EXPLAINABLE AI

AI BASED ALGORITHMS ARE ALGORITHMS THAT USE AI MODELS } HAVE BECOME VERY WIDESPREAD ↝ DEAL WITH A LOT OF COMPLEX PROBLEMS

BUT

IF WE DON'T KNOW HOW THEY WORK WE CAN'T USE THEM IN CRITICAL SYSTEMS (AS BLACK BOXES)

↝ WHY? ↝ BECAUSE MODELS ARE NOT "ALGORITHMS", THEIR ANSWER IS NOT PERFECT

≈ PROBABILISTIC SAT SOLVER

EXPLAINABLE AI TRIES TO OPEN AND DESCRIBE THIS BOX ↝ well, WHITE BOXES EXIST TOO → DECISION TREES, → KNN, → LOGISTIC REGRESSION ...

IN PARTICULAR, GIVEN A MODEL XAI TRIES TO IMPROVE:

- **TRANSPARENCY**
  - SIMULATABILITY ↝ How easily can we simulate it with pen & paper?
  - DECOMPOSABILITY ↝ Do we have a clear understanding of how the aspects of the model influence it?
  - ALGORITHMIC TRANSPARENCY ↝ Can we understand how the output was produced? VIA MATHS
- **INTERPRETABILITY** ↝ Does the model provide an explanation of the process + output justification
  ≈ **EXPLAINABILITY** ↝ INTERPRETABILITY BUT NOT FROM AN EXPERT
- **TRUSTWORTHINESS/CONFIDENCE** ↝ Given the model and the explainable aspects, still, how much do users trust it?

