Locally Sparse Neural Networks for Tabular Biomedical Data

Junchen Yang, Ofir Lindenbaum, Yuval Kluger

ICML

2022

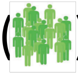
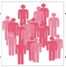
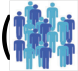
Heterogeneity in biomedical data

Biological population is often diverse

Heterogeneity in biomedical data

Biological population is often diverse

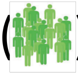
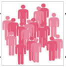
Different subgroups require different models

f()=y, g()=y, h()=y

Heterogeneity in biomedical data

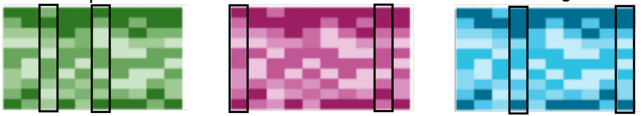
Biological population is often diverse

Different subgroups require different models

f()=y, g()=y, h()=y

The idea: find set of informative features locally

S_1 S_2 S_3



Heterogeneity in biomedical data

Consider 5 dimensional Gaussian Mixture Model, and define the response

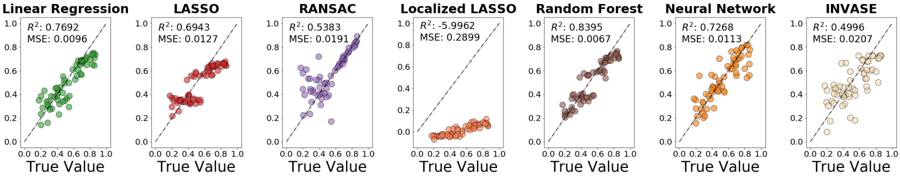
$$y = \begin{cases} -2\boldsymbol{x}_1 + \boldsymbol{x}_2 - 0.5\boldsymbol{x}_3, & \text{if in group 1} \\ -0.5\boldsymbol{x}_3 + \boldsymbol{x}_4 - 2\boldsymbol{x}_5, & \text{if in group 2} \end{cases}$$

Heterogeneity in biomedical data

Consider 5 dimensional Gaussian Mixture Model, and define the response

$$y = \begin{cases} -2\boldsymbol{x}_1 + \boldsymbol{x}_2 - 0.5\boldsymbol{x}_3, & \text{if in group 1} \\ -0.5\boldsymbol{x}_3 + \boldsymbol{x}_4 - 2\boldsymbol{x}_5, & \text{if in group 2} \end{cases}$$

Diversity in informative features makes the prediction challenging

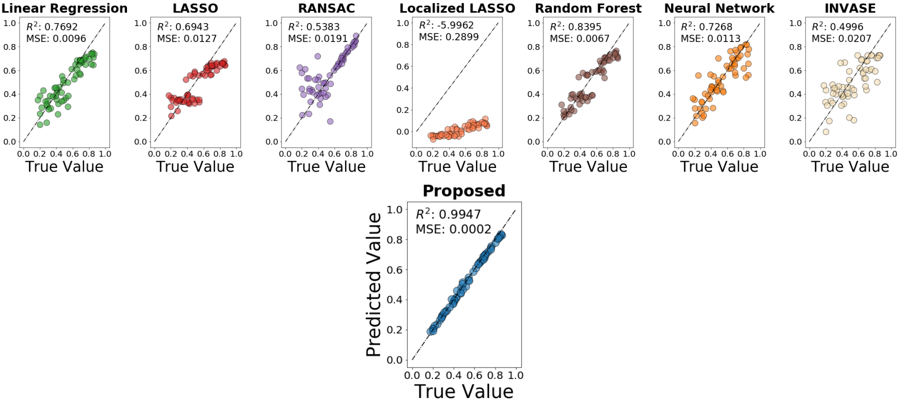


Heterogeneity in biomedical data

Consider 5 dimensional Gaussian Mixture Model, and define the response

$$y = \begin{cases} -2x_1 + x_2 - 0.5x_3, & \text{if in group 1} \\ -0.5x_3 + x_4 - 2x_5, & \text{if in group 2} \end{cases}$$

Diversity in informative features makes the prediction challenging

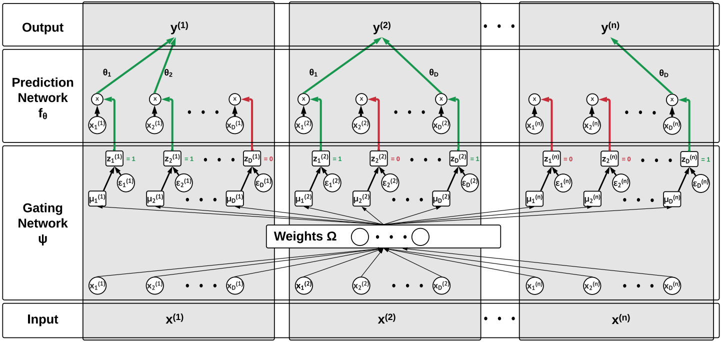


Locally **SP**arse **I**nterpretable **N**N (**LSPIN**)

