Human-centric Artificial Intelligence

2024/2025

Worksheet #6:
Interpretable and Explainable AI (XAI):
Interpretable Models

Luís Macedo, Hugo Oliveira

6.1 Topics

- Interpretable / Explainable AI
  - Interpretable Models
  - Model-agnostic Methods for XAI

6.2 Pre-class Materials

- Course slides on XAI:
  - Introduction
  - Interpretable Models
  - Model-Agnostic Interpretable Models
  - Explainable AI: Other Methods
- Short videos:
  - What is Explainable AI?: https://www.youtube.com/watch?v=jFHPEQi55Ko (7 min)
  - From Explainable AI to Human-centered AI: https://www.youtube.com/watch?v=UuiV0icAlRs (14 min)

6.3 Complementary Materials

- Readings: Molnar [2019] (Chapters 1–9)
- Videos:
  - Explainable AI Cheat Sheet: https://www.youtube.com/watch?v=Yg3q5x7yDeM&t=472s
  - Explainable AI explained!
    1. Introduction: https://www.youtube.com/watch?v=OZJ1IgSgP9E&t=2s
    2. By-design interpretable models: https://www.youtube.com/watch?v=qPn9m30ojfc
    3. LIME: https://www.youtube.com/watch?v=d6j6bofhj2M
    4. SHAP: https://www.youtube.com/watch?v=9haIOplEIGM

5. Counterfactual explanations and adversarial attacks: https://www.youtube.com/watch?v=UUZxRct8rIk
6. Layerwise Relevance Propagation: https://www.youtube.com/watch?v=PDRewtcqmaI

– Videos from *A Data Odyssey*:

* Introduction to Explainable AI: https://www.youtube.com/watch?v=YuDijSIR9iM
* The 6 Benefits of Explainable AI: https://www.youtube.com/watch?v=UGhKIcQUJ54
* Model Agnostic Methods for XAI: https://www.youtube.com/watch?v=EWD9jsIzY80
* An introduction to LIME for local interpretations: https://www.youtube.com/watch?v=dQ_jvRkzN1Q
* SHAP values for beginners: https://www.youtube.com/watch?v=MQ6fFDwjuco

- Tutorials, video-classes:

  – https://www.youtube.com/watch?v=SDSYiACS3xs (XAI- why?)

  – https://www.youtube.com/watch?v=YSsYXAn_L00 (David Aha - XAI)

  – https://www.youtube.com/watch?v=AFC8yWzypss (XAI - Dr. Wojciech Samek - ODSC Europe 2019)

  – https://www.youtube.com/watch?v=NOzn6ip2anc (Applied Explainable AI - ODSC London 2019 talk)

6.4 Theoretic-Practical Exercises

Question 6.1

1. What is an explanation? What does it mean to interpret? What are their differences? What is explainability/interpretability? What are their differences?

2. Why and when do we need XAI? Do we really need it? For what? Where is it critical? Provide examples of situations in which you need XAI. Do the same for situations in which you do not need XAI. Think of what makes you asking for explanations in some situations and not in others. Give examples of explanations.

3. Concerning Machine Learning models, is accuracy enough? Is it always necessary to develop a specific model for providing explainations?

4. Regarding the Machine Learning pipeline, explain the concept of XAI.

5. Can you think of different categories of XAI? What are their differences?

6. What are the challenges of XAI?

7. What are the challenging problem areas?

8. Provide an illustrative example of a XAI System.

Question 6.2 Consider the problem of predicting how many bikes of a company will be rented, depending on the weather and calendar using linear regression. The full data set is available at http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset, but its first 31 instances are depicted in Figure 6.1.

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	01/01/2011	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
2	02/01/2011	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
3	03/01/2011	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
4	04/01/2011	1	0	1	0	2	1	1	0.2	0.212122	0.590435	0.160296	108	1454	1562
5	05/01/2011	1	0	1	0	3	1	1	0.226957	0.22927	0.436957	0.1869	82	1518	1600
6	06/01/2011	1	0	1	0	4	1	1	0.204348	0.233209	0.518261	0.0895652	88	1518	1606
7	07/01/2011	1	0	1	0	5	1	2	0.196522	0.208839	0.498696	0.168726	148	1362	1510
8	08/01/2011	1	0	1	0	6	0	2	0.165	0.162254	0.535833	0.266804	68	891	959
9	09/01/2011	1	0	1	0	0	0	1	0.138333	0.116175	0.434167	0.36195	54	768	822
10	10/01/2011	1	0	1	0	1	1	1	0.150833	0.150888	0.482917	0.223267	41	1280	1321
11	11/01/2011	1	0	1	0	2	1	2	0.169091	0.191464	0.686364	0.122132	43	1220	1263
12	12/01/2011	1	0	1	0	3	1	1	0.172727	0.160473	0.599545	0.304627	25	1137	1162
13	13/01/2011	1	0	1	0	4	1	1	0.165	0.150883	0.470417	0.301	38	1368	1406
14	14/01/2011	1	0	1	0	5	1	1	0.16087	0.188413	0.537826	0.126548	54	1367	1421
15	15/01/2011	1	0	1	0	6	0	2	0.233333	0.248112	0.49875	0.157963	222	1026	1248
16	16/01/2011	1	0	1	0	0	0	1	0.231667	0.234217	0.48375	0.188433	251	953	1204
17	17/01/2011	1	0	1	1	1	0	2	0.175833	0.176771	0.5375	0.194017	117	883	1000
18	18/01/2011	1	0	1	0	2	1	2	0.216667	0.232333	0.861667	0.146775	9	674	683
19	19/01/2011	1	0	1	0	3	1	2	0.292174	0.298422	0.741739	0.208317	78	1572	1650
20	20/01/2011	1	0	1	0	4	1	2	0.261667	0.25505	0.538333	0.195904	83	1844	1927
21	21/01/2011	1	0	1	0	5	1	1	0.1775	0.157833	0.457083	0.353242	75	1468	1543
22	22/01/2011	1	0	1	0	6	0	1	0.0591304	0.0790696	0.4	0.17197	93	888	981
23	23/01/2011	1	0	1	0	0	0	1	0.0965217	0.0988391	0.436522	0.2466	150	836	986
24	24/01/2011	1	0	1	0	1	1	1	0.0973913	0.11793	0.491739	0.15833	86	1330	1416
25	25/01/2011	1	0	1	0	2	1	2	0.223478	0.234526	0.616957	0.129796	186	1799	1985
26	26/01/2011	1	0	1	0	3	1	3	0.2175	0.2036	0.8625	0.29385	34	472	506
27	27/01/2011	1	0	1	0	4	1	1	0.195	0.2197	0.6875	0.113837	15	416	431
28	28/01/2011	1	0	1	0	5	1	2	0.203478	0.223317	0.793043	0.1233	38	1129	1167
29	29/01/2011	1	0	1	0	6	0	1	0.196522	0.212126	0.651739	0.145365	123	975	1098
30	30/01/2011	1	0	1	0	0	0	1	0.216522	0.250322	0.722174	0.0739826	140	956	1096
31	31/01/2011	1	0	1	0	1	1	2	0.180833	0.18625	0.60375	0.187192	42	1459	1501