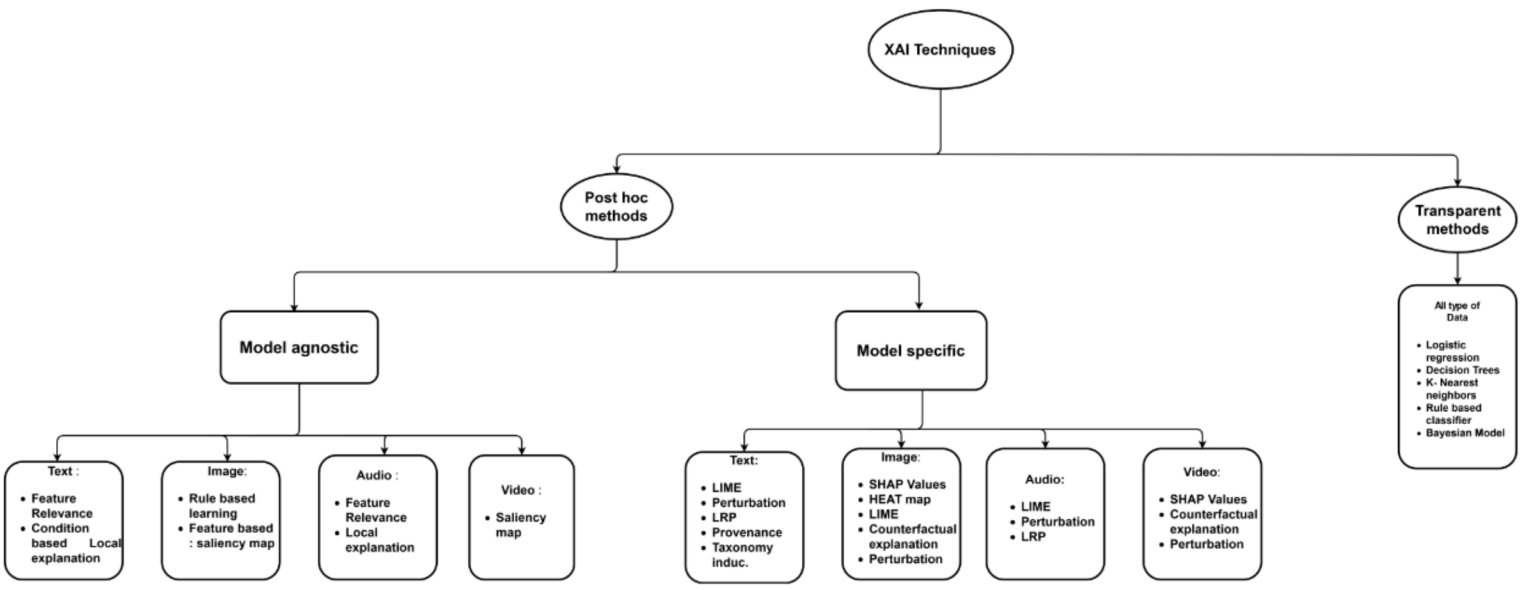
So, BLACK

BLACK BOXES EXIST, BUT ALSO WHITE BOXES DO

'EXPLAINABLE BY DEFAULT MODELS' ↝ WITH THESE, NOT MUCH else is needed, we ALREADY HAVE TRANSPARENCY

⇓

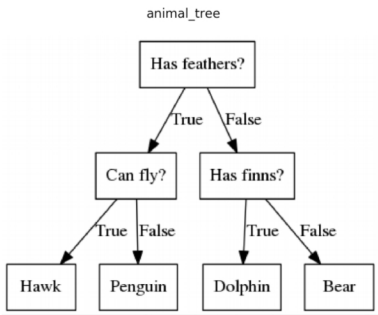
XAI TAXONOMY

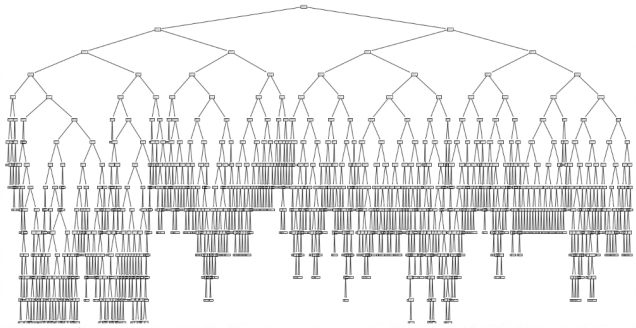


FIRST DISTINCTION

- POST HOC METHODS
- TRANSPARENT METHODS → WHITE BOXES → THEY HAVE TRANSPARENCY, BUT TRANSPARENCY ≠ EXPLAINABLE
  - KNN, TREES, LOGISTIC REGRESSION...

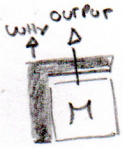
(AND WHEN THE MODEL GROWS SCORE TRANSPARENCY MAY BE LOST)





IN MANY SCENARIOS THEY MAY LEAD TO LOW ACCURACY/OVERFIT...

SOLUTION, USE NEURAL NETS, BUT THEY ARE BLACK BOXES



SO, WE 'ATTACH' POST HOC METHODS

ANOTHER CLASSIFICATION

- MODEL SPECIFIC
  - WE AIM TO EXPLAIN A SPECIFIC CLASS OF MODELS (ALSO LEVERAGING THEIR STRUCTURE)
- MODEL AGNOSTIC
  - MORE GENERAL, ONE OF THE THINGS WE CAN DO IS ANALYZE INPUT/OUTPUT PAIRS
    - AN EXAMPLE:
      - LIME (LOCAL INTERPRETABLE MODEL-AGN. EXPLANATIONS)
        - SEE AUTOMATIC DIFFERENTIATION (... PROGRAM DERIVATIVE)

AN ORTHOGONAL SUBDIVISION

EXPLANATIONS ARE

- LOCAL → WE EXPLAIN WHY A SAMPLE GIVES SUCH RESULTS } DON'T SCALE IN GENERAL
- GLOBAL → WE EXPLAIN THE WHOLE MODEL OR CLASSES (EG: IN CLASSIFICATION)

