Coverage-based explanations for classifiers

Martin Cooper   Leila Amgoud

IRIT, CNRS, University of Toulouse III, Toulouse

ARDM Workshop — June 2022

Outline

- New definition of prime-implicant explanations in the presence of constraints
- Complexity is a real issue for neural network classifiers, so we can use the dataset or a sample rather than an exhaustive search over the whole of feature space. Dataset-based explanations provide a trade-off between efficiency and consistency
- We now have a catalogue of different types of explanations with different complexities and different formal guarantees

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE

Prime-implicant abductive explanations

Classifiers

A *classifier* is a function $\kappa : \mathbb{F} \to \mathcal{K}$, where $\mathbb{F}$ is feature-space and $\mathcal{K}$ a set of classes.

Examples:

1. Should we accept a student on a Master course?
2. Should we prescribe this medecine for a patient?
3. Should the bank grant a loan to a customer?
4. Who should be president/prime minister?

Explaning decisions: $\kappa, \mathbf{v}, c, \mathcal{C} \longrightarrow E$
Find a set of features which explains the decision $\kappa(\mathbf{v}) = c$, knowing that feature vectors are subject to the constraints $\mathcal{C}$.

Constraints on feature space

There are often constraints between features:

- physical constraints
- functional dependencies
- constraints learnt from analysis of data

Example

- years of work $<$ age
- pregnant $\rightarrow$ woman
- social security number $\rightarrow$ surname
- Computer Science degree $\rightarrow$ has studied Programming
- California always votes Democratic

Abductive explanations under constraints

A feature vector **v** can be viewed as a set of literals.
An explanation can be viewed as a set of literals/a set of features/a predicate.

Definition

A weak abductive explanation (*weak AXp*) E of $\kappa(\mathbf{v}){=}c$ is a subset of **v** which is sufficient to guarantee the same decision. Viewing E as a predicate,

$$\forall x \in \mathbb{F} \;\; (\; E(x) \wedge \mathcal{C}(x) \rightarrow \kappa(x) = c \;)$$

An *AXp* is a subset-minimal weak AXp.

Example (pregnant woman)

$\kappa(x_1, x_2) = x_1 \wedge x_2 \qquad \mathbf{v} = (1, 1) \qquad \mathcal{C}\colon x_2 \rightarrow x_1$
There are 2 weak AXp's: $\{x_2\}$, $\{x_1, x_2\}$
and 1 AXp: $\{x_2\}$.

Abductive explanations under constraints

A feature vector **v** can be viewed as a set of literals.
An explanation can be viewed as a set of literals/a set of features/a predicate.

Definition

A weak abductive explanation (*weak AXp*) E of $\kappa(\mathbf{v}){=}c$ is a subset of **v** which is sufficient to guarantee the same decision. Viewing E as a predicate,

$$\forall x \in \mathbb{F} \ \ (\ E(x) \wedge \mathcal{C}(x) \rightarrow \kappa(x) = c \ )$$

An *AXp* is a subset-minimal weak AXp.

Example (Master degree $\rightarrow$ Bachelor degree)

$\kappa(x_1, x_2) = x_1 \qquad \mathbf{v} = (1, 1) \qquad \mathcal{C}\colon x_2 \rightarrow x_1$
There are 3 weak AXp's: $\{x_1\}$, $\{x_2\}$, $\{x_1, x_2\}$
and 2 AXp's: $\{x_1\}$, $\{x_2\}$.

Applying constraints in the definition of prime implicant

Example (pregnant woman)

$\kappa(x_1, x_2) = x_1 \land x_2 \qquad \mathbf{v} = (1, 1) \qquad \mathcal{C}\colon x_2 \to x_1$
There are 2 weak AXp's: $\{x_2\}$, $\{x_1, x_2\}$
and 1 AXp: $\{x_2\}$.

Applying constraints allows us to reduce the size of an AXp.

Example (Master degree $\to$ Bachelor degree)

$\kappa(x_1, x_2) = x_1 \qquad \mathbf{v} = (1, 1) \qquad \mathcal{C}\colon x_2 \to x_1$
There are 3 weak AXp's: $\{x_1\}$, $\{x_2\}$, $\{x_1, x_2\}$
and 2 AXp's: $\{x_1\}$, $\{x_2\}$.

The AXp $\{x_2\}$ is redundant. We can eliminate this redundancy by also applying constraints in the definition of prime implicant.

