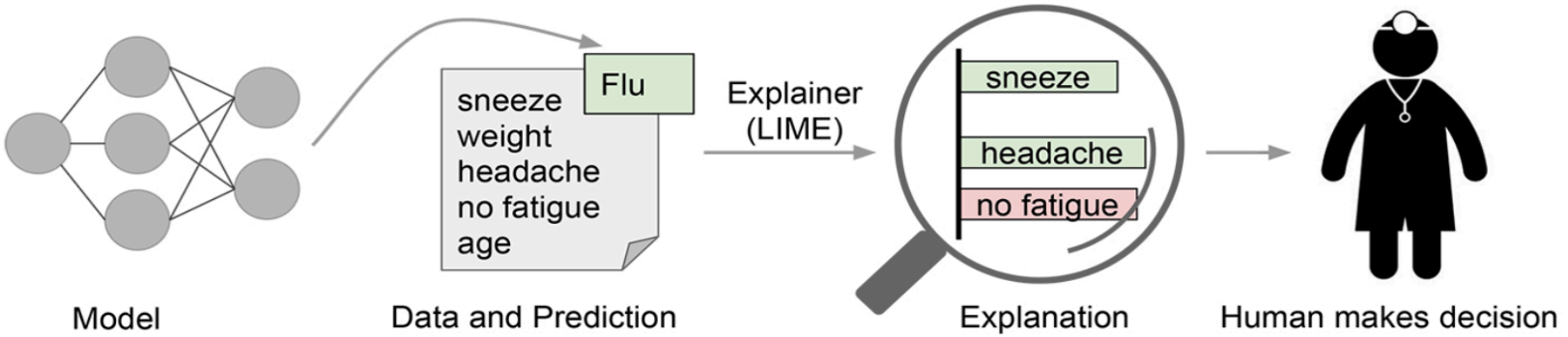
To many people, advanced machine learning model is a black box. Understanding reasons behind the prediction of a model can be very important if one (e.g., doctor) takes actions according to prediction results

- **Explainable machine learning** is a methodology to help people understand the rationale of the model prediction by generating interpretable artifacts such as **reason codes**, charts, etc.

This presentation will introduce one of explainable machine learning methods/algorithms/techniques: Local Interpretable Model-Agnostic Explanations (**LIME)**

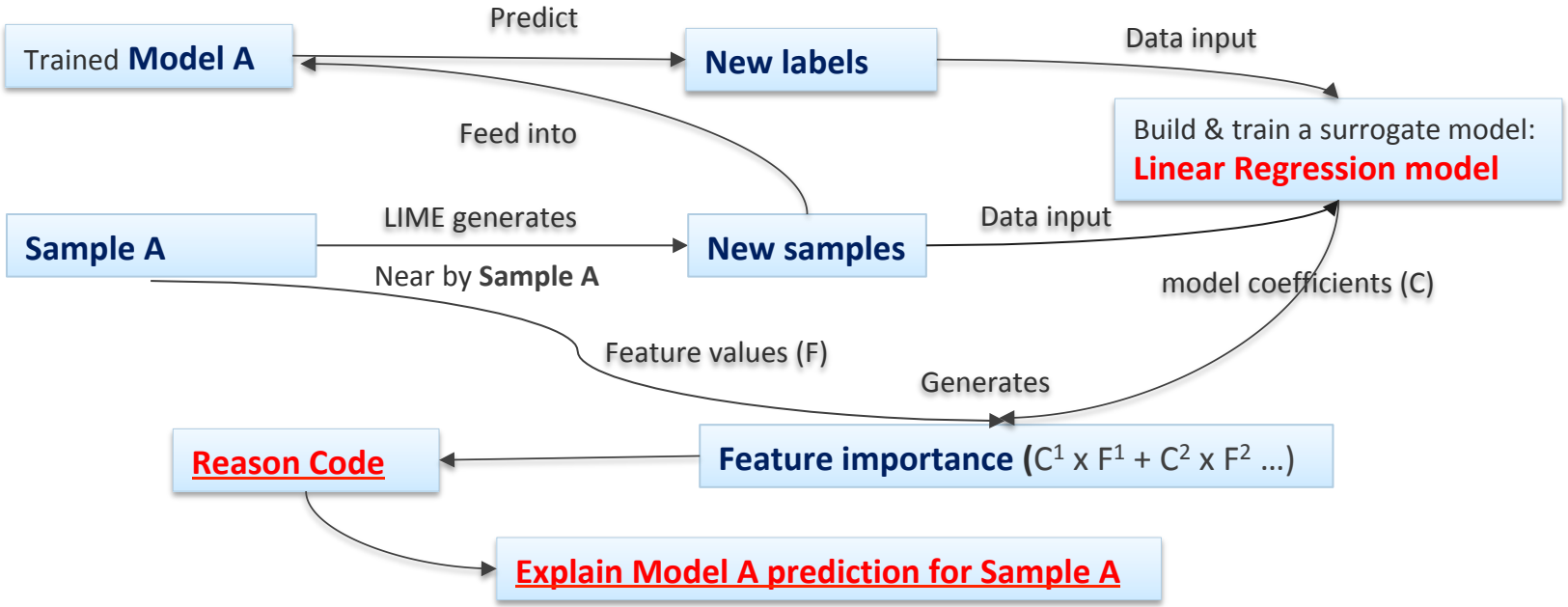
What is explainable machine learning (1-2)