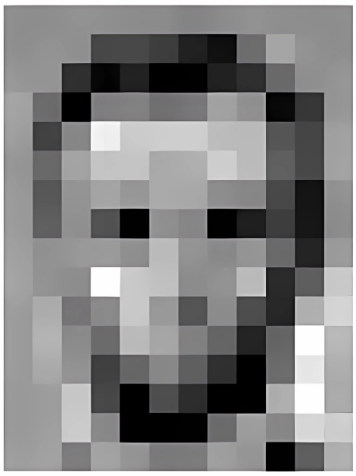
THOUGH EXPLAINING THE OUTPUT IN INPUT TERMS MAY NOT BE SUFFICIENT

⇓

THE INPUT FEATURES MAY NOT BE HUMAN UNDERSTANDABLE

USE → CONCEPT BASED METHODS/NETWORKS

157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	191	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
196	206	123	207	177	121	123	200	175	13	96	218

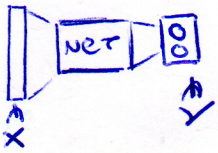


157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	216
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

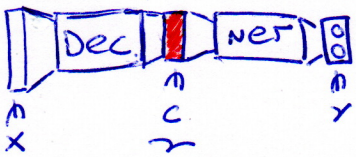
CONCEPT BASED METHODS → THEY TAKE AS INPUT CONCEPTS (HIGHER ABSTRACTION) → SO THEY CAN EXPLAIN THE RESULT IN HUMAN MEANINGFUL WAY → CONCEPT = HIGH LEVEL FEATURE

CONCEPT BASED CLASSIFIERS → IN A CLASSIFICATION TASK WE HAVE A CLASSIFIER TAKING CONCEPTS AS INPUTS → I

→ INSTEAD OF $f: X \to Y$



WE FIRST USE A DECODER/ FEATURE EXTRACTOR $g: X \to C$ AND THEN $h: C \to Y$



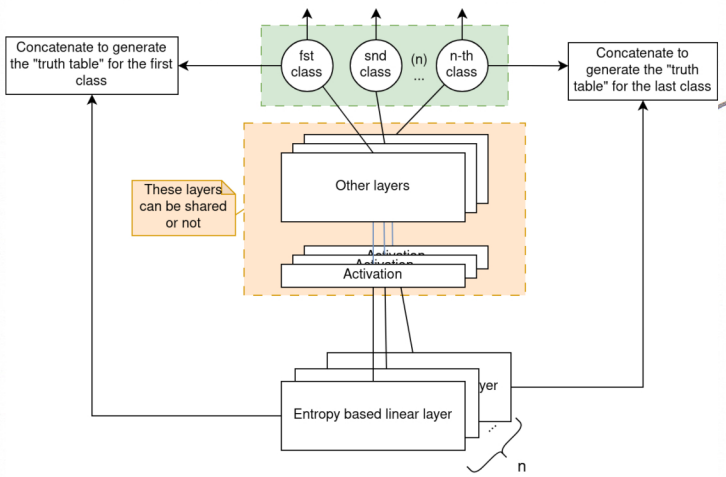
IN GENERIC MUCH SMALLER THAN AN ELEMENT IN X

ENTROPY BASED EXPLANATION OF N.N.

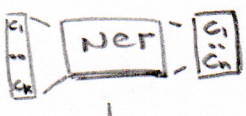
OF A CONCEPT BASED CLASSIFIER...

