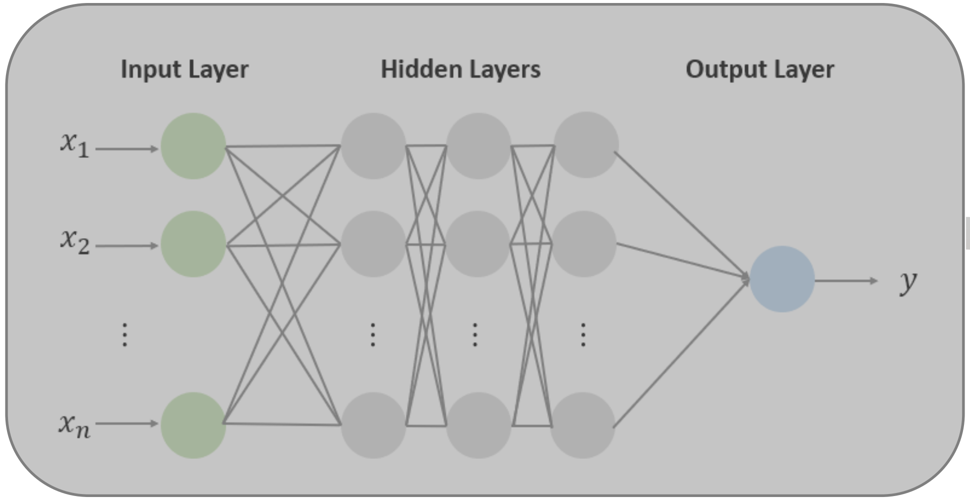
CS 4501/6501 Interpretable Machine Learning

Post-hoc explanations: perturbation-based methods

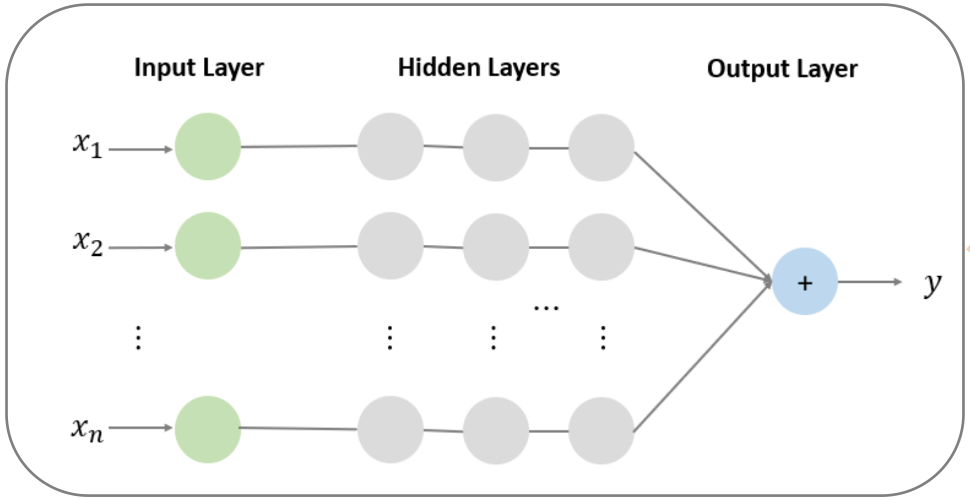
Hanjie Chen, Yangfeng Ji
Department of Computer Science
University of Virginia
{hc9mx, yangfeng}@virginia.edu

Trade-off

Black-box
Neural Network



Interpretable GAM



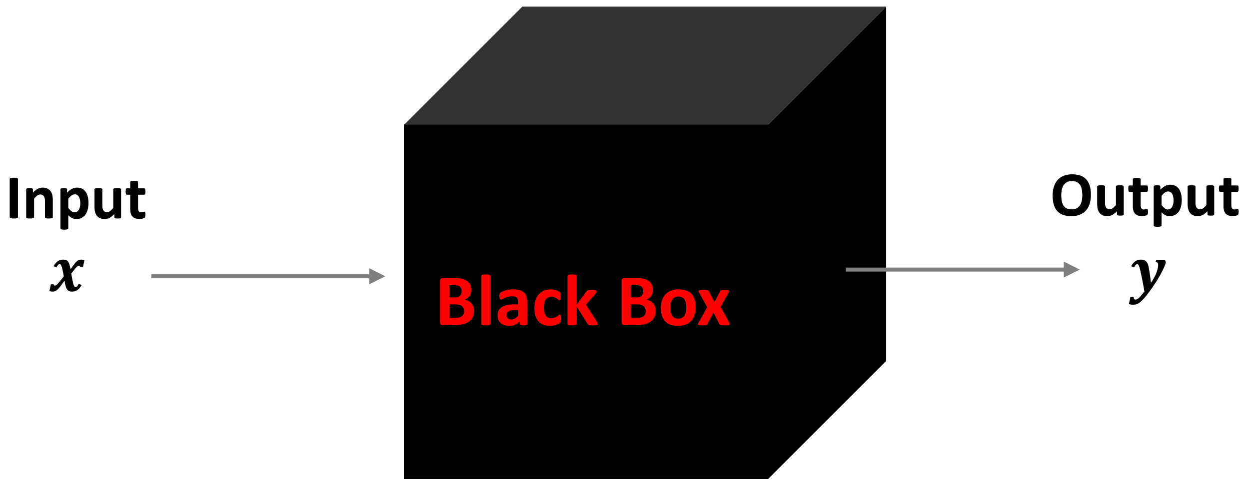
Limitations

- Ignoring complex feature interactions
- Performance drop

Explaining Black-box Model

How to improve model interpretability?

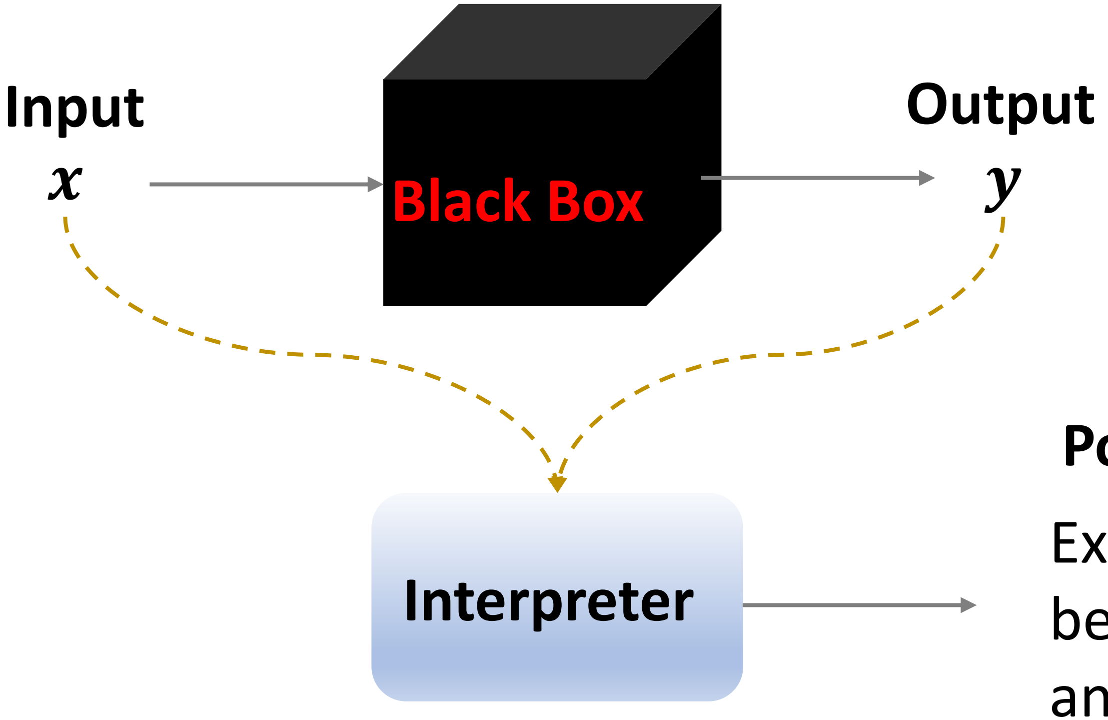
Model's inner working and decision making are hidden in the black box



Explaining Black-box Model

How to improve model interpretability?

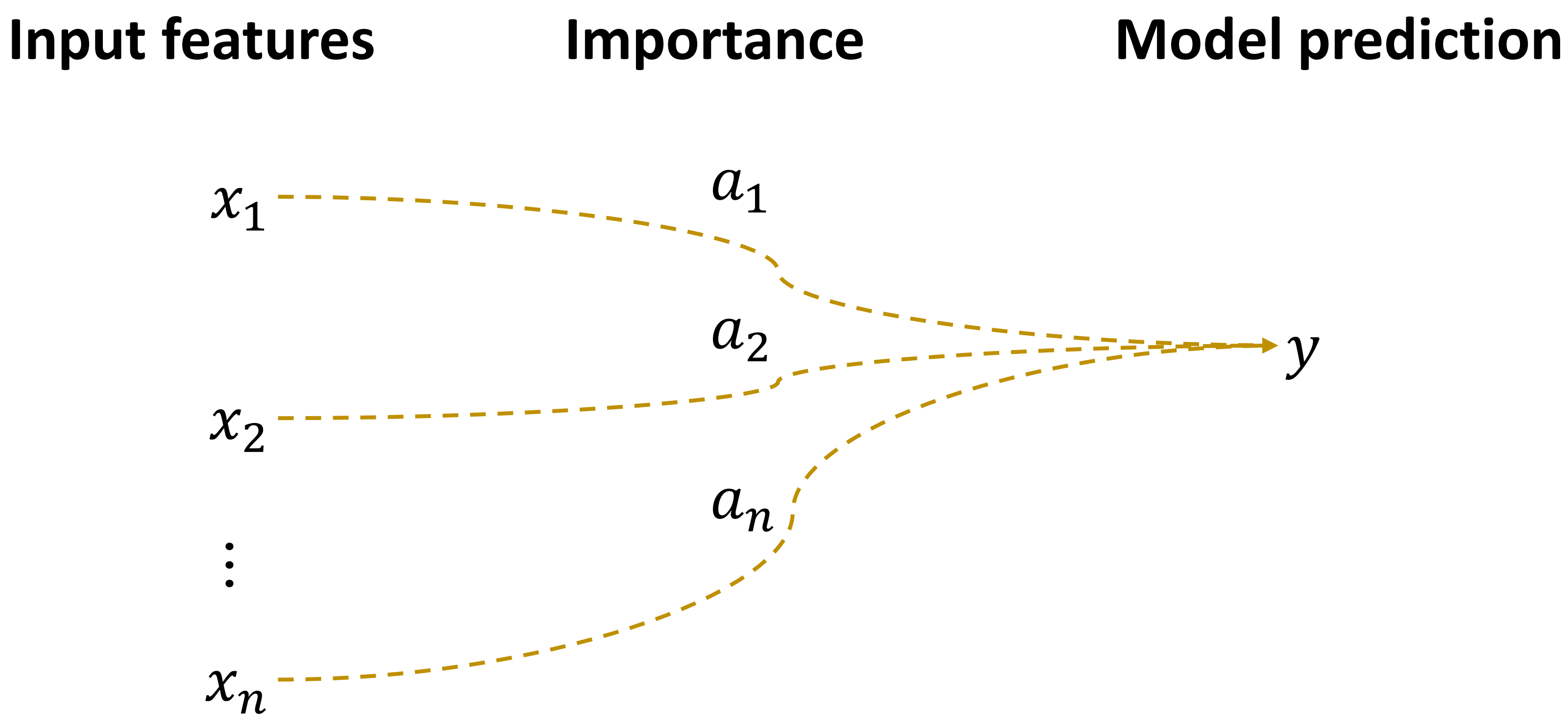
Explaining model predictions from the post-hoc manner



Post-hoc explanation

Extracting relationships between input features and the model prediction

Post-hoc Explanation



Identifying important features

Post-hoc Explanation

- Example 1: tabular data
  - Mushroom dataset
  - **Task**: predicting if a mushroom is edible or poisonous

Feature
Odor
gill size
stalk surface above ring
Spore print color
stalk surface below ring

Post-hoc Explanation

- Example 1: tabular data
  - Mushroom dataset
  - **Task**: predicting if a mushroom is edible or poisonous

Input

Feature	Value
Odor=foul	$x_1 = 1$ (true)
gill size=broad	$x_2 = 1$
stalk surface above ring=silky	$x_3 = 1$
Spore print color=chocolate	$x_4 = 1$
stalk surface below ring=silky	$x_5 = 1$