Figure 6.1: 31 first instances of the bike data set

1. Table 6.1 shows the weights, SE, and t-statistic for this data set, as computed by a Linear Regression toolkit. Make interpretations for each feature.

2. Consider the same data set of the previous question and again the linear regression interpretable model. Consider also the feature effect

	Weight	SE	|t|
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

Table 6.1: Estimated weight, standard error of the estimate (SE), and absolute value of the t-statistic ($|t|$)

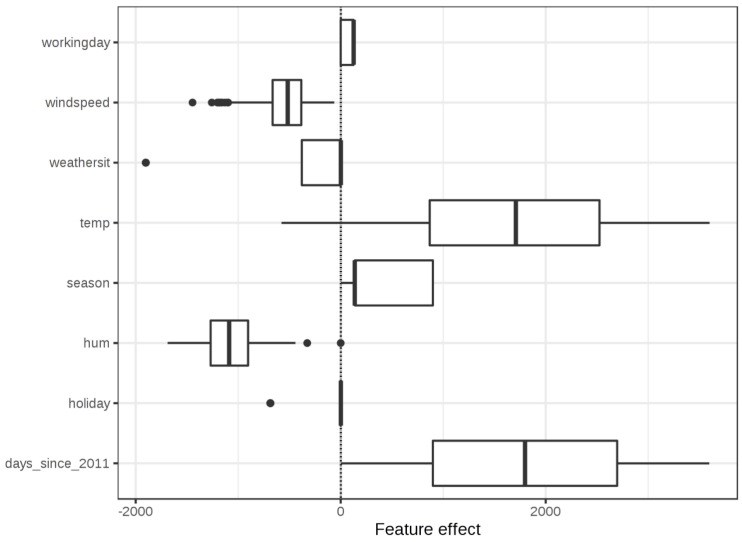

Figure 6.2: The feature effect plot - the distribution of effects (i.e., feature value times feature weight) across the data per feature

plot in Figure 6.2, which shows the distribution of effects (feature value $\times$ feature weight) across the data per feature.

Which features have the largest contributions to the expected number of rented bicycles?

3. For the same domain and data set of the previous problems, the 6th instance is described in Table 6.2

Feature	Value
season	SPRING
yr	2011
mnth	JAN
holiday	NO HOLIDAY
weekday	THU
workingday	WORKING DAY
weathersit	GOOD
temp	1.604356
hum	51.8261
windspeed	6.000868
cnt	1606
days_since_2011	5

Table 6.2: Description of instance 6 of the bike rental dataset.

Assuming that the average predictions for the training data are 4504, and that the prediction for the 6th instance is 1571 bikes, how much has each feature of this instance contributed to such prediction? You can draw the effect plot (Figure 6.2) and mark the local effect of this instance.

4. Discuss to what extent do linear models provide good explanations?

Question 6.3 Consider the model-agnostic methods for XAI and answer the following questions.

1. What exactly are model-agnostic methods? Describe their main features and differences among them.

2. When are model-agnostic methods relevant?

Question 6.4 A surrogate model is a simple model that is used to explain a complex model. Surrogate models are usually created by training a linear regression or decision tree on the original inputs and predictions of a complex model. Coefficients, variable importance, trends, and interactions displayed in the surrogate model are then assumed to be indicative of the internal mechanisms of the complex model.

1. What is the scope of interpretability for surrogate models? Global or local?
2. What complexity of functions can surrogate models help explain?
3. How do surrogate models enhance understanding?
4. How do surrogate models enhance trust?
5. What are the main differences between LIME and SHAP?

Question 6.5
Consider other methods for XAI.

1. Give examples of situations where humans rely on previous examples for making decisions.
2. Which other methods do you know for XAI? Discuss their advantages and their limitations.

Bibliography

Christoph Molnar. *Interpretable Machine Learning*. 2019. https://christophm.github.io/interpretable-ml-book/.