→ CONTEXT: CLASSIFICATION TASK, n CLASSES, K CONCEPTS

THE MODEL



→ IT RECALLS A CLASSIC CLASSIFIER



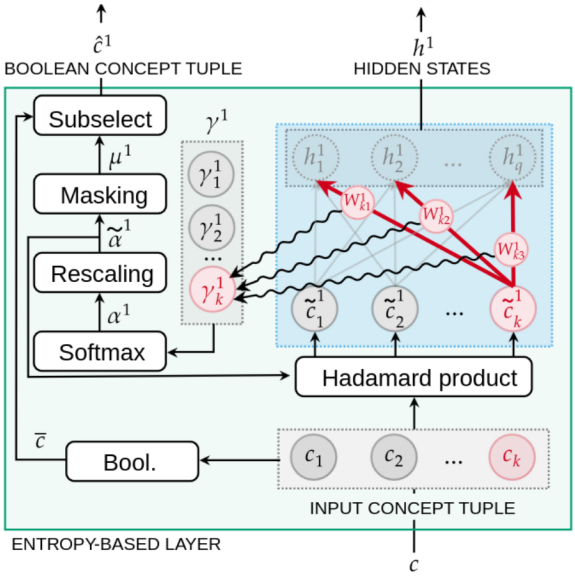
→ NOW WE HAVE n TIMES MORE INPUTS THOUGH WEIGHTS CAN BE SHARED

THOUGH NOW WE HAVE "n" COMPONENTS CONCATENATING BOOLEAN CONCEPT TUPLES AND THE OUTPUT OF A SPECIFIC OUTPUT NEURON

ONE OUTPUT OF THE ENTROPY LAYER

THEY DRIVE THE MODEL EXPLAINABILITY

↳ WE CREATE ONE 'TRUTH TABLE' FOR EACH CLASS



GOAL:

MAKE THE MODEL CHOOSE A SMALL NO. OF CONCEPTS TO DESCRIBE EACH CLASS AND PROVIDE AN EXPLANATION

THESE LAYERS PRODUCE TWO OUTPUTS:

- BOOLEAN CONCEPT TUPLE
- HIDDEN STATE (PASSED TO THE REST OF THE NET)

THESE n LAYERS CHARACTERIZE THE FIRST LAYER OF THE NETWORK

THEY ARE LINEAR AND INDEPENDENT SO WE NEED TO STORE A WEIGHT MATRIX AND A BIAS VECTOR

W^i b^i

TO COMPUTE SUCH OUTPUTS WE NEED TO COMPUTE THE 'ATTENTION' EACH CLASS i HAS TOWARDS EACH CONCEPT j

$$\gamma_j^i = \|W_j^i\|_1$$

YES, IT'S A VECTOR

THEN $\forall$ CLASS WE NORMALIZE THEM, GETTING A PROBABILITY VECTOR

$$\alpha_j^i = \frac{e^{\gamma_j^i/\tau}}{\sum_{l=1}^{k} e^{\gamma_l^i/\tau}} \quad \text{with } \tau \in \mathbb{R}^+$$

τ: TEMPERATURE (HOW TO WEIGHT THE CONCEPT DIFFERENCE)

$\tau \to \infty$ ALL WEIGHT THE SAME

$\tau \to 0^+$ ONE MUCH MORE RELEVANT

RESCALING FOR NUMERICAL REASONS

$$\tilde{\alpha}_j^i = \frac{\alpha_j^i}{\max_u \alpha_u^i}$$

"THE ATTENTION VALUES FOR A CLASS"

WITH THESE VALUES WE COMPUTE THE HADAMARD PRODUCT WITH THE INPUT

$\tilde{c}^i = c \odot \tilde{\alpha}^i$

THEN WE COMPUTE THE LINEAR FUNCTION AND THE REST OF THE NETWORK CARRIES ON

$(h^i = W^i \tilde{c}^i + b^i)$

e.g.: $[1, 50, 49]$

$\tau = 1$ (SOFTMAX) $\to [\sim 0, 0.731, 0.268]$

$\tau = 0.001 \to [\sim 0, \sim 1, \sim 0]$

$\tau = 1000 \to [0.322, 0.338, 0.338]$

THE BOOLEAN CONCEPT TUPLE?

