Towards using Splines as Activation in Neural Networks

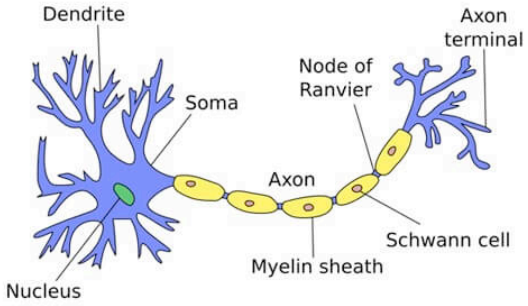
STAT690STA: Semi-Parametric Regression

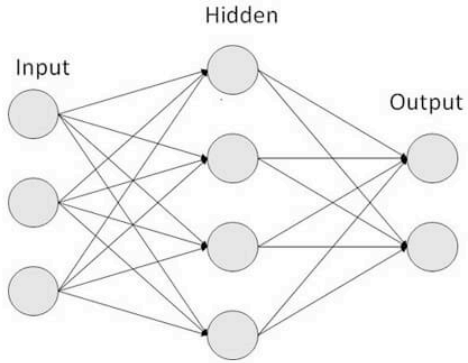
University of Massachusetts Amherst

• By Animesh Sengupta

What is a Neural Network ?

1. A neural network is a type of machine learning model that is inspired by the structure and function of the brain.
2. It is composed of layers of interconnected processing nodes (or neurons) that can learn to recognize patterns and relationships in data by adjusting the strength of the connections between them based on training data.
3. The input layer receives the raw data, the output layer produces the predictions or classifications, and the hidden layers in between perform transformations on the data to extract higher-level features.
4. Think of each neuron as a separate Linear model

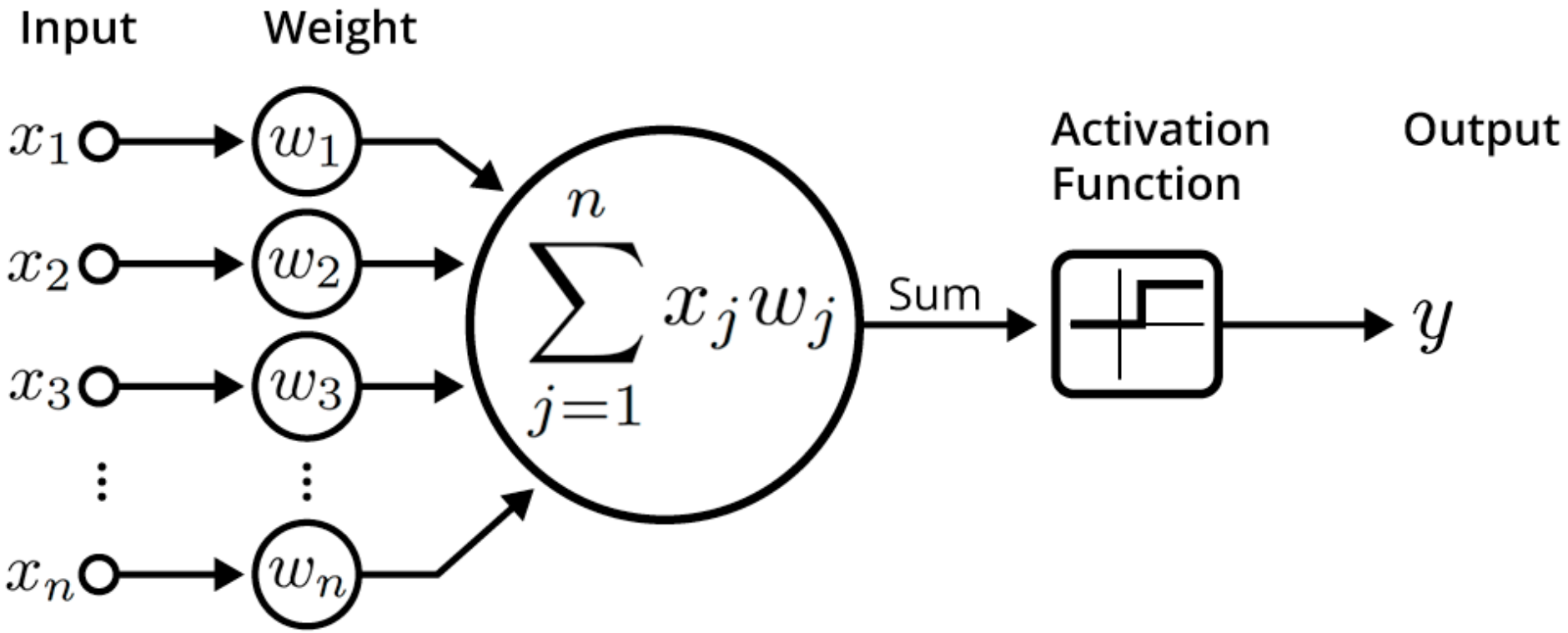




University of Massachusetts Amherst

Recipe for a Neural Network

1. We need some linear models called as neurons.
2. Some Activation functions
3. An efficient and super fast optimizer
4. Bunch of Hyper Parameters
5. **LOTS** of computation power



An illustration of an artificial neuron. Source: Becoming Human.

University of Massachusetts Amherst

Motivation of Study

During this study we aim to answer few of these questions:

