Facets of Data

St. Joseph's University, Bengaluru

Basics of Data Science

Structured Data

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S
12	13	0	3	Saundercock, Mr. William Henry	male	20.0	0	0	A/5. 2151	8.0500	NaN	S
13	14	0	3	Andersson, Mr. Anders Johan	male	39.0	1	5	347082	31.2750	NaN	S
14	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.0	0	0	350406	7.8542	NaN	S
15	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.0	0	0	248706	16.0000	NaN	S
16	17	0	3	Rice, Master. Eugene	male	2.0	4	1	382652	29.1250	NaN	Q
17	18	1	2	Williams, Mr. Charles Eugene	male	NaN	0	0	244373	13.0000	NaN	S
18	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vande...	female	31.0	1	0	345763	18.0000	NaN	S
19	20	1	3	Masselmani, Mrs. Fatima	female	NaN	0	0	2649	7.2250	NaN	C

Unstructured Data

SHAP (SHapley Additive exPlanations)

#SHAP #shapley #interpretableai #LIME #KernelSHAP #TreeSHAP [[Shapley]]

The goal is to explain the prediction of an instance x by computing the contribution of each feature to the prediction. The players can also be groups of features. For example in images, pixels can be grouped into super pixels and the payoff distributed aming them. The extra thing here is an additive feature attribution method; a linear model. $g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z_{j'}$ where g is the explanation model, $(z' \in 0, 1^M)$ is the coalition vector M is the maximum coalition size and $\phi_j \in \mathbb{R}$ is the feature attributioni for a feature j , the Shapely values. Terminoligy: "coalition vector" = "simplified features"

For x, the instance of interest, the coalition vector x' is a vector of all 1's, i.e., all feature values are "present". The formula simplifies to $g(x')\phi_0 + \sum_{j=1}^{M} \phi_j$ '

Properties

* SHAP satisfies Efficiency, Dummy, Symmetry, and Additivity

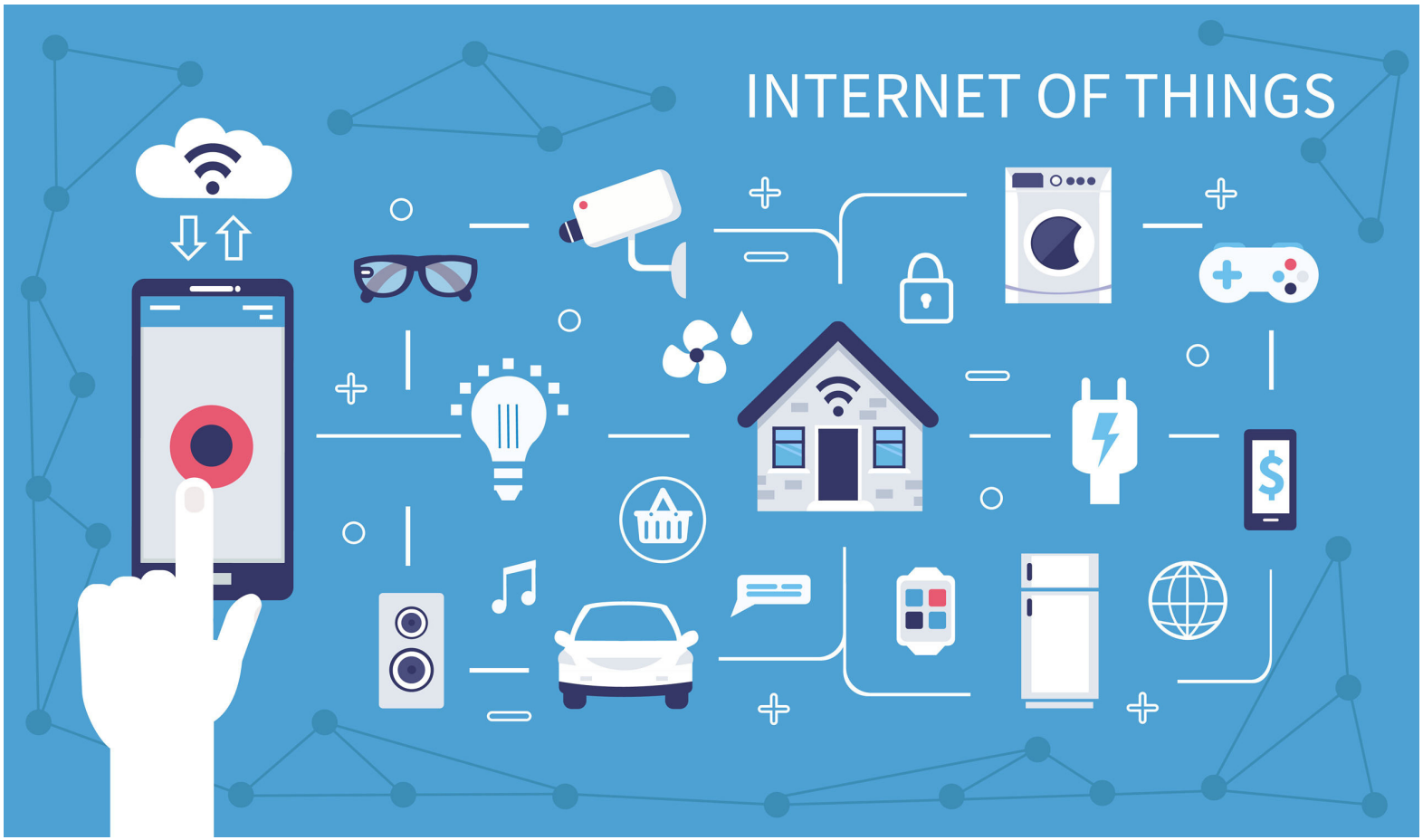
1. **Local Accuracy** $f(x) = g(x') = \phi_0 + \sum_{j=1}^{M} \phi_j x_{j'}$
2. **Missingness** $x_{j'} = 0 \implies \phi_j = 0$ This means that missing feature gets an attribution of 0.
3. **Consistency** Let $f_x(z') = f(h_x(z'))$ and $z'_{\setminus j}$ indicate that $z_{j'} = 0$. For any two models f and f' that satisfy: $f_{x'}(z') - f_{x'}(z'_{\setminus j}) \geq f_x(z') - f_x(z'_{\setminus j}) \forall z \in 0, 1^M$ then $\phi_j(f', x) \geq \phi_j(f, x)$

Natural Language

NLP

This is an example of NL data. NLP problems are usually hard to solve. Human language tends to be ambigous by nature. Different language and dialects used by humans across the globe further complicate this.

Machine-generated Data



Examples

Web server logs, Call detail records, Network event logs,

Graph-Based or Network Data

Audio, Image and Video