❖ Explaining individual predictions[1,3]



- The model predicts that a patient has a flu based on the input features (the patient's symptoms).
- LIME highlights the patient's symptoms that support or against the prediction.
- "Sneeze" and "headache" in green indicate contributing to the "flu", and "no fatigue" in red indicates evidence against it.
- With this information, a doctor can make a decision if the model's prediction is trustable based on his/her domain knowledge. A trustable model can assist a doctor for diagnosis when a patient has complicated conditions.

LIME algorithm - explaining individual predictions by a surrogate model. How it works (2-1)



1. **Model A**: is a trained model to be explained
2. **Sample A**: is a data sample . Its prediction is to be explained
3. **New samples**: LIME randomly generates a list of new data samples nearby data **Sample A** (by np.normal (mean, std.)).
4. **New labels**: Use **Model A** (1) to predict new labels for the **New samples**(3)
5. **Linear Regression model**: trained by the **New** (data) **samples** and **New labels** (3,4) which is a surrogate of the Model A.
6. **Feature importance:** equal to the **coefficients** (of the Linear Regression model) **X** the **feature values** of **Sample A** (2)
7. **Reason Code**: use the learned **features importance** (6) to explain **model A** (1) prediction for **Sample A** (2)

LIME algorithm - explaining individual predictions by a surrogate model. How it works (2-2)

- Build and train a machine learning model (*Model A*) to be explained
- Have LIME randomly generate a list of new data samples nearby a given data *Sample A* to be predicted (np.normal (mean, std.))
- Use *Model A* to predict new labels for the new samples
- Use the new data samples and new labels to build and train a Linear Regression model (a surrogate model)
- Use the learned linear model coefficients (Beta) to calculate features importance ($Beta^1$ x $feature^1$ + $Beta^2$ x $feature^2$…)
- Use the learned features importance (reason codes) to explain *Model A* prediction result of *Sample A*

How to use LIME to explain individual prediction (3-1) Outline

Prepare dataset → **Train a model** to be explained

Use LIME to generate reason codes

Creating Pipeline:
An object that combines a trained model with testing data preprocessing steps.

Selecting LIME Explainer:
There are different explainers (e.g., text, image, tabular, etc.) . In this case, tabular explainer is selected since the data in our example is in the tabular format.

Explaining Model Prediction:
(1). The explain_instance() function of the Explainer needs a trained model with a function to predict probability of the labels.

(2). The function also requires three parameters: raw feature vector, trained model with probability function, number of features.

How to use LIME to explain individual prediction (3-2)

Example

For implementation details, please check out my article and GitHub links in the reference section[5, 6].

- **Prepare Dataset:** Wisconsin Breast Cancer Diagnosis dataset (WBCD) [5] is used for this presentation. The following 9 features were used in model training.

