Neural and Evolutionary Computation (NEC)

A1: Prediction with Back-Propagation and Linear Regression

Report

Sergi Salido Cubero

sergi.salido@estudiants.urv.cat

Contents

Description of the implementation (languages, tools used, etc.)

To code **BP** I have used **Julia**.

To code **MLR** I have used **Python**.

I have used multiple libraries for both languages. To create the MLR model and KFold I have used **sklearn**.

I have used VS Code as IDE.

Execution instructions

Open the folder in VS Code.

BP execution

Run **BackPropagation.jl** file. The file automatically reads the selected parameters_file.txt

MLR execution

Run **MultipleLinearRegression_turbine.py** file.

Run **MultipleLinearRegression_synthetic.py** file.

Run **MultipleLinearRegression_realestate.py** file.

Implementation decisions

I followed all the recommendations from the Dr.

Scaling into the range [0.1,0.9] instead of [0,1] means that a smaller number of epochs are required to obtain good predictions.

Description and link to the selected dataset

I have selected the **Real estate valuation** dataset from UCI Machine Learning Repository.

Description: The real estate valuation is a regression problem. The market historical data set of real estate valuation are collected from Sindian Dist., New Taipei City, Taiwan.

Data Set Characteristics:	Multivariate	**Number of Instances:**	414	**Area:**	Business
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	7	**Date Donated**	2018-08-18
Associated Tasks:	Regression	**Missing Values?**	N/A	**Number of Web Hits:**	137731

Data Set Information: The market historical data set of real estate valuation are collected from Sindian Dist., New Taipei City, Taiwan. The real estate valuation is a regression problem. The data set was randomly split into the training data set (2/3 samples) and the testing data set (1/3 samples).

Attribute Information:

The inputs are as follows

- X1=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)
- X2=the house age (unit: year)
- X3=the distance to the nearest MRT station (unit: meter)
- X4=the number of convenience stores in the living circle on foot (integer)
- X5=the geographic coordinate, latitude. (unit: degree)
- X6=the geographic coordinate, longitude. (unit: degree)

The output is as follow

- Y= house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)

Link: https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set

Comments on cross-validation (method used, parameters space searched, etc.) and results

Method used: K-fold

Number of folds used: 5

BP cross-validation results, turbine dataset

The errors for each fold are:

- 1.2113754716058258
- 0.8922224689073492
- 0.9009978284304881
- 1.3892146672505115
- 1.191516420690105

The average prediction error of cross-validation is E(%)= 1.117065371376856

MLR cross-validation results, turbine dataset

The errors for each fold are:

- 4.856129583260225
- 3.9566191208368293
- 4.428582752144864
- 4.9701807719307585
- 4.416999012160889

The average prediction error of cross-validation is E(%)= 4.954852188087313

BP cross-validation results, synthetic dataset

The errors for each fold are:

- 5.553410199009076
- 5.380249240150564
- 5.0667360053933495
- 6.232409526840523
- 5.648959063657876

The average prediction error of cross-validation is E(%)= 5.576352807010278

MLR cross-validation results, synthetic dataset

The errors for each fold are:

- 7.252785747740151
- 9.26335695589951
- 7.917336015721033
- 7.271650847718419
- 9.285765010451811

The average prediction error of cross-validation is E(%)= 8.198178915506185

BP cross-validation results, Real estate valuation dataset

The errors for each fold are:

- 12.389787056499548
- 12.702896141625839
- 11.753676332059689
- 12.826441965840552
- 10.685818367252404

The average prediction error of cross-validation is E(%)= 12.071723972655606

MLR cross-validation results, Real estate valuation dataset

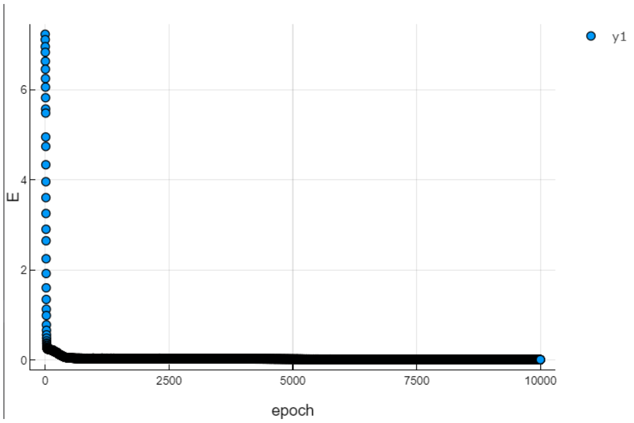
The errors for each fold are:

- 14.99235719014876
- 16.641514804540503
- 15.327778273818698
- 16.036396390704077
- 20.39330927940822

