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

Representation Space & Explanation Space

Consistency

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 $\underline{\beta\text{-consistent}}$ 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

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

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 <u>γ-explainable</u> 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 <u>second-order γ-explainable</u> 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

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

Theorem (Lipschitz $\circ$ 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 $\circ$ 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

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 $\underline{\epsilon\text{-valid}}$ 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.

Outline: Next Topic

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 $\underline{(\epsilon, \alpha)\text{-complete}}$ 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

content...