Number	Features	Abbreviations
1	Clump Thickness	CT
2	Uniformity of Cell Size	UCShape
3	Uniformity of Cell Shape	UCShape
4	Marginal Adhesion	MA
5	Single Epithelial Cell Size	SECSize
6	Bare Nuclei	BN
7	Bland Chromatin	BC
8	Normal Nucleoli	NN
9	Mitoses	

How to use LIME to explain individual prediction (3-3) Example

- Train a model to be explained: Random Forest is used in this presentation.
- Use LIME to generate reason codes:

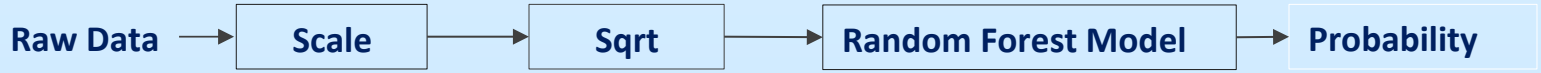
1. **Creating Pipeline**

```
from sklearn.pipeline import make_pipeline
from lime.lime_tabular import LimeTabularExplainer

scale = Scale()
sqrt  = Sqrt()

machine_learning_pipeline = make_pipeline(scale, sqrt, rfc) # rfc is trained model
```

Pipeline: combine a trained model with testing data preprocessing steps to be used in LIME explainer



How to use LIME to explain individual prediction (3-4)

Example

2. Selecting LIME Explainer:

There are different explainers (e.g., text, image, tabular, etc.) . In this case, tabular is selected.

```
class_names = ['Benign', 'Malignant']
explainer = LimeTabularExplainer(feature_names=feature_names,
                                 class_names=class_names,
                                 training_data=X_train_values)
```

3. Explaining Model Prediction:

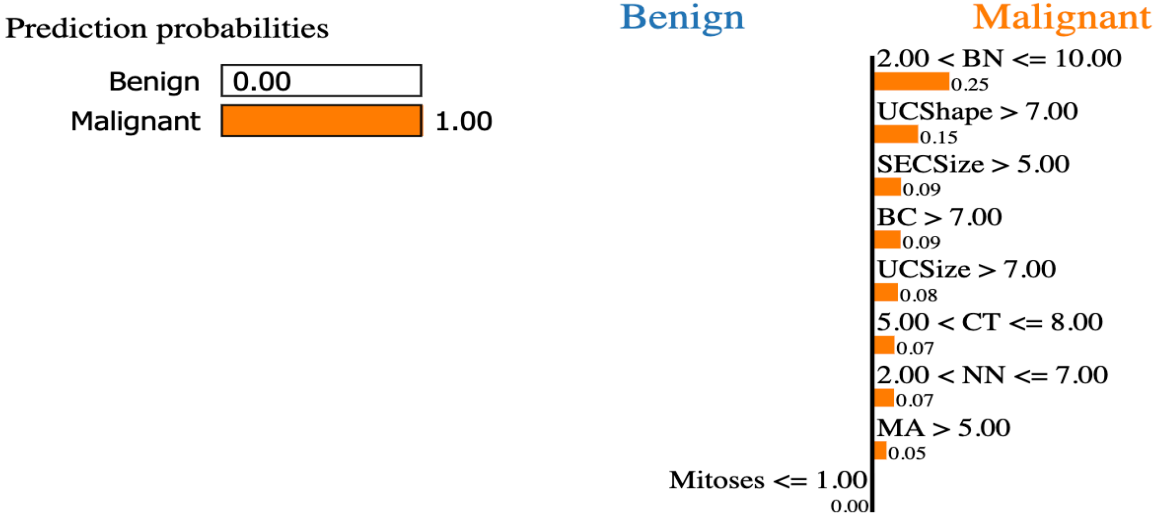
Explainer instance needs a trained model with function to predict probability of labels.

```
def explain(feature_vector, machine_learning_pipeline, label=1):
    exp = explainer.explain_instance(feature_vector,
                                     machine_learning_pipeline.predict_proba,
                                     num_features=9)
    fig = plot(exp, label) # explaining prediction as pyplot figure
    exp.show_in_notebook(show_table=True, show_all=false)
```