1. Can spline be used as activation function of neural nets?

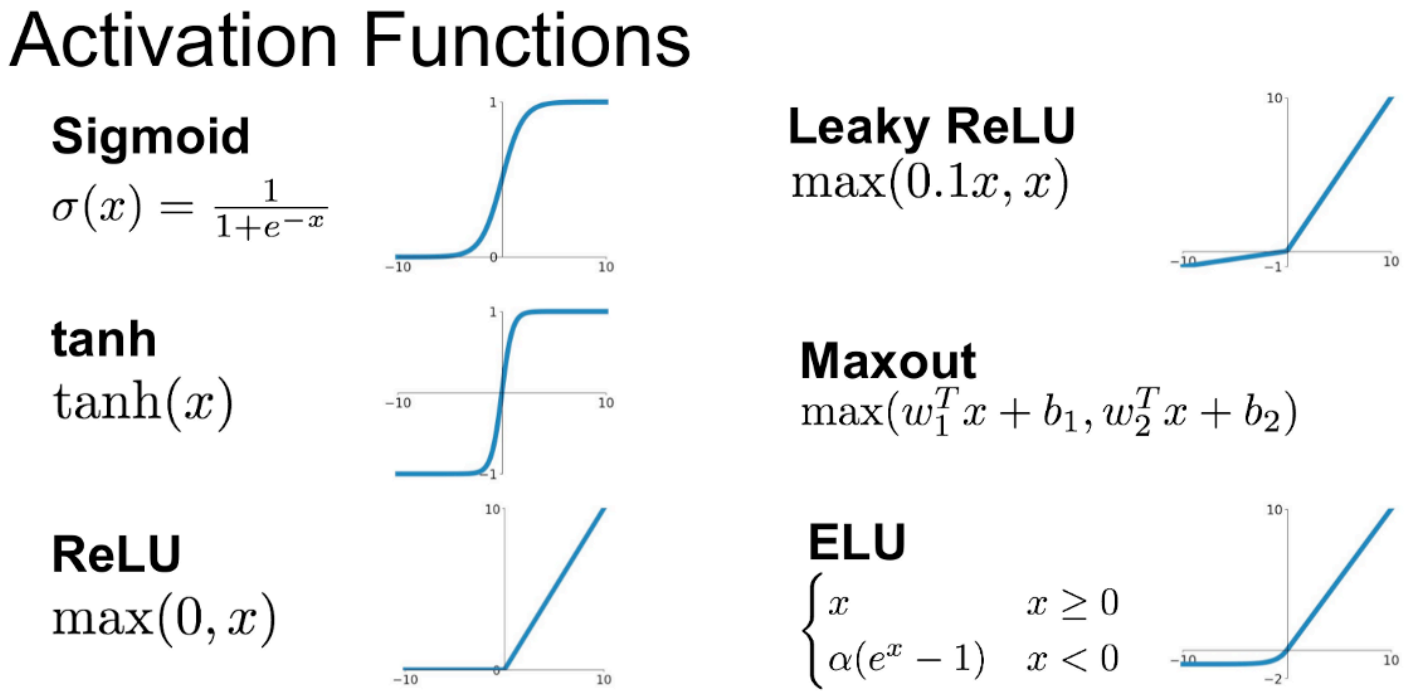
2. Will it be computationally efficient ?

3. Would spline work well to introduce non linearity ?

4. Can these models be used for predictions and inferential analysis ?

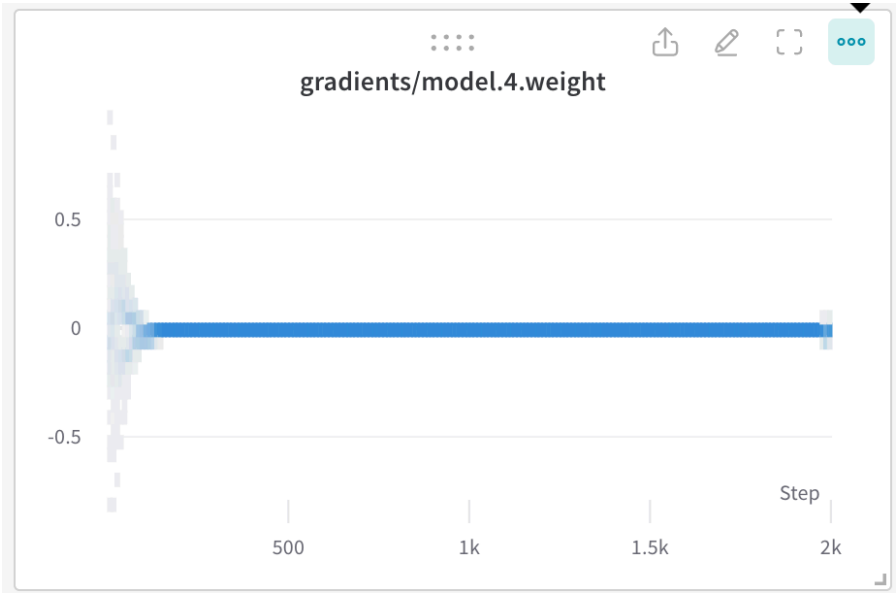
What is an activation function ?

1. In neural networks, activation functions are used to introduce nonlinearity into the model.
2. They are applied to the output of each neuron in the hidden layers, transforming it into a more complex representation that can capture more complex patterns in the data.
3. Activation functions can be categorized into two types: linear and nonlinear.
4. f(W'x+b)



Issues with Activation functions

1. Sigmoidal functions have vanishing gradients issue
2. ReLU have issues with dead neurons
3. We still need to control number of layers and units to add complexity and non linearity to the model



Motivation Behind Splines as Activation function

1. Curiosity. Why not splines be used as activation function ?
2. ReLU is the most basic linear splines
3. Cubic Splines are better smoother and non linear function than linear splines
4. If we can introduce better smoothing and non linearity using splines, we can reduce the number of layers and units.
5. Smaller architecture means smaller training time and lesser computational consumption
6. Continuous and piecewise linear CPWL

Splines Equation

$$\varphi_k(x) = \beta^1(x-k), \text{ for } k_{\min} < k < k_{\max},$$

$$\beta^1(x) = (x+1)_+ - 2(x)_+ + (x-1)_+$$

$$= \begin{cases} 1-|x|, & x \in [-1,1] \\ 0, & \text{otherwise.} \end{cases}$$

$$\varphi_{k_{\min}-1}(x) = (-x+k_{\min})_+ = \begin{cases} k_{\min}-x, & x < k_{\min} \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

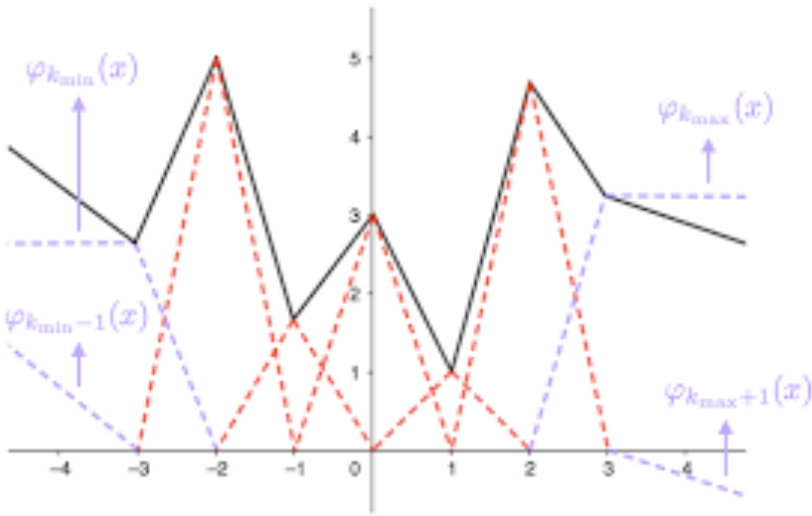
$$\varphi_{k_{\min}}(x) = (-x+k_{\min}+1)_+ - (-x+k_{\min})_+$$

