



# LIME

Local Interpretable Model-Agnostic Explanations

統計 111 鄭佳鈴  
統計 110 黃思媛  
統計 109 謝宜均

# Topic

Introduction

01

---

02

Different data types

01



# Introduction

---

# Example by a video

---

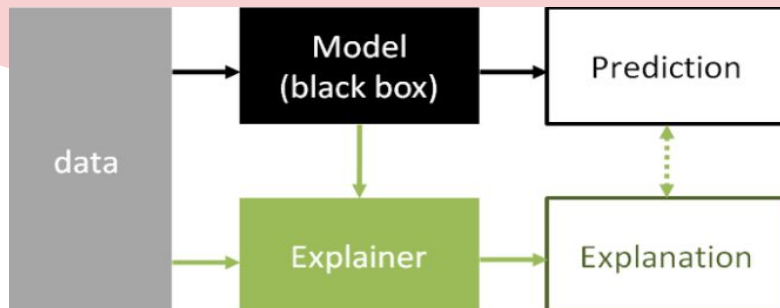
Sometimes you don't know if you can trust a machine learning prediction...



# Background

現今的數據挖掘領域中，機器學習模型被廣泛使用。然而，機器學習模型的「黑盒」屬性導致了其內部工作原理難以被理解，輸入與輸出之間往往存在極其複雜的函數關係。

因此，應用複雜的機器學習模型時，我們需要構造一個「解釋器」，對模型的預測結果進行事後歸因解析，而LIME便是一個很好的事後解釋法。



# What is LIME ?

---

**L**ocal

LIME基於想要解釋的觀測值及其附近的樣本，建構**局部**的線性/其他模型。

**I**nterpretable

LIME作出的解釋易於理解。

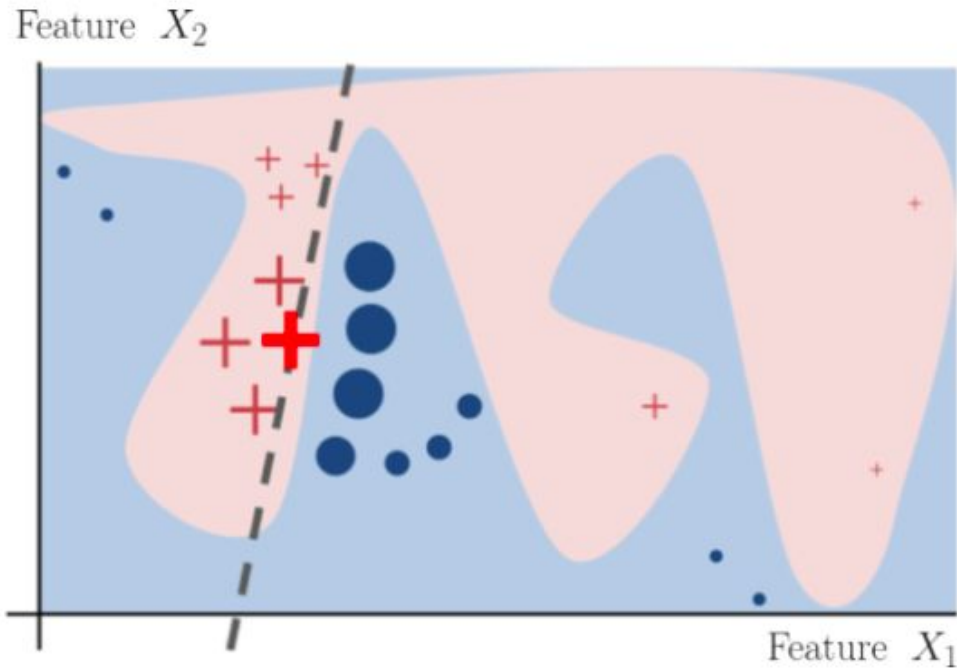
**m**odel - Agnostic

LIME解釋方法與模型無關。

**E**xplanation

LIME是一種事後的解釋方法。

# The Core Value of LIME



LIME 的名字完整說明了對於這個問題的核心精神

1. 在每一個觀察值附近 (Local)
2. 找出一個簡單 / 容易理解的 (Interpretable)
3. 決策準則 (Explanation)
4. 而且對於任何的模型 都能夠適用 (Model-Agnostic)

