Approximating Probabilistic Explanations via Supermodular Minimization (Supplementary Material)

Louenas Bounia[1] **Frederic Koriche**[1]

[1]CRIL UMR CNRS 8188, Université d'Artois, France

A ADDITIONAL EXPERIMENTS

In this section, we present all experimental results obtained for all benchmarks. Recall that for each benchmark b, an explanation task consists a tuple $(\mathcal{T}, \boldsymbol{x}, I, k)$, where

- $\mathcal{T}$ is a decision tree learned, using the CART algorithm;
- $\boldsymbol{x}$ is a data instance from the test set, which is binarized according to the features in $\mathcal{T}$;
- I is a path-explanation for $\boldsymbol{x}$ and $\mathcal{T}$;
- k is the size limit.

Let h denote the hypothesis associated with $\mathcal{T}$. As indicated in the main paper, the performance of explainers on a benchmark b is measured by drawing uniformly at random m instances $\boldsymbol{x}$ from the test set of b, and averaging the resulting error $\epsilon_{h,\boldsymbol{x}}(S)$ and size $|S|$ of the output $S \subseteq I$. In our experiments, $m = \min\{s, 150\}$, where s is the size of the test set of b. The URL addresses of all benchmarks is given in Table 4.

Results for $k = 7$. In Table 2 are reported the results for all the 50 benchmarks considered in this study. The leftmost column gives the name of the dataset b, and the next two columns report statistics about the decision tree accuracy (acc), and the number of its internal nodes (d). The fourth column gives the average size of the path-explanation I for a random (d-dimensional) test instance $\boldsymbol{x}$. The next two groups of three columns report the average errors ($\epsilon_{h,\boldsymbol{x}}(S)$) and sizes ($|S|$) for the approximation algorithms GA and GD, and the SAT-based method. Finally, the average runtime of the SAT-method is provided in the last column. The benchmarks colored in blue indicate that, for some explanation tasks, the SAT-method has reached the timeout before completing binary search, which results in a degradation of precision (and conciseness). For the benchmarks in magenta, the solver could not perform a single run of binary search before reaching the timeout.

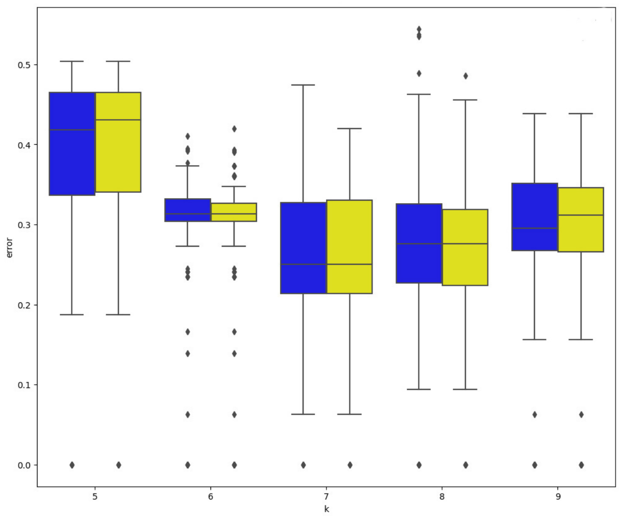
Results for $k = 7 \pm 2$. In Figure 4 are merged 12 bar plots reporting the error results for GA and GD, when k ranges from 5 to 9. Note that for most of these explanation tasks, the SAT solver could not complete binary search before reaching the timeout. We can observe that the performances of GA and GD are very similar. A notable exception is the dataset *compas*, where GA is slightly better than GD when k increases.

Results for $I = [d]$ (and $k = 7$). As emphasized by Izza et al. 2022, path-explanations are not necessarily abductive, since decision trees can include irrelevant features. Nevertheless, the above results indicate that GA and GD are robust enough to reduce such input explanations. Through further analysis, one may wonder whether these approximation algorithms are able to find probabilistic explanations when I is no longer a path-explanation for $\boldsymbol{x}$ and $\mathcal{T}$, but includes all features occurring in $\boldsymbol{x}$ (i.e. $I = [d]$). The corresponding results are reported in Table 3. Unsurprisingly, the performance of the SAT-based approach degrades when d increases. We can also observe that the performance of GD is slightly worse than the performance of GA. Starting from $S_n = [d]$, GD must perform d iterations of the main loop, which increases the chance of reaching a bad local optimum. On the other hand, GA only performs $\mathcal{O}(k \ln(1/\gamma))$ iterations of the main loop in order to select a solution that is close to the optimum found by the SAT approach.