$$= \begin{cases} 1, & x \le k_{\min} \\ 1-(x-k_{\min}), & x \in (k_{\min}, k_{\min}+1) \\ 0, & x \ge k_{\min}+1 \end{cases}$$

$$\varphi_{k_{\max}}(x) = (x-k_{\max}+1)_+ - (x-k_{\max})_+$$

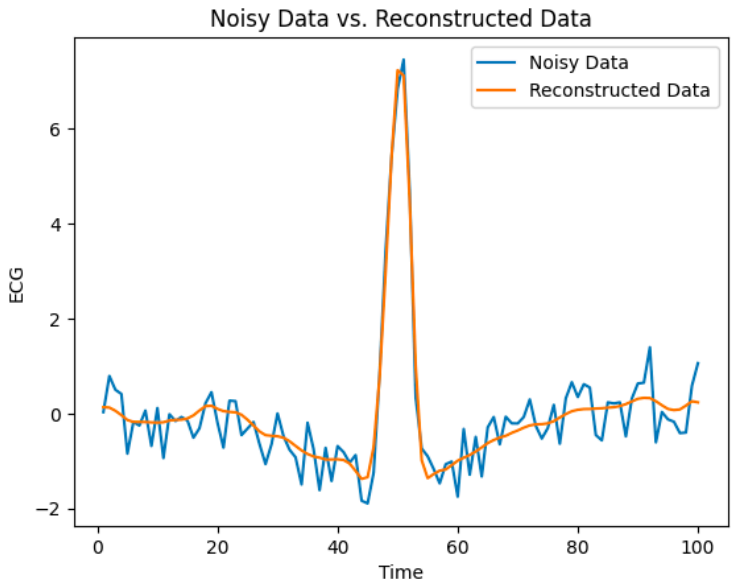
$$= \begin{cases} 0, & x \le k_{\max}-1 \\ x-k_{\max}+1, & x \in (k_{\max}-1, k_{\max}) \\ 1, & x \ge k_{\max} \end{cases}$$

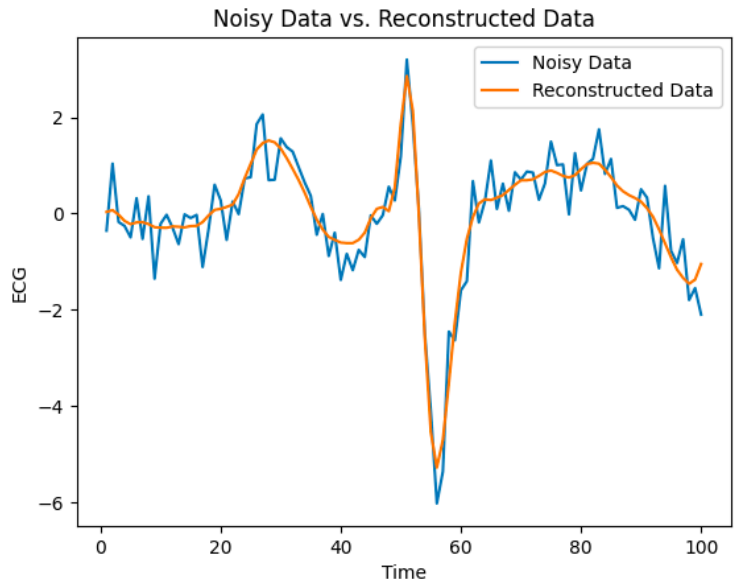
$$\varphi_{k_{\max}+1}(x) = (x-k_{\max})_+ = \begin{cases} 0, & x \le k_{\max} \\ x-k_{\max}, & x > k_{\max}. \end{cases} \quad (15)$$

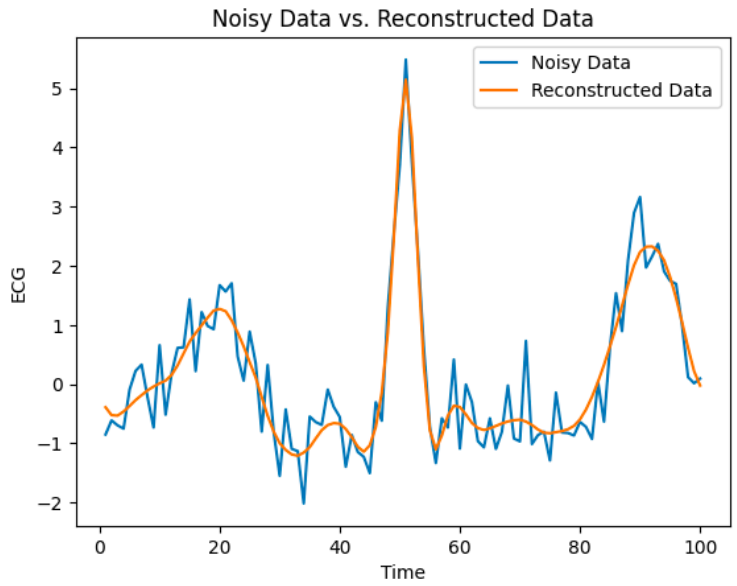


University of Massachusetts Amherst

Smoothing Effectiveness of Spline







We experminted on smoothing noisy ECG signals using Splines based deep neural nets. We chose 21 knots and size as 5. After 2500 iterations , we got MSE loss of 0.06.

University of Massachusetts Amherst

California Housing Data

The data pertains to the houses found in a given California district and some summary stats about them based on the 1990 census data.A block group is the smallest geographical unit for which the U.S.
Census Bureau publishes sample data (a block group typically has a population
of 600 to 3,000 people).