→ Model →

Prediction

poisonous

Explanation

Importance
$a_1 = 0.26$
$a_2 = -0.13$
$a_3 = 0.11$
$a_4 = 0.08$
$a_5 = 0.06$

Post-hoc Explanation

- Example 1: tabular data
  - Mushroom dataset
  - **Task**: predicting if a mushroom is edible or poisonous

Input

Feature	Value
Odor=foul	$x_1 = 1$ (true)
gill size=broad	$x_2 = 1$
stalk surface above ring=silky	$x_3 = 1$
Spore print color=chocolate	$x_4 = 1$
stalk surface below ring=silky	$x_5 = 1$

→ Model →

Prediction

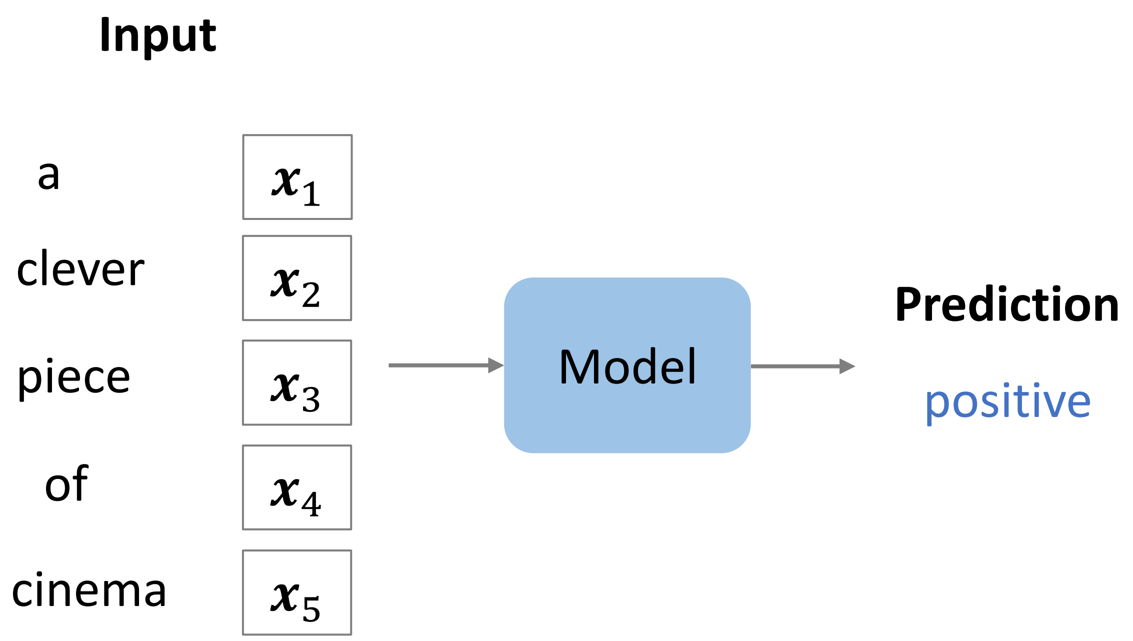
poisonous

Explanation

Importance	
$\boxed{a_1 = 0.26}$	(the most important feature)
$\boxed{a_2 = -0.13}$	(indicating edible)
$a_3 = 0.11$	
$a_4 = 0.08$	
$a_5 = 0.06$	

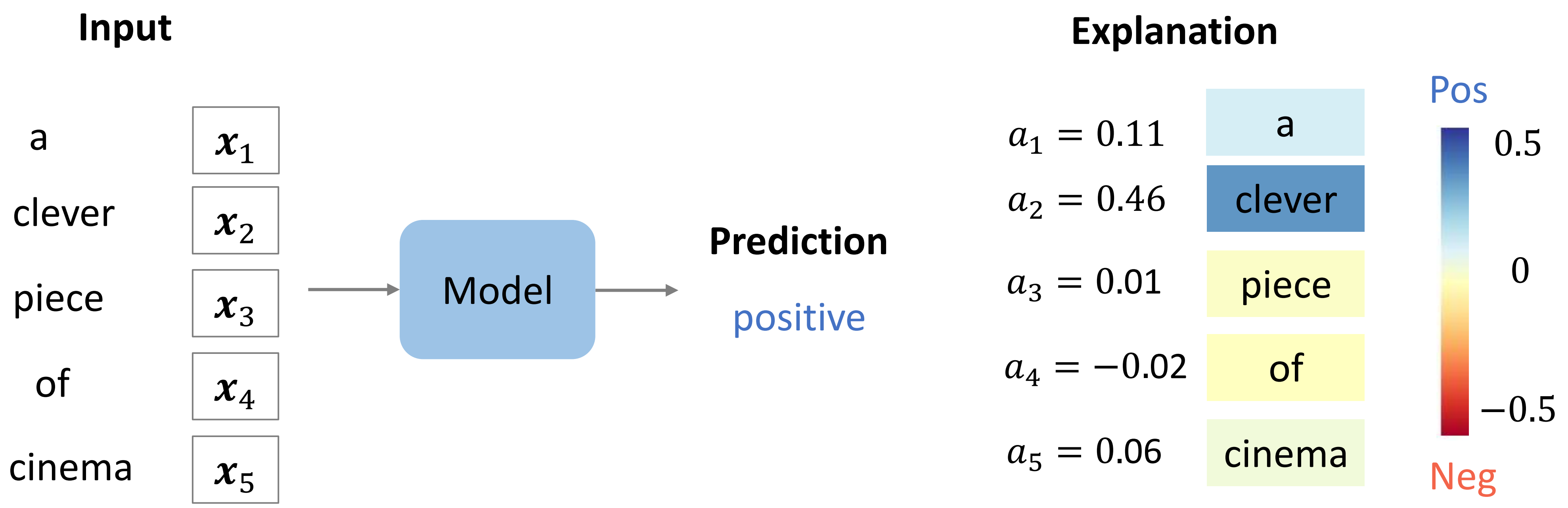
Post-hoc Explanation

- Example 2: text data
  - Movie review
  - **Task**: predicting the sentiment of a text (positive or negative)



Post-hoc Explanation

- Example 2: text data
  - Movie review
  - **Task**: predicting the sentiment of a text (positive or negative)



(Word saliency map)

Post-hoc Explanation

- Example 3: image data

Task: Object recognition

Feature: a pixel $\boldsymbol{x}_{ij}$

(color, intensity…)

Source: https://medium.datadriveninvestor.com/visualizing-neural-networks-using-saliency-maps-in-pytorch-289d8e244ab4

Post-hoc Explanation

- Example 3: image data

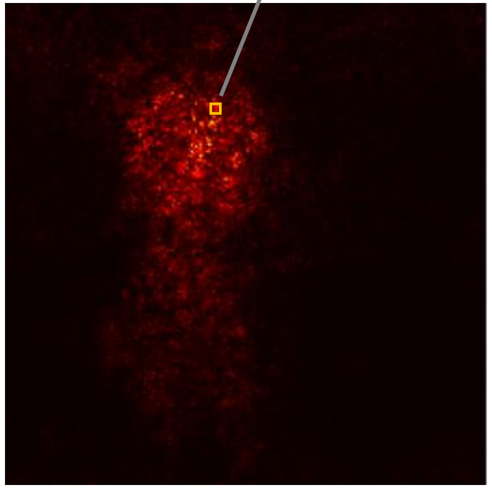
Task: Object recognition
Prediction: Dog

Input

$\boldsymbol{x}_{ij}$

Explanation

$\boldsymbol{a}_{ij}$



Saliency map: the lighter color, the larger value

Source: https://medium.datadriveninvestor.com/visualizing-neural-networks-using-saliency-maps-in-pytorch-289d8e244ab4

Post-hoc Explanation

How to learn feature importance?

Perturbation-based methods

- Model-agnostic (black-box): not requiring access to model inner working
- Local: explaining model prediction per example

Perturbation-based methods

- LIME (Ribeiro et al., KDD, 2016)
- SHAP (Lundberg and Lee, NIPS, 2017)

LIME

"Why Should I Trust You?"
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

(KDD, 2016)

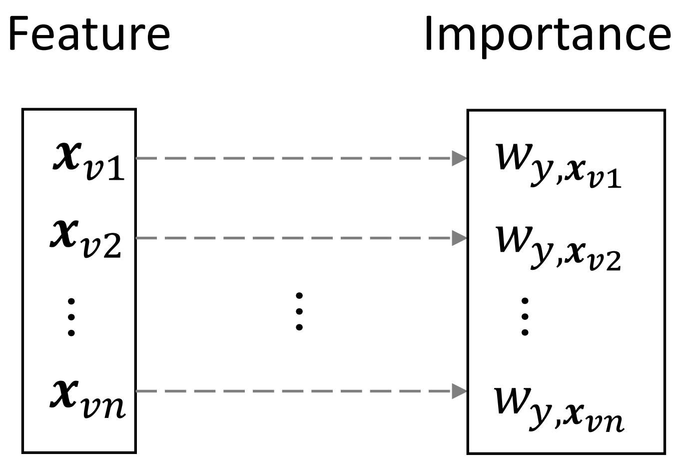
Interpretable Model

- Linear model

$$h_y(\boldsymbol{x}) = \boldsymbol{w_y}^T \boldsymbol{x} \qquad \boldsymbol{x} \in \{0,1\}^n$$

  - $w_{y,j}$: the contribution of x_j
  - Higher weights indicate more important features

Global interpretation



Interpretable Model

- Linear model

$$h_y(\boldsymbol{x}) = \boldsymbol{w}_y{}^T \boldsymbol{x} \qquad \boldsymbol{x} \in \{0, 1\}^n$$

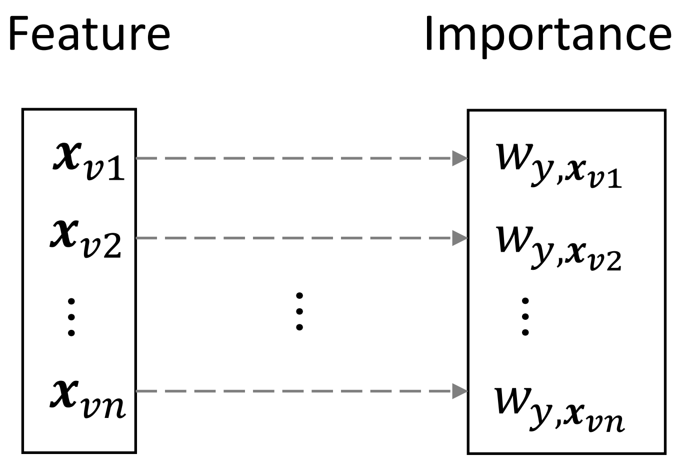
- $w_{y,j}$: the contribution of x_j
- Higher weights indicate more important features

Logistic regression

	"It"	"is"	"a"	"fantastic"	"movie"	
[Neg] $\boldsymbol{w}_0$	0.89	0.72	1.13	-1.92	0.34	1.16
[Pos] $\boldsymbol{w}_1$	0.85	0.82	1.05	2.21	0.26	5.19

Prediction: positive

Global interpretation



Neural Networks

Global interpretation is not capable of explaining each specific model prediction

- Neural networks can capture complex relationships between features and the response

- The meaning of a feature may vary across different examples

“good” $\xrightarrow{\textit{adjective}}$ Morally excellent

“good” $\xrightarrow{\textit{noun}}$ Possessions

Neural Networks

Global interpretation is not capable of explaining each specific model prediction

- Neural networks can capture complex relationships between features and the response