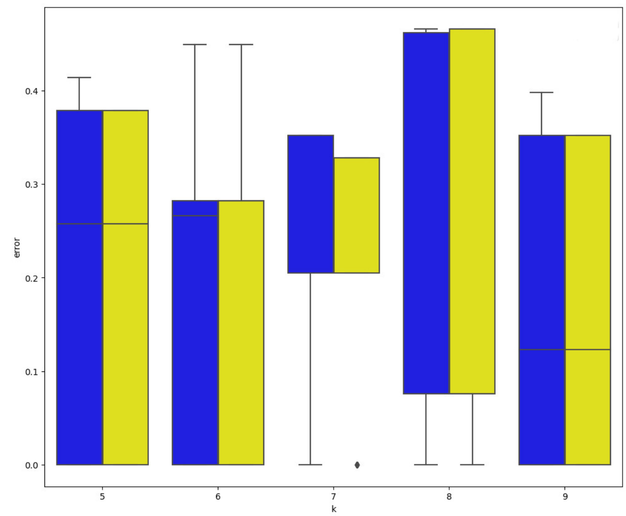
References

Yacine Izza, Alexey Ignatiev, and João Marques-Silva. On tackling explanation redundancy in decision trees. *Journal of Artificial Intelligence Research*, 75:261–321, 2022.

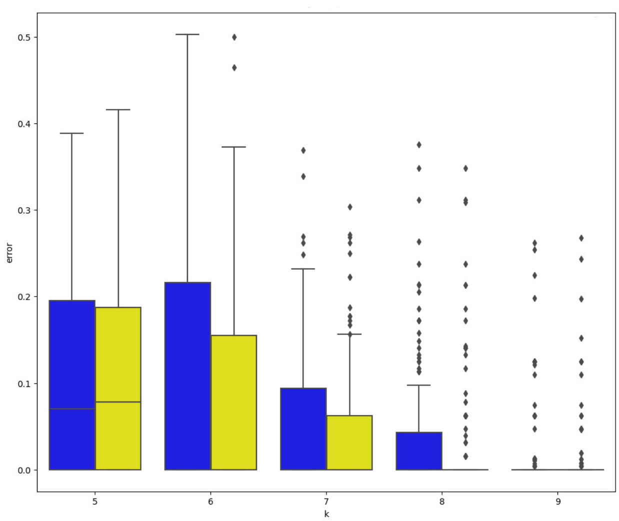
Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).