02



**Data Type**



# Installation

---

The lime package is on [PyPI](#). Simply run:

```
pip install lime
```

Or clone the repository and run:

```
pip install .
```

We dropped python2 support in `0.2.0`, `0.1.1.37` was the last version before that.

```
In [3]: pip install lime  
import lime
```

# Text

當參數 `num_features = k`, 則顯示前 `k` 重要之判斷依據

```
exp = explainer.explain_instance(newsgroups_test.data[idx], c.predict_proba, num_features=6)
```

- 使用約2000份文檔, 其中依內容分為無神論及基督教兩類(約各半)
- 以隨機森林分類, 判斷文檔是哪一類, 再以 `lime` 解釋單一篇文檔中, 判別依據為何

```
Text with highlighted words
From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu
```

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish. This is the same question I have and I have not seen an answer on the net. If anyone has a contact please post on the net or email me.

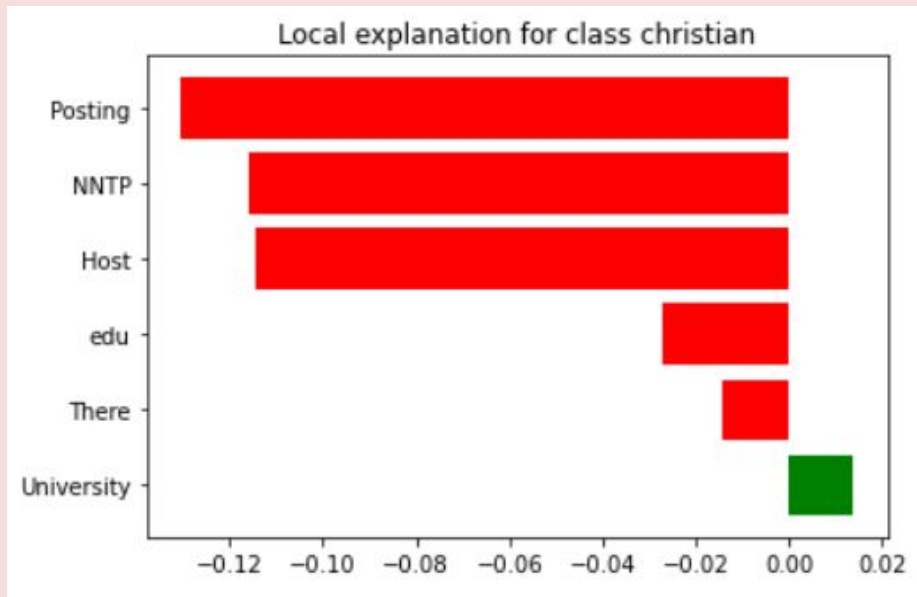
Thanks,

john chadwick  
johnchad@triton.unm.edu

# Text

- 可發現被選出的單字在常理上來看，與是否為基督教之間關聯不大，也就是說以這篇文檔來說，隨機森林分類得不是很好。

## Prediction probabilities



以0為原點，數值為負則表示這個詞語是該文檔應該分到「非基督教(無神論)」的依據；數值為正則表示這個詞語是該文檔應該分到「基督教」的依據。而離原點越遠，則代表它是越重要的判斷依據。

# Tabular data

---

```
In [11]: explainer = lime.lime_tabular.LimeTabularExplainer(train, feature_names=boston.feature_names,
class_names=['price'], categorical_features=categorical_features, verbose=True, mode='regression')
```

## Explainer : two mode

1. Regression
2. Classification

# Tabular data

---

## Data Set Characteristics:

Number of Instances: 506

Number of Attributes: 13 numeric/categorical predictive. Median Value (attribute 14) is usually the target.