- The meaning of a feature may vary across different examples

"good" $\xrightarrow{\textit{adjective}}$ Morally excellent

"good" $\xrightarrow{\textit{noun}}$ Possessions

Local interpretation

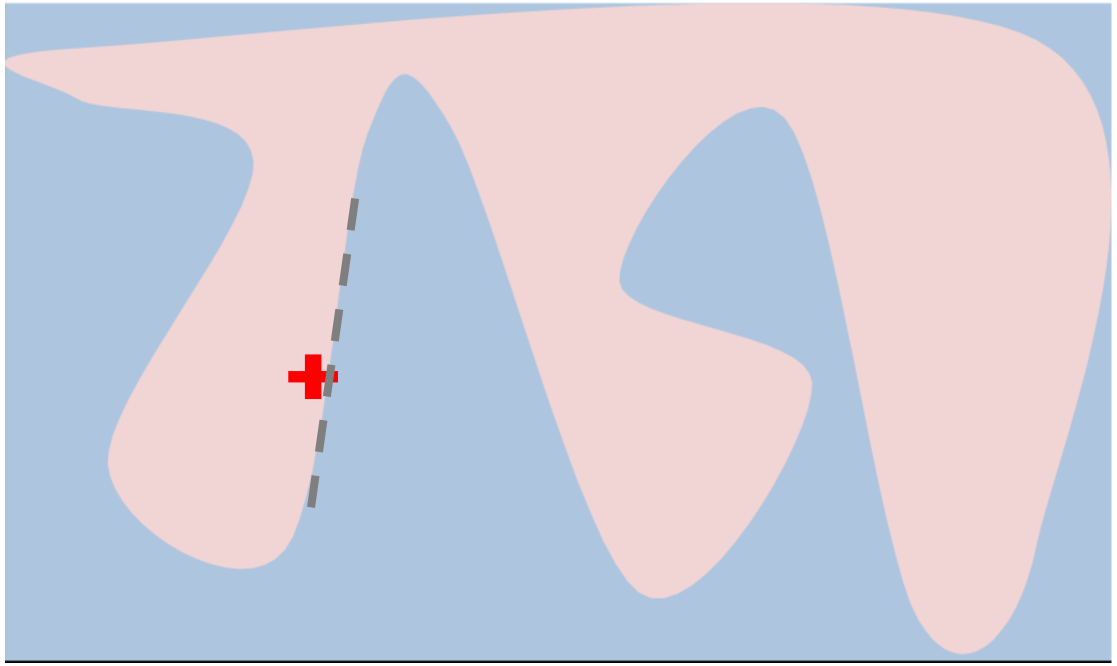
Explaining model prediction per example by identifying local feature importance

LIME: *Local Interpretable Model-Agnostic Explanations*

The way that explains model predictions or the generated explanations are understandable to humans

LIME: *Local Interpretable Model-Agnostic Explanations*

Idea: using local linear model to approximate neural network for each example



- Decision boundary of a neural network f
- Blue/pink background represents negative (-) /positive (+) class
- Bold red cross: the instance $\boldsymbol{x}$ being explained
- Dashed line: local linear model g

$$g \approx f$$

LIME: *Local Interpretable Model-Agnostic Explanations*

- Interpretable data representations

Neural network f

$$\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]$$

Feature representation
$\boldsymbol{x}_i \in \mathbb{R}^{\boldsymbol{d}}$ is uninterpretable

Linear model g

$$\boldsymbol{x}' = [x'_1, x'_2, \cdots, x'_N]$$

Feature representation
$x'_i \in \{0, 1\}$ is interpretable

- n: the number of features in the example
- N: the number of all features

LIME: *Local Interpretable Model-Agnostic Explanations*

- Interpretable data representations

Neural network f

$$\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]$$

Feature representation
$\boldsymbol{x}_i \in \mathbb{R}^{\boldsymbol{d}}$ is uninterpretable

Image

$\boldsymbol{x}_i$: a tensor with three color channels per pixel

Text

$\boldsymbol{x}_i$: a high-dimensional vector (word embedding)

Linear model g

$$\boldsymbol{x}' = [x'_1, x'_2, \cdots, x'_N]$$

Feature representation
$x'_i \in \{0, 1\}$ is interpretable

Image

0/1 indicates the absence/presence of a patch of pixels

Text

0/1 indicates the absence/presence of a word
(bag-of-words representation)

LIME: *Local Interpretable Model-Agnostic Explanations*

- Interpretable data representations

Neural network f $\approx$ Linear model g

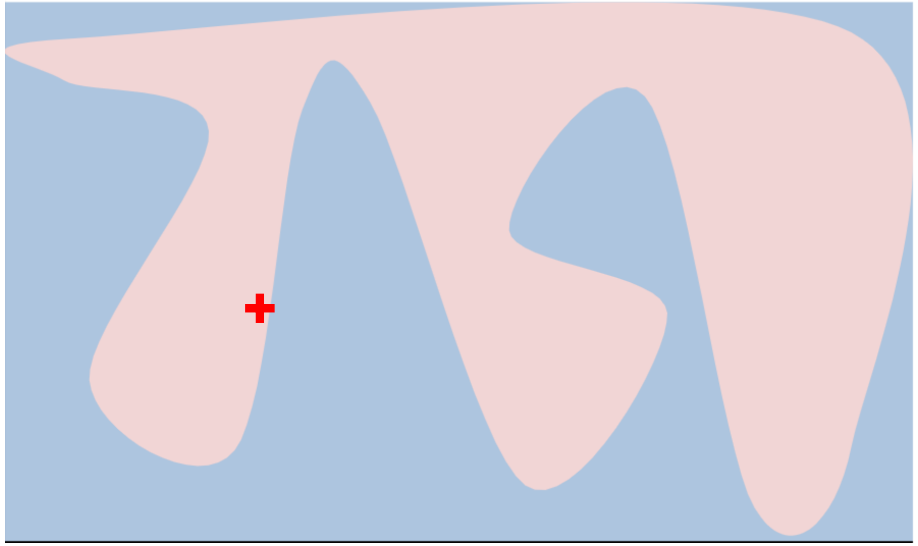
$$\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n] \longrightarrow \boldsymbol{x}' = [x'_1, x'_2, \cdots, x'_N]$$

Text	$\boldsymbol{x}$
a	$\boldsymbol{x}_1$
good	$\boldsymbol{x}_2$
movie	$\boldsymbol{x}_3$

Vocab	$\boldsymbol{x}'$
⋮	⋮ (0)
a	1
⋮	⋮
good	1
⋮	⋮
movie	1
⋮	⋮

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sampling for local exploration



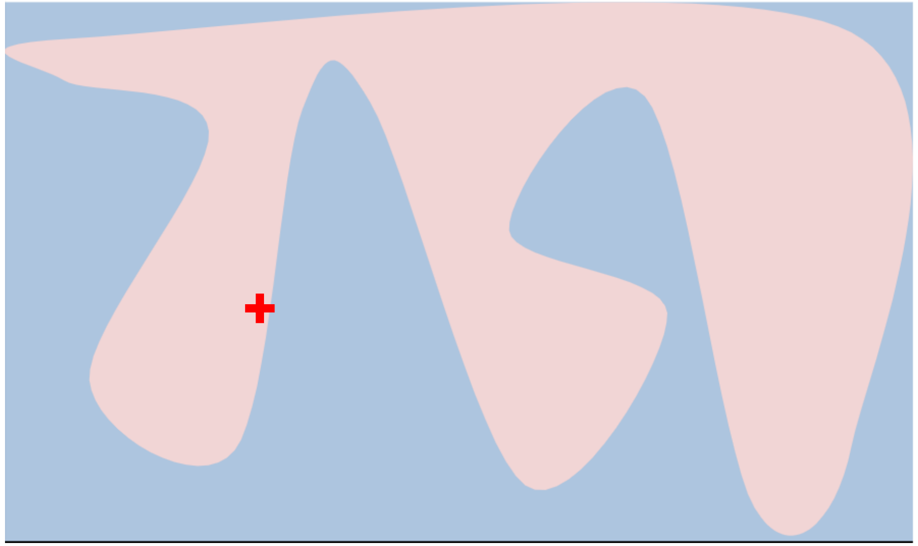
Need more samples to fit a local linear model

It is a fantastic movie

$$\boldsymbol{x}' = [0, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 0, 1, \cdots, 0]_N$$

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sampling for local exploration



Need more samples to fit a local linear model

It is a fantastic movie

$$\boldsymbol{x}' = [0, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 0, 1, \cdots, 0]_N$$

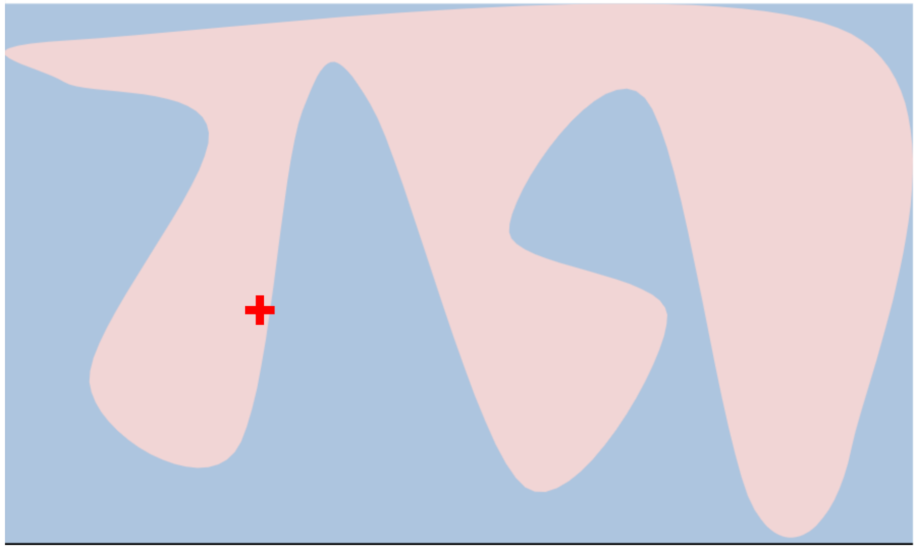
Randomly sample nonzero elements

a movie

$$\boldsymbol{z}_1{}' = [0, \cdots, 0, \cdots, 0, \cdots, 1, \cdots, 0, \cdots, 0, 1, \cdots, 0]_N$$

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sampling for local exploration



Need more samples to fit a local linear model

It is a fantastic movie

$$\boldsymbol{x}' = [0, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 0, 1, \cdots, 0]_N$$

Randomly sample nonzero elements

a movie

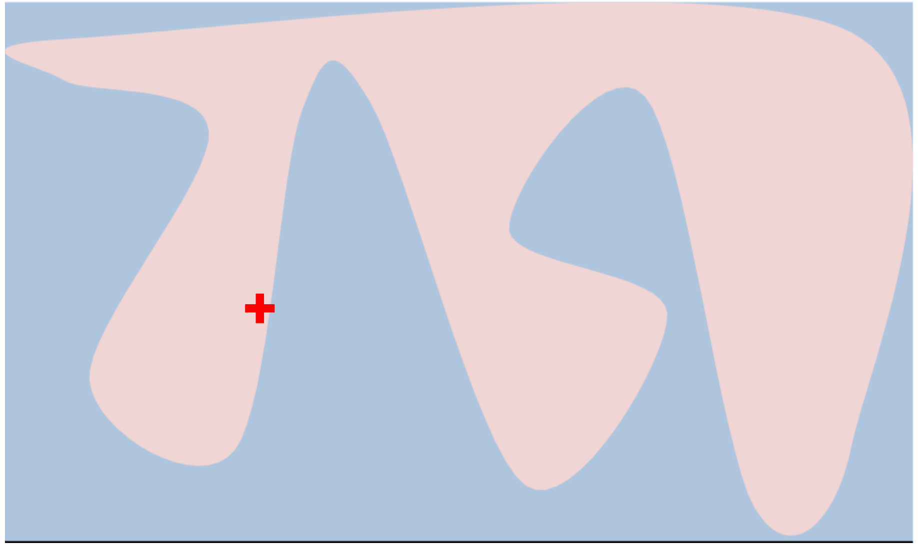
$$\boldsymbol{z}_1{}' = [0, \cdots, 0, \cdots, 0, \cdots, 1, \cdots, 0, \cdots, 0, 1, \cdots, 0]_N$$

fantastic movie

$$\boldsymbol{z}_2{}' = [0, \cdots, 0, \cdots, 0, \cdots, 0, \cdots, 1, \cdots, 0, 1, \cdots, 0]_N$$

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sampling for local exploration



Need more samples to fit a local linear model

It is a fantastic movie

$$\boldsymbol{x}' = [0, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 0, 1, \cdots, 0]_N$$