The average prediction error of cross-validation is E(%)= 16.6782711877

Evaluation of the predictions: cross-validation error, and test error

BP evaluation of predictions, turbine dataset



Cross-validation error

The average prediction error of cross-validation is 1.117065371376856

Test error

The prediction percentage error is E(%)= 1.007188206363206

MLR evaluation of predictions, turbine dataset

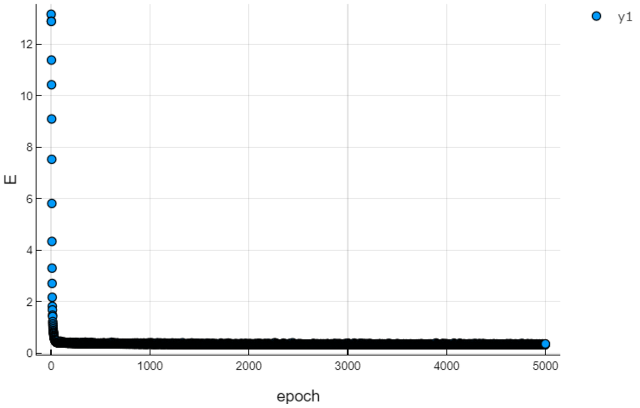
Cross-validation error

The average prediction error of cross-validation is E(%)= 4.525702248066713

Test error

The prediction percentage error is E(%)= 4.954852188087313

BP evaluation of predictions, synthetic dataset



Cross-validation error

The average prediction error of cross-validation is E(%)= 5.576352807010278

Test error

The prediction percentage error is E(%)= 5.119689677713887

MLR evaluation of predictions, synthetic dataset

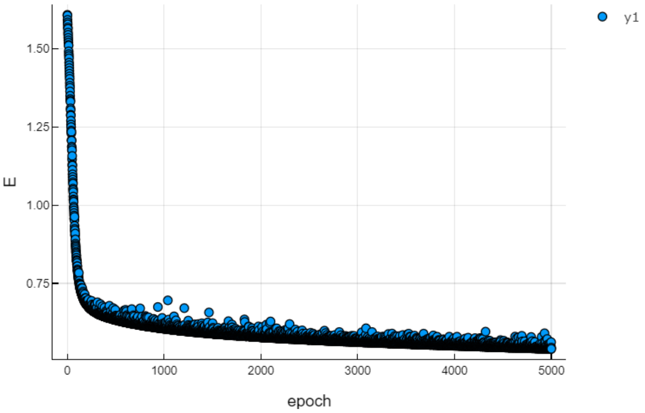
Cross-validation error

The average prediction error of cross-validation is E(%)= 8.198178915506185

Test error

The prediction percentage error is E(%)= 6.890516763670864

BP evaluation of predictions, Real estate valuation dataset



Cross-validation error

The average prediction error of cross-validation is E(%)= 12.071723972655606

Test error

The prediction percentage error is E(%)= 12.242946509095079

MLR evaluation of predictions, Real estate valuation dataset

Cross-validation error

The average prediction error of cross-validation is E(%)= 16.6782711877

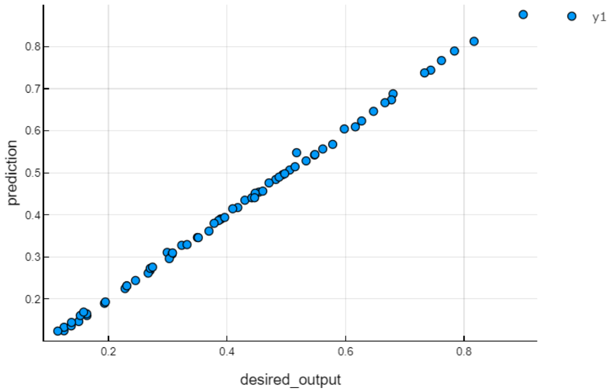
Test error

The prediction percentage error is E(%)= 15.262074482672215

Scatter plots of the prediction versus real value for both BP and MLR on the Test subsets

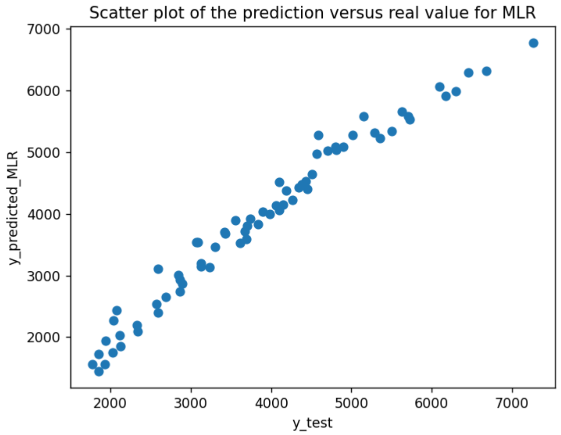
BP predictions versus real values, turbine dataset

