EE6222: MACHINE VISION

Project Report (Assignment 1)

Author:

Teo Chen Ning (U1820456K)

Date of submission:

06 November 2021

Question 1

What is the effect of direct links from the input layer to the output layer (i.e. with and without)?

Due to limitations regarding hardware, datasets above 20000 rows will not be chosen. Instead, a sampling of 10 datasets were picked out of the 121 datasets of the UCI repository, which remain constant throughout the remaining report. The 10 datasets, along with their shapes are:

Dataset name	Shape (including index and label columns)
abalone	(4177, 10)
bank	(4521, 18)
car	(1728, 8)
dermatology	(366, 36)
echocardiogram	(131, 12)
glass	(214, 11)
parkinsons	(195, 24)
seeds	(210, 9)
titanic	(2201, 5)
wine	(178, 15)

All datasets chosen comes with three files:

1. **<name>_R.mat** (The actual columnar data with labels)
2. **<name>_conxuntos.mat** (Indexes for train-test partitioning)
3. **<name>_conxuntos_kfold.mat** (Indexes for 4-fold cross validation)

I will be testing various options for the Random Vector Functional Link (RVFL) model, a classical single layer feedforward neural network with randomized weights. The RVFL model will be trained on the training set and tested on the test set to check its accuracy. A hyperparameter sweep will be conducted, whereby the optimal values will undergo 4-fold cross validation to check that the model is performing decently.

Input features are first normalized by subtracting from the mean and dividing them by the standard deviation. Hyperparameter tuning is then performed on the following options:

1. **N:** Number of neurons in the hidden layer
2. **C:** Regularization parameter strength
3. **S:** Linear scale of random variables before feeding into non-linear activation function

Other default options (unless changed) for RVFL include:

1. **Scalemode:** Using option 3 for scaling the range of the randomization of weights for uniform distribution
2. **Bias:** Set to 1
3. **ActivationFunction:** Radbas, unless otherwise specified
4. **Mode:** Ridge regression (regularized least square), unless otherwise specified

Firstly, I will be testing the effect of direct links from the input layer to the output layer (i.e., with and without direct links). Running both RVFL models on the datasets chosen gives the following accuracy scores:

Dataset name	Bias, without direct link	Bias, with direct link
abalone	0.66020115 ± 0.000174035448	0.66211686 ± 0.000117151374
bank	0.89911504 ± 0.0000129219203	0.89955752 ± 0.0000103766936
car	0.9525463 ± 0.00000133959191	0.9525463 ± 0.0000120563272
dermatology	0.96703297 ± 0.0000603791813	0.98076923 ± 0.0000830213742
echocardiogram	0.81818182 ± 0.00275482	0.85606061 ± 0.00017218
glass	0.641509434 ± 0.0048059808	0.6698113208 ± 0.0015129939
parkinsons	0.908163 ± 0.000521	0.887755 ± 0.001145
seeds	0.951923 ± 0.000832	0.947115 ± 0.000624
titanic	0.785909 ± 0.0000122	0.789545 ± 0.00000227
wine	0.98295455 ± 0.000355113636	0.98295455 ± 0.000355113636
Average	**0.8567536264 ± 0.0009529790578**	**0.8628231391 ± 0.0004034163305**

Each column shows the mean accuracy score as well as the variance as computed from the accuracy outputs of the 4-fold cross validation under the corresponding RVFL options, using the optimal N, C and S values from the hyperparameter sweep.

In general, there is a small increase in the mean accuracy when there are direct links from the input to the output layer compared to when there are no direct links. The variance also generally drops, resulting in a more consistent output.

Therefore, direct links may generally be helpful to improve results from an RVFL model. However, depending on the dataset, it may be prudent to experiment with either having them or not on the specific dataset to determine whether to include such an option. This is because other studies show that direct links and bias do not play an important role in improving RVFL accuracy for typical nonlinear regression systems[1].

[1] https://arxiv.org/abs/2003.13090

Question 2

Do performance comparisons of 2 activation functions: one from “relu, sigmoid, radbas, sine” and one from “hardlim, tribas”.

