Vicinity

Defining the kernel

Defining the vicinity

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.

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)

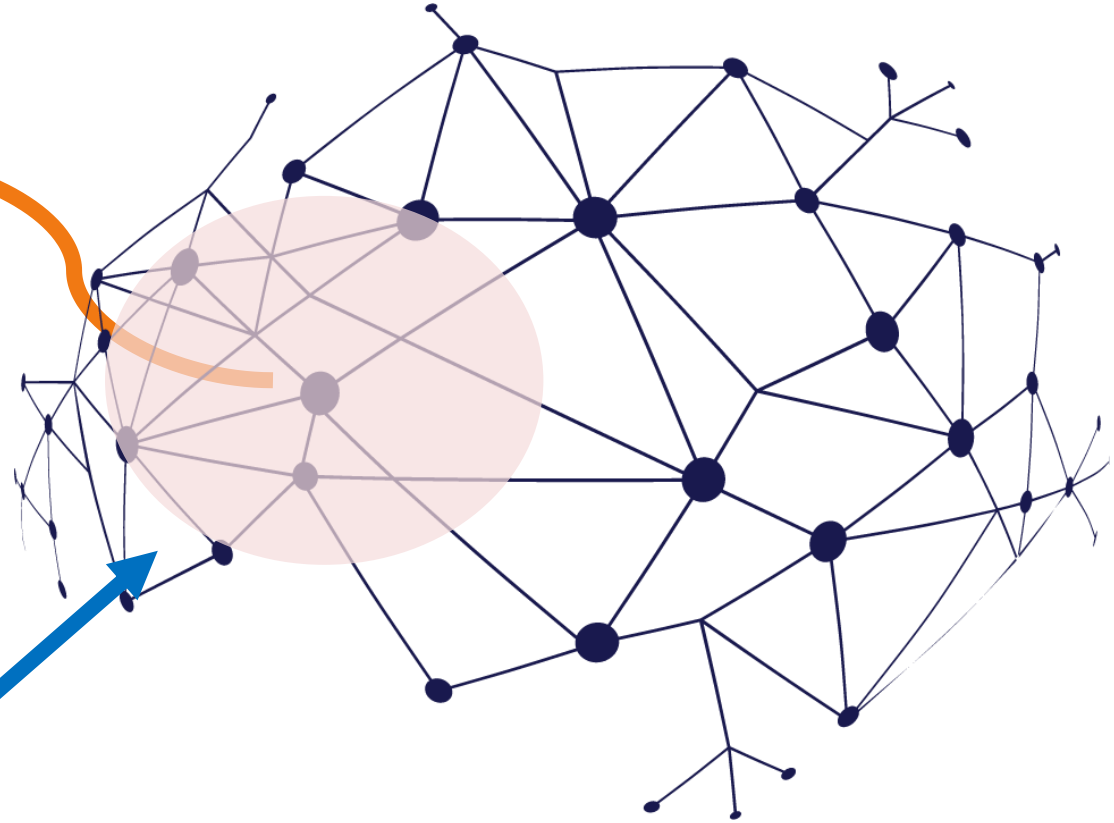
π **is the weight**

Ω is the complexity constraint

- number of features
- depth of the tree

Locality

Why does it make this prediction?



Local surrogate

How far away are we allowed to sample?

How similar is one point to another?

Similarity

=

?

How similar are these images or texts?

British scientists unveil the ‘world’s first’ laptop powered by light. Mobile computer owners are looking forward for first batch of the laptops.

=

British unveil the ‘world’s first’ laptop powered by light. Mobile computer owners are looking forward for first batch of the laptops.

?

Images from: https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/

Vicinity or locality matters

$$W = e^{-D(x,z)^2 / \sigma^2}$$

D is a distance metric

- Euclidean for tables
- Cosine for images / text

x is the original data point

z is the synthetic data point

σ is some arbitrary kernel

Vicinity or locality matters

Tabular data ➔ kernel_width = np.sqrt(training_data.shape[1]) * 0.75

Images ➔ kernel_width = 0.25

Text ➔ 25

Numbers are totally arbitrary.

Train In Data

THANK YOU

www.trainindata.com