Use a hyper-network to predict informative features

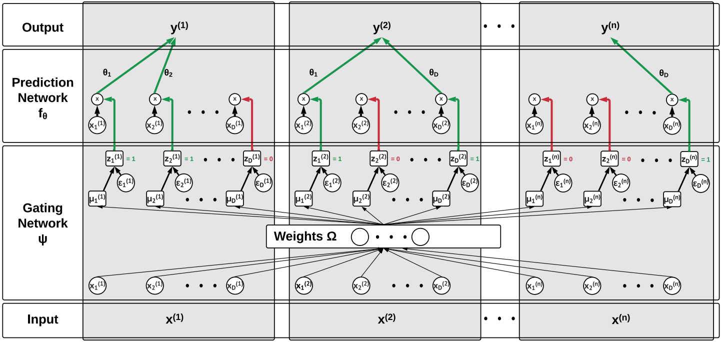
Locally **SP**arse **I**nterpretable **N**N (**LSPIN**)

Use a hyper-network to predict informative features



Locally **SP**arse **I**nterpretable **NN** (**LSPIN**)

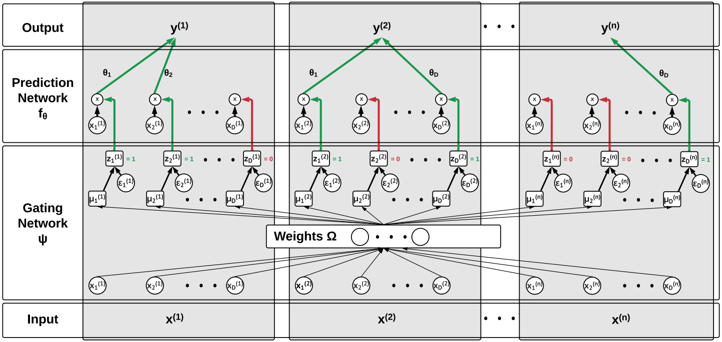
Use a hyper-network to predict informative features



- Minimize $\mathbb{E}[\mathcal{L}(\boldsymbol{f_\theta}(\boldsymbol{x}^{(i)} \odot \mathbf{z}^{(i)}), y^{(i)}) + \mathcal{R}(\mathbf{z}^{(i)})]$

Locally **SP**arse **I**nterpretable **NN** (**LSPIN**)

Use a hyper-network to predict informative features

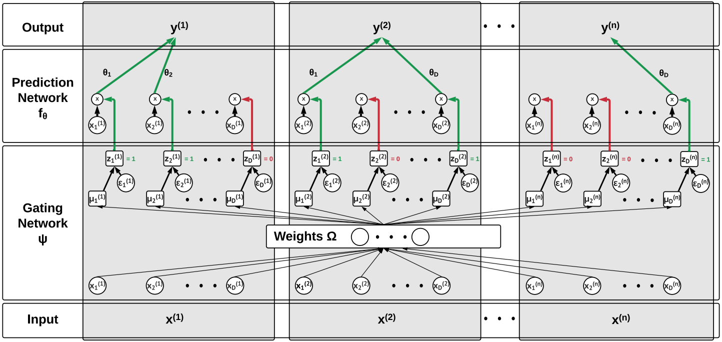


- Minimize $\mathbb{E}[\mathcal{L}(\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)} \odot \mathbf{z}^{(i)}), y^{(i)}) + \mathcal{R}(\mathbf{z}^{(i)})]$
- With $\mathcal{R}(\mathbf{z}^{(i)}) = \lambda_1 \|\mathbf{z}^{(i)}\|_0 + \lambda_2 \sum_j K_{i,j} \|\mathbf{z}^{(i)} - \mathbf{z}^{(j)}\|_2^2$.