Attribute Information (in order):

- CRIM per capita crime rate by town
- ZN proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS proportion of non-retail business acres per town
- CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX nitric oxides concentration (parts per 10 million)
- RM average number of rooms per dwelling
- AGE proportion of owner-occupied units built prior to 1940
- DIS weighted distances to five Boston employment centres
- RAD index of accessibility to radial highways
- TAX full-value property-tax rate per 10,000 US dollars.
- PTRATIO pupil-teacher ratio by town
- B  $1000(B_k - 0.63)^2$  where  $B_k$  is the proportion of blacks by town
- LSTAT % lower status of the population
- MEDV Median value of owner-occupied homes in 1000's US dollars.

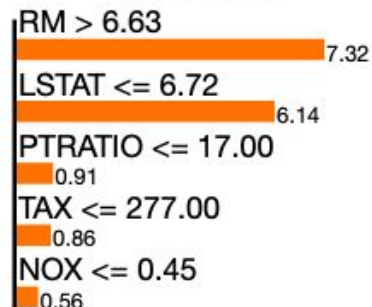
# Tabular data

Predicted value



negative

positive



Feature Value

RM	7.82
LSTAT	3.76
PTRATIO	14.90
TAX	216.00
NOX	0.44

RM : average number of rooms per dwelling

LSTAT : % lower status of the population

PTRATIO: pupil-teacher ratio by town

TAX : full-value property-tax rate per 10,000 US dollars.

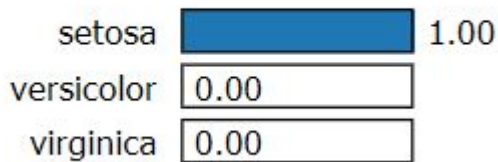
NOX : nitric oxides concentration (parts per 10 million)

# Tabular data

```
explainer = lime.lime_tabular.LimeTabularExplainer(train, feature_names=iris.feature_names, class_names=iris.target_names,  
|discretize_continuous=True, mode = "classification")
```

分為setosa及非setosa兩類，且預測十分準確

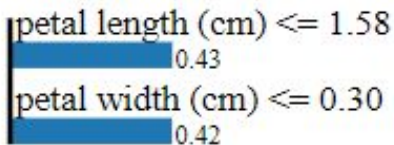
## Prediction probabilities



判斷setosa時，最大判斷依據是花瓣長度  $\leq 1.58$  公分

NOT setosa

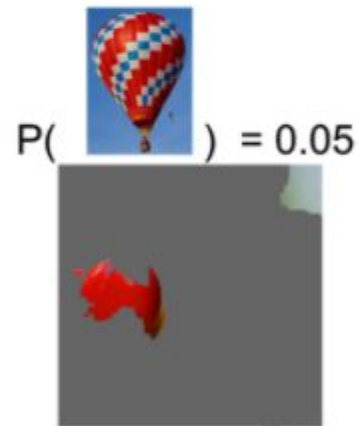
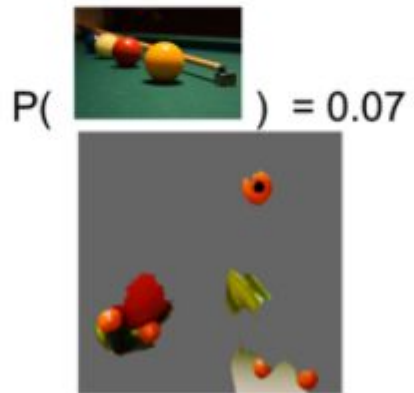
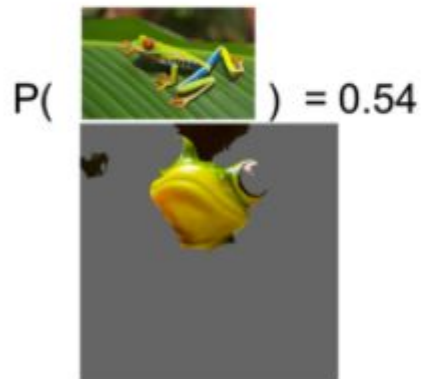
setosa



