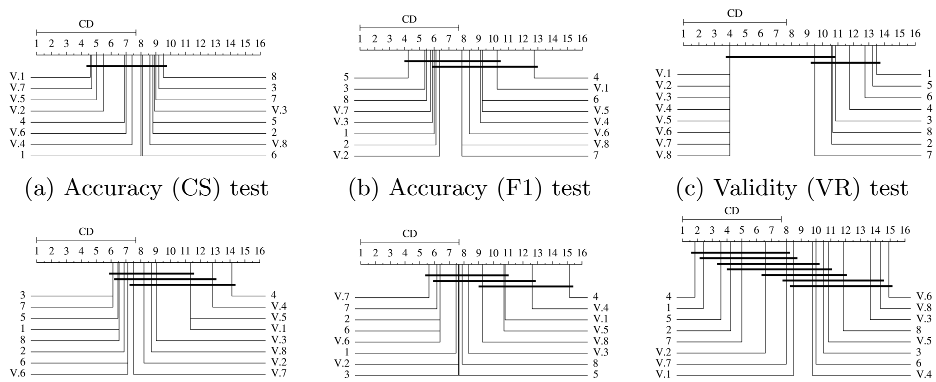
The Importance of Local Perturbations on Quality Measures of Model Explanations

Alejandro Kuratomi ⊠, Tony Lindgren ⊠, and Panagiotis Papapetrou ⊠

Stockholm University, Borgarfjordsgatan 12, 16455 Kista, Sweden
{alejandro.kuratomi,tony,panagiotis}@dsv.su.se

1 Performance Evaluation Results for All Algorithms

To compare the generation algorithms performance, a critical difference evaluation is carried out [1]. The results are shown in Figure 1 for each metric. Figure 1, shows



(a) Accuracy (CS) test (b) Accuracy (F1) test (c) Validity (VR) test

(d) Fidelity (LR) test (e) Fidelity (DT) test (f) Execution time test

Fig. 1: Critical Differences test with respect to different criteria.

no significant difference among algorithms in numerical explanation accuracy, though it is important to note that 6 out of 8 validity-constrained algorithms are located on the left half of the diagram. There is a difference between algorithms *5*, *3*, *8*, *V.7* and *4* in binary explanation accuracy, the former 4 being statistically better, even though algorithms *5* and *3* did not present the best F1 score for any dataset. In terms of *VR* all validity-constrained algorithms, together with algorithms *7*, *2*, *8* and *3* present the highest performance. Algorithms *4*, *6*, *5* and *1* create more examples outside the original feature space manifold. Moreover, with regard to Numerical *FR*, algorithms *3* and *7* perform statistically better than LEAP-based algorithms (*4* and *V.4*). In Binary *FR*, the single best algorithm is *V.7*. In terms of execution time *4* is the best by far. Note that 5 out of 8 validity-constrained algorithms are on the right half of the diagram, and that the top 5 methods are not validity-constrained (methods *4*, *1*, *5*, *2* and *7)*, indicating that validity comes at a cost in execution time.

Table 1: Explanation Accuracy (**CS** and **F1**), Fidelity Ratio (**LR** for logistic regression model and **DT** for decision tree model) and Execution time (**T**) measured in seconds (s) for all generation algorithms and datasets.