Scatter plot of the prediction versus real value for BP



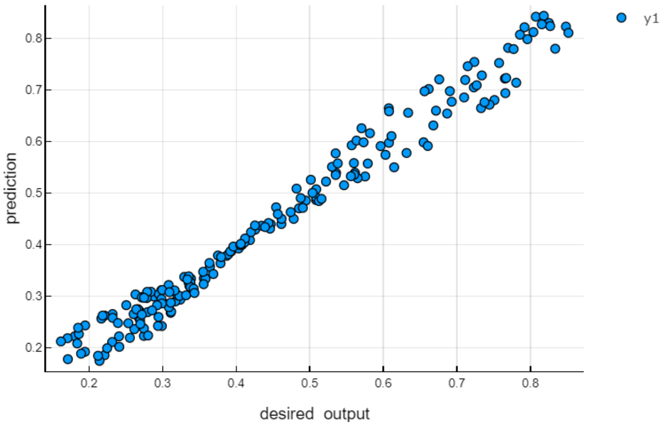
MLR prediction versus real value, turbine dataset

Scatter plot of the prediction versus real value for MLR



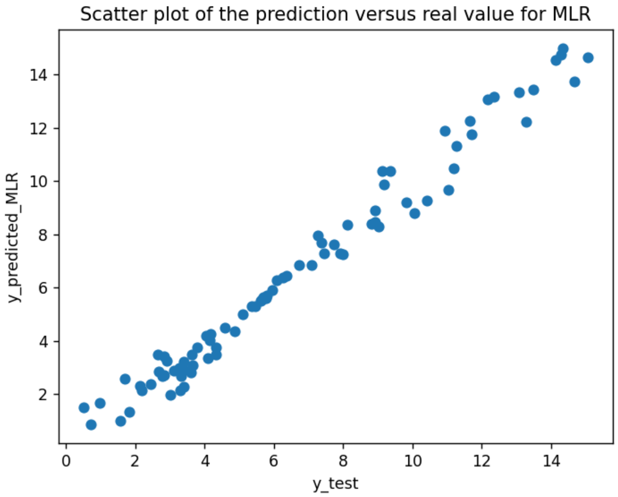
BP predictions versus real values, synthetic dataset

Scatter plot of the prediction versus real value for BP



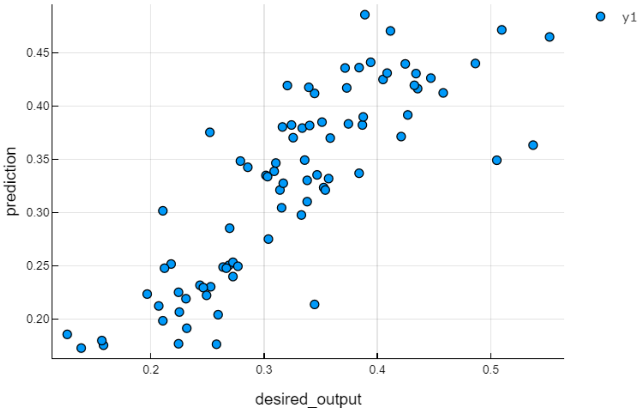
MLR prediction versus real value, turbine dataset

Scatter plot of the prediction versus real value for MLR



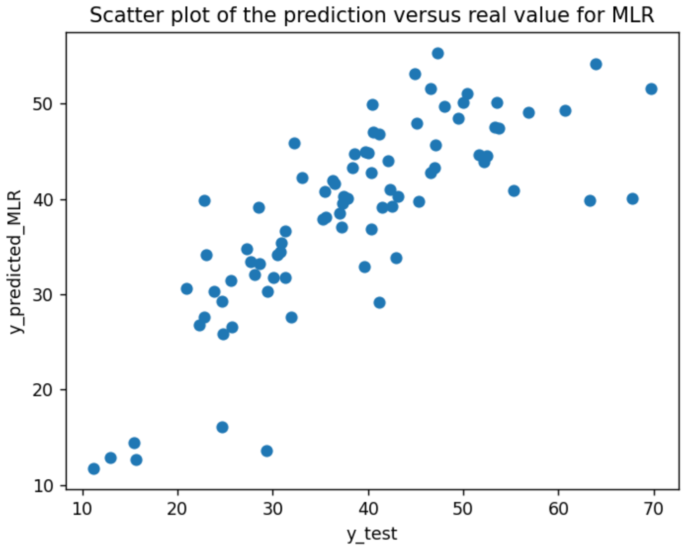
BP predictions versus real values, Real estate valuation dataset

Scatter plot of the prediction versus real value for BP



MLR prediction versus real value, Real estate valuation dataset

Scatter plot of the prediction versus real value for MLR



Discussion and interpretation of the results

In the scatter plots we can see the actual values in data against the values predicted by the models.

We can see the results are good because the closer the points are to the diagonal line, the more accurate the model is. The BP model is more accurate than the MLR model for all the datasets.

Results by dataset

Test error	**turbine**	**synthetic**	**Real estate valuation**
BP	1.01	5.12	12.24
MLR	4.95	6.89	15.26