

EXPLAINABLE AI

AI BASED ALGORITHMS ARE ALGORITHMS THAT USE AI MODELS

} HAVE BECOME VERY WIDESPREAD \rightsquigarrow DEAL WITH A LOT OF COMPLEX PROBLEMS

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

BUT

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

\rightsquigarrow PROBABILISTIC SAT SOLVER

EXPLAINABLE AI TRIES TO OPEN AND DESCRIBE THIS BOX

\rightsquigarrow well, white boxes exist too

DECISION TREES

KNN

LOGISTIC REGRESSION

IN PARTICULAR, GIVEN A MODEL XAI TRIES TO IMPROVE:

• TRANSPARENCY

\rightsquigarrow SIMULABILITY \rightsquigarrow HOW EASILY CAN WE SIMULATE IT WITH PENS & PAPER?
 \rightsquigarrow DECOMPOSABILITY \rightsquigarrow DO WE HAVE A CLEAR UNDERSTANDING OF HOW THE ASPECTS OF THE MODEL INFLUENCE IT?
 \rightsquigarrow ALGORITHMIC TRANSPARENCY \rightsquigarrow CAN WE UNDERSTAND HOW THE OUTPUT WAS PRODUCED? VIA MATHS

• INTERPRETABILITY

\rightsquigarrow Does the model provide AN EXPLANATION OF THE PROCESS + OUTPUT JUSTIFICATION

EXPLAINABILITY

\rightsquigarrow INTERPRETABILITY BUT NOT FROM AN EXPERT

• TRUSTWORTHINESS/CONFIDENCE

\rightsquigarrow GIVEN THE MODEL AND THE EXPLAINABLE ASPECTS, STILL, HOW MUCH DO USERS TRUST IT?