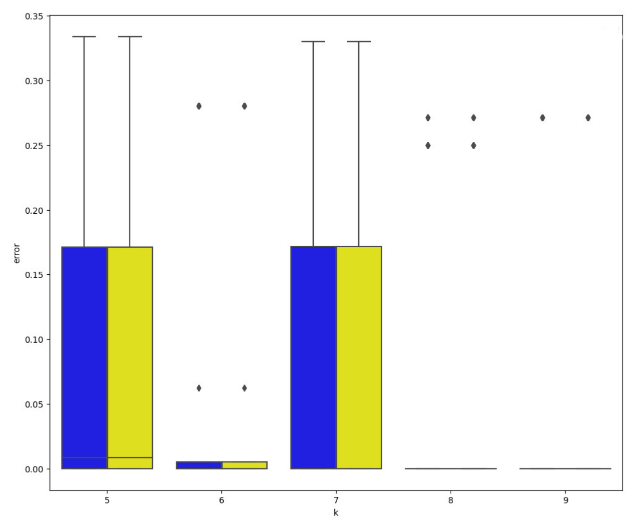
(a) *adult*



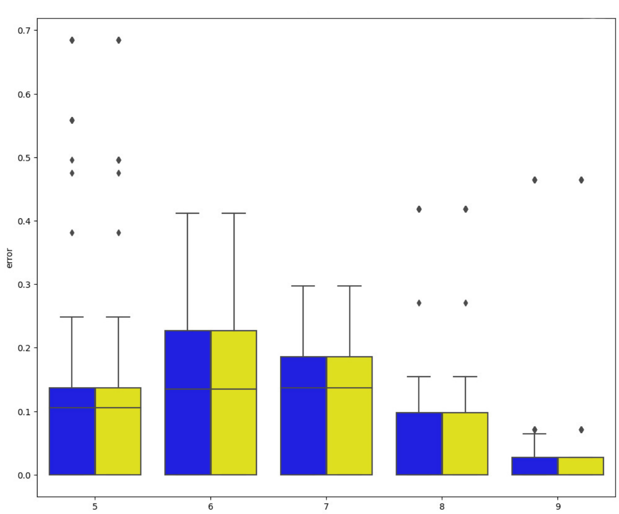
(b) *cnae*



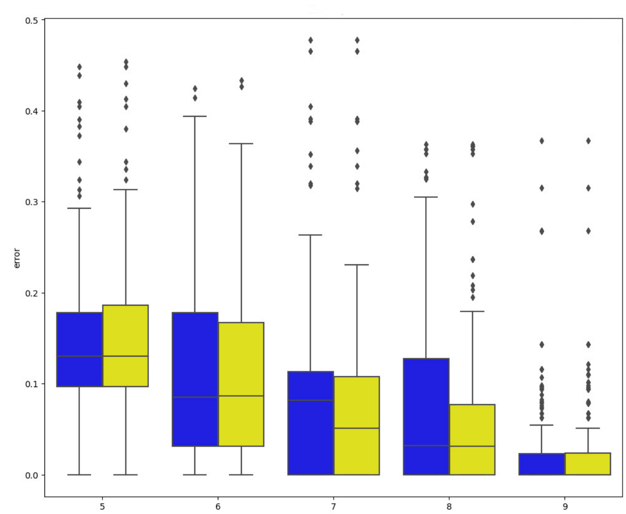
(c) *compas*



(d) *dexter*

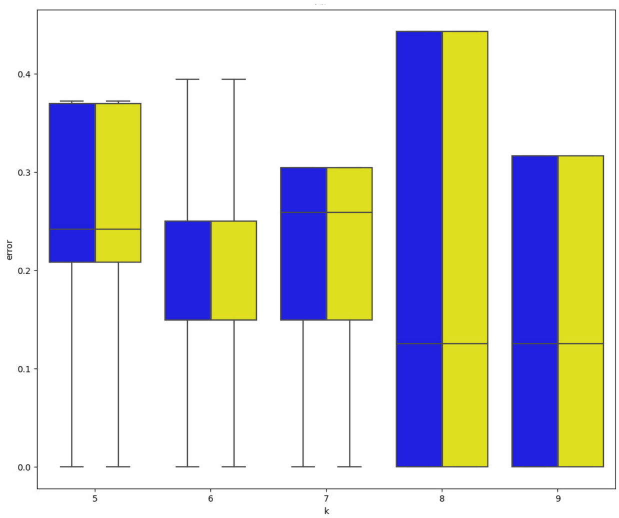


(e) *email spam*

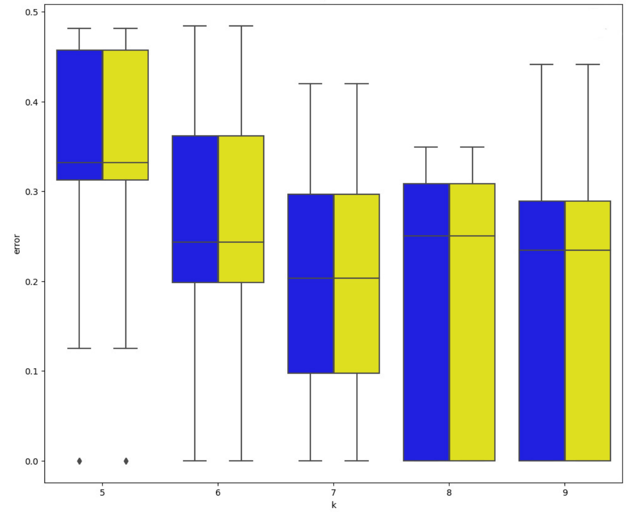


(f) *employee attr.*

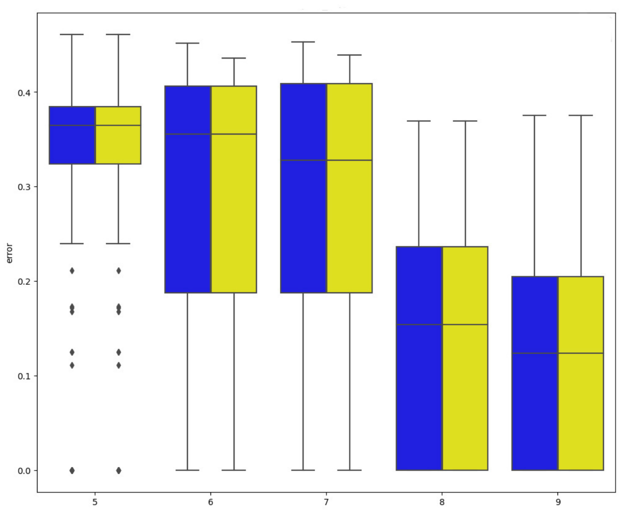
Figure 4: Bar plots for the errors of GA (yellow) and GD (blue), using $k = 7 \pm 2$.