		1	**V.1**	**2**	**V.2**	**3**	**V.3**	**4**	**V.4**	**5**	**V.5**	**6**	**V.6**	**7**	**V.7**	**8**	**V.8**
S1	**CS**	0.61	0.63	0.62	0.62	0.64	0.63	0.47	0.47	0.60	0.62	0.53	0.53	0.61	0.64	0.63	0.64
	LR	100	100	100	100	100	100	97.8	97.8	100	100	100	100	100	100	100	100
	DT	100	100	100	100	100	100	91.1	90.0	100	100	100	100	100	100	100	100
	T	0.53	0.65	0.56	0.64	0.68	0.76	0.17	0.42	0.55	0.66	0.72	0.85	0.56	0.64	0.66	0.74
S2	**CS**	0.37	0.39	0.39	0.39	0.39	0.38	0.33	0.34	0.38	0.39	0.33	0.33	0.39	0.39	0.39	0.39
	LR	100	100	100	100	100	100	84.4	85.6	100	100	100	100	100	100	100	100
	DT	100	100	100	100	100	100	88.9	90.0	100	100	100	100	100	100	100	100
	T	0.92	1.33	0.98	1.35	1.59	1.97	0.34	1.62	0.93	1.36	1.21	1.75	0.97	1.35	1.60	1.96
S3	**CS**	0.85	0.84	0.84	0.84	0.85	0.84	0.77	0.78	0.84	0.84	0.81	0.82	0.84	0.84	0.84	0.84
	LR	100	100	100	100	100	100	97.8	98.9	100	100	100	100	100	100	100	100
	DT	100	100	100	100	100	100	95.6	95.6	100	100	100	100	100	100	100	100
	T	0.70	0.84	0.72	0.82	0.83	0.94	0.21	0.48	0.69	0.86	0.91	1.09	0.74	0.84	0.86	0.95
S4	**CS**	0.72	0.71	0.70	0.71	0.68	0.69	0.70	0.69	0.71	0.71	0.76	0.75	0.71	0.70	0.69	0.70
	LR	100	100	100	100	100	100	100	100	100	100	98.9	98.9	100	100	100	100
	DT	100	100	100	100	100	100	91.1	91.1	100	100	100	100	100	100	100	100
	T	0.41	0.48	0.42	0.47	0.47	0.51	0.15	0.24	0.41	0.48	0.54	0.61	0.42	0.47	0.48	0.52
S5	**F1**	0.68	0.65	0.67	0.65	0.68	0.66	0.57	0.61	0.66	0.68	0.65	0.65	0.67	0.65	0.67	0.65
	LR	100	95.6	100	98.9	100	98.9	95.6	96.7	100	95.6	100	100	100	100	100	98.9
	DT	100	93.3	100	96.7	100	97.8	73.3	100	100	87.8	100	100	100	100	100	97.8
	T	0.42	0.48	0.44	0.45	0.68	0.70	0.19	1.13	0.43	0.59	0.56	1.02	0.45	0.52	0.69	0.75
S6	**F1**	0.85	0.76	0.85	0.78	0.85	0.83	0.69	0.76	0.85	0.77	0.84	0.79	0.85	0.80	0.85	0.83
	LR	100	85.6	100	96.7	100	91.1	87.8	88.9	100	81.1	100	100	100	96.7	100	92.2
	DT	100	81.1	100	96.7	100	91.1	64.4	88.9	100	85.6	100	100	100	100	100	94.4
	T	0.46	0.52	0.48	0.50	1.06	1.10	0.20	1.24	0.46	0.52	0.59	0.94	0.49	0.51	1.07	1.10
U1	**CS**	0.76	0.96	0.73	0.78	0.70	0.76	0.75	0.81	0.77	0.96	0.71	0.82	0.73	0.79	0.7	0.76
	F1	0.56	0.33	0.64	0.50	0.47	0.36	0.28	0.36	0.56	0.33	0.33	0.30	0.64	0.46	0.51	0.35
	LR	100	100	100	100	100	100	60.9	100	100	100	100	100	100	100	100	100
	DT	65.2	100	78.3	100	65.2	100	82.6	100	65.2	100	100	100	78.3	100	65.2	100
	T	0.15	0.21	0.16	0.17	0.27	0.29	0.18	0.56	0.15	0.21	0.21	0.28	0.17	0.17	0.28	0.29
U2	**CS**	0.53	0.47	0.55	0.54	0.53	0.49	0.61	0.58	0.52	0.48	0.55	0.53	0.55	0.55	0.53	0.49
	F1	0.53	0.52	0.47	0.40	0.47	0.45	0.63	0.59	0.55	0.52	0.55	0.53	0.46	0.43	0.47	0.45
	LR	100	100	100	100	100	100	98.2	98.7	100	100	100	100	100	100	100	100
	DT	78.1	76.8	96.4	97.3	77.7	78.6	71.0	78.1	78.6	77.7	98.2	98.2	97.8	97.8	78.1	77.7
	T	2.16	2.32	2.21	2.28	2.36	2.50	0.33	1.02	3.77	4.07	4.70	5.02	2.33	2.49	3.97	4.01
U3	**CS**	0.43	0.74	0.43	0.59	0.43	0.46	0.60	0.48	0.43	0.73	0.45	0.48	0.43	0.60	0.43	0.46
	F1	0.63	0.71	0.63	0.82	0.63	0.70	0.62	0.68	0.63	0.70	0.62	0.64	0.63	0.81	0.63	0.70
	LR	100	78.3	98.8	94.0	100	91.6	95.2	94.0	98.8	78.3	98.8	98.8	100	97.6	100	92.8
	DT	100	78.3	100	98.8	100	95.2	38.6	63.9	100	75.9	100	100	100	100	100	94.0
	T	0.71	0.90	0.74	0.85	1.15	1.27	0.31	2.03	0.71	1.03	0.92	2.83	0.72	0.88	1.11	1.31
U4	**CS**	0.52	0.78	0.47	0.64	0.48	0.59	0.59	0.57	0.52	0.78	0.56	0.58	0.46	0.61	0.48	0.59
	F1	0.77	0.72	0.77	0.83	0.77	0.78	0.77	0.81	0.77	0.75	0.77	0.77	0.77	0.83	0.77	0.77
	LR	99.4	78.9	99.4	99.4	100	90.4	89.1	94.9	100	87.8	100	100	100	99.4	99.4	95.5
	DT	100	80.1	100	99.4	99.4	94.2	50.0	86.5	99.4	81.4	99.4	99.4	100	100	99.4	94.2
	T	0.77	0.89	0.77	0.80	1.17	1.22	0.18	1.79	0.78	2.87	1.01	5.03	0.78	1.34	1.20	2.58
U5	**CS**	0.49	0.63	0.50	0.55	0.50	0.47	0.70	0.64	0.50	0.60	0.56	0.55	0.50	0.55	0.50	0.47
	F1	0.94	0.92	0.94	0.94	0.93	0.94	0.89	0.91	0.95	0.93	0.93	0.95	0.93	0.95	0.93	0.94
	LR	100	83.3	100	100	100	100	91.7	91.7	100	83.3	100	100	100	100	100	100
	DT	100	100	100	100	100	100	54.2	83.3	100	100	100	100	100	100	100	100
	T	0.15	0.39	0.16	0.25	0.64	0.76	0.18	0.60	0.16	0.39	0.21	1.27	0.17	0.24	0.65	0.76
U6	**CS**	0.26	0.25	0.25	0.25	0.25	0.25	0.36	0.35	0.27	0.25	0.31	0.30	0.25	0.25	0.25	0.26
	F1	0.86	0.86	0.86	0.86	0.87	0.87	0.84	0.86	0.87	0.86	0.86	0.86	0.86	0.87	0.87	0.87
	LR	100	99.1	100	100	100	100	96.2	98.1	100	99.1	100	100	100	100	100	100
	DT	99.1	77.1	100	100	100	100	79.1	85.7	99.1	73.3	99.1	99.1	100	100	99.1	99.1
	T	0.53	1.15	0.61	0.79	1.43	1.61	0.19	0.96	0.60	1.15	0.82	1.58	0.60	0.79	1.43	1.60

References

1. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research **7**, 1–30 (2006)