Prime-implicant explanations under constraints

E_1 *subsumes* E_2 if $E_2 \wedge \mathcal{C} \rightarrow E_1$ (where $\mathcal{C}$ are the constraints).
Alternative definition: Define the *coverage* of E to be

$$cov(E) = \{x \mid E(x) \wedge \mathcal{C}(x) \wedge (\kappa(x) = c)\}.$$

Then E_1 subsumes E_2 if $cov(E_2) \subseteq cov(E_1)$.

E_1 *strictly subsumes* E_2 if E_1 subsumes E_2 but E_2 does not subsume E_1.

Definition

A coverage-based prime-implicant explanation (*CPI-Xp*) is a weak AXp not strictly subsumed by any other weak AXp.

Example (Master degree $\rightarrow$ Bachelor degree)

$\kappa(x_1, x_2) = x_1$ $\mathbf{v} = (1, 1)$ $\mathcal{C}$: $x_2 \rightarrow x_1$
The only CPI-Xp is $\{x_1\}$, since $x_2 \rightarrow x_1$ but $x_1 \not\rightarrow x_2$.

Student example

Example

A student is accepted on a CS Masters course if $\kappa = 1$, where

$$\kappa = (CS \vee M \vee EE) \wedge (X \geq 60 \vee W \geq 1) \wedge (P \vee A)$$

where CS, M, EE indicates whether they have a degree in CS, Maths, EEng; X is the final exam mark, W is years of work experience; P, A indicate whether they have taken classes in Programming, Algorithmics.

Constraints $\mathcal{C}$:

- $CS \;\rightarrow\; (P \wedge A)$
- $(X \geq 60 \wedge P \wedge A) \;\rightarrow\; (CS \vee M \vee EE)$

Abductive and prime-implicant explanations

Definition

An abductive explanation (*AXp*) is a subset-minimal set of features that are sufficient to explain the decision $\kappa(v) = c$.

Example

The AXp's of $\kappa(1, 0, 0, 65, 0, 1, 1) = 1$ are $\{CS, X\}$, $\{X, P, A\}$

Abductive and prime-implicant explanations

Definition

An abductive explanation (*AXp*) is a subset-minimal set of features that are sufficient to explain the decision $\kappa(v) = c$.

Example

The AXp's of $\kappa(1, 0, 0, 65, 0, 1, 1) = 1$ are $\{CS, X\}$, $\{X, P, A\}$

Definition

A coverage-based prime-implicant explanation (*CPI-Xp*) is a weak AXp not strictly subsumed by any other weak AXp.

Example

The only CPI-Xp of $\kappa(1, 0, 0, 65, 0, 1, 1) = 1$ is $\{X, P, A\}$

Explanation	Complexity of testing	Complexity of finding one
AXp	co-NP-complete	$\mathrm{FP}^{\mathrm{NP}}$
CPI-Xp	Π_2^P-complete	$\mathrm{FP}^{\Sigma_2^{\mathrm{P}}}$

We assume a white box, i.e. κ is an arbitrary but *known* function. $\mathrm{FP}^{\mathcal{L}}$ is the class of function problems that can be solved by a polynomial number of calls to an oracle for the language $\mathcal{L}$.

Optimal abductive explanations

There are two criteria for choosing an optimal AXp/CPI-Xp:

- smallest explanation
- maximum coverage

Explanation	Complexity of testing	Complexity of finding one
smallest AXp max-coverage AXp	Π_2^P-complete $\#$P-hard	$\mathrm{FP}^{\Sigma_2^P}$ $\mathrm{FP}^{\mathrm{NP}^{\#\mathrm{P}}}$
smallest CPI-Xp	Π_2^P-hard	$\mathrm{FP}^{\Sigma_3^P}$

Dataset-based explanations

Dataset-based explanations

If κ is a *black-box function*, then testing whether E is an AXp requires exhaustive search which is prohibitively expensive.
$\Rightarrow$ dataset-based explanations

Let $\mathcal{T}$ be the dataset. It can be the actual training data or a random sample of feature space (possibly of points close to $\mathbf{v}$). We may filter the training data so that we only keep points where the training data agrees with the model κ.
For technical reasons, we assume $\mathbf{v} \in \mathcal{T}$ and that all vectors in $\mathcal{T}$ satisfy the constraints $\mathcal{C}$.