Dependent Variables

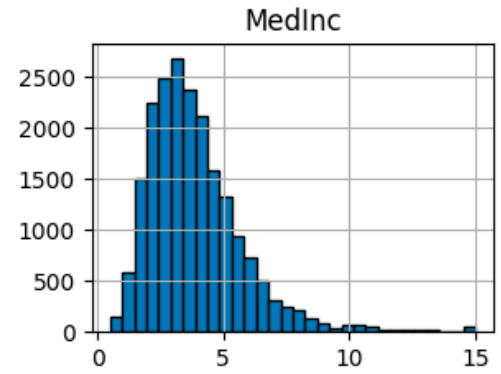
- MedInc median income in block group. Measured per 100k USD
- HouseAge median house age in block group.
- AveRooms average number of rooms per household
- AveBedrms average number of bedrooms per household
- Population block group population
- AveOccup average number of household members
- Latitude block group latitude
- Longitude block group longitude

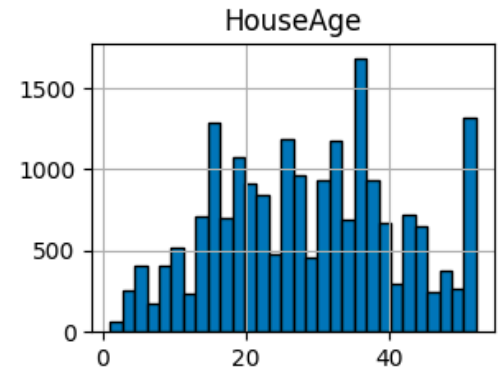
Independent Variable

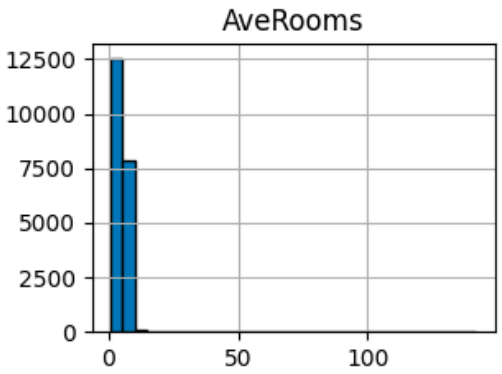
- MedianHouseVal - the median house value for California districts, expressed in hundreds of thousands of dollars ($100,000).