↳ WELL, DEFINE A FUNCTION $\mathbb{I}_{\gg\varepsilon}(x): \mathbb{R}^m \to \{0,1\}^m$ (A FILTER) THAT GIVEN x, RETURNS A VECTOR IN WHICH EVERY l IS 0 IF EVERY l IN x WAS $< \varepsilon$, 1 ELSE

eg: $\mathbb{I}_{\geq 0.2}([0.1, 0.05, 0.85]) = [0, 0, 1]$

LASTLY, TO GET THE BOOLEAN INTERPRETATION FOR C

$\hat{C}^i := \xi(\mathbb{I}_{\geq\varepsilon}(c), \mathbb{I}_{\geq\varepsilon}(\tilde{\alpha}^i))$

- THIS SHOULD BE A X1
- SUBSELECT
- SELECT THE MOST RELEVANT CONCEPTS

THERE'S ONE $\forall$ CLASS (THE WEIGHTS ARE ≠)

WE USE THESE TO SELECT THE COMPONENTS OF THE FIRST ARGUMENT (0/1) WHOSE POSITION CORRESPOND TO A 1 IN THE SECOND ARGUMENT

IT'S A SORT OF MASK

eg: $\xi([0,1,1], [0,0,1]) = [0,0,1]$

THE INPUT HAS STRONG 2ND & 3RD CONCEPTS

BUT FOR THIS CLASS ONLY THE THIRD IS RELEVANT

WE DON'T CARE ABOUT EVEN IF IT WAS HIGH

DURING TRAINING WE COMPUTE THE WEIGHTS

↳ AND WE COMPUTE THE 'TRUTH' TABLES → THIS I 'GUESS' ONLY AT THE END... → SUPPOSIZIONE

HOW? ↳ GIVEN A SAMPLE C, WE CALCULATE $\hat{C}^i$ AND WE CONCATENATE COLUMNWISE THE DISCRETIZED OUTPUT OF THE NETWORK FOR THE CLASS i, NAMELY

$$\bar{R}^i(c) = \mathbb{I}_{\geq\varepsilon}(R^i(c))$$

OUTPUT OF THE i^{TH} NEURON

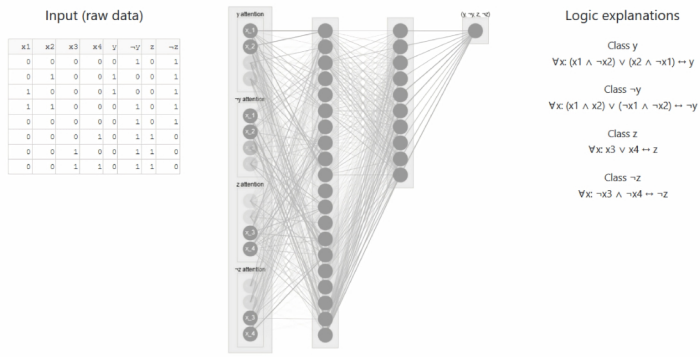
c_1	c_2	c_3	...	$\bar{R}^i$

THE LOSS FUNCTION

$$\mathcal{L}(f, y, \alpha^1, \alpha^2, ..\alpha^n) = L(f, y) + \lambda \sum_{i=1}^{n} \mathcal{H}(\alpha^i)$$

Where $L(f, y)$ is any loss function which used in a multi-label classification task (e.g. cross-entropy or KL divergence) and λ is an hyperparameter balancing the accuracy of the model and the succinctness of the explanation given.

↳ cc_i IS A VECTOR (CONCEPT i)



- Local (single sample) explanation: as mentioned above an entry in the truth table is a column wise concatenation of $\hat{c}^i$ and $\overline{f}^i(c)$. To derive an explanation for a sample need to conjoin the true concepts and the negated counterparts of the false ones. Repeating this process for all the classes will give us a n conjuctions explaining why the sample belongs or not to each class.
- Class explanation: provides an explanation of a whole class, namely it characterizes all the samples that will be classified as belonging to such class. To derive such explanation for a generic class i consider T_i and put in a disjunction all the local explanations for the samples positively labelled (i.e $\overline{f}^i(c) = 1$).

THAT'S IT ↳ well IT'S IMPORTANT AND WORKS, BUT THEY COMPARED IT WITH MOSTLY WHITE BOX METHODS, WHAT IF THE NET USED BECOMES BIGGER, DO EXPLANATIONS STILL HOLD (THEY CONSIDER ONLY THE FST LAYER)... ↳ THIS IS A DOUBT OF MINE