So, BLACK BLACK BOXES EXIST, BUT ALSO WHITE BOXES DO

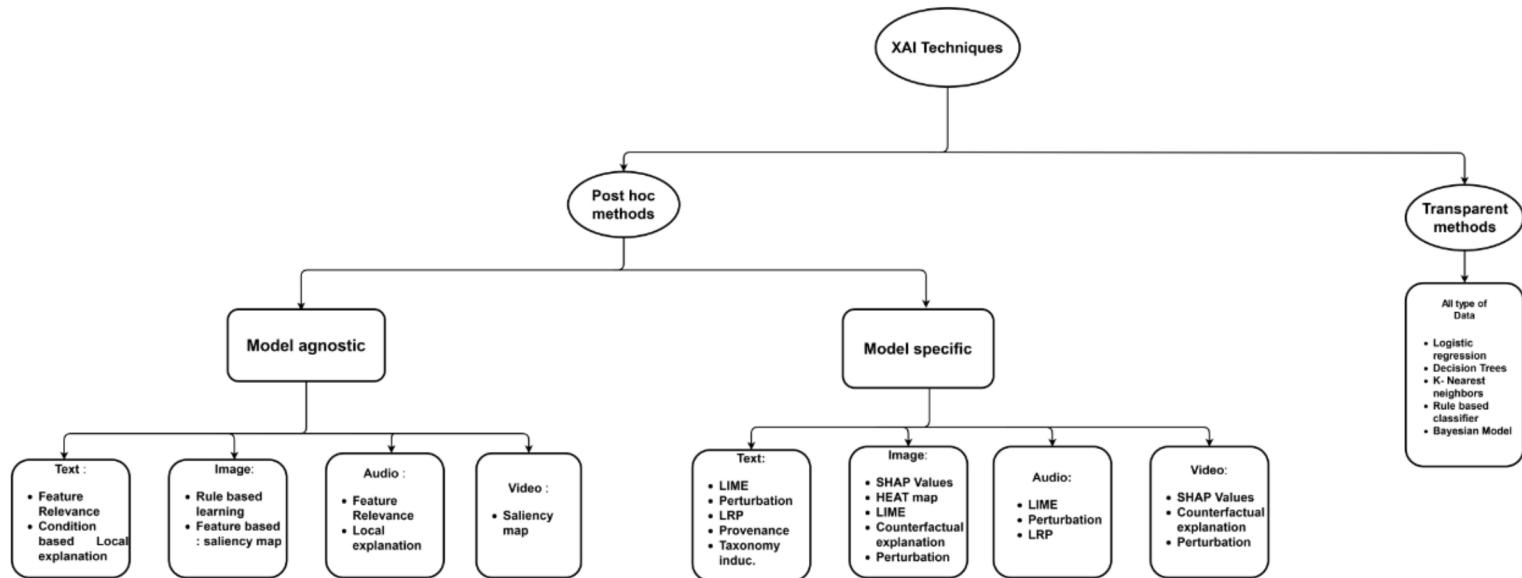


'EXPLAINABLE

BY DEFAULT MODELS'

\rightsquigarrow WITH THESE, NOT MUCH ELSE IS NEEDED, WE ALREADY HAVE TRANSPARENCY

XAI TAXONOMY



FIRST DISTINCTION

↳ POST HOC METHODS

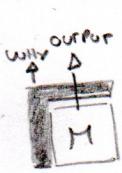
↳ TRANSPARENT METHODS

↓
KNN, trees,
LOGISTIC REGRESSION...

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

→ WHITE BOXES → THEY HAVE
TRANSPARENCY, BUT
TRANSPARENCY $\not\Rightarrow$ EXPLAINABLE

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

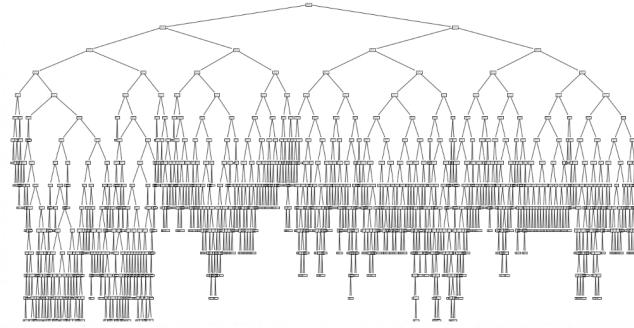
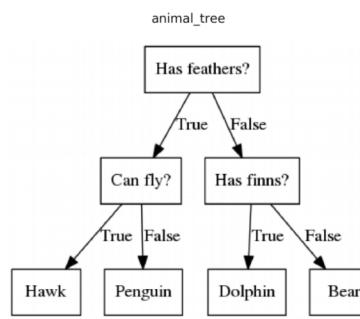


SC, we
'ATTACH' POST
HOC METHODS

↓
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:

↳ See AUTOMATIC DIFFERENTIATION
(... PROGRAM DERIVATIVE)

LIME (LOCAL INTERPRETABLE
Model-AGN. EXPLANATIONS)

AN ORTHOGONAL
SUBDIVISION

↳ EXPLANATIONS
ARE

↳ LOCAL → we EXPLAIN
WHY A SAMPLE GIVES SUCH RESULTS } DON'T SCALE IN GENERAL

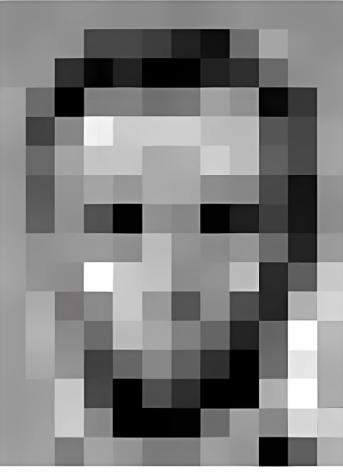
↳ GLOBAL → we EXPLAIN
THE WHOLE MODEL OR CLASSES (eg: IN CLASSIFICATION)