The default model implemented for the RVFL for this question include having both bias and direct links. Ridge regression is also used across both comparison models. I then examined the performance comparisons of two different activation functions chosen: **radbas** vs **hardlim.**

The corresponding function implementations are as follows:

Activation function	Formula
radbas	$y = \exp(-x^2)$
hardlim	$y = (\mathrm{sign}(x) + 1)/2)$

The results are as shown below:

Dataset name	Radbas activation function, bias, with direct link	Hardlim activation function, bias, with direct link
abalone	0.66211686 ± 0.000117151374	0.64727011 ± 0.0000451287966
bank	0.89955752 ± 0.0000103766936	0.89358407 ± 0.0000154182003
car	0.9525463 ± 0.0000120563272	0.86226852 ± 0.0000388481653
dermatology	0.98076923 ± 0.0000830213742	0.98351648 ± 0.0000301895906
echocardiogram	0.85606061 ± 0.00017218	0.87121212 ± 0.00017218
glass	0.6698113208 ± 0.0015129939	0.6981132075 ± 0.0030259879
parkinsons	0.887755 ± 0.001145	0.897959 ± 0.001041
seeds	0.947115 ± 0.000624	0.947115 ± 0.000624
titanic	0.789545 ± 0.00000227	0.788636 ± 0.0000105
wine	0.98295455 ± 0.000355113636	0.99431818 ± 0.0000968491736
Average	**0.8628231391 ± 0.0004034163305**	**0.8583992688 ± 0.0005100101826**

We can see that generally, the radbas function results in a higher mean accuracy then the hardlim function. Variance is also generally lower. Therefore, it may be that radbas generally performs better than hardlim in terms of using it as an activation function for RVFL.

Question 3

Compare the performance of Moore-Penrose pseudoinverse and ridge regression (or regularized least square solutions) for the computation of the output weights.

The default model implemented for the RVFL for this question include having both bias and direct links. It also uses radbas as the activation function since this combination of options has performed well from the previous questions. I then examined the performance comparisons between using Moore-Penrose pseudoinverse vs ridge regression for computation of output weights. The results are as follows:

Dataset name	**Ridge regression, radbas activation function, bias, with direct link**	**Moore-Penrose pseudoinverse, radbas activation function, bias, with direct link**
abalone	0.66211686 ± 0.000117151374	0.6566092 ± 0.0000644533257
bank	0.89955752 ± 0.0000103766936	0.89513274 ± 0.00000959354687
car	0.9525463 ± 0.0000120563272	0.9525463 ± 0.0000120563272
dermatology	0.98076923 ± 0.0000830213742	0.97527473 ± 0.0000830213742
echocardiogram	0.85606061 ± 0.00017218	0.87121212 ± 0.00017218
glass	0.6698113208 ± 0.0015129939	0.4433962264 ± 0.0036489854
parkinsons	0.887755 ± 0.001145	0.857143 ± 0.000208
seeds	0.947115 ± 0.000624	0.961538 ± 0.000185
titanic	0.789545 ± 0.00000227	0.789545 ± 0.00000227
wine	0.98295455 ± 0.000355113636	0.95454545 ± 0.000774793388
Average	**0.8628231391 ± 0.0004034163305**	**0.8356942766 ± 0.0005160353362**

On average, the Moore-Penrose pseudoinverse solution performed worse than the one using ridge regression, dropping in mean accuracy by almost 3% while having a higher variance. For generalized models, it may be preferred to simply use ridge regression for computation of the output weights.

Conclusion

While this report provides some analysis on different options that a generalized RVFL model can choose as sensible defaults, it is in no way conclusive due to the small sample size. As with all machine learning applications, it is best to re-test the model's accuracy with data pursuant to one's own requirements and choose the optimal options from there.

For others looking for a generalized guideline for RVFL models, it appears (at least from this experiment) that a sensible option would be to have bias, direct links, use the radbas activation function and use ridge regression.