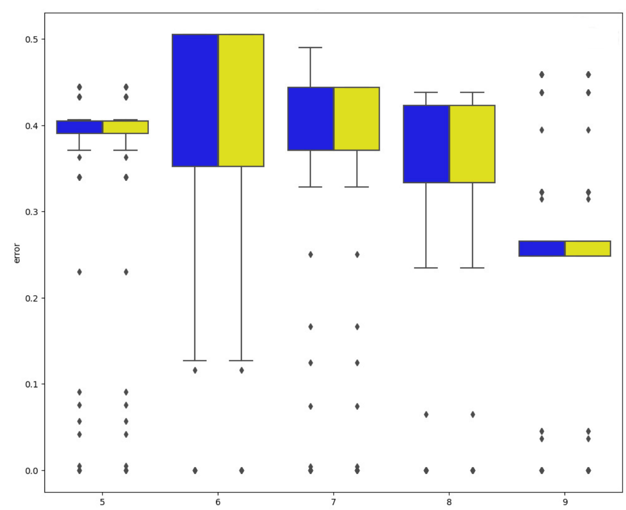
(g) *farm ads*



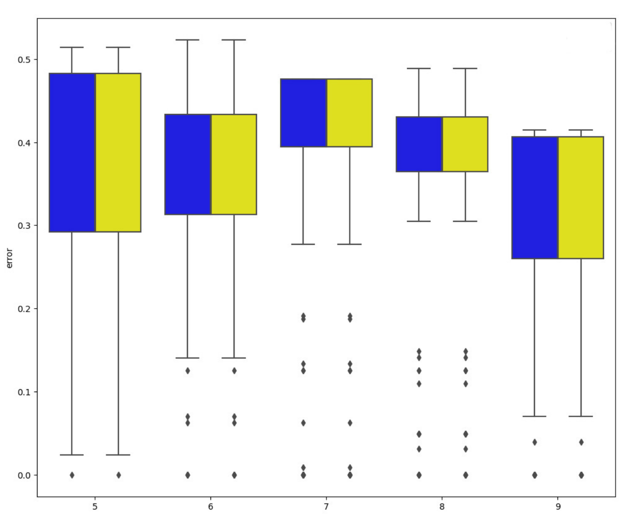
(h) *gina agnostic*



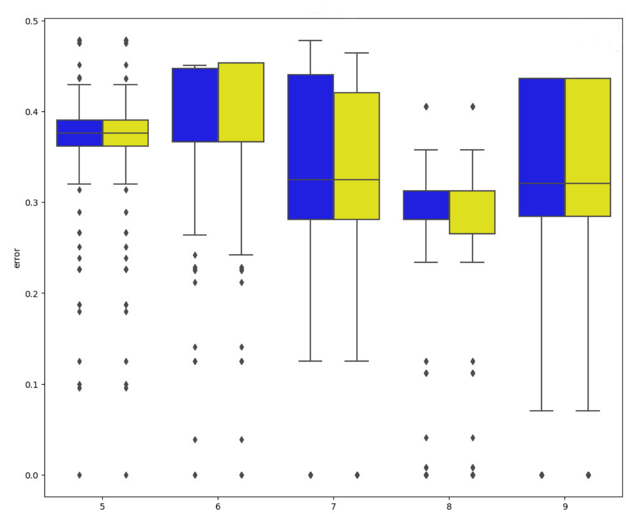
(i) *gina prior*



(j) *gisette*



(k) *mnist38*



(l) *mnist49*

Figure 4: Bar plots for the errors of GA (yellow) and GD (blue), using $k = 7 \pm 2$ (*cont.*).