THOUGH EXPLAINING THE
OUTPUT IN INPUT TERMS MAY
NOT BE SUFFICIENT

↓
THE INPUT FEATURES
MAY NOT BE HUMAN
UNDERSTANDABLE

↳ USE → CONCEPT BASED
METHODS/NETWORKS

167	163	174	168	160	162	129	161	172	161	166	166
165	162	163	74	75	62	93	17	110	210	180	154
180	180	50	14	34	6	10	93	48	106	199	181
206	199	5	124	131	111	120	204	166	16	56	180
194	68	137	261	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
189	88	179	209	185	215	211	168	139	76	20	169
189	97	166	84	10	168	184	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	96	190
205	174	155	262	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	103	227	210	127	122	36	101	265	224
190	214	173	66	103	143	95	50	2	189	249	215
187	196	205	75	1	81	47	0	6	217	265	211
183	202	237	145	0	0	12	168	200	188	243	236
195	206	123	287	177	121	123	200	175	13	96	218



157	153	174	168	150	152	129	151	172	161	165	166
155	182	163	74	75	62	83	17	110	210	180	154
180	180	50	14	84	6	10	83	48	105	159	181
205	169	5	124	191	111	120	204	165	15	56	180
144	68	197	251	297	289	299	228	227	67	71	201
172	105	207	233	293	214	220	239	228	91	74	206
188	88	179	209	165	215	211	158	199	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	105	36	190
205	174	155	282	256	251	149	178	228	43	95	234
190	216	116	149	206	167	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	258	224
190	214	173	68	103	143	95	60	2	109	249	215
187	156	205	78	1	81	47	0	6	217	285	211
183	202	237	146	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

157	153	174	168	150	152	129	161	172	161	166	156
155	182	163	74	75	62	83	17	110	210	180	154
180	180	50	14	84	6	10	83	48	105	159	181
206	169	5	124	191	111	120	204	165	15	56	180
194	68	187	251	237	239	228	227	87	71	201	
172	105	207	233	293	214	220	239	228	98	74	206
188	88	179	209	165	215	211	158	199	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	282	256	251	149	178	228	43	95	234
190	216	116	149	206	167	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	258	224
190	214	173	68	103	143	95	60	2	109	249	215
187	156	205	78	1	81	47	0	6	217	285	211
183	202	237	146	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

CONCEPT BASED METHODS

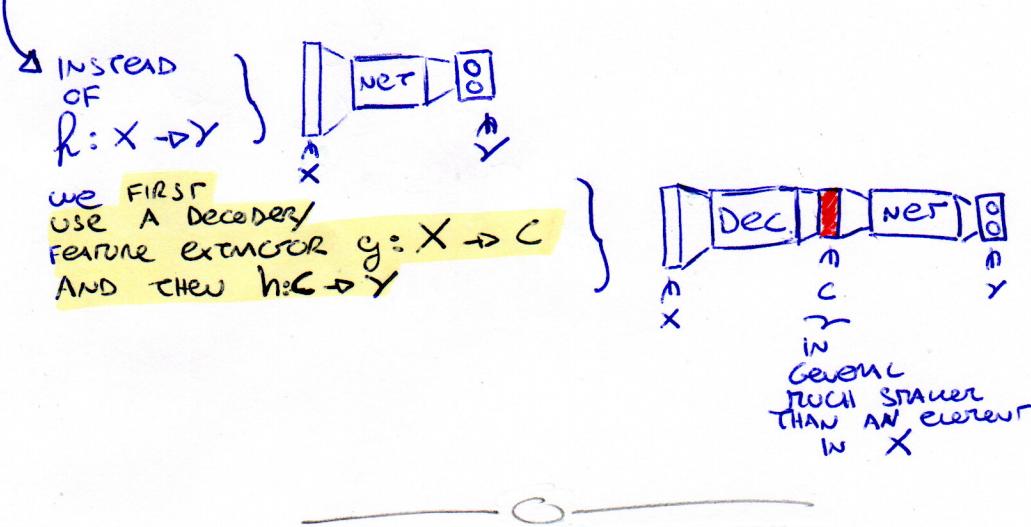
→ THEY TAKE AS INPUT CONCEPTS (HIGHER ABSTRACTION)

