LIME
Surrogate model

LIME – surrogate model

1. Choose the data point to explain.
2. Generate synthetic data in its proximity.
3. Obtain the black box predictions for the data from 2.
4. Obtain the distance between synthetic data and original data point.
5. **Train a white box with the perturbed data (2) to predict the black box predictions (3), weighted by their locality (4).**
6. Interpret the white box.

Surrogate requisites

- **Intrinsically explainable:**
  - Linear / logistic regression
  - Generalized linear models
  - Decision trees

- **Interpretable ➔ Ω**
  - Limited number of features
  - Limited depth

LIME - mathematically

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

$\boldsymbol{\varepsilon}$ is the explanation (LIME)

$\mathfrak{l}$ is the loss (weighted sum of squares)

f is the black box model

g is the surrogate (tree, linear regression)

π is the weight

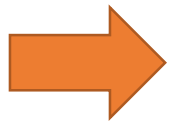
Ω is the complexity constraint

- number of features
- depth of the tree

4. Train surrogate

Synthetic data

Colour	Age	Income	Car make	Nr. Cards	Predictions	weights
0	65	51000	1	5	y1	W1
1	63	59000	0	5.2	y2	W2
0	60	55000	1	5.9	y3	W3
0	58	45000	0	6	y4	W4
0	55	47000	1	4	y5	W5



Explainable model (tree, linear regression)

```
from sklearn.linear_model import Lasso

reg = Lasso(alpha=0.1).fit(X, y, weights)

reg.coef_
```

LIME - mathematically

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

$\boldsymbol{\varepsilon}$ is the explanation (LIME)

$\mathfrak{l}$ is the loss (weighted sum of squares)

f is the black box model

g is the surrogate (**tree, linear regression**)

π **is the weight**

Ω is the **complexity constraint**

- number of features
- depth of the tree

LIME - mathematically

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

There is more than 1 possible explanation.

G is the family of possible explanations.

g is one of the possible explanations.

Train In Data

THANK YOU

www.trainindata.com