A Formal Approach to Explainability

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

Phaphontee Yamchote

July 29, 2024

Outline

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 $g(x, h(x))$

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 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 <u>γ-explainable</u> w.r.t. g if

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

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 <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

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

Representation Space & Explanation Space

Consistency representation and Lipschitz classifier

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!