[1] Yamada et al., 2020

Locally SParse Interpretable NN (LSPIN)

Use a hyper-network to predict informative features



- Minimize $\mathbb{E}[\mathcal{L}(\boldsymbol{f_\theta}(\boldsymbol{x}^{(i)} \odot \mathbf{z}^{(i)}), y^{(i)}) + \mathcal{R}(\mathbf{z}^{(i)})]$
- With $\mathcal{R}(\mathbf{z}^{(i)}) = \lambda_1 \|\mathbf{z}^{(i)}\|_0 + \lambda_2 \sum_j K_{i,j} \|\mathbf{z}^{(i)} - \mathbf{z}^{(j)}\|_2^2$.
- $K_{i,j} \geq 0$ is an affinity kernel

LSPIN Accurately recovers informative features

We generate data from $P(\boldsymbol{x}_j = 1) = P(\boldsymbol{x}_j = -1) = 0.5, j = 1, ..., 20$

E5: $$\text{y} = \begin{cases} \boldsymbol{x}_1 \times \boldsymbol{x}_2 + 2\boldsymbol{x}_{21}, & \text{if } \boldsymbol{x}_{21} = -1, \\ \boldsymbol{x}_2 \times \boldsymbol{x}_3 + 2\boldsymbol{x}_{21}, & \text{if } \boldsymbol{x}_{21} = 0, \\ \boldsymbol{x}_3 \times \boldsymbol{x}_4 + 2\boldsymbol{x}_{21}, & \text{if } \boldsymbol{x}_{21} = 1. \end{cases}$$

LSPIN Accurately recovers informative features

We generate data from $P(\boldsymbol{x}_j = 1) = P(\boldsymbol{x}_j = -1) = 0.5, j = 1, ..., 20$

E5:

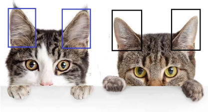
$$\mathrm{y} = \begin{cases} \boldsymbol{x}_1 \times \boldsymbol{x}_2 + 2\boldsymbol{x}_{21}, & \text{if } \boldsymbol{x}_{21} = -1, \\ \boldsymbol{x}_2 \times \boldsymbol{x}_3 + 2\boldsymbol{x}_{21}, & \text{if } \boldsymbol{x}_{21} = 0, \\ \boldsymbol{x}_3 \times \boldsymbol{x}_4 + 2\boldsymbol{x}_{21}, & \text{if } \boldsymbol{x}_{21} = 1. \end{cases}$$

	E1		**E2**		**E3**		**E4**		**E5**	
	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	MSE
LASSO	0.5000	52.00	0.5000	74.50	0.6250	71.50	0.1290	64.00	0.3704	1.0190
SVC	0.4444	51.50	0.5000	74.50	0.6667	71.50	0.2353	68.00	NA	NA
RF	0.5333	88.50	0.5333	88.50	0.6250	87.00	0.0769	86.00	0.2500	0.2499
XGBoost	0.5333	93.00	0.5333	**95.00**	0.6250	86.50	0.0769	96.00	0.2500	0.0118
MLP	NA	78.00	NA	88.50	NA	84.50	NA	64.00	NA	0.6526
Linear STG	0.4000	55.50	0.4000	76.00	0.3750	69.00	0.6667	70.00	0.5000	1.0067
Nonlinear STG	0.7272	84.50	0.7272	90.00	0.7143	86.00	0.6667	76.00	0.7500	**0.0004**
INVASE	0.5390	89.00	0.7000	88.00	0.6923	86.00	0.6667	94.00	0.1526	3.1264
L2X	0.7986	88.00	0.6050	94.50	0.2450	87.00	0.5000	92.00	0.6081	0.5134
TabNet	0.4789	54.50	0.5426	65.50	0.6905	78.50	0.0036	60.00	0.4454	1.0317
REAL-x	0.8306	85.00	0.7089	88.50	0.7823	86.00	0.8511	90.00	NA*	NA*
LLSPIN	0.3337	80.50	0.7216	86.50	0.4741	73.50	0.9458	90.00	0.6815	0.4927
LSPIN	**0.9761**	**94.00**	**0.8600**	**95.00**	**0.9296**	**89.00**	**0.9615**	**98.00**	**1**	0.0019