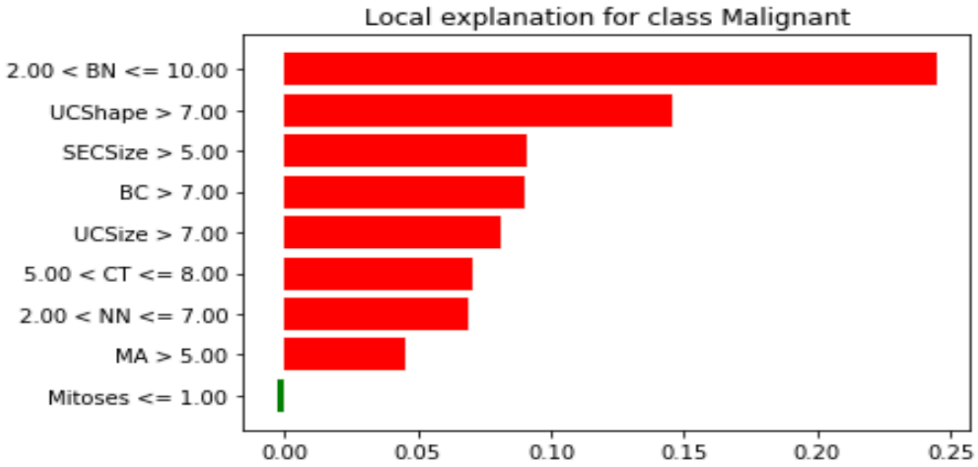
```
explain(sample_M, machine_learning_pipeline, label=1)
# 1: explaining Malignant, sample_M is its feature vector.
# 0: explaining Benign, sample_B is its feature vector.
```

Reason Code: Model prediction explanations (4-1)

Sample 1: Prediction of Malignant

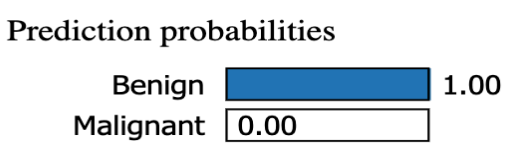


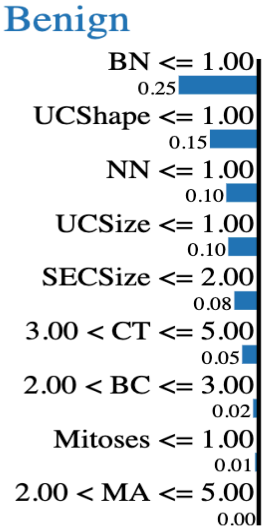
Feature	Value
BN	10.00
UCShape	10.00
SECSize	7.00
BC	9.00
UCSize	10.00
CT	8.00
NN	7.00
MA	8.00
Mitoses	1.00



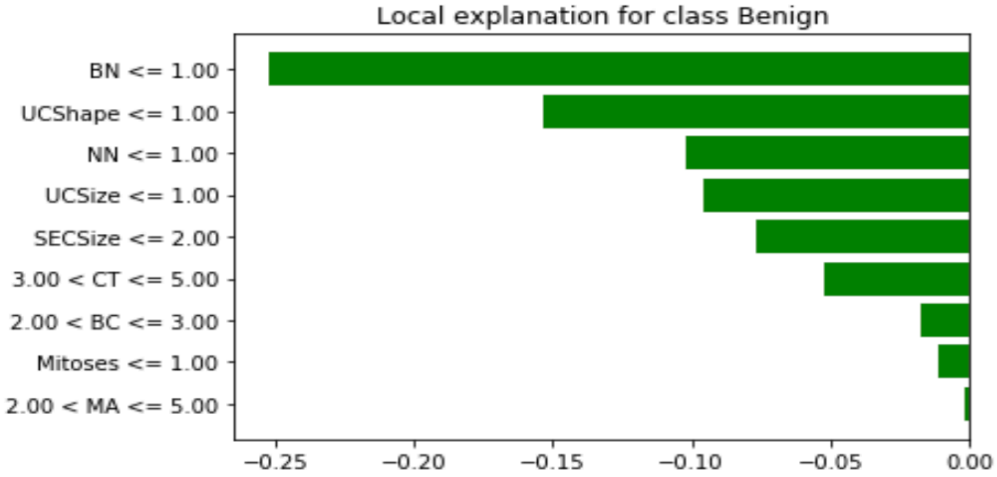
Reason Code: Model prediction explanations (4-2)