Randomly sample nonzero elements

a movie

$$\boldsymbol{z}_1{}' = [0, \cdots, 0, \cdots, 0, \cdots, 1, \cdots, 0, \cdots, 0, 1, \cdots, 0]_N$$

fantastic movie

$$\boldsymbol{z}_2{}' = [0, \cdots, 0, \cdots, 0, \cdots, 0, \cdots, 1, \cdots, 0, 1, \cdots, 0]_N$$

$$\vdots$$

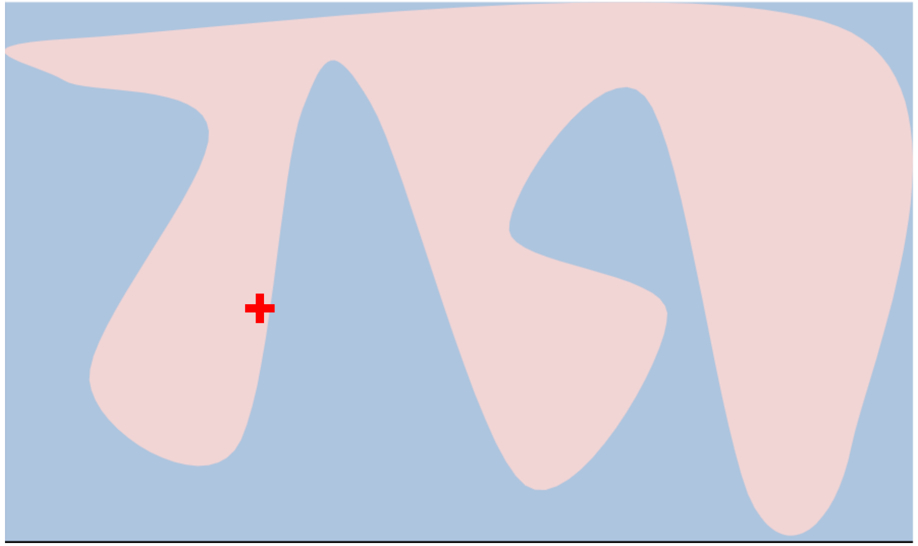
fantastic

$$\boldsymbol{z}_M{}' = [0, \cdots, 0, \cdots, 0, \cdots, 0, \cdots, 1, \cdots, 0, 0, \cdots, 0]_N$$

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sampling for local exploration

Need more samples to fit a local linear model



It is a fantastic movie

$$\boldsymbol{x}' = [0, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 0, 1, \cdots, 0]_N$$

Randomly sample nonzero elements

a movie

$$\boldsymbol{z}_1{}' = [0, \cdots, 0, \cdots, 0, \cdots, 1, \cdots, 0, \cdots, 0, 1, \cdots, 0]_N$$

fantastic movie

$$\boldsymbol{z}_2{}' = [0, \cdots, 0, \cdots, 0, \cdots, 0, \cdots, 1, \cdots, 0, 1, \cdots, 0]_N$$

$$\vdots$$

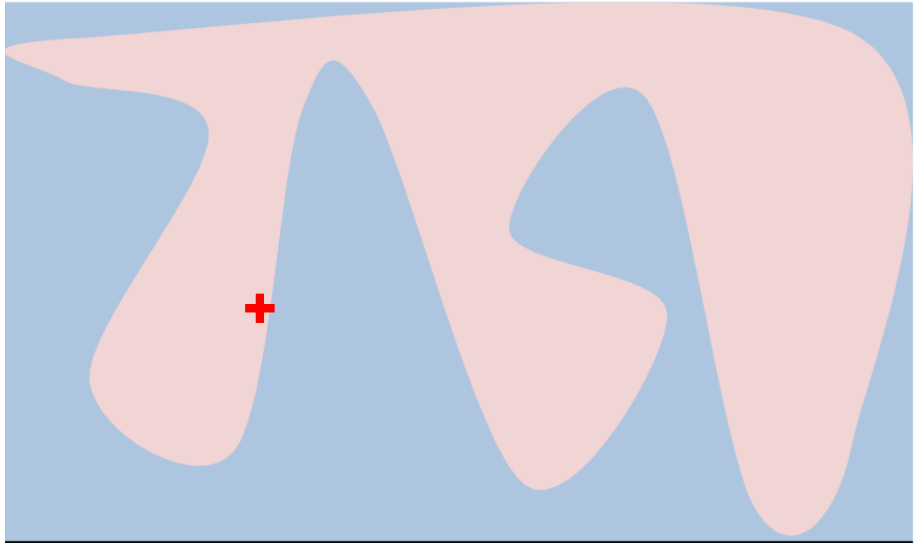
fantastic

$$\boldsymbol{z}_M{}' = [0, \cdots, 0, \cdots, 0, \cdots, 0, \cdots, 1, \cdots, 0, 0, \cdots, 0]_N$$

What are the labels of these pseudo examples?

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sampling for local exploration



Labeling pseudo examples with neural network f

$\boldsymbol{z}_1{}' \longrightarrow \boldsymbol{z}_1 \longrightarrow f(\boldsymbol{z}_1) \longrightarrow$ Negative ●

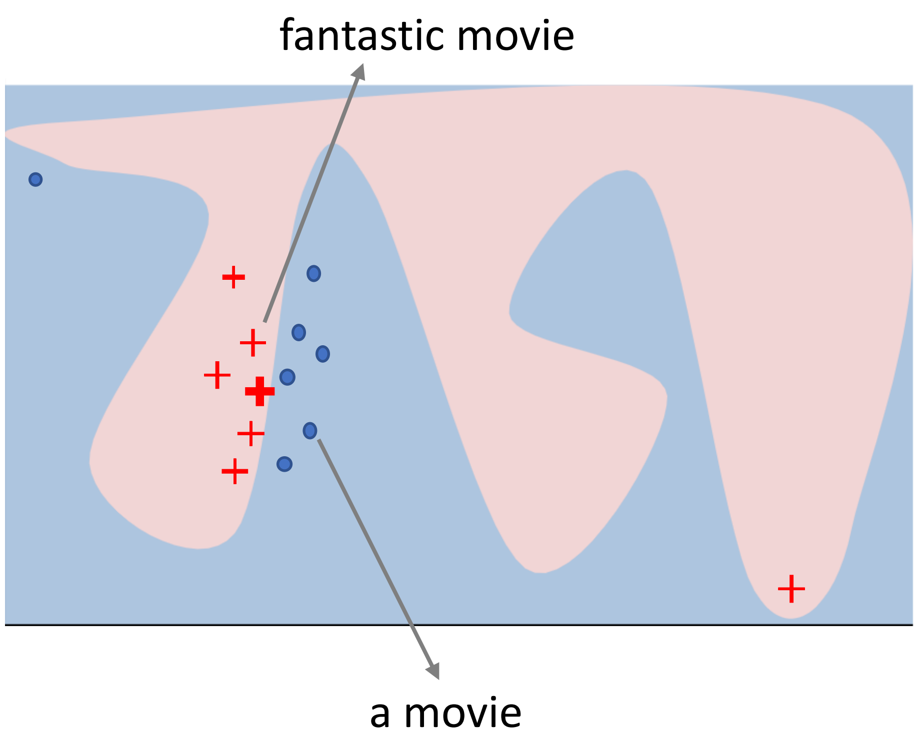
$\boldsymbol{z}_2{}' \longrightarrow \boldsymbol{z}_2 \longrightarrow f(\boldsymbol{z}_2) \longrightarrow$ Positive +

$\vdots \qquad \vdots$

$\boldsymbol{z}_M{}' \longrightarrow \boldsymbol{z}_M \longrightarrow f(\boldsymbol{z}_M) \longrightarrow$ Positive +

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sampling for local exploration



Labeling pseudo examples with neural network f

$\boldsymbol{z}_1{}' \longrightarrow \boldsymbol{z}_1 \longrightarrow f(\boldsymbol{z}_1) \longrightarrow$ Negative ●

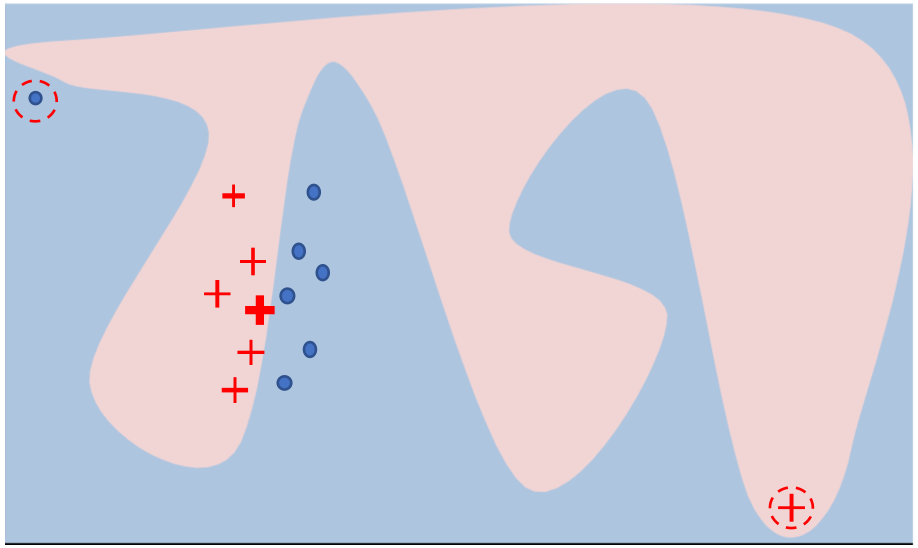
$\boldsymbol{z}_2{}' \longrightarrow \boldsymbol{z}_2 \longrightarrow f(\boldsymbol{z}_2) \longrightarrow$ Positive +

$\vdots \quad \vdots$

$\boldsymbol{z}_M{}' \longrightarrow \boldsymbol{z}_M \longrightarrow f(\boldsymbol{z}_M) \longrightarrow$ Positive +

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sampling for local exploration



Penalize noisy examples

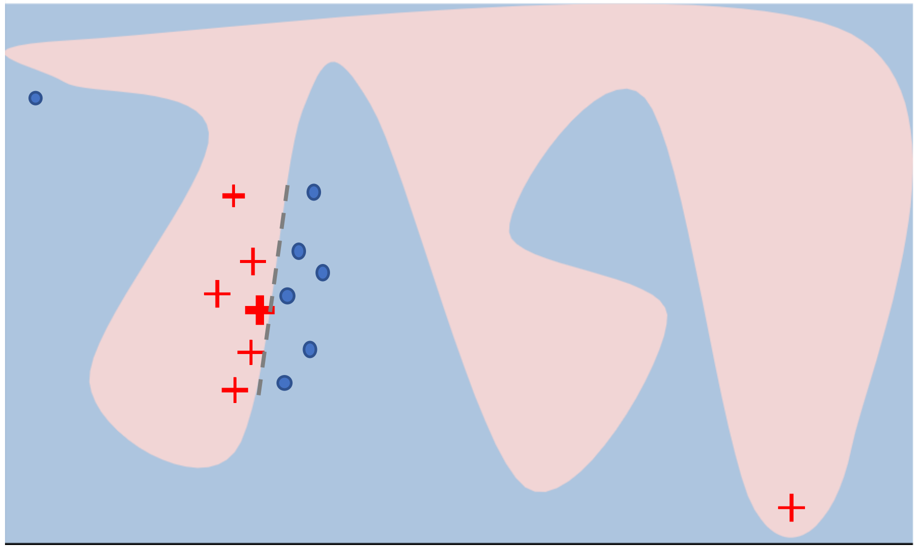
Distance between $\boldsymbol{x}$ and $\boldsymbol{z}_m$

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

D : cosine distance (for text), L_2 distance (for image)

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sparse linear explanation



Fitting a local linear model

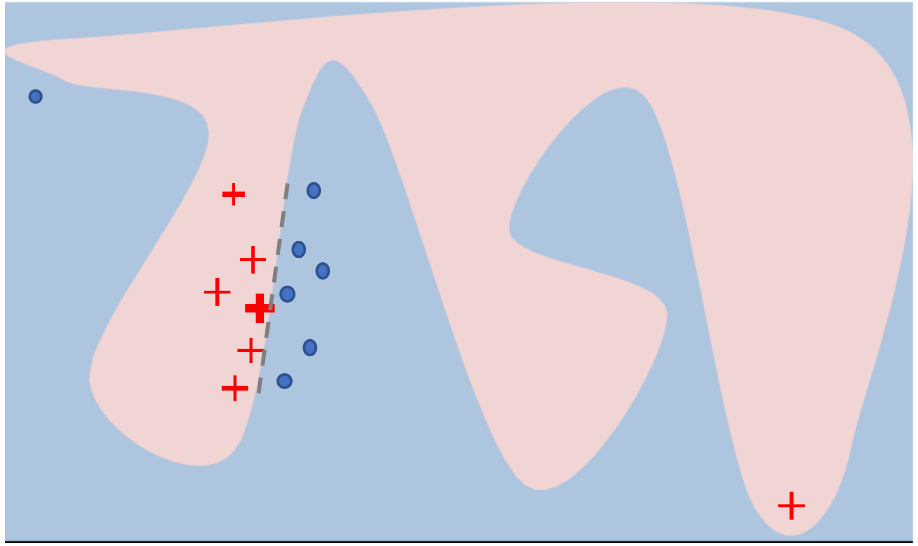
$$\{(\boldsymbol{z}_m', f(\boldsymbol{z}_m))\}_{m=1,\cdots,M}$$

