Neural Basis Models for Interpretability (NeurIPS, 2022)

SeongSik Choi

January 22, 2024

Seoul National University

Generalized Additive Model (GAM)

Given a input $\boldsymbol{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D$, a label $y \in \mathbb{R}$, a link function $g : \mathbb{R}^D \to \mathbb{R}$, $g(\boldsymbol{x})$ can be expressed as

$$\text{GAM} : g(\boldsymbol{x}) = f_0 + \sum_{i=1}^{D} f_i\left(x_i\right)$$

$$\text{GA}^2\text{M} : g(\boldsymbol{x}) = f_0 + \sum_{i=1}^{D} f_i\left(x_i\right) + \sum_{i=1}^{D} \sum_{j>i} f_{ij}(x_i, x_j)$$

for some bias $f_0 \in \mathbb{R}$, univariate functions f_i, and bivariate functions $f_{ij} : \mathbb{R} \to \mathbb{R}$.

NAM and NBM

$$\text{GAM} : g(\boldsymbol{x}) = f_0 + \sum_{i=1}^{D} f_i\left(x_i\right)$$

Neural Additive Model (NAM): each f_i is parametrized by DNN.

Neural Basis Model (NBM): each f_i is represented as $f_i(x_i) = \sum_{k=1}^{B} h_k(x_i) a_{ik}$. And basis functions $(h_1, \ldots, h_B) : \mathbb{R} \to \mathbb{R}^B$ are parametrized by DNN.

NBM Extension(Multi-class)

$$\text{Multiclass GAM} : g_l(\boldsymbol{x}) = f_{0l} + \sum_{i=1}^{D} f_i\left(x_i\right) w_{il}$$

Neural Basis Model (NBM): each f_i is represented as $f_i(x_i) = \sum_{k=1}^{B} h_k(x_i) a_{ik}$. And basis functions $(h_1, \ldots, h_B) : \mathbb{R} \rightarrow \mathbb{R}^B$ are parametrized by DNN.

NBM Extension(NB^{2}M)

$$\text{GA}^2\text{M} : g(\boldsymbol{x}) = f_0 + \sum_{i=1}^{D} f_i\left(x_i\right) + \sum_{i=1}^{D}\sum_{j>i} f_{ij}(x_i, x_j)$$

NB^{2}M: each f_{ij} is represented as $f_{ij}\left(x_i, x_j\right) = \sum_{k=1}^{B} u_k\left(x_i, x_j\right) b_{ijk}$.

And additional basis functions $(u_1, \ldots, u_B) : \mathbb{R}^2 \to \mathbb{R}^B$ are parametrized by DNN.

Extension to multi-class setting can be done in the similar way as for NBM.

Selecting the number of bases

If all f_is are in an RKHS, then risk converges to 0 as $n \to 0$.
$\Rightarrow B = O(\log D)$ bases are sufficient.
The proof seems a little awkward to me.

Rather than tuning this hyperparameter, they recommend setting B = 100 for NBM and B = 200 for NB^{2}M as it performs well across a large variety of datasets they experimented with.

NAM vs NBM(Overview)

(1) Number of parameters : Number of weight parameters needed to learn the model. When the input dimension is large, NBM has far fewer parameters than NAM.

(2) Throughput : The number of data instances processed per second, which directly affects the training speed. NBM are much more efficient than NAM.

(3) Performance : NBM outperform NAM and NODE-GAM(state of the art) on most datasets.

(4) Stability : the functions f_i of NBM are much more stable than those of NAM.

NAM vs NBM

(1) Number of parameters and (2) Throughput

Model	CA Housing		FICO		CoverType		Newsgroups		iNat. Birds	
	#par.	x/sec	#par.	x/sec	#par.	x/sec	#par.	x/sec	#par.	x/sec
NAM	54K	0.5M	262K	123K	363K	80K	984M	23	2.3M	15K
NBM	65K	3.4M ×6.8	68K	821K ×6.7	70K	530K ×6.6	18M	†9K ×391	0.5M	74K ×4.9
NA^{2}M	243K	119K	5.3M	6K	10M	3K	–	–	320M	99
NB^{2}M	161K	641K ×5.4	0.3M	30K ×5.0	0.5M	15K ×5.0	–	–	66M	374 ×3.8

When the input dimension is large, NBM has far fewer parameters than NAM.

NBM are much more efficient than NAM.

NAM vs NBM

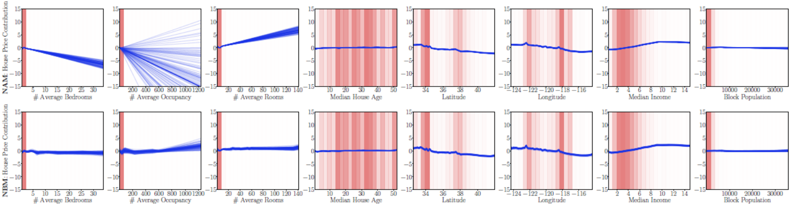
(3) Performance

Model	MIMIC-II	Credit	Click	Epsilon	Higgs	Microsoft	Yahoo	Year
	AUROC ↑	AUROC ↑	Error ↓	Error ↓	Error ↓	MSE ↓	MSE ↓	MSE ↓
NAM	0.8539 ±0.0004	0.9766 ±0.0027	0.3317 ±0.0005	0.1079 ±0.0002	0.2972 ±0.0001	0.5824 ±0.0002	0.6093 ±0.0003	85.25 ±0.01
NODE GAM	0.8320 ±0.0110	0.9810 ±0.0110	0.3342 ±0.0001	0.1040 ±0.0003	0.2970 ±0.0001	0.5821 ±0.0004	0.6101 ±0.0006	**85.09** ±0.01
NBM	**0.8549** ±0.0004	**0.9829** ±0.0014	**0.3312** ±0.0002	**0.1038** ±0.0002	**0.2969** ±0.0001	**0.5817** ±0.0001	**0.6084** ±0.0001	85.10 ±0.01
NA^2M	0.8639 ±0.0011	0.9824 ±0.0032	0.3290 ±0.0005	—	0.2555 ±0.0003	0.5622 ±0.0003	—	79.80 ±0.05
NODE GA^2M	0.8460 ±0.0110	**0.9860** ±0.0100	0.3307 ±0.0001	0.1050 ±0.0002	0.2566 ±0.0003	**0.5618** ±0.0003	**0.5807** ±0.0004	79.57 ±0.12
NB^2M	**0.8690** ±0.0010	0.9856 ±0.0017	**0.3286** ±0.0002	—	**0.2545** ±0.0002	**0.5618** ±0.0002	—	**79.01** ±0.03

NBM outperform NAM and NODE-GAM(SOTA) on most datasets.

NAM vs NBM

(4) Stability



The functions f_i of NBM are much more stable than those of NAM.