→ SO THEY CAN EXPLAIN THE RESULT IN HUMAN MEANINGFUL WAY

CONCEPT = HIGH LEVEL FEATURE

CONCEPT BASED CLASSIFIERS

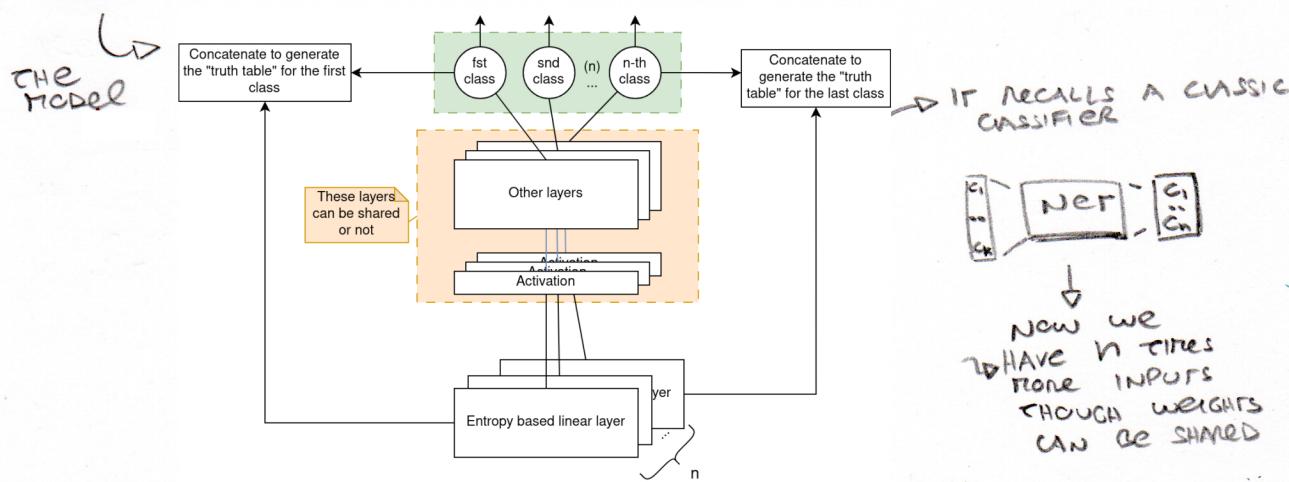
→ IN A CLASSIFICATION TASK WE HAVE A CLASSIFIER TAKING CONCEPTS AS INPUTS



ENTROPY BASED EXPLANATION OF N.N.

⇒ CONTEXT: CLASSIFICATION TASK, n CLASSES, k CONCEPTS

OF A CONCEPT BASED CLASSIFIER...



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

IV

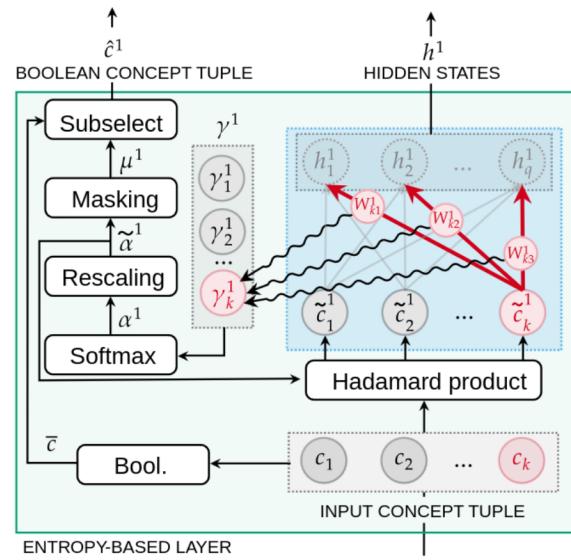
ONE OUTPUT
OF THE
ENTROPY
LAYER

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 capture such outputs we need to capture the 'Attention' each class has
towards each concept

j

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

Yes, it's a vector

Then \forall class
we normalize
then, getting
a probability
vector

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

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

With these values we compute the HADAMARD PRODUCT WITH THE INPUT

RESCALING
FOR NUMERICAL
REASONS

$$\tilde{a}_j^i = \frac{a_j^i}{\max_u a_u^i}$$

τ : TEMPERATURE
(How to weight the
concept difference)