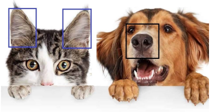
LSPIN leads to intepretable results

We use the following metrics to evaluate interpretability

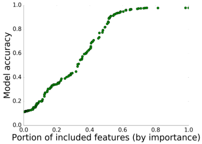
Stability



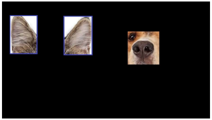
Diversity



Faithfulness



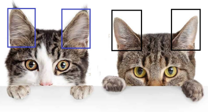
Generalizability



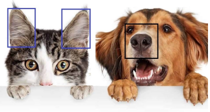
LSPIN leads to intepretable results

We use the following metrics to evaluate interpretability

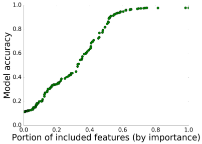
Stability



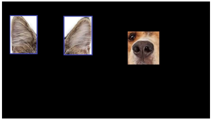
Diversity



Faithfulness



Generalizability

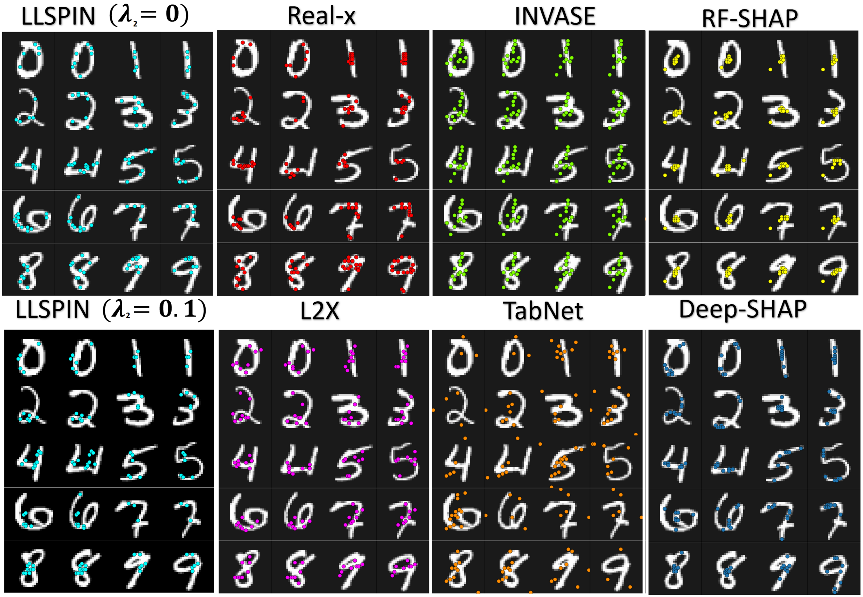


Using MNIST

Method	ACC	# Feat	Stability	Diversity	Faithfulness	Gen-SVM	Gen-k-means
RF-SHAP	96.81	8	0.418	41.15	0.531	52.42	40.15
Deep-SHAP	97.86	8	0.634	88.10	0.780	83.56	56.31
REAL-x	96.95	10	0.415	80.55	0.885	94.04	87.94
L2X	89.11	8	0.268	94.79	0.791	94.18	89.56
INVASE	85.07	11	0.162	13.43	0.864	69.67	43.02
TabNet	96.79	6	0.265	89.17	0.759	54.42	43.22
LLSPIN ($\lambda_2 = 0$)	98.26	7	0.294	99.14	0.950	98.22	97.50
LLSPIN ($\lambda_2 = 0.1$)	98.18	6	0.098	99.05	0.926	96.23	96.84
LSPIN ($\lambda_2 = 0$)	**98.45**	7	0.256	**99.84**	**0.987**	**98.43**	**97.99**
LSPIN ($\lambda_2 = 0.1$)	98.29	6	**0.065**	99.39	0.917	98.42	97.57

Making the model intepretable **by design** improves all metrics

LSPIN leads to intepretable results

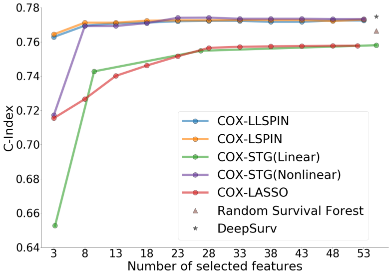


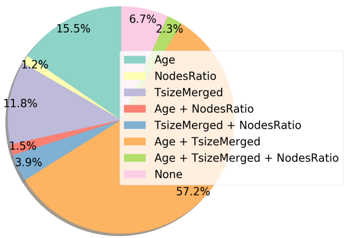
LSPIN is highly predictive on real-world tabular data