Benchmark				$\epsilon_{h,\boldsymbol{x}}(S)$			$\lvert S\rvert$			Time (s)
name	*acc*	*d*	|*I*|	GA	GD	SAT	GA	GD	SAT	SAT
voting	94.66	16	3.05	0.07 (±0.09)	0.07 (±0.09)	0.07 (±0.09)	2.00	2.00	2.00	0.37
placement	95.38	18	3.62	0.14 (±0.20)	0.14 (±0.20)	0.14 (±0.20)	2.02	2.02	2.02	0.55
anneal2	98.52	14	4.03	0.19 (±0.11)	0.19 (±0.11)	0.19 (±0.11)	2.10	2.10	2.10	0.21
cars	98.36	21	4.33	0.16 (±0.11)	0.16 (±0.11)	0.16 (±0.11)	2.08	2.08	2.08	0.74
hepatitis	83.72	13	4.47	0.05 (±0.10)	0.05 (±0.10)	0.05 (±0.10)	3.00	3.00	3.00	0.41
autos	87.10	21	4.50	0.12 (±0.12)	0.12 (±0.12)	0.12 (±0.12)	3.53	3.53	3.53	0.72
backache	88.89	18	4.83	0.24 (±0.11)	0.24 (±0.11)	0.24 (±0.11)	2.22	2.22	2.22	0.66
cleveland2	75.82	35	4.98	0.05 (±0.08)	0.05 (±0.08)	0.05 (±0.08)	4.10	4.10	4.10	4.42
mushroom	100	17	4.99	0.28 (±0.15)	0.28 (±0.15)	0.28 (±0.15)	3.02	3.02	3.02	0.55
meta-data	87.42	44	5.09	0.08 (±0.11)	0.08 (±0.11)	0.08 (±0.11)	3.10	3.10	3.10	12.14
vehicle	96.85	23	5.19	0.21 (±0.10)	0.21 (±0.10)	0.21 (±0.10)	2.02	2.02	2.02	1.23
ionosphere	95.28	19	5.22	0.54 (±0.40)	0.54 (±0.40)	0.54 (±0.40)	1.03	1.03	1.03	0.39
kr-vs-kp	99.79	32	5.29	0.21 (±0.15)	0.21(±0.15)	0.21 (±0.15)	1.07	1.07	1.07	2.08
balance	85.64	17	5.37	0.06 (±0.08)	0.06 (±0.08)	0.06 (±0.08)	4.03	4.03	4.03	1.82
glass	78.46	31	5.38	0.26 (±0.11)	0.26 (±0.11)	0.26 (±0.11)	2.14	2.14	2.14	2.36
student perf.	91.79	30	5.41	0.26 (±0.11)	0.26 (±0.11)	0.26 (±0.11)	2.00	2.00	2.00	2.16
biomed	96.83	15	5.68	0.25 (±0.10)	0.25 (±0.10)	0.25 (±0.10)	2.32	2.32	2.32	0.45
tae	65.22	32	5.78	0.12 (±0.19)	0.13 (±0.20)	0.12 (±0.18)	3.96	3.91	3.94	3.22
primary tumor	84.31	23	6.23	0.09 (±0.09)	0.09 (±0.09)	0.09 (±0.08)	4.22	4.22	4.22	3.58
liver disorders	75.96	58	6.38	0.18 (±0.09)	0.18 (±0.08)	0.18 (±0.08)	4.00	4.00	4.00	27.33
schizophrenia	80.39	33	6.39	0.37 (±0.24)	0.37 (±0.24)	0.37 (±0.24)	1.27	1.27	1.27	4.79
tic-tac-toe	92.36	9	6.46	0.24 (±0.13)	0.24 (±0.13)	0.23 (±0.10)	2.96	2.87	2.87	1.43
australian	84.06	70	6.48	0.24 (±0.13)	0.24 (±0.14)	0.24 (±0.13)	3.56	3.55	3.55	63.49
hungarian	62.92	13	6.65	0.12 (±0.12)	0.12 (±0.12)	0.11 (±0.10)	3.58	3.56	3.56	1.68
horse colic	75.68	40	6.73	0.14 (±0.07)	0.13 (±0.07)	0.13 (±0.07)	4.03	4.06	4.06	11.56
madelon	66.41	500	7.33	0.06 (±0.09)	0.06 (±0.09)	–	6.50	6.50	–	–
haberman	67.39	53	8.18	0.73 (±0.11)	0.76 (±0.09)	0.75 (±0.10)	2.96	3.02	3.08	29.43
indian liver	64.57	84	8.21	0.10 (±0.09)	0.10 (±0.09)	0.16 (±0.12)	5.08	4.89	6.12	176.28
pima indians	75.32	97	8.30	0.15 (±0.14)	0.15 (±0.14)	0.16 (±0.12)	5.85	5.84	6.58	484.6
dexter	86.11	20000	8.32	0.06 (±0.08)	0.06 (±0.08)	–	5.95	5.96	–	–
loan eligibility	74.31	68	8.47	0.19 (±0.13)	0.18 (±0.13)	0.20 (±0.14)	5.60	5.70	6.82	42.87
patient treat.	66.01	10	8.92	0.05 (±0.09)	0.03 (±0.06)	0.03 (±0.08)	5.63	5.94	5.94	24.08
wine	69.58	11	9.03	0.09 (±0.10)	0.09 (±0.09)	0.09 (±0.12)	5.59	5.64	5.62	36.32
christine	50.8	1023	9.47	0.28 (±0.10)	0.28 (±0.10)	–	7.00	6.99	–	–
gina agnostic	85.31	970	9.69	0.20 (±0.15)	0.20 (±0.15)	–	6.90	6.90	–	–
email spam	93.88	3000	10.46	0.09 (±0.13)	0.09 (±0.13)	–	5.22	5.22	–	–
gina prior	85.59	784	10.53	0.27 (±0.16)	0.27 (±0.16)	–	6.89	6.89	–	–
employee attr.	82.45	63	10.56	0.06 (±0.09)	0.06 (±0.09)	0.20 (±0.11)	6.41	6.39	6.98	1017.24
contraceptive	51.36	90	10.84	0.06 (±0.08)	0.06 (±0.08)	0.39 (±0.17)	4.27	4.26	5.95	1096.07
compas	67.60	40	10.95	0.03 (±0.07)	0.04 (±0.08)	0.05 (±0.09)	5.68	5.83	6.78	1082.32
fetal health	91.85	93	11.33	0.12 (±0.06)	0.12 (±0.06)	0.23 (±0.11)	5.59	5.59	6.00	930.61
dorothea	91.88	100000	12.90	0.25 (±0.10)	0.25 (±0.10)	–	6.70	6.70	–	–
bank market.	89.49	882	13.11	0.29 (±0.08)	0.29 (±0.07)	–	6.99	6.99	–	–
mnist49	95.99	784	15.57	0.37 (±0.14)	0.37 (±0.14)	–	6.97	6.89	–	–
spambase	92.11	236	16.09	0.24 (±0.11)	0.23 (±0.09)	–	6.87	6.87	–	–
adult	81.16	2433	16.43	0.33 (±0.12)	0.33 (±0.12)	–	6.87	6.87	–	–
mnist38	96.42	784	17.89	0.37 (±0.13)	0.38 (±0.14)	–	6.93	6.93	–	–
cnae	92.59	856	19.07	0.32 (±0.25)	0.32 (±0.25)	–	5.97	5.97	–	–
gisette	94.10	5000	21.42	0.32 (±0.11)	0.32 (±0.11)	–	6.88	6.88	–	–
farm ads	80.78	54877	23.15	0.13 (±0.17)	0.13 (±0.17)	–	6.31	6.31	–	–

Table 2: Experimental results on 50 benchmarks for decision tree explanations, using $k = 7$.