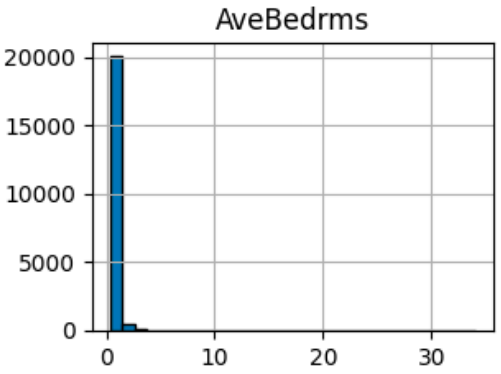
University of Massachusetts Amherst

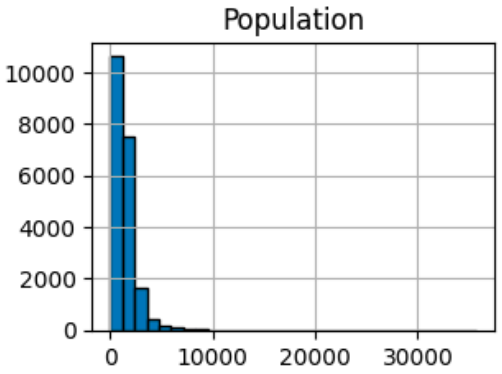
Histograms

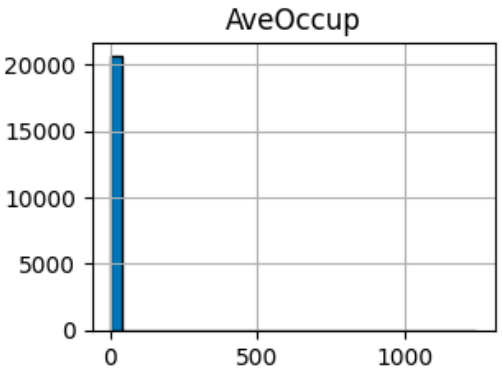


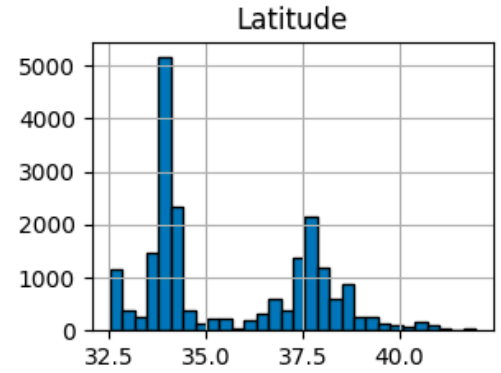


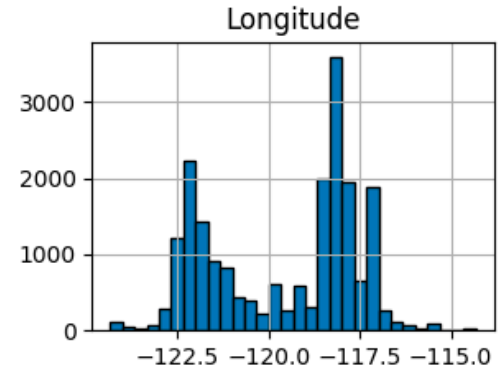


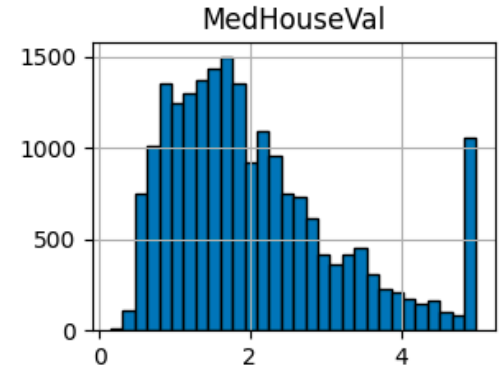






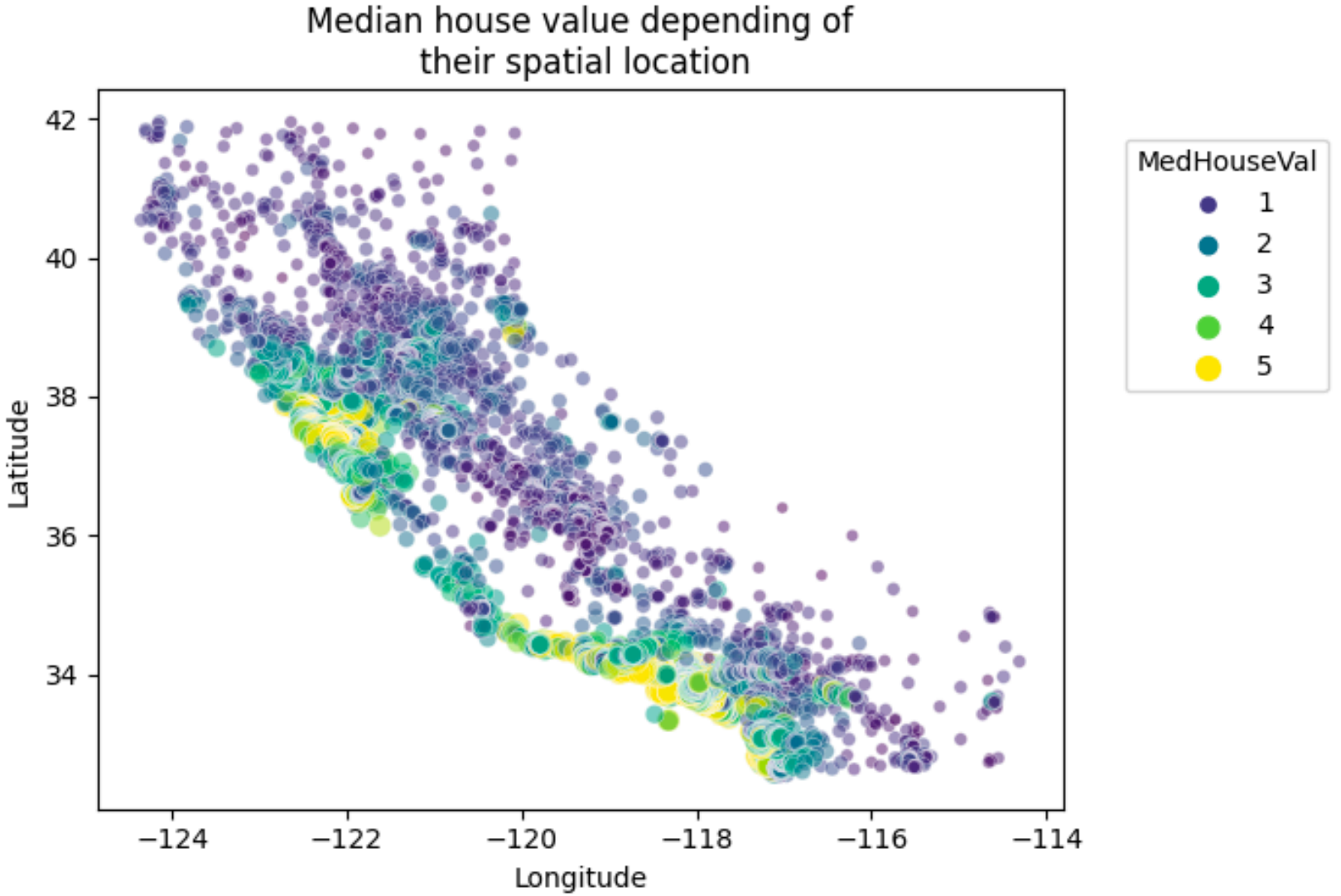




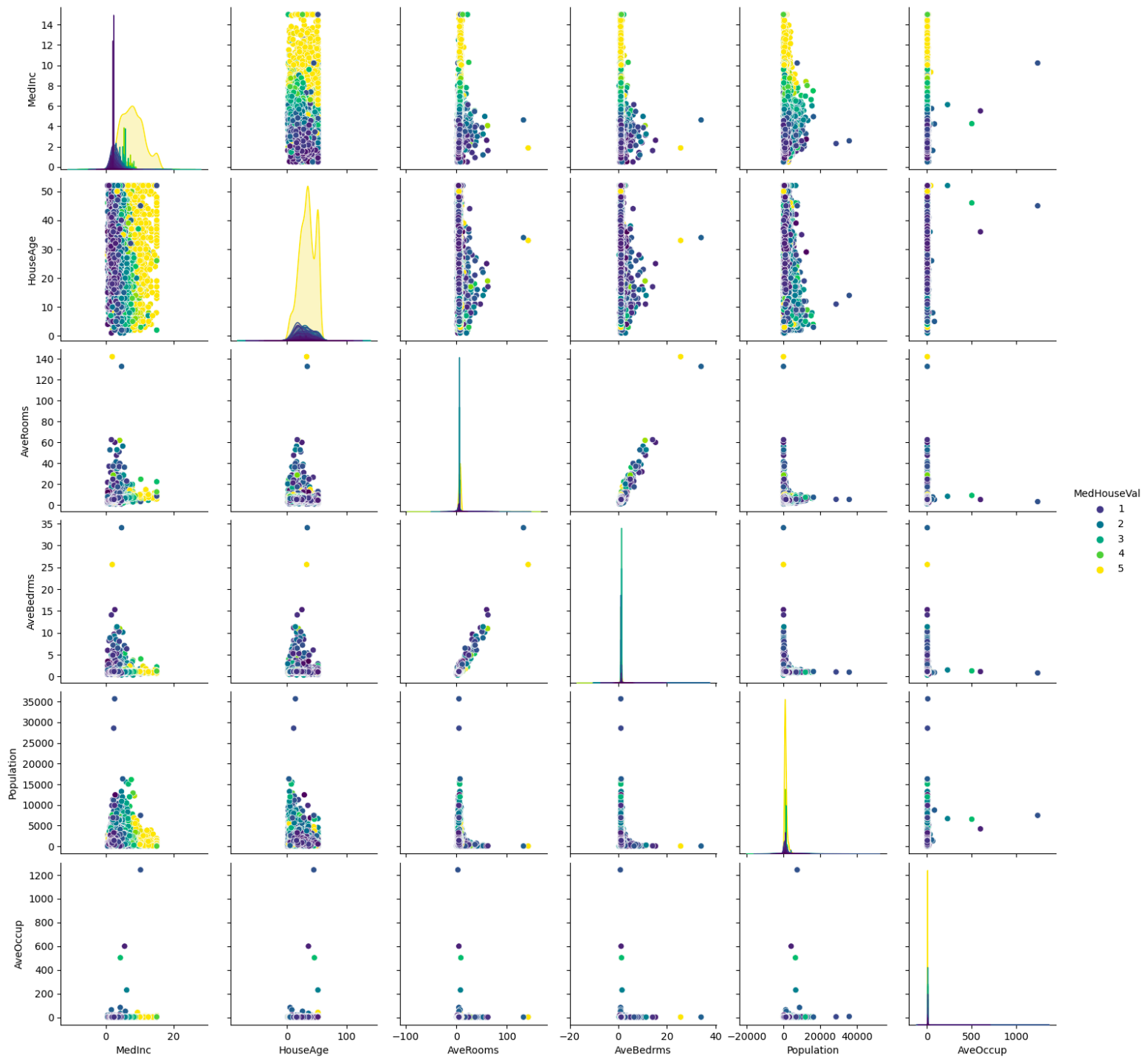


iversity of
ssachusetts
herst

Median House Value based on their target location



University of Massachusetts Amherst



University of Massachusetts Amherst

Neural Network Architecture Components

The Components of the Neural Net architecture are as follows