$$g(\boldsymbol{z}') \approx f(\boldsymbol{z})$$

$$g(\boldsymbol{z}') = \boldsymbol{w}^T \boldsymbol{z}'$$

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sparse linear explanation



Fitting a local linear model

$$\{(\boldsymbol{z}_m', f(\boldsymbol{z}_m))\}_{m=1,\cdots,M}$$

$$g(\boldsymbol{z}') \approx f(\boldsymbol{z})$$

$$g(\boldsymbol{z}') = \boldsymbol{w}^T\boldsymbol{z}'$$

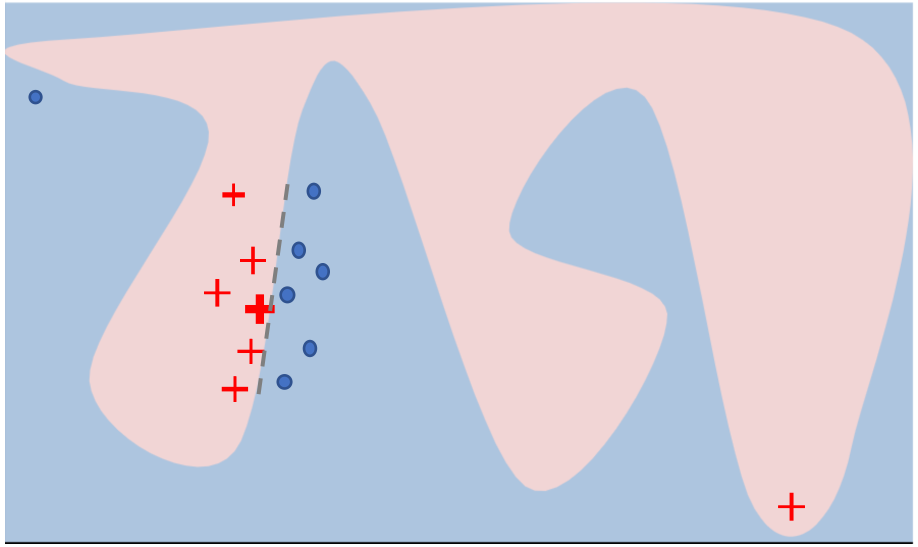
Objective

$$\min \mathcal{L}(f, g)$$

$$\mathcal{L}(f, g) = \sum \pi_{\boldsymbol{x}}(\boldsymbol{z})(f(\boldsymbol{z}) - g(\boldsymbol{z}'))^2$$

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sparse linear explanation



Fitting a local linear model

$$\{(\boldsymbol{z}_m{}', f(\boldsymbol{z}_m))\}_{m=1,\cdots,M}$$

$$g(\boldsymbol{z}') \approx f(\boldsymbol{z})$$

$$g(\boldsymbol{z}') = \boldsymbol{w}^T\boldsymbol{z}'$$

Objective

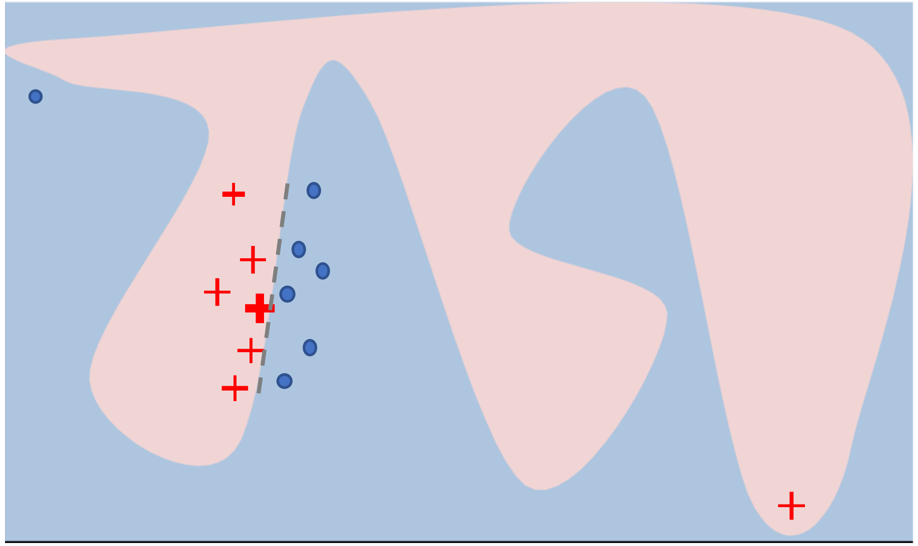
$$\min \mathcal{L}(f,g) + \Omega(g)$$

Restricting complexity (the number of nonzero weights)

$$\mathcal{L}(f,g) = \sum \pi_{\boldsymbol{x}}(\boldsymbol{z})(f(\boldsymbol{z}) - g(\boldsymbol{z}'))^2$$

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sparse linear explanation



Extracting feature importance scores

$$\boldsymbol{w}_{\hat{y}}{}^{T}$$

- $\hat{y}$: model prediction on the original example
- Local explanation: $\{w_{\hat{y},\boldsymbol{x}_1}, \cdots, w_{\hat{y},\boldsymbol{x}_n}\}$

Question?

LIME: *Local Interpretable Model-Agnostic Explanations*

- Explaining each example individually, not the whole dataset (locally faithful)
- May not work for highly non-linear models

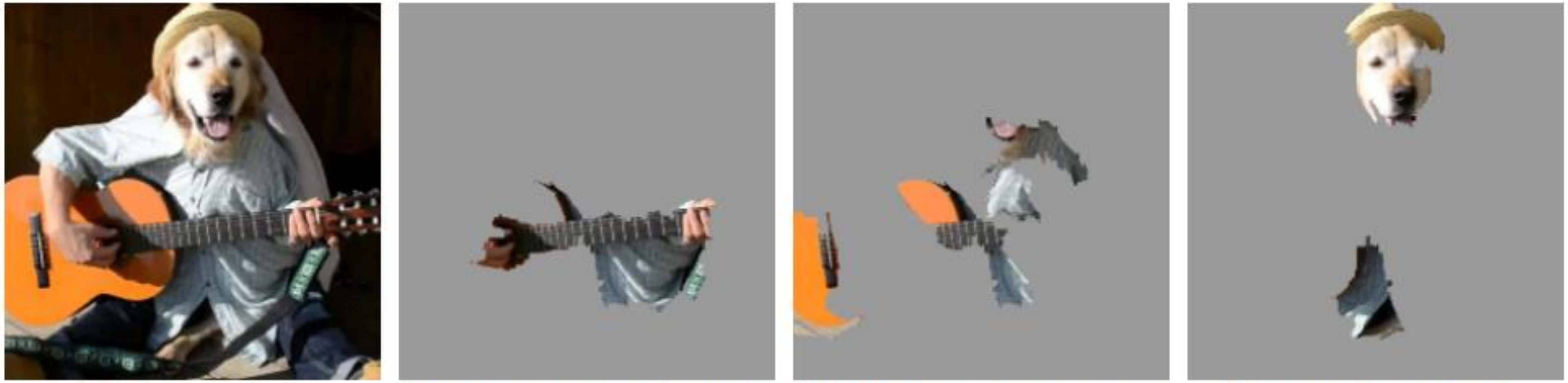
LIME: *Local Interpretable Model-Agnostic Explanations*

- Example: Deep networks for image classification

Model: Google's pre-trained Inception neural network

Top 3 predicted classes

Electric guitar | Acoustic guitar | Labrador



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

The explanations enhance trust in the model, as it acts in a reasonable manner

Question?

Submodular Pick for Explaining Models

Single explanation is not sufficient to evaluate and assess trust in the model as a whole

➡

Providing a global understanding of the model by explaining a set of individual instances

Submodular Pick for Explaining Models

Single explanation is not sufficient to evaluate and assess trust in the model as a whole

Providing a global understanding of the model by explaining a set of individual instances

How to select these instances judiciously?

Submodular Pick for Explaining Models

- Budget B: the number of explanations users are willing to look at in order to understand a model

A set of instances X $\xrightarrow{\text{select}}$ B instances (diverse, representative)

Submodular Pick for Explaining Models

- Budget B: the number of explanations users are willing to look at in order to understand a model

A set of instances X $\xrightarrow{\text{select}}$ B instances (diverse, representative)

➢ **Submodular pick (SP) algorithm**

$\boldsymbol{x}^{(1)}$ contains two features

Explanation matrix W

	x_1	x_2	x_3	x_4	x_5
$\boldsymbol{x}^{(1)}$	$w_{1,1}$	$w_{1,2}$			
$\boldsymbol{x}^{(2)}$		$w_{2,2}$	$w_{2,3}$		
$\boldsymbol{x}^{(3)}$		$w_{3,2}$	$w_{3,3}$		
$\boldsymbol{x}^{(4)}$		$w_{4,2}$		$w_{4,4}$	
$\boldsymbol{x}^{(5)}$				$w_{5,4}$	$w_{5,5}$

- Each row represents an instance
- Each column represents a feature
- Each value represents a local importance

Submodular Pick for Explaining Models

- Budget B: the number of explanations users are willing to look at in order to understand a model

A set of instances X $\xrightarrow{\text{select}}$ B instances (diverse, representative)

➢ **Submodular pick (SP) algorithm**

Explanation matrix W

	x_1	x_2	x_3	x_4	x_5
$\boldsymbol{x}^{(1)}$	$w_{1,1}$	$w_{1,2}$			
$\boldsymbol{x}^{(2)}$		$w_{2,2}$	$w_{2,3}$		
$\boldsymbol{x}^{(3)}$		$w_{3,2}$	$w_{3,3}$		
$\boldsymbol{x}^{(4)}$		$w_{4,2}$		$w_{4,4}$	
$\boldsymbol{x}^{(5)}$				$w_{5,4}$	$w_{5,5}$

I_2: global importance of x_2

❑ Select instances that cover important features (x_2)

Submodular Pick for Explaining Models

- Budget B: the number of explanations users are willing to look at in order to understand a model

A set of instances X $\xrightarrow{\text{select}}$ B instances (diverse, representative)

➢ **Submodular pick (SP) algorithm**

Explanation matrix W

	x_1	x_2	x_3	x_4	x_5
$\boldsymbol{x}^{(1)}$	$w_{1,1}$	$w_{1,2}$			
$\boldsymbol{x}^{(2)}$		$w_{2,2}$	$w_{2,3}$		
$\boldsymbol{x}^{(3)}$		$w_{3,2}$	$w_{3,3}$		
$\boldsymbol{x}^{(4)}$		$w_{4,2}$		$w_{4,4}$	
$\boldsymbol{x}^{(5)}$				$w_{5,4}$	$w_{5,5}$

If $\boldsymbol{x}^{(2)}$ is selected, there is no need to select $\boldsymbol{x}^{(3)}$

❑ Select instances that cover important features (x_2)

❑ Avoid selecting instances with similar explanations (redundant features)

Submodular Pick for Explaining Models

- Budget B: the number of explanations users are willing to look at in order to understand a model

A set of instances X $\xrightarrow{\text{select}}$ B instances (diverse, representative)

➢ **Submodular pick (SP) algorithm**

Explanation matrix W

	x_1	x_2	x_3	x_4	x_5
$\boldsymbol{x}^{(1)}$	$w_{1,1}$	$w_{1,2}$			
$\boldsymbol{x}^{(2)}$		$w_{2,2}$	$w_{2,3}$		
$\boldsymbol{x}^{(3)}$		$w_{3,2}$	$w_{3,3}$		
$\boldsymbol{x}^{(4)}$		$w_{4,2}$		$w_{4,4}$	
$\boldsymbol{x}^{(5)}$				$w_{5,4}$	$w_{5,5}$

- ❑ Select instances that cover important features (x_2)
- ❑ Avoid selecting instances with similar explanations (redundant features)
- ❑ Select less instances, while covering more features ($\boldsymbol{x}^{(2)}$ and $\boldsymbol{x}^{(5)}$)

Submodular Pick for Explaining Models

- **Submodular pick (SP) algorithm**

Algorithm Submodular pick (SP) algorithm

Require: Instances X, Budget B

for all $\boldsymbol{x}^{(i)}$ in X **do**

$\boldsymbol{w}^{(i)} \leftarrow LIME(\boldsymbol{x}^{(i)})$

Construct the explanation matrix

Submodular Pick for Explaining Models

- **Submodular pick (SP) algorithm**

Algorithm Submodular pick (SP) algorithm

Require: Instances X, Budget B

for all $\boldsymbol{x}^{(i)}$ in X **do**

$\boldsymbol{w}^{(i)} \leftarrow LIME(\boldsymbol{x}^{(i)})$

Construct the explanation matrix

for $j \in \{1, \cdots, N\}$ **do**

$I_j \leftarrow \sqrt{\sum_{i=1}^{|X|} w_{i,j}}$

Compute global feature importance

Submodular Pick for Explaining Models

- **Submodular pick (SP) algorithm**

Algorithm Submodular pick (SP) algorithm

Require: Instances X, Budget B

for all $\boldsymbol{x}^{(i)}$ in X **do**

$\boldsymbol{w}^{(i)} \leftarrow LIME(\boldsymbol{x}^{(i)})$

Construct the explanation matrix

for $j \in \{1, \cdots, N\}$ **do**

$I_j \leftarrow \sqrt{\sum_{i=1}^{|X|} w_{i,j}}$

Compute global feature importance

$V \leftarrow \{\}$

while $|V| < B$ **do**

$i^* = \underset{i}{\operatorname{argmax}}\, c\left(V \cup \boldsymbol{x}^{(i)}, W, I\right)$