Sample 2: Prediction of Benign



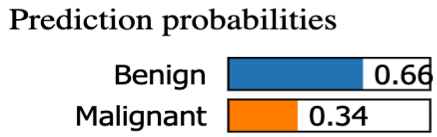


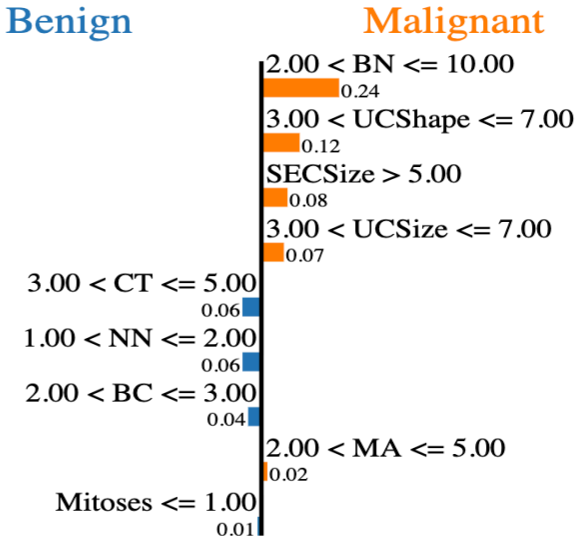
Feature	Value
BN	1.00
UCShape	1.00
NN	1.00
UCSize	1.00
SECSize	2.00
CT	4.00
BC	3.00
Mitoses	1.00
MA	3.00



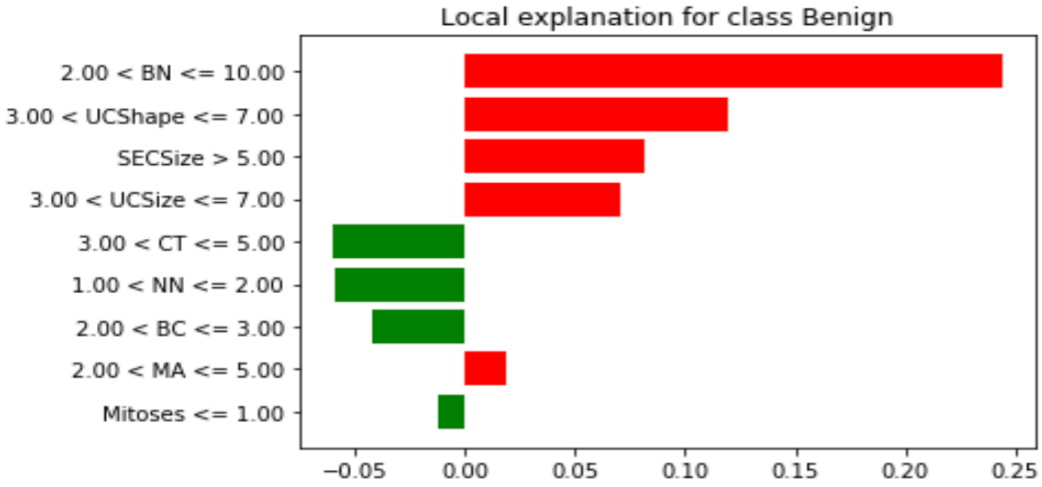
Reason Code: Model prediction explanations (4-3)

Sample 3: Prediction of Benign





Feature	Value
BN	10.00
UCShape	4.00
SECSize	7.00
UCSize	4.00
CT	5.00
NN	2.00
BC	3.00
MA	5.00
Mitoses	1.00



Summary

- In this presentation I discussed LIME algorithm/method [1,3] and used it to demonstrate how to explain the individual prediction results in a machine learning model in breast cancer diagnosis.

- Understanding the reasons behind the machine learning model predictions is very important in assessing trust, e.g. if a doctor plans to take an action to treat the cancer based on a prediction of diagnosis. Such understanding can also help doctors with domain expertise to detect errors with the model predictions so that the model in use can be further improved.

References

1. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?" Explaining the Predictions of Any Classifier: https://arxiv.org/pdf/1602.04938.pdf
2. C. Molnar, Interpretable Machine Learning: https://christophm.github.io/interpretable-ml-book/
3. M. T. Ribeiro, S. Singh, and C. Guestrin, Local Interpretable Model-Agnostic Explanations (LIME): An Introduction: https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime
4. P. Hall and N. Gill, An Introduction to Machine Learning Interpretability, An Applied Perspective on Fairness, Accountability, Transparency, and Explainable AI, Second Edition, O'Reilly Media, Inc., 2019
5. WBCD dataset and description: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/
6. Y. Huang, Explainable Machine Learning for Healthcare: https://towardsdatascience.com/explainable-machine-learning-for-healthcare-7e408f8e5130
7. Y. Huang, Github: https://github.com/yuhuang3/machine-learning/tree/master/lime

Thank you!

APS AI/ML workshop on 1/21/2020