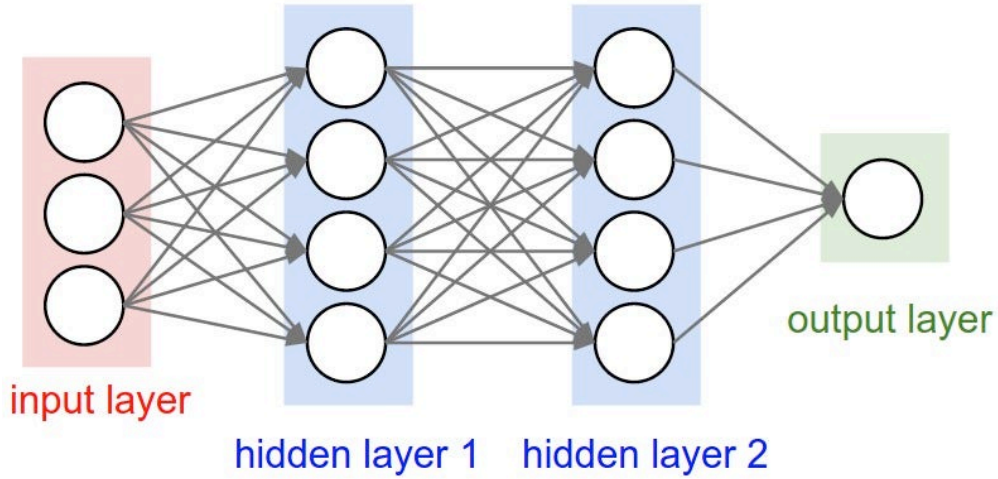
1. Input dimension - 8 for each units

2. Hidden layers - 2 Hidden layers with 24 and 12 as the dimensions respectively

3. Adam Optimizer with learning rate as 0.01 and random initialisation of parameters. De Facto method

4. Activation functions - Tanh vs ReLU vs Cubic Splines

5. 1000-2500 Epochs - one epoch is one iteration through the whole dataset.



University of Massachusetts Amherst

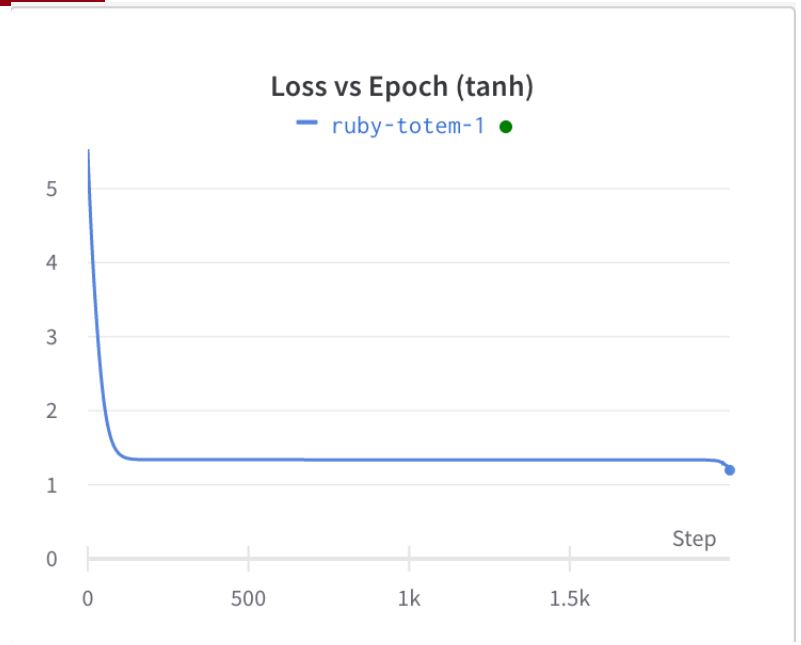
HyperParameter Selection

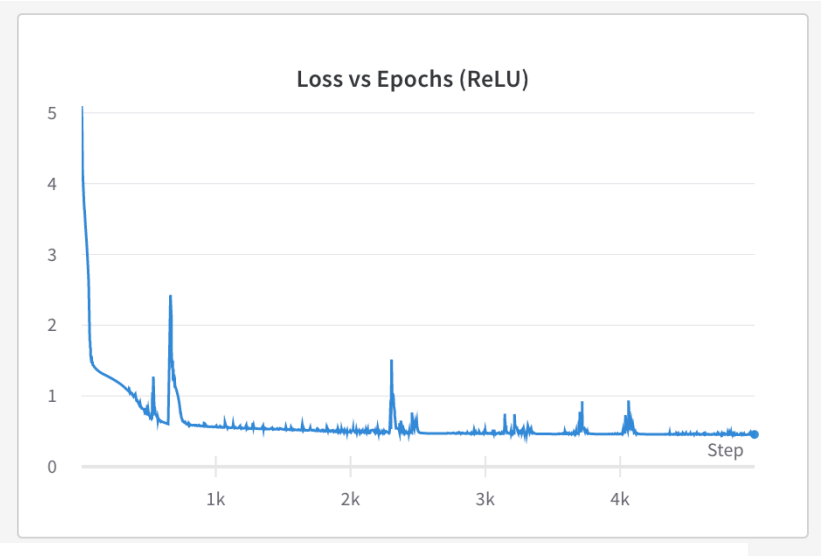
The following hyperparameter were selected using grid search and cross validation.