$V \leftarrow V \cup \boldsymbol{x}^{(i^*)}$

Return V

Greedily add examples that maximize the coverage gain

$c(V, W, I) = \sum_{j=1}^{N} 1_{[\exists i \in V : w_{i,j} > 0]} I_j$

The coverage computes the total global importance of the features that appear in at least one instance in a set V

SP-LIME

- Compute global feature importance based on local feature importance from LIME

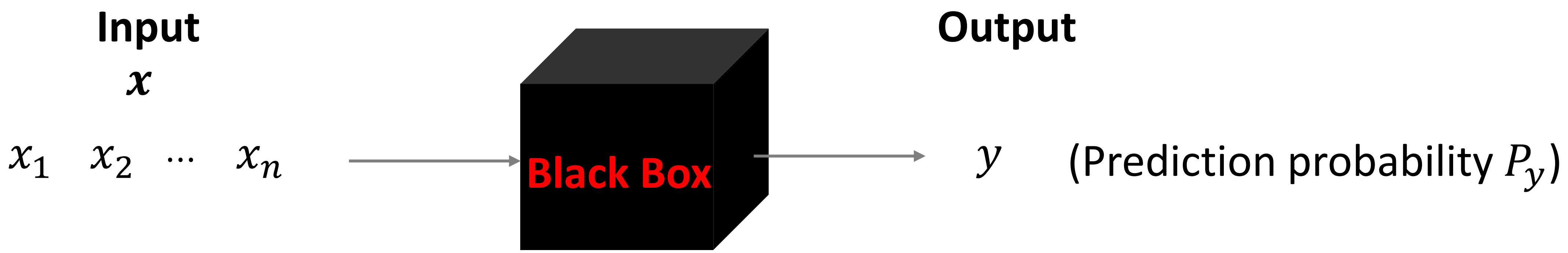
- Provide a global understanding of the model by selecting a set of representative instances

Question?

Perturbation-based methods

- LIME (Ribeiro et al., KDD, 2016)
- SHAP (Lundberg and Lee, NIPS, 2017)

Explaining Black-box Model

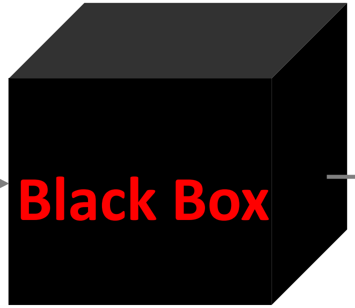


Explaining Black-box Model

Input

$\boldsymbol{x}$

$x_1 \quad x_2 \quad \cdots \quad x_n$



Output

y (Prediction probability P_y)

Importance of x_i

$\cancel{x_1} \quad x_2 \quad \cdots \quad x_n$

$P_y{}'$

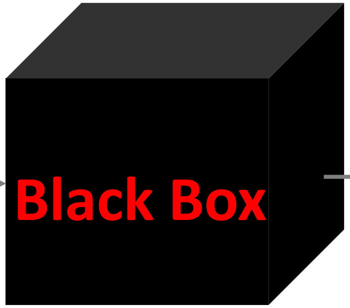
$P_y - P_y{}'$

Explaining Black-box Model

Input

$\boldsymbol{x}$

$x_1 \quad x_2 \quad \cdots \quad x_n$



Output

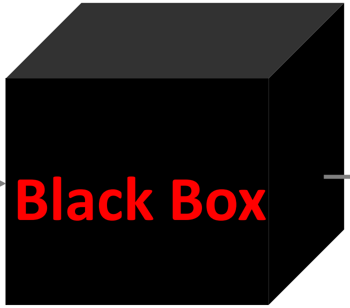
y (Prediction probability P_y)

			Importance of x_i
$\cancel{x_1} \quad x_2 \quad \cdots \quad x_n$		$P_y{}'$	$P_y - P_y{}'$
$x_1 \quad \cancel{x_2} \quad \cdots \quad x_n$		$P_y{}''$	$P_y - P_y{}''$

Explaining Black-box Model

Input

$\boldsymbol{x}$

$x_1 \quad x_2 \quad \cdots \quad x_n$ → 

→ y (Prediction probability P_y)

Output

		Importance of x_i
$\cancel{x_1} \quad x_2 \quad \cdots \quad x_n$	$P_y{}'$	$P_y - P_y{}'$
$x_1 \quad \cancel{x_2} \quad \cdots \quad x_n$	$P_y{}''$	$P_y - P_y{}''$
$\vdots$	$\vdots$	$\vdots$

[Leave-one-out, (Li et al., 2016)]

Leave-one-out

- Sentiment classification

Model prediction: positive

Text	**Confidence**	**Word importance**	
The movie is interesting	0.98		
~~The~~ movie is interesting	0.95	The	0.03
The ~~movie~~ is interesting	0.87	movie	0.11
The movie ~~is~~ interesting	0.96	is	0.02
The movie is ~~interesting~~	0.61	interesting	0.37

Leave-one-out

- Leave **ONE** feature out at each step

Feature importance may be misleading

Text	**Confidence**	**Word importance**	
The movie is interesting and impressive	0.97		
The movie is ~~interesting~~ and impressive	0.95	interesting	0.02
The movie is interesting and ~~impressive~~	0.96	impressive	0.01

Leave-one-out

- Leave **ONE** feature out at each step

Feature importance may be misleading

Text	**Confidence**	**Word importance**	
The movie is interesting and impressive	0.97		
The movie is ~~interesting~~ and impressive	0.95	interesting	0.02
The movie is interesting and ~~impressive~~	0.96	impressive	0.01

Need a better way to quantify feature importance

SHAP

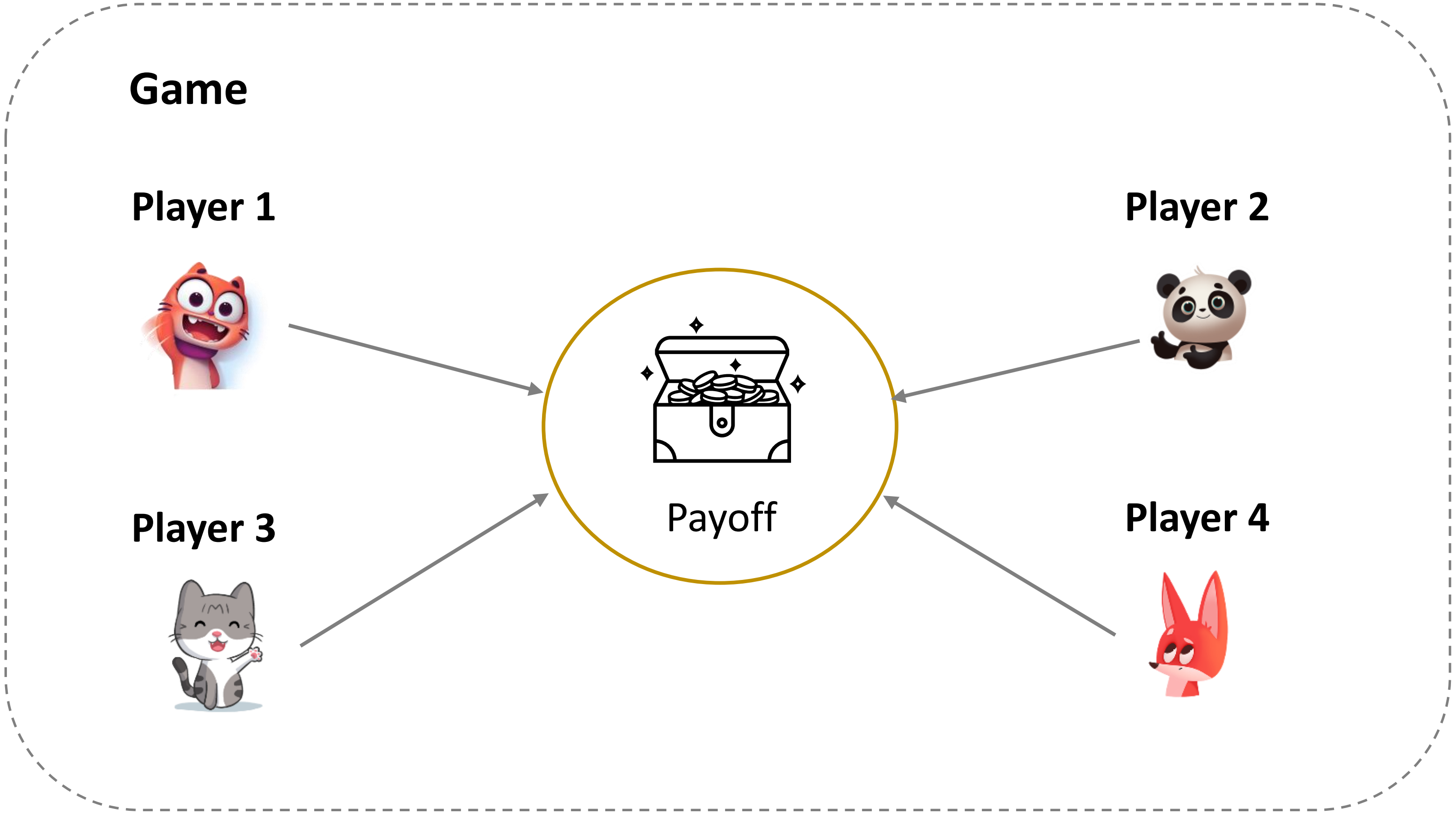
A unified approach to interpreting model predictions

Scott M. Lundberg, Su-In Lee

(NIPS, 2017)

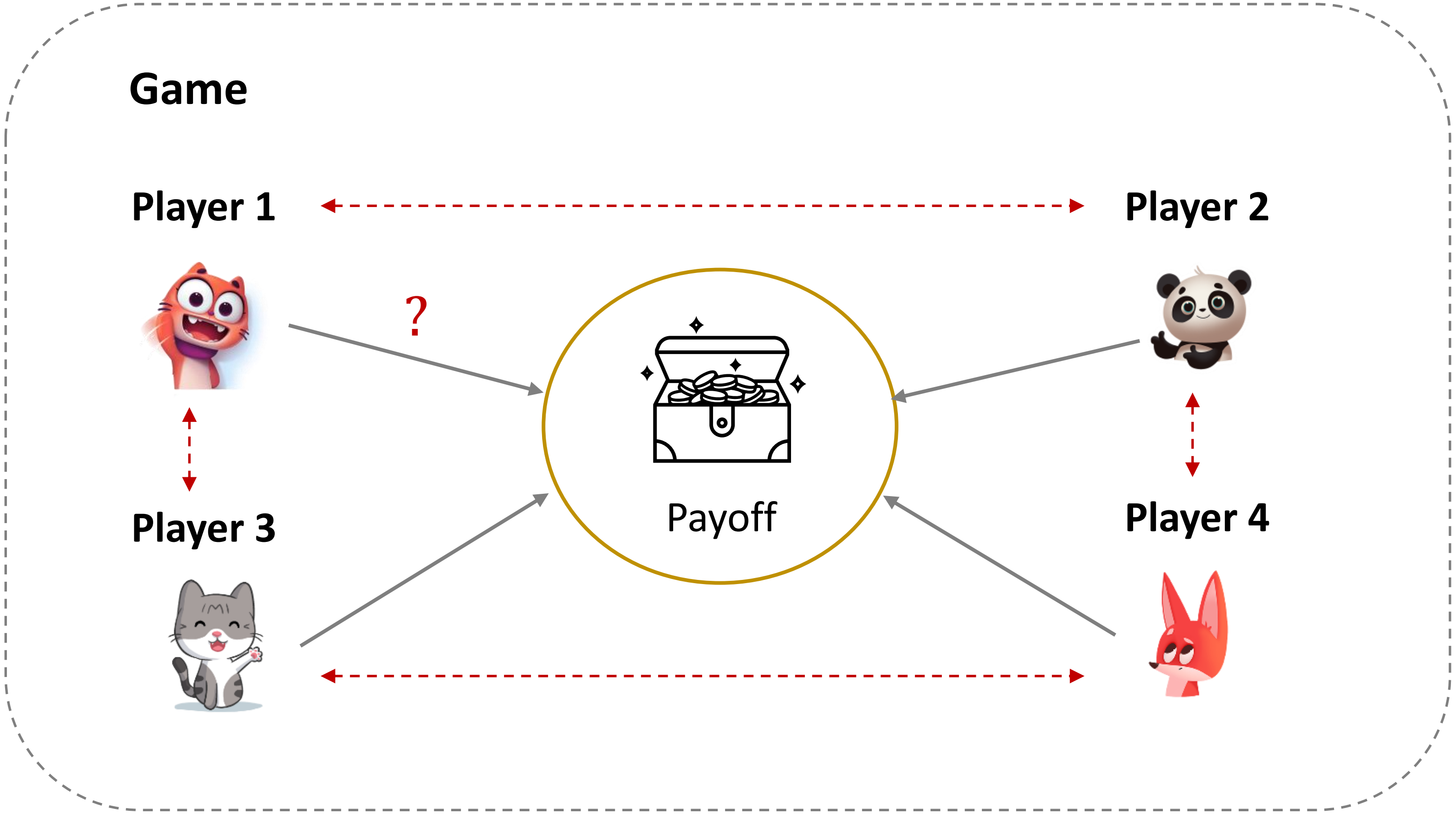
SHAP

- Shapley value [Shapley, 1953]