$\tau \rightarrow \infty$ ALL WEIGHT THE SAME
 $\tau \rightarrow 0^+$ ONE MUCH MORE RELEVANT

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

$$\tau = 1 \rightarrow [0, 0.731, 0.268] \quad (\text{softmax})$$

$$\tau = 0.001 \rightarrow [~0, ~1, ~0]$$

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

"THE ATTENTION
VALUES FOR
A CLASS"

THE BOOLEAN CONCEPT TUPLE?

↳ well, define a function
that given x , returns a vector in which element l
is 0 if element l in x was $< \epsilon$, 1 else

$$I_{>\epsilon}(x) : \mathbb{R}^m \rightarrow \mathbb{R}^m \quad (\text{A FILTER})$$

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

LASTLY, TO GET THE BOOLEAN INTERPRETATION FOR C

$$\hookrightarrow \hat{\Sigma}^i := \Sigma(\mathbb{I}_{\geq \varepsilon}(c), \mathbb{I}_{\geq \varepsilon}(\tilde{c}))$$

\hookrightarrow THIS SHOULD BE A x_1 SUBSELECT \hookrightarrow SELECT THE MOST RELEVANT CONCEPTS



we use these to SELECT THE COMPONENTS OF THE FIRST ARGUMENT (α^1) WHOSE POSITION CORRESPONDS TO A 1 IN THE SECOND ARGUMENT
IT'S A SORT OF MASK

THERE'S ONE CLASS!
(THE WEIGHTS ARE +)

THE INPUT HAS 2ND & 3RD CONCEPTS BUT FOR THIS CLASS ONLY THE 3RD IS RELEVANT

$$\text{eg: } \Sigma([0, 1, 1], [0, 0, 1])$$

$$= [0, 0, 1]$$

we DON'T CARE ABOUT EVEN IF IT WAS HIGH

DURING TRAINING
WE COMPUTE THE WEIGHTS

\hookrightarrow AND we compute the 'truth' tables

\hookrightarrow THIS I 'GUESS' ONLY AT THE END... \hookrightarrow SUPPOSITIONE

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

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

OUTPUT OF THE i^{th} NEURON

$c_1 c_2 c_3 \dots \bar{f}^i$

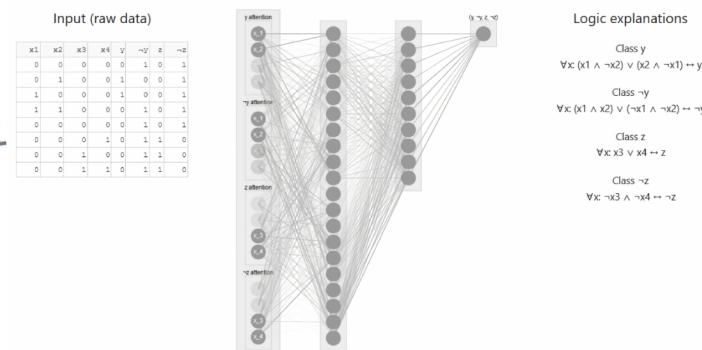
THE LOSS FUNCTION



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

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.

$\hookrightarrow \alpha^i$ IS A VECTOR (CONCEPT i)



- Local (single sample) explanation: as mentioned above an entry in the truth table is a column wise concatenation of $\hat{\Sigma}^i$ and $\bar{f}^i(c)$. 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 conjunctions 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 $\bar{f}^i(c) = 1$).

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