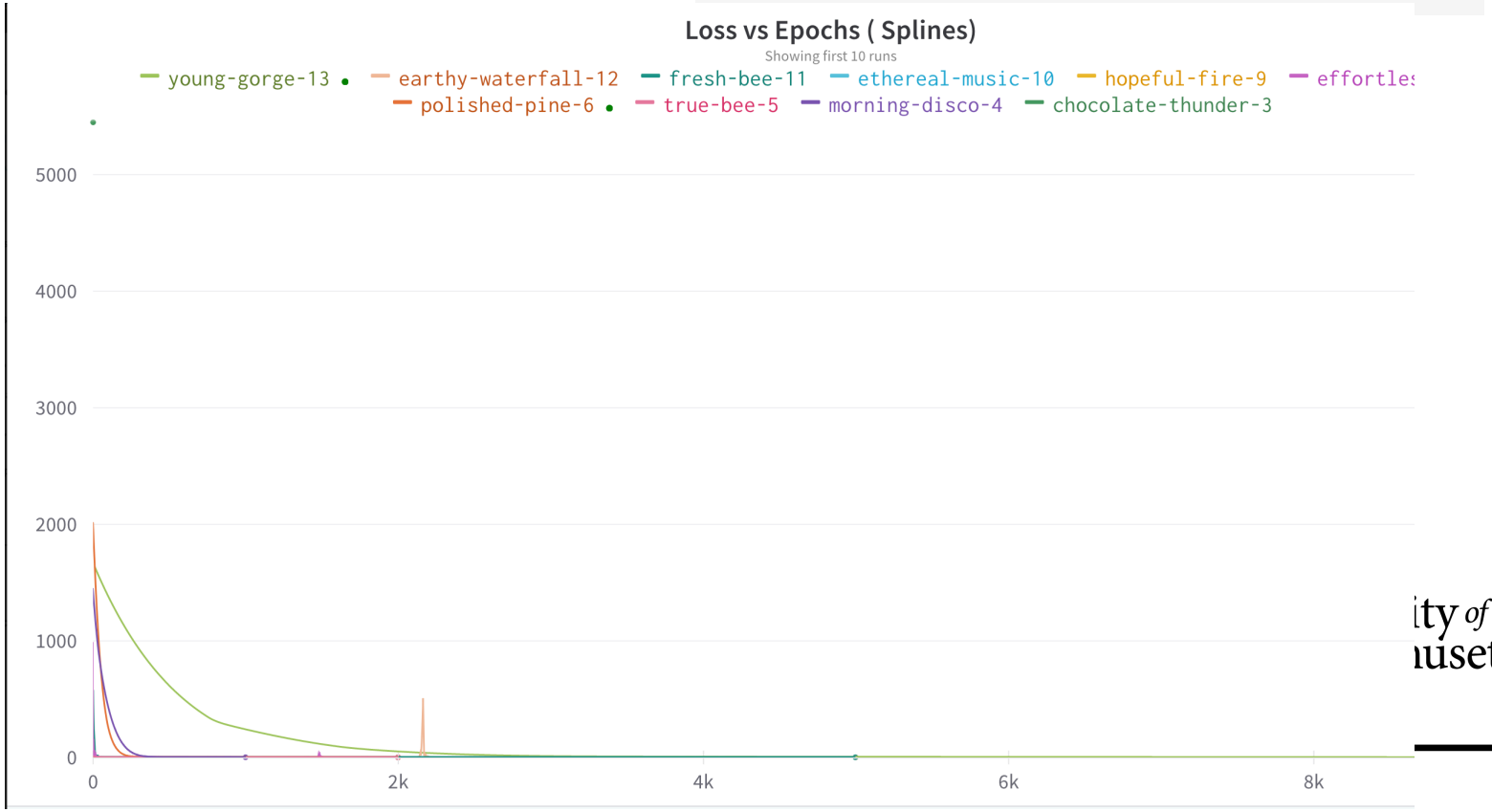
1. **Optimizer : Adam - De Facto Method**
2. **Learning rate: 0.01 for faster convergence**
3. **Epochs - 2500 since by 2500 epochs the returns are diminished and there is no learning**
4. **No. Of layers and Units - Kept constant for each experiment**
5. **Spline knot - [11,21,31,41,51,61,71] grid search**
6. **Spline Range - [1,7]**
7. **Cross Validation using Test and Train split.**

University of Massachusetts Amherst

Error vs Loss for experimentation.





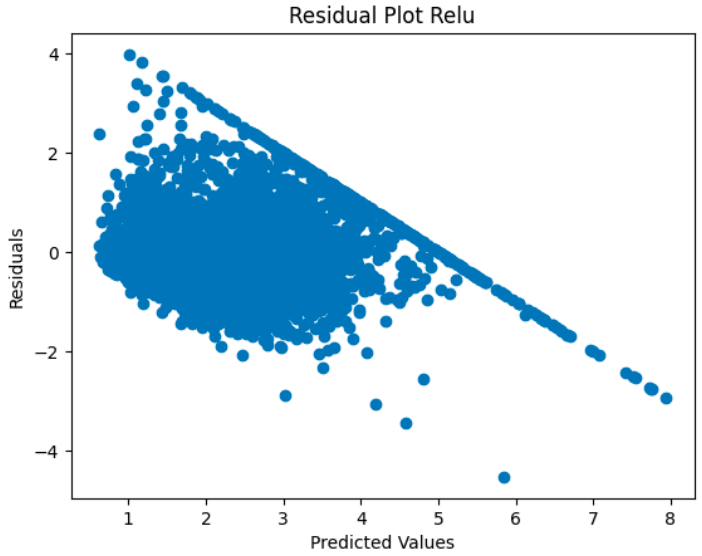


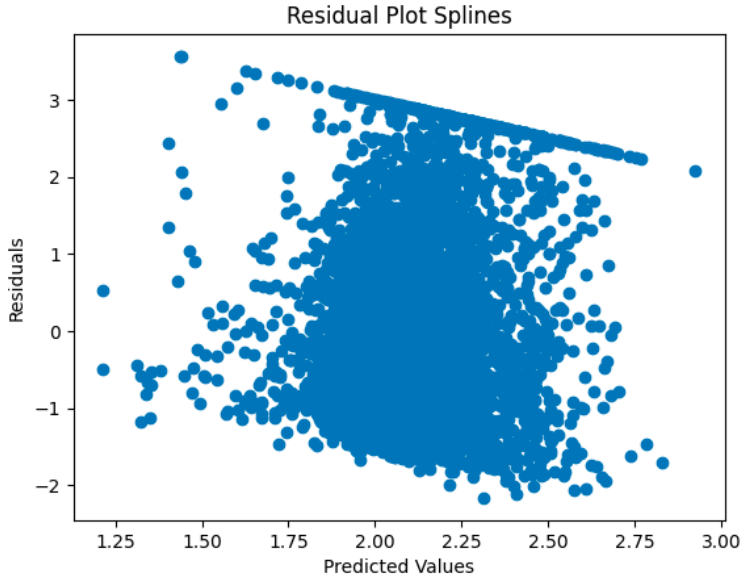
Experimentation Results

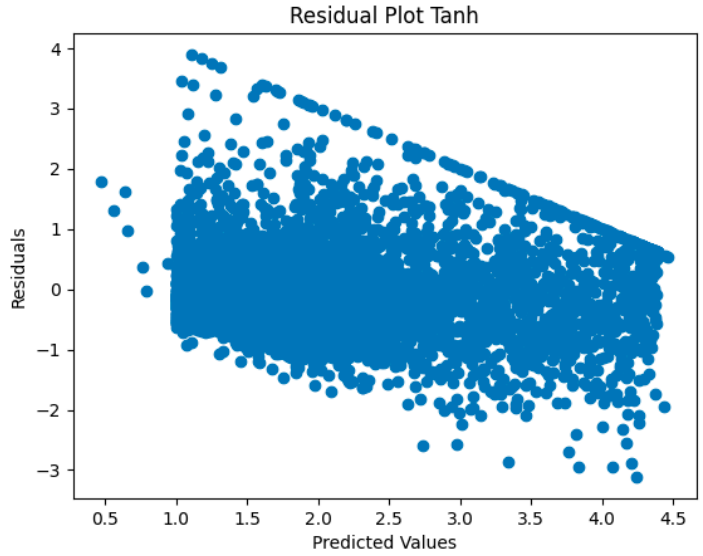
Model Name	Activation	Epoch	MSE Loss	Time(s)
Neural Net	Tanh	2000	1.48	20s
Neural Net	Relu	2000	0.49	11s
Neural Net	Splines	2000	0.6	90s
pyGAM	Splines	1	2.2	433s