SHAP

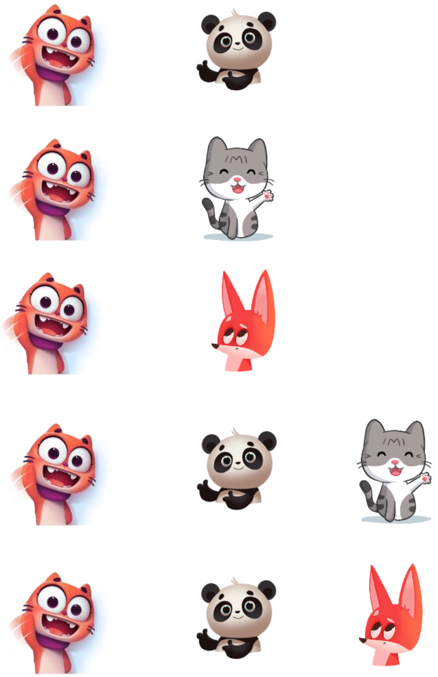
- Shapley value [Shapley, 1953]



SHAP

- Shapley value [Shapley, 1953]

Coalitions



$\vdots$

(2^3)

Payoff

P_1

P_2

P_3

P_4

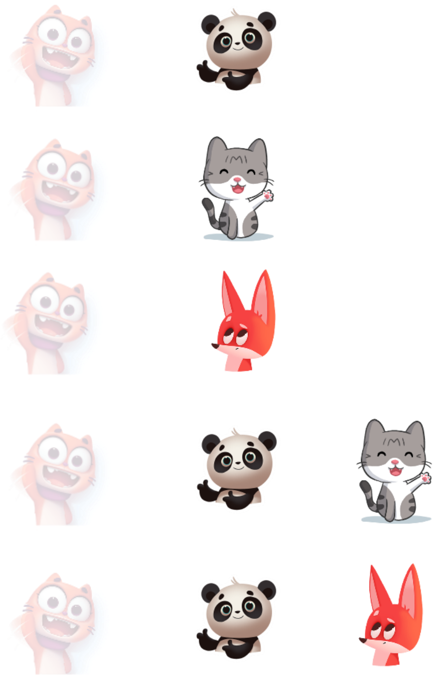
P_5

$\vdots$

SHAP

- Shapley value [Shapley, 1953]

Coalitions



⋮

(2^3)

Payoff

P_1 $\color{red}{P_1'}$

P_2 $\color{red}{P_2'}$

P_3 $\color{red}{P_3'}$

P_4 $\color{red}{P_4'}$

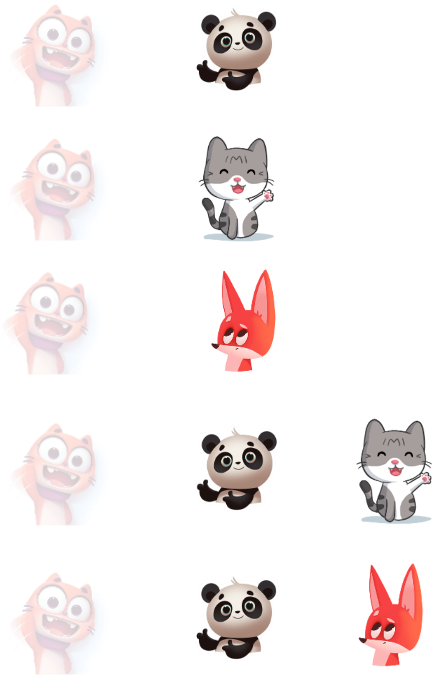
P_5 $\color{red}{P_5'}$

⋮

SHAP

- Shapley value [Shapley, 1953]

Coalitions



⋮

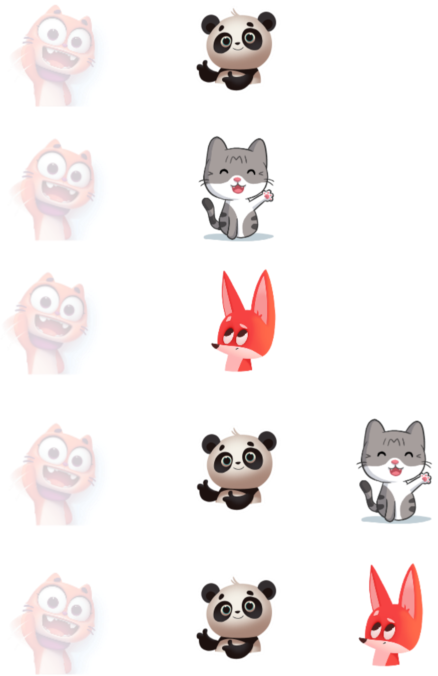
(2^3)

Payoff	**Marginal contribution**
$P_1 - {\color{red}P_1'}$	ΔP_1
$P_2 - {\color{red}P_2'}$	ΔP_2
$P_3 - {\color{red}P_3'}$	ΔP_3
$P_4 - {\color{red}P_4'}$	ΔP_4
$P_5 - {\color{red}P_5'}$	ΔP_5
⋮	

SHAP

- Shapley value [Shapley, 1953]

Coalitions



$\vdots$

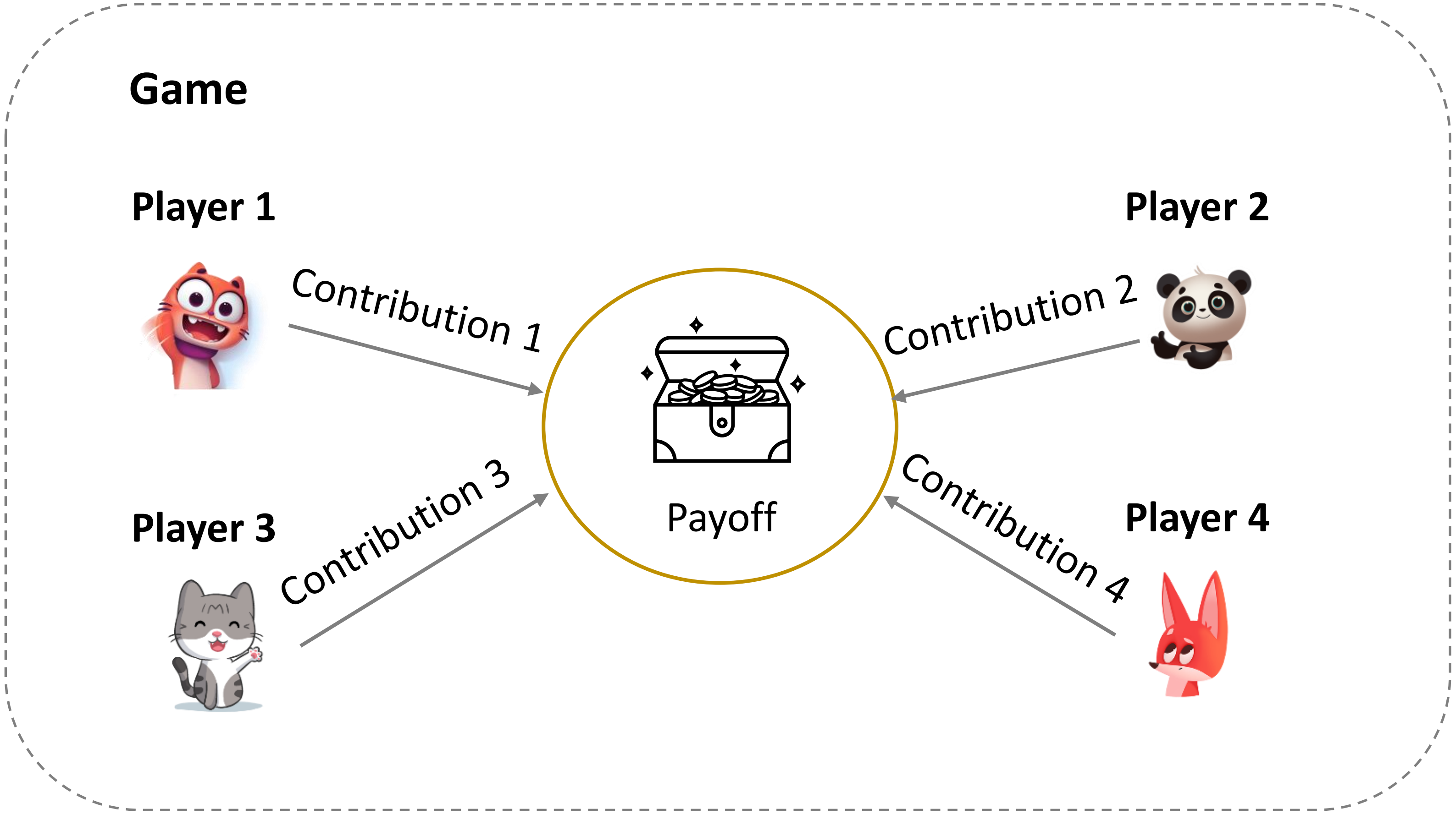
(2^3)

Payoff	**Marginal contribution**
$P_1 - P_1'$	ΔP_1
$P_2 - P_2'$	ΔP_2
$P_3 - P_3'$	ΔP_3
$P_4 - P_4'$	ΔP_4
$P_5 - P_5'$	ΔP_5
$\vdots$	