- 使用iris dataset
- 使用隨機森林

# Image

---





# Image

x: 每一個super pixel 的partition是否存在(0,1 的向量)  
y: 被遮住部分的影像丟進模型出來的機率值

$X'$

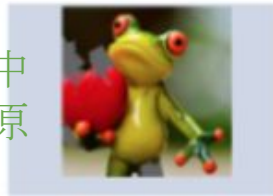
SP <sub>1</sub>	SP <sub>2</sub>	SP <sub>3</sub>		SP <sub>k</sub>
1	0	0	...	1

SP <sub>1</sub>	SP <sub>2</sub>	SP <sub>3</sub>		SP <sub>k</sub>
0	1	0	...	0

SP <sub>1</sub>	SP <sub>2</sub>	SP <sub>3</sub>		SP <sub>k</sub>
1	1	1	...	1






super pixel

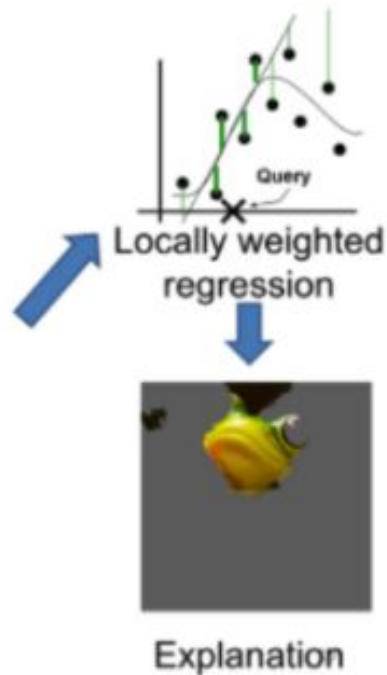


隨機遮住其中  
幾塊來擾動原  
本的樣本

# Image



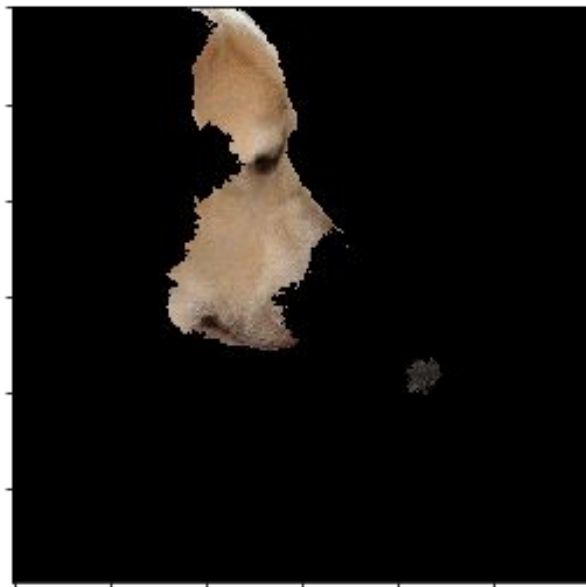
Perturbed Instances	$P(\text{tree frog})$
	<div><div></div></div> 0.85
	<div><div></div></div> 0.00001
	<div><div></div></div> 0.52



Forward selection  
Lasso regression  
Ridge regression

# Example

---





Thanks

# References

---

- Paper: “Why Should I Trust You?”: Explaining the Predictions of Any Classifier
- LIME Package (Python)
- LIME - Local Interpretable Model-Agnostic Explanation 技術介紹
- Local Interpretable Model-Agnostic Explanations (LIME): An Introduction
- lime package — lime 0.1 documentation
- <https://youtu.be/hUnRCxnydCc>
- <https://kknews.cc/zh-tw/tech/5j8nrk2.html>
- <https://kknews.cc/zh-tw/code/q4j4eky.html>



# Appendix

---

LIME Package (R)

Example (R)