Benchmark			$\epsilon_{h,x}(S)$			|S|			Time (s)		
name	acc	d	GA	GD	SAT	GA	GD	SAT	GA	GD	SAT
shuttle	94.87	9	0.0019 (±0.0229)	0.0017 (±0.0203)	0.0017 (±0.0203)	2.15	2.75	4.26	0.0002	0.0004	1.02
tic-tac-toe	95.58	9	0.0008 (±0.0102)	0.0092 (±0.0326)	0.0008 (±0.0101)	4.68	4.97	5.85	0.0006	0.0008	0.79
patient treat.	66.77	10	0.0154 (±0.0555)	0.0157 (±0.0664)	0.0152 (±0.0574)	5.67	6.37	6.52	0.0023	0.0034	35.30
wine	70.0	11	0.0150 (±0.0406)	0.0308 (±0.0487)	0.0153 (±0.0383)	5.98	6.66	6.90	0.0025	0.0037	25.83
hepatitis	81.74	13	0.0038 (±0.0377)	0.0087 (±0.0318)	0.0034 (±0.0187)	2.86	4.07	3.85	0.0002	0.0002	0.63
hungarian	80.90	13	0.0075 (±0.0503)	0.0116 (±0.0271)	0.0075 (±0.0442)	5.46	5.80	6.90	0.0009	0.0013	1.53
anneal2	99.63	14	0.0248 (±0.0498)	0.0332 (±0.0450)	0.0245 (±0.0490)	3.66	3.98	5.00	0.0002	0.0003	0.32
biomed	93.65	15	0.0000 (±0.0000)	0.0000 (±0.0000)	0.0009 (±0.0000)	4.52	4.59	4.52	0.0004	0.0005	0.47
voting	93.89	16	0.0026 (±0.0188)	0.0048 (±0.0269)	0.0026 (±0.0164)	3.01	3.78	5.02	0.0003	0.0006	0.54
balance	81.38	17	0.0032 (±0.0118)	0.0121 (±0.0183)	0.0020 (±0.0084)	5.45	5.77	5.85	0.0014	0.0025	27.22
mushroom	100.0	17	0.0009 (±0.0066)	0.0052 (±0.0222)	0.0005 (±0.0015)	4.55	4.87	5.45	0.0004	0.0007	2.59
backache	85.19	18	0.0100 (±0.0357)	0.0535 (±0.0512)	0.0080 (±0.0237)	3.43	4.43	4.92	0.0003	0.0006	1.16
placement	92.31	18	0.0000 (±0.0000)	0.0010 (±0.0291)	0.0000 (±0.0000)	3.66	4.15	5.10	0.0002	0.0002	0.32
ionosphere	96.23	19	0.0121 (±0.0490)	0.0301 (±0.0493)	0.0075 (±0.0355)	4.36	5.00	5.85	0.0005	0.0001	1.51
autos	85.48	21	0.0499 (±0.0741)	0.0838 (±0.0650)	0.0045 (±0.0219)	4.48	4.90	6.12	0.0005	0.0009	5.71
cars	91.80	21	0.0329 (±0.0864)	0.0701 (±0.1120)	0.0238 (±0.0550)	3.47	3.71	5.05	0.0006	0.0013	4.70
primary tumour	81.37	23	0.0206 (±0.0493)	0.0464 (±0.0459)	0.0074 (±0.0221)	5.37	5.74	6.08	0.0013	0.0030	47.23
vehicle	96.06	23	0.0304 (±0.0458)	0.0619 (±0.0608)	0.0143 (±0.0344)	4.49	4.77	4.90	0.0008	0.0014	14.31
breast cancer	92.98	30	0.0091 (±0.0293)	0.0244 (±0.0439)	0.0066 (±0.0237)	4.43	4.94	5.52	0.0005	0.0009	4.84
student perf	90.77	30	0.0077 (±0.0316)	0.0219 (±0.0394)	0.0005 (±0.0010)	4.23	5.00	5.75	0.0007	0.0016	10.50
glass	87.69	31	0.1168 (±0.0791)	0.1410 (±0.0906)	0.0772 (±0.0758)	4.75	4.85	5.50	0.0010	0.0019	56.90
kr-vs-kp	99.79	32	0.0096 (±0.0299)	0.0300 (±0.0573)	0.0095 (±0.0285)	3.07	3.52	4.40	0.0015	0.0047	34.30
tae	69.57	32	0.0339 (±0.0634)	0.0385 (±0.0950)	0.0236 (±0.0492)	5.20	5.37	5.82	0.0011	0.0022	72.75
schizophrenia	90.20	33	0.0022 (±0.0142)	0.0024 (±0.0089)	0.0019 (±0.0070)	4.29	4.48	6.80	0.0014	0.0037	46.06