Definition

Definitions of the dataset versions of AXp and CPI-Xp (*d-AXp*, *d-CPI-Xp*) are obtained by replacing the constraints $\mathcal{C}$ by $\mathcal{T}$ i.e. assuming (wrongly) that the only possible feature vectors are those in the dataset.

Complexity of testing/finding d-AXp's/d-CPI-Xp's

Explanation	Complexity of testing	Complexity of finding one
d-AXp smallest d-AXp max-coverage d-AXp	$O(mn^2)$ co-NP-complete co-NP-complete	$O(mn^2)$ $\mathrm{FP}^{\mathrm{NP}}$ $\mathrm{FP}^{\mathrm{NP}}$
d-CPI-Xp smallest d-CPI-Xp	$O(m^2 n)$ co-NP-complete	$O(m^2 n^2)$ $\mathrm{FP}^{\mathrm{NP}}$

where $m = |\mathcal{T}|$ and n is the number of features.

Properties of explanations

Properties of explanations

Definition

$\mathbb{F}[\mathcal{C}]$ denotes the set of feature vectors x that satisfy $\mathcal{C}$.

Let $\mathbf{E}(\mathbf{v})$ be the set of explanations of $\kappa(\mathbf{v}) = c$. We can define the following properties of $\mathbf{E}$.

- (Consistency) For any $v \in \mathbb{F}[\mathcal{C}]$, each $E \in \mathbf{E}(v)$ satisfies the constraints $\mathcal{C}$.
- (Coherence) For all $v, v' \in \mathbb{F}[\mathcal{C}]$ s.t. $\kappa(v) \neq \kappa(v')$, $\forall E \in \mathbf{E}(v), \forall E' \in \mathbf{E}(v'), \nexists v'' \in \mathbb{F}[\mathcal{C}]$ s.t. $(E \cup E')(v'')$.
- (Irreducibility) For any $v \in \mathbb{F}[\mathcal{C}]$, $\forall E \in \mathbf{E}(v)$, $\forall \ell \in E$, $\exists v' \in \mathbb{F}[\mathcal{C}]$ such that $\kappa(v') \neq \kappa(v)$ and $(E \setminus \{\ell\})(v')$.
- (Irredundance) For any $v \in \mathbb{F}[\mathcal{C}]$, $\forall E, E' \in \mathbf{E}(v)$, $E \not\approx E'$, where $E \approx E'$ if they subsume each other.

Properties satisfied by each explanation

	AXp	CPI-Xp	d-AXp	d-CPI-Xp
Consistency	•	•	•	•
Coherence	•	•		
Irreducibility	•		•	•
Irredundance				

• means the property is satisfied

Examples

Example (of incoherence of dataset-based explanations)

- A mouse is a mammal because it milks its young
- An eagle is not a mammal because it lays eggs
- **but** a platypus ($\notin$ dataset) milks its young and lays eggs!

Example (of reducibility of CPI-Xp's)

In the student example, if we have the constraint $CS \leftrightarrow P \wedge A$ then the explanations $\{CS, X\}$, $\{X, P, A\}$ and $\{CS, X, P, A\}$ are all equivalent (they have the same coverage) **but** $\{CS, X, P, A\}$ is reducible (i.e. not subset-minimal).

Example (of redundance of AXp's (and CPI-Xp's))

In the same student example, $\{CS, X\}$, $\{X, P, A\}$ are equivalent, hence listing them both is redundant.

Properties satisfied by each explanation

Definition

A preferred coverage-based PI-explanation (*pCPI-Xp*) is a **representative** of an equivalence class of CPI-Xp's which is **minimal** for inclusion.

	AXp	CPI-Xp	pCPI-Xp	d-AXp	d-CPI-Xp
Consistency	•	•	•	•	•
Coherence	•	•	•		
Irreducibility	•		•	•	•
Irredundance			•		

Complexities for testing and finding pCPI-Xp and CPI-Xp's coincide.

Conclusion

- New definition of prime-implicant explanations in the presence of constraints, **but** this increases complexity.
- Complexity is a real issue for black-box classifiers, so we can search over a dataset rather than exhaustively over the whole of feature space, **but** this can lead to incoherent pairs of explanations.
- We have a catalogue of different types of explanations with different complexities and different formal guarantees.
- Dataset-based explanations provide a trade-off between efficiency and coherence.
- pCPI-Xp's satisfy all the desired properties but are expensive to find.