Contribution$= \sum \Delta P_i$

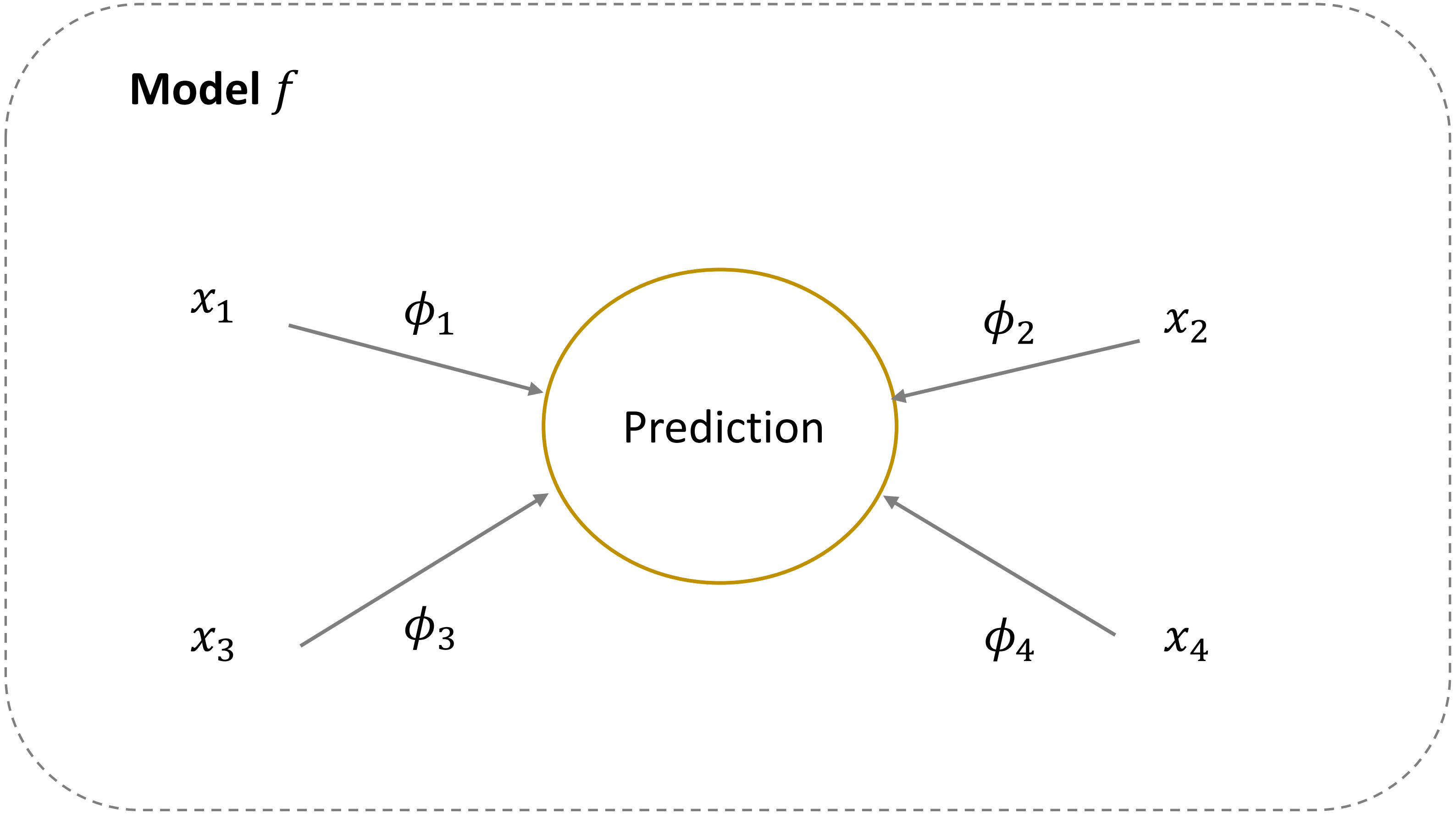
SHAP

- Shapley value [Shapley, 1953]



SHAP

- Shapley value [Shapley, 1953]

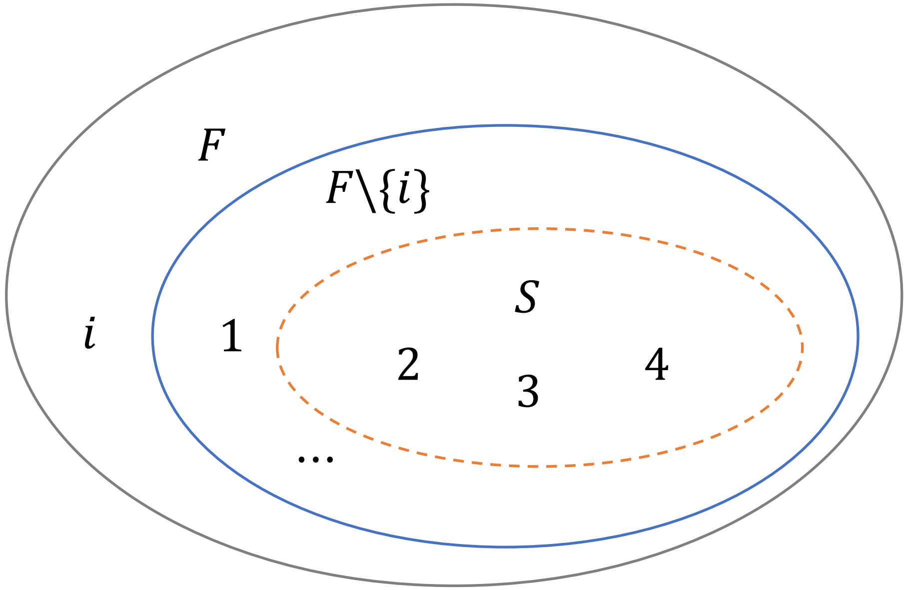


SHAP

- Shapley value [Shapley, 1953]

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S(x_S) \right]$$

Marginal contribution of x_i given S

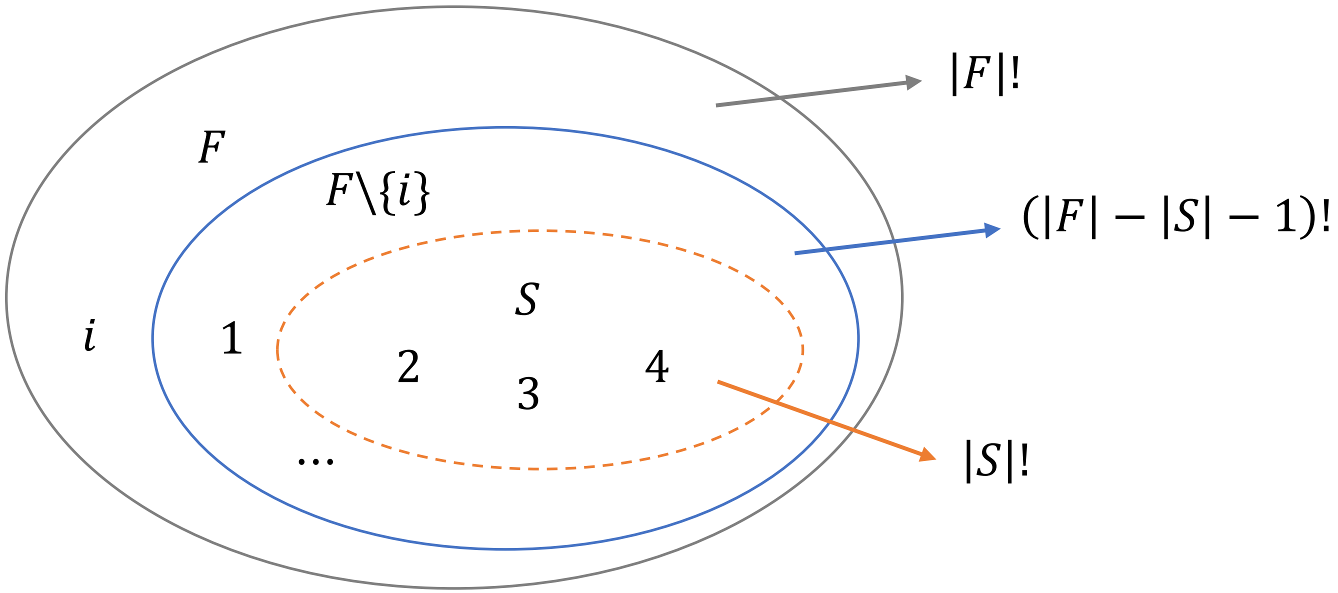


SHAP

- Shapley value [Shapley, 1953]

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

Weighted by the permutations of features



SHAP

- SHapley Additive exPlanation (SHAP)

Additive feature attribution method

$$g(z') \approx f\left(h_x(z')\right)$$

$$g(z') = \phi_0 + \sum_{i=1}^{N} \phi_i {z_i}'$$

$z' \approx x'$ $\underline{x} = h_x(\underline{x'})$

Original input Interpretable input

SHAP

- SHapley Additive exPlanation (SHAP)

Additive feature attribution method

$$g(z') \approx f\left(h_x(z')\right)$$

$$g(z') = \phi_0 + \sum_{i=1}^{N} \phi_i {z_i}'$$

$$z' \approx x' \qquad \underline{x} = h_x(\underline{x'})$$

Original input Interpretable input

LIME is a special case, but not optimal

$$g(z') = \sum_{i=1}^{N} w_i {z_i}'$$

SHAP

- SHapley Additive exPlanation (SHAP)

Additive feature attribution method

$$g(z') \approx f\big(h_x(z')\big)$$

$z' \approx x'$ $\quad$ $\underline{x} = h_x(\underline{x'})$

Original input $\quad$ Interpretable input

$$g(z') = \phi_0 + \sum_{i=1}^{N} \phi_i {z_i}'$$

❑ **Property 1: Local accuracy**

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{N} \phi_i {x_i}'$$

$$\phi_0 = h_x(0)$$

SHAP

- SHapley Additive exPlanation (SHAP)

Additive feature attribution method

$$g(z') \approx f\left(h_x(z')\right)$$

$$g(z') = \phi_0 + \sum_{i=1}^{N} \phi_i {z_i}'$$

$z' \approx x'$ $\quad$ $\underline{x} = h_x(\underline{x'})$

Original input $\quad$ Interpretable input

❑ **Property 2: Missingness**

$$x_i' = 0 \quad \Longrightarrow \quad \phi_i = 0$$

Missingness constrains features missing in the original input to have no attributed impact

SHAP

- SHapley Additive exPlanation (SHAP)

Additive feature attribution method

$$g(z') \approx f\big(h_x(z')\big)$$

$z' \approx x'$ $\underline{x} = h_x(\underline{x'})$

Original input Interpretable input

$$g(z') = \phi_0 + \sum_{i=1}^{N} \phi_i z_i{}'$$

❑ **Property 3: Consistency**

For any two models f_1 and f_2, if $f_1\big(h_x(z')\big) - f_1\big(h_x(\underline{z'\backslash i})\big) \geq f_2\big(h_x(z')\big) - f_2\big(h_x(z'\backslash i)\big)$

$z_i' = 0$

for all inputs $z' \in \{0, 1\}^N$, then $\phi_i(f_1, x) \geq \phi_i(f_2, x)$

SHAP

- SHapley Additive exPlanation (SHAP)

Additive feature attribution method

$$g(z') \approx f\left(h_x(z')\right)$$

$$z' \approx x' \qquad \underline{x} = h_x(\underline{x'})$$

Original input   Interpretable input

$$g(z') = \phi_0 + \sum_{i=1}^{N} \phi_i {z_i}'$$

Only Shapley value satisfies all the three properties

$$\phi_i(f, x) = \sum_{\underline{z'} \subseteq x'} \frac{|z'|!\,(N - |z'| - 1)!}{N!}\left[f\left(h_x(z')\right) - f\left(h_x(z' \backslash i)\right)\right]$$

Contains a subset of non-zero entries in x'

SHAP

- SHapley Additive exPlanation (SHAP)

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(N - |z'| - 1)!}{N!}\left[f\left(h_x(z')\right) - f\left(h_x(z' \backslash i)\right)\right]$$

Challenge

Computational complexity

$O(2^n)$

SHAP

- SHapley Additive exPlanation (SHAP)

Model-agnostic approximations

- Shapley sampling values
- Kernel SHAP

Model-type-specific approximations

- Linear SHAP
- Low-Order SHAP
- Max SHAP
- Deep SHAP

SHAP

- SHapley Additive exPlanation (SHAP)

Model-agnostic approximations

- Shapley sampling values
- Kernel SHAP

Model-type-specific approximations

- Linear SHAP
- Low-Order SHAP
- Max SHAP
- Deep SHAP

Initialize the number of samples M

$\phi_i \longleftarrow 0$

for $m \in \{1, \cdots, M\}$ **do**

Sample $z' \subseteq x'$

$$\phi_i \longleftarrow \phi_i + \frac{|z'|!(N-|z'|-1)!}{N!}\left[f\left(h_x(z')\right) - f\left(h_x(z' \backslash i)\right)\right]$$

SHAP

- SHapley Additive exPlanation (SHAP)

Model-agnostic approximations

- Shapley sampling values
- Kernel SHAP   Linear LIME + Shapley values

Model-type-specific approximations

- Linear SHAP
- Low-Order SHAP
- Max SHAP
- Deep SHAP

The solutions would be consistent with properties 1-3

$$\Omega(g) = 0$$

$$\pi_{x'}(z') = \frac{(N-1)}{(N\ choose\ |z'|)|z'|(N-|z'|)}$$

$$\mathcal{L}(f,g) = \sum \pi_{x'}(z')\left(f\left(h_x(z')\right) - g(z')\right)^2$$

SHAP

- SHapley Additive exPlanation (SHAP)

Model-agnostic approximations

- Shapley sampling values
- Kernel SHAP

Model-type-specific approximations

- Linear SHAP
- Low-Order SHAP
- Max SHAP
- Deep SHAP

Faster model-specific methods

SHAP values can be approximated directly from the model's weight coefficients

Question?

Reference

- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.
- Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Proceedings of the 31st international conference on neural information processing systems*. 2017.
- Li, Jiwei, Will Monroe, and Dan Jurafsky. "Understanding neural networks through representation erasure." *arXiv preprint arXiv:1612.08220* (2016).
- Lloyd S Shapley. 1953. A value for n-person games. Contributions to the Theory of Games, 2(28).