A Formal Approach to Explainability

(by Lior Wolf, Tomer Galanti, and Tamir Hazan)

Phaphontee Yamchote

July 29, 2024

Outline

Outline: Next Topic

Setting

Let

- an input space $\mathcal{X}$
- an output space $\mathcal{Y}$
- a representation space $\mathcal{R}$
- an explanation space G
- a representation function $f : \mathcal{X} \to \mathcal{R}$
- a classifier function $c : \mathcal{R} \to \mathcal{Y}$

We want to explain a model $h = c \circ f$
by an explanation function $g : \mathcal{X} \times \mathcal{Y} \to G$
in terms of $g(x, h(x))$

Outline: Next Topic

Definition (Consistent Representation)

Given a function $\beta : (0, \infty) \to (0, \infty)$ mapping distance in $\mathcal{R}$ into distance in G.
A representation f is β-consistent w.r.t. g if

$$\forall \epsilon > 0 \forall x_1, x_2 \in \mathcal{X}, |g(x_1, h(x_1)) - g(x_2, h(x_2))| \leqslant \epsilon \Rightarrow |f(x_1) - f(x_2)| \leqslant \beta(\epsilon)$$

Representation Space & Explanation Space

Explainability

Definition (Explainable Representation)

Given a function $\gamma : (0, \infty) \to (0, \infty)$ mapping distance in $\mathcal{R}$ into distance in G.
A representation f is γ-explainable w.r.t. g if

$$\forall \epsilon > 0 \forall x_1, x_2 \in \mathcal{X}, |f(x_1) - f(x_2)| \leqslant \epsilon \Rightarrow |g(x_1, h(x_1)) - g(x_2, h(x_2))| \leqslant \gamma(\epsilon)$$

Representation Space & Explanation Space

Explainability

Definition (Explainable Representation)

Given a function $\gamma : (0, \infty) \to (0, \infty)$ mapping distance in $\mathcal{R}$ into distance in G.
A representation f is γ-explainable w.r.t. g if

$$\forall \epsilon > 0 \forall x_1, x_2 \in \mathcal{X}, |f(x_1) - f(x_2)| \leqslant \epsilon \Rightarrow |g(x_1, h(x_1)) - g(x_2, h(x_2))| \leqslant \gamma(\epsilon)$$

Definition (Second-order Explainable Representation)

Given a function $\gamma : (0, \infty) \times (0, \infty) \to (0, \infty)$.
A representation f is second-order γ-explainable w.r.t. g if

$$\forall \epsilon_0 \epsilon_1 > 0 \forall x_1, x_2 \in \mathcal{X}, |f(x_1) - f(x_2)| \leqslant \epsilon_0 \wedge |f_x(x_1) - f_x(x_2)| \leqslant \epsilon_1$$

$$\Downarrow$$

$$|g(x_1, h(x_1)) - g(x_2, h(x_2))| \leqslant \gamma(\epsilon_0, \epsilon_1)$$

Representation Space & Explanation Space

Consistency Recall

$$\forall \epsilon > 0 \forall x_1, x_2 \in \mathcal{X}, |g(x_1, h(x_1)) - g(x_2, h(x_2))| \leqslant \epsilon \Rightarrow |f(x_1) - f(x_2)| \leqslant \beta(\epsilon)$$

- What if the representation of our machine learning model is consistent, i.e. $h(x) = c(f(x))$ where f is consistent?
- Let try: $|h(x_1) - h(x_2)| = |c(f(x_1)) - c(f(x_2))|$.
- What can connect between $|c(f(x_1)) - c(f(x_2))|$ and $|f(x_1) - f(x_2)|$?

Definition (l-Lipschitz continuous)

A function L is l-Lipschitz continuous if

$$\forall x_1, x_2, |F(x_1) - F(x_2)| \leqslant l\,|x_1 - x_2|$$

Representation Space & Explanation Space

Consistency representation and Lipschitz classifier

Theorem (Lipschitz ∘ Consistent is Consistent)

Given a model $h = c \circ f : \mathcal{X} \to \mathcal{Y}$ with an explanation function $g : \mathcal{X} \times \mathcal{Y} \to G$,
if f is β-consistent w.r.t. g and c is l-Lipschitz continuous,
then h is $l\beta$-consistent w.r.t. g.

Let's prove!

Representation Space & Explanation Space

Explainable representation and Lipschitz classifier

$$\forall \epsilon > 0 \forall x_1, x_2 \in \mathcal{X}, |f(x_1) - f(x_2)| \leqslant \epsilon \Rightarrow |g(x_1, h(x_1)) - g(x_2, h(x_2))| \leqslant \gamma(\epsilon)$$

Theorem (upstream function in Lipschitz ∘ Consistent is consistent)

Given a model $h = c \circ (f_2 \circ f_1) : \mathcal{X} \to \mathcal{Y}$ with an explanation function $g : \mathcal{X} \times \mathcal{Y} \to G$,
if f is γ-explainable w.r.t. g and c is l-Lipschitz continuous,
then f_1 is $\hat{\gamma}$-explainable w.r.t. g where $\hat{\gamma}(\epsilon) := \gamma(l\epsilon)$.