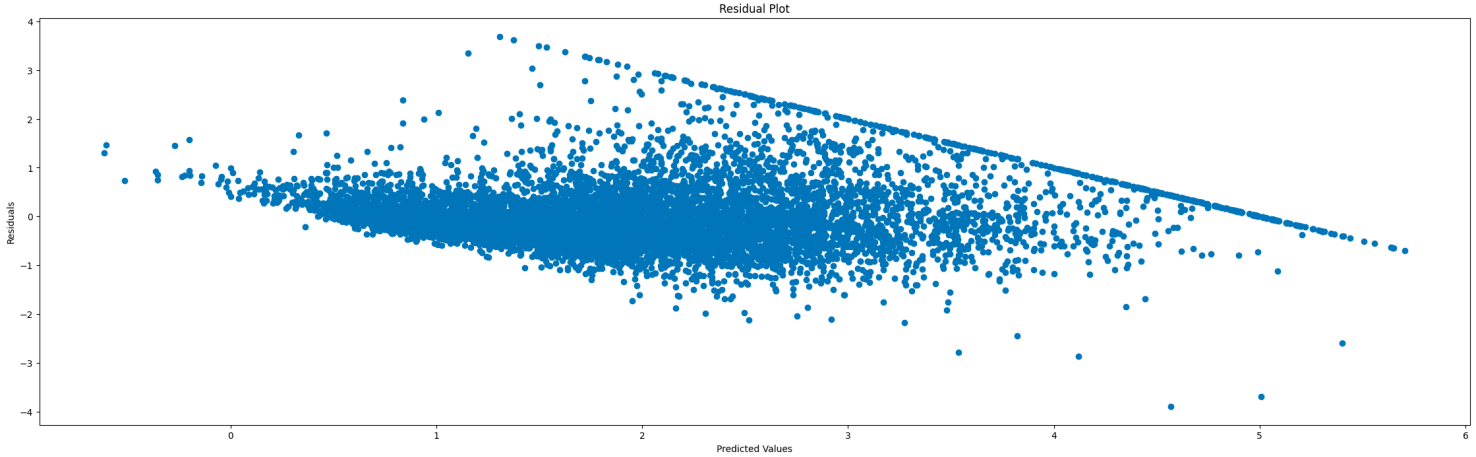
University of Massachusetts Amherst

Residual plots



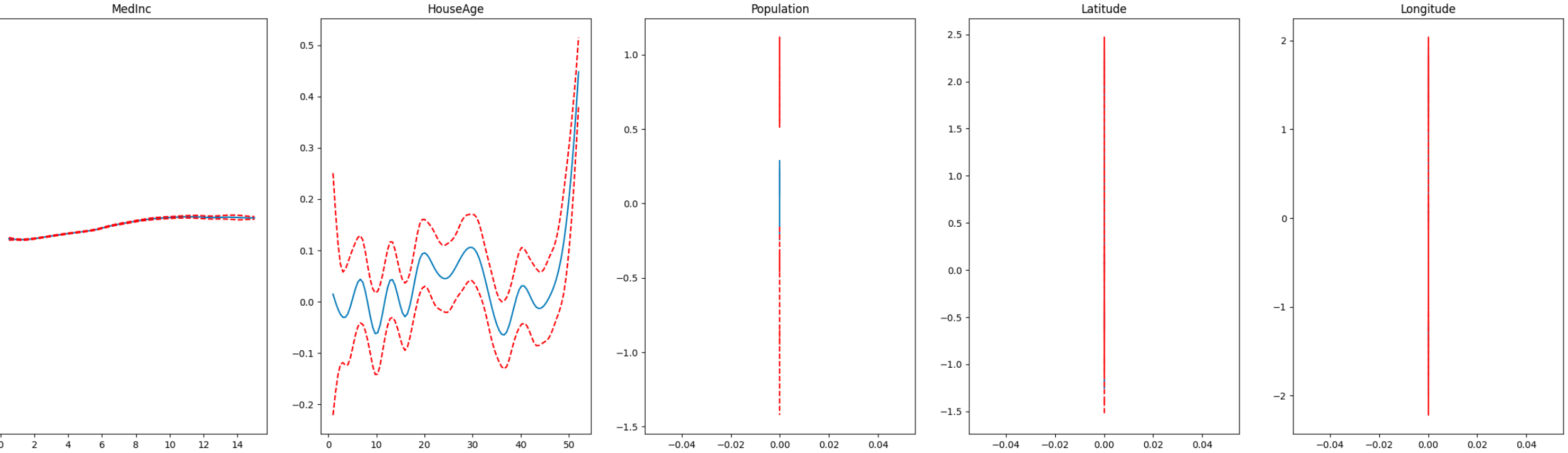






University of Massachusetts Amherst

Partial Dependency plots



University of Massachusetts Amherst

Conclusion

1. Splines can indeed be used in neural network to construct deep neural net techniques.
2. It is much faster to process larger datasets than the traditional gam function.
3. It can very well act as a smoothing function to bring in non linearity in neural network architecture

References

1. Dataset: Pace, R. Kelley, and Ronald Barry. "Sparse spatial autoregressions." Statistics & Probability Letters 33.3 (1997): 291-297.

2. Bohra, J. Campos, H. Gupta, S. Aziznejad, M. Unser, "Learning Activation Functions in Deep (Spline) Neural Networks," IEEE Open Journal of Signal Processing, vol. 1, pp.295-309, November 19, 2020.

3. Aziznejad, H. Gupta, J. Campos, M. Unser, "Deep Neural Networks with Trainable Activations and Controlled Lipschitz Constant," IEEE Transactions on Signal Processing, vol. 68, pp. 4688-4699, August 10, 2020.

4. Deepsplines by : https://github.com/joaquimcampos/DeepSplines

Questions and Answers

Thanks You !

University of Massachusetts Amherst