Survival analysis on the SEER data

LSPIN is highly predictive on real-world tabular data

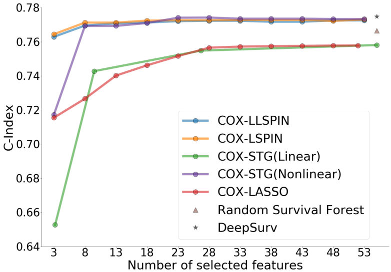
Survival analysis on the SEER data

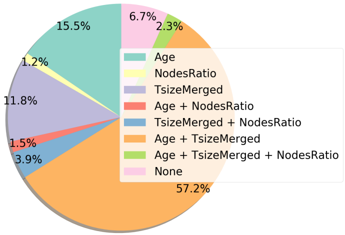




LSPIN is highly predictive on real-world tabular data

Survival analysis on the SEER data

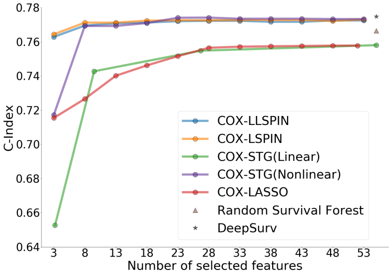


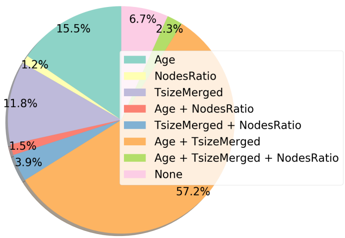


Classification on low-sample-size (LSS) tabular data

LSPIN is highly predictive on real-world tabular data

Survival analysis on the SEER data





Classification on low-sample-size (LSS) tabular data

	BASEHOCK	RELATHE	PCMAC	PBMC	COLON	TOX-171	Median Rank
LASSO	74.46 ± 5.19 [34]	58.69 ± 1.59 [18]	68.09 ± 4.08 [21]	90.30 ± 0.36 [31]	81.54 ± 9.85 [24]	87.71 ± 4.62 [49]	6.5
SVC	74.46 ± 3.37 [22]	56.48 ± 3.00 [6]	67.41 ± 3.72 [12]	89.02 ± 0.74 [30]	76.15 ± 9.39 [25]	81.14 ± 7.47 [38]	8.5
RF	64.46 ± 4.52 [10]	71.42 ± 3.50 [50]	67.44 ± 7.00 [9]	48.56 ± 6.18 [10]	79.23 ± 9.76 [47]	53.71 ± 9.96 [42]	11.5
XGBoost	90.37 ± 1.05 [45]	76.75 ± 1.67 [32]	83.93 ± 0.67 [43]	76.58 ± 0.72 [64]	76.15 ± 12.14 [7]	67.43 ± 5.60 [38]	6
MLP	56.51 ± 1.43	55.44 ± 2.38	54.38 ± 1.27	61.57 ± 1.45	81.54 ± 7.84	62.59 ± 8.03	12.5
Linear STG	89.36 ± 1.40 [27]	69.94 ± 5.05 [16]	$\mathbf{85.11 \pm 1.07}$ **[42]**	88.22 ± 0.82 [27]	74.62 ± 11.44 [14]	71.14 ± 5.78 [16]	7
Nonlinear STG	89.24 ± 1.18 [20]	74.83 ± 3.95 [27]	84.16 ± 0.90 [32]	86.29 ± 1.31 [19]	76.15 ± 13.95 [8]	67.43 ± 7.25 [14]	6.5
INVASE	84.02 ± 0.81 [42]	70.81 ± 1.56 [43]	77.06 ± 1.01 [48]	86.34 ± 0.81 [30]	76.92 ± 12.40 [6]	76.86 ± 7.39[26]	7.5
L2X	88.48 ± 2.01 [1]	77.10 ± 5.19 [10]	78.69 ± 3.62 [10]	70.77 ± 11.24 [10]	78.46 ± 8.28 [8]	71.71 ± 10.42[9]	6.5
TabNet	88.21 ± 2.00 [3]	67.84 ± 15.40 [10]	69.35 ± 10.49 [4]	$\mathbf{92.13 \pm 0.59}$ **[3]**	64.62 ± 12.02 [28]	30.00 ± 6.29 [34]	9.5
REAL-x	89.80 ± 1.96 [5]	80.61 ± 1.31 [3]	80.98 ± 3.05 [6]	83.39 ± 2.19 [24]	75.38 ± 12.78 [15]	77.71 ± 7.65 [42]	5
LSPIN	89.37 ± 1.48 [3]	80.59 ± 1.95 [3]	78.51 ± 1.48 [3]	88.67 ± 0.64 [15]	71.54 ± 6.92 [1]	90.29 ± 5.45[1]	4.5
LLSPIN	$\mathbf{91.56 \pm 1.51}$ **[4]**	$\mathbf{82.01 \pm 2.20}$ **[11]**	81.48 ± 1.74 [3]	90.43 ± 0.6 [18]	$\mathbf{83.85 \pm 5.38}$ **[7]**	$\mathbf{92.57 \pm 6.41}$ **[6]**	**1**

Thank You!

Code: https://github.com/jcyang34/lspin