Case Study: Image Classification

I still don't understand this topic right now,
it requires background in image processing, which I'm not familiar with

The following theorem states that if our model is of the form $h(x) = \arg\max_{i \in \mathcal{Y}}(m_i^\top \cdot p(x))$ and our EF has the form $g(x, h(x)) = \frac{\partial(m_{h(x)}^\top \cdot p(x))}{\partial x}$, where $p = c \circ f$ such that c, f and the derivative of c are Lipschitz continuous functions, then, f is explainable with respect to g.

THEOREM 4.3. *Let $\mathcal{Y} = [K]$ and $h : \mathbb{R}^n \to \mathcal{Y}$ a model of the form, $h(x) = \arg\max_{i \in \mathcal{Y}} m_i^\top \cdot p(x)$, where $p : \mathbb{R}^n \to \mathbb{R}^d$ and $m_i \in \mathbb{R}^d$, for $i \in [K]$. Let $g(x, h(x)) = \frac{\partial(m_{h(x)}^\top \cdot p(x))}{\partial x}$ be an EF. Assume that for all $i \in [K]$, $p = c \circ f$, such that: c, $\frac{\partial c(x)}{\partial x}$, $\frac{\partial p(x)}{\partial x}$ and f are Lipschitz continuous functions. Additionally, assume that: $\forall i \neq j \in [K], x \in \mathcal{X} : m_i^\top \neq m_j^\top$ and $\forall x \in \mathcal{X} : |p(x)| \geq \Delta$, for some constant $\Delta > 0$. Then, f is second-order $O(\epsilon_0 + \epsilon_1)$-explainable with respect to g.*

Outline: Next Topic

Properties of Explanation Functions

Validity

Definition (Valid Explanation Functions)

Given a fixed constant $\epsilon > 0$ and $x \sim \mathcal{D}$.
An explanation function g is ϵ-valid w.r.t. a model h
if there is a function $t : G \to \mathcal{Y}$ s.t.

$$\mathbb{E}_{x \sim \mathcal{D}} \left[\ell \left(t \left(g \left(x, h \left(x \right) \right) \right), h \left(x \right) \right) \right] \leqslant \epsilon,$$

where ℓ is a loss function.

Properties of Explanation Functions

Completeness

Definition (Complete Explanation Functions)

Given a fixed constant $\alpha, \epsilon > 0$ and $x \sim \mathcal{D}$.
An explanation function g is (ϵ, α)-complete w.r.t. a model h if every $\bar{g} : \mathcal{X} \to \mathbb{R}^d$ s.t. $I(g(x, h(x)); \bar{g}(x)) \leqslant \epsilon$ and every $s : \mathbb{R}^d \to \mathcal{Y}$

$$\mathbb{E}_{x \sim \mathcal{D}} \left[\ell \left(s(\bar{g}(x)), h(x) \right) \right] \geqslant \alpha,$$

where ℓ is a loss function.

Properties of Explanation Functions

if we are able to recover $h(x)$ from $\bar{g}(x)$ and from $g(x, h(x))$, then, $\bar{g}(x)$ and $g(x, h(x))$ cannot be independent of each other.

Theorem (Valid $\Rightarrow$ Complete)

Let $h : \mathbb{R}^n \to \mathcal{Y}$ be a model, $g : \mathcal{Z} \to G$ an ϵ_0-valid EF for some constant $\epsilon_0 \in (0, 0.5)$ and $x \sim D$.
Assume that $Y = \{\pm 1\}$ and denote, $p := \mathbb{P}[h(x) = 1]$.
Then, g is (ϵ, α)-complete with respect to h, with $\alpha := \frac{\sqrt{1+H(p)(H(p)-\epsilon-2\sqrt{\epsilon_0})}-1}{H(p)}$ and any $\epsilon > 0$ that satisfies, $H(p) > \epsilon + 2\sqrt{\epsilon_0}$.
In particular, if $p = 1/2$, we have: $\alpha = \sqrt{2 - \epsilon - 2\sqrt{\epsilon_0}} - 1$.

Need a lot of lemmas from other works

Outline: Next Topic

Intersection and Union of RVs

Definition

Let $x \sim \mathcal{D}$ on $\mathcal{X}$, a constant $\epsilon > 0$ and
given two functions $f_1 : \mathcal{X} \to \mathcal{X}_1$ and $f_2 : \mathcal{X} \to \mathcal{X}_2$.
If there are two invertible functions $r_1 : \mathcal{X}_1 \to \mathcal{V}_1$ and $r_2 : \mathcal{X}_2 \to \mathcal{V}_2$ such that

$$r_1(f_1(x)) = (e_1(x), u(x))$$
$$r_2(f_2(x)) = (e_2(x), u(x)),$$

where mutual information $I(e_i(x); f_j(x)) \leqslant \epsilon$ for $i \neq j \in \{1, 2\}$

- the RV $u(x)$ is called ϵ-intersection of f_1 and f_2
- the RV $(e_1(x), u(x), e_2(x))$ is ϵ-union of f_1 and f_2

(Still don't get its idea why define like this)
(this work as well proved that they are unique up to invertible transformation)