cleveland2	65.93	35	0.1217 (±0.0797)	0.1594 (±0.0923)	0.0626 (±0.0750)	4.87	4.93	5.04	0.0017	0.0048	301.59
haberman	67.39	53	0.7467 (±0.1123)	0.7941 (±0.1074)	0.7430 (±0.1252)	2.96	3.02	3.08	0.0034	0.0041	31.50
compas	66.41	40	0.0569 (±0.0697)	0.0694 (±0.0783)	0.1586 (±0.0879)	5.65	6.55	7.00	0.0298	0.0931	1486.50
horse colic	79.28	40	0.0503 (±0.0699)	0.0342 (±0.0543)	0.0133 (±0.0176)	6.43	6.70	7.00	0.0028	0.0059	671.60
meta-data	88.68	44	0.0223 (±0.0821)	0.0290 (±0.0572)	0.0199 (±0.0771)	3.62	5.54	5.96	0.0017	0.0089	171.91
employee attr.	81.88	63	0.0708 (±0.0610)	0.0960 (±0.0723)	0.2212 (±0.1525)	6.25	6.45	7.00	0.0958	0.1032	1340.60
loan eligibility	70.14	68	0.0629 (±0.0616)	0.0870 (±0.0668)	0.0956 (±0.0500)	6.10	6.00	6.50	0.0090	0.0378	856.10
australian	79.23	70	0.0000 (±0.0000)	0.0311 (±0.0592)	0.0137 (±0.0286)	4.70	6.40	6.10	0.0041	0.0308	573.94
liver disorders	71.43	84	0.0939 (±0.1087)	0.1101 (±0.0891)	0.1278 (±0.0878)	6.03	6.90	7.00	0.0150	0.0744	897.81
contraceptive	51.36	90	0.0600 (±0.0800)	0.0600 (±0.0800)	0.3910 (±0.1700)	4.27	4.26	5.95	0.0452	0.0892	1096.07
fetal health	91.85	93	0.1200 (±0.0600)	0.1200 (±0.0600)	0.2310 (±0.1100)	5.59	5.59	6.25	0.0568	0.0723	930.61
pimas indians	72.56	97	0.1181 (±0.0834)	0.1897 (±0.1072)	0.1812 (±0.0622)	6.75	7.00	6.95	0.0183	0.1675	709.57
spambase	92.03	236	0.1268 (±0.0645)	0.1845 (±0.0813)	–	6.89	6.99	–	0.0974	0.1828	–
madelon	71.03	500	0.1257 (±0.1198)	0.1971 (±0.0930)	–	6.85	6.98	–	0.0709	0.2611	–
gina-prior	87.13	784	0.2073 (±0.0837)	0.2496 (±0.0874)	–	6.97	6.99	–	0.1087	1.3443	–
mnist38	95.82	784	0.1561 (±0.0743)	0.2675 (±0.0698)	–	6.95	7.00	–	0.1082	1.4738	–
mnist49	95.02	784	0.1857 (±0.0731)	0.2867 (±0.0970)	–	6.97	6.98	–	0.2019	1.3611	–
bank market.	89.49	882	0.3245 (±0.1201)	0.3890 (±0.1267)	–	6.99	6.99	–	0.2631	1.4288	–
gina-agnostic	83.86	970	0.1593 (±0.0792)	0.2102 (±0.1084)	–	6.96	7.00	–	0.1306	1.7953	–
christine	50.37	1023	0.2710 (±0.0678)	0.3207 (±0.0816)	–	7.00	7.00	–	0.6179	1.9402	–
adult	81.16	2433	0.3882 (±0.1521)	0.3925 (±0.1877)	–	6.87	6.87	–	0.1461	0.9339	–
email spam	94.07	3000	0.1090 (±0.1494)	0.1761 (±0.2169)	–	4.90	6.78	–	0.0704	1.0079	–
gisette	93.81	5000	0.1309 (±0.0706)	0.2006 (±0.1056)	–	6.87	6.94	–	0.0746	2.8047	–
dexter	83.89	20000	0.0321 (±0.0532)	0.0614 (±0.0525)	–	5.74	6.63	–	0.1250	0.2330	–
farm ads	80.78	54877	0.1300 (±0.1700)	0.1300 (±0.1700)	–	6.31	6.31	–	0.2710	0.9892	–
dorothea	93.04	10^5	0.2383 (±0.0743)	0.2514 (±0.0758)	–	6.87	6.97	–	0.2027	1.2040	–

Table 3: Experimental results on 50 benchmarks, using $I = [d]$ and $k = 7$.

Benchmark	Source	Link
adult	UCI	`https://archive.ics.uci.edu/ml/datasets/adult`
anneal2	UCI	`https://archive.ics.uci.edu/ml/datasets/Annealing`
australian	OpenML	`https://www.openml.org/search?type=data&status=active&id=40981`
autos	UCI	`https://archive.ics.uci.edu/ml/machine-learning-databases/autos/`
backache	Kaggle	`www.kaggle.com/datasets/sammy123/lower-back-pain-symptoms-dataset`
balance	UCI	`https://archive.ics.uci.edu/ml/datasets/balance+scale`
bank market.	UCI	`https://archive.ics.uci.edu/ml/datasets/bank+marketing`
biomed	OpenML	`https://www.openml.org/search?type=data&status=active&id=481`
cars	OpenML	`www.openml.org/search?type=data&status=active&id=967`
christine	OpenML	`https://www.openml.org/search?type=data&status=active&id=41142`
cleveland2	Kaggle	`https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci`
cnae	UCI	`https://archive.ics.uci.edu/ml/datasets/cnae-9`
compas	OpemML	`www.openml.org/search?type=data&sort=runs&id=42193&status=active`
contraceptive	UCI	`https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice`
dexter	UCI	`https://archive.ics.uci.edu/ml/datasets/dexter`
dorothea	OpenML	`www.openml.org/search?type=data&status=active&id=4137`
email spam	Kaggle	`https://www.kaggle.com/datasets/veleon/ham-and-spam-dataset`
employee attr.	Kaggle	`https://www.kaggle.com/datasets/HRAnalyticRepository/employee-attrition-data`
farm-ads	UCI	`https://archive.ics.uci.edu/ml/datasets/Farm+Ads`
fetal health	Kaggle	`www.kaggle.com/datasets/andrewmvd/fetal-health-classification`
gina agnostic	OpenML	`www.openml.org/search?type=data&sort=runs&id=1038&status=active`
gina prior	OpenML	`https://www.openml.org/search?type=data&status=active&id=1042`
gisette	UCI	`https://archive.ics.uci.edu/ml/datasets/Gisette`
glass	UCI	`https://archive.ics.uci.edu/ml/datasets/glass+identification`
haberman	UCI	`https://archive.ics.uci.edu/ml/datasets/haberman's+survival`
hepatisis	UCI	`https://archive.ics.uci.edu/ml/datasets/hepatitis`
horse colic	Kaggle	`www.kaggle.com/datasets/uciml/horse-colic?select=horse.csv`
hungarian	OpenML	`www.openml.org/search?type=data&status=active&id=858`
indian liver	Kaggle	`www.kaggle.com/datasets/uciml/indian-liver-patient-records`
ionosphere	UCI	`https://archive.ics.uci.edu/ml/datasets/ionosphere`
kr-vs-kp	UCI	`https://archive.ics.uci.edu/ml/datasets/Chess+(King-Rook+vs.+King-Pawn)`
liver disorders	UCI	`https://archive.ics.uci.edu/ml/datasets/Liver+Disorders`
loan eligibility	Kaggle	`www.kaggle.com/datasets/devzohaib/eligibility-prediction-for-loan`
madelon	UCI	`https://archive.ics.uci.edu/ml/datasets/madelon`
meta-data	UCI	`https://archive.ics.uci.edu/ml/datasets/Meta-data`
mnist38	Kaggle	`https://www.kaggle.com/datasets/oddrationale/mnist-in-csv`
mnist49	Kaggle	`https://www.kaggle.com/datasets/oddrationale/mnist-in-csv`
mushroom	UCI	`https://archive.ics.uci.edu/ml/datasets/mushroom`
patient treat.	Kaggle	`www.kaggle.com/datasets/manishkc06/patient-treatment-classification`
pima indians	Kaggle	`www.kaggle.com/datasets/uciml/pima-indians-diabetes-database`
placement	Kaggle	`https://www.kaggle.com/datasets/ahsan81/job-placement-dataset`
primary tumor	UCI	`https://archive.ics.uci.edu/ml/datasets/primary+tumor`
schizophrenia	UCI	`http://archive.ics.uci.edu/ml/datasets/mhealth+dataset`
shuttle	UCI	`https://archive.ics.uci.edu/ml/datasets/Shuttle+Landing+Control`
spambase	UCI	`https://archive.ics.uci.edu/ml/datasets/spambase`
student perf.	OpenML	`www.openml.org/search?type=data&sort=runs&id=42351&status=active`
tae	UCI	`archive.ics.uci.edu/ml/datasets/teaching+assistant+evaluation`
tic-tac-toe	UCI	`https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame`
vehicle	Kaggle	`www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho`
voting	Kaggle	`www.kaggle.com/datasets/devvret/congressional-voting-records`
wine	Kaggle	`www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009`

Table 4: List of benchmarks used in the experiments. Notably, *anneal2* is a binarized version of *annealing*, where the goal is to separate the label 2 from all other labels; *mnist49* (resp. *mnist38*) is a subset of *mnist* restricted to the instances labeled as 4 and 9 (resp. 3 and 8).