Neural Networks

Predictive modeling

ITESO, Universidad
Jesuita de Guadalajara

Gregorio Alvarez

Introduction

This report aims to evaluate the predictive power, benefits, and drawbacks of using a neural network model in a regression problem that requires a complex solution. The primary objective is to analyze the input variables statistically, test multiple linear models as benchmarks, compare their performance with a non-optimized neural network model, and ultimately utilize a grid search to identify the optimal hyperparameters for the optimizer function. The technical procedures involved in finding a regression model that can effectively fit the "Appliances Energy Prediction" dataset will be outlined.

Dataset description

The data used for this analysis is the Appliances Energy Prediction dataset from the UC Irvine Machine Learning Repository. The dataset contains 19,735 observations and 29 variables, including the target variable. The target variable is a continuous variable representing the energy consumption in Wh of appliances in a low energy building. The remaining variables are a mix of continuous and categorical variables, and are described in the table below.

Variable	Description	Type
date	date in format "yyyy-mm-dd hh:mm:ss"	categorical
Appliances	energy consumption in Wh of appliances	continuous
lights	energy consumption in Wh of light fixtures	continuous
T1	temperature in kitchen area in Celsius	continuous
RH_1	humidity in kitchen area, in percentage	continuous
T2	temperature in living room area in Celsius	continuous
RH_2	humidity in living room area, in percentage	continuous
T3	temperature in laundry room area in Celsius	continuous
RH_3	humidity in laundry room area, in percentage	continuous
T4	temperature in office room in Celsius	continuous
RH_4	humidity in office room, in percentage	continuous
T5	temperature in bathroom in Celsius	continuous
RH_5	humidity in bathroom, in percentage	continuous
T6	temperature outside the building (north side) in Celsius	continuous
RH_6	humidity outside the building (north side), in percentage	continuous
T7	temperature in ironing room in Celsius	continuous
RH_7	humidity in ironing room, in percentage	continuous
T8	temperature in teenager room 2 in Celsius	continuous
RH_8	humidity in teenager room 2, in percentage	continuous
T9	temperature in parents room in Celsius	continuous
RH_9	humidity in parents room, in percentage	continuous
T_out	temperature outside (from Chievres weather station) in Celsius	continuous

Variable	Description	Type
Press_mm_hg	pressure (from Chievres weather station), in mm Hg	continuous
RH_out	humidity outside (from Chievres weather station), in percentage	continuous
Windspeed	wind speed (from Chievres weather station), in m/s	continuous
Visibility	visibility (from Chievres weather station), in km	continuous
Tdewpoint	dew point temperature (from Chievres weather station) in Celsius	continuous
rv1	random variable 1, unrelated to other variables	continuous
rv2	random variable 2, unrelated to other variables	continuous

For this report, the date and random variables were removed from the dataset. The target variable was constructed from the addition of the Appliances and lights variables. The remaining variables were used as input variables for the models.

Methods

Analysis

A pairplot was generated to examine the distribution of variables and their relationships with each other. The pairplot revealed that the majority of the variables had a unimodal distribution, several independent variables exhibited high correlation, while the dependent variable was highly skewed and showed little to no correlation with the input variable.

Pairplot

Figure 1: Pairplot of all variables

To gain a more accurate understanding of the degree of correlation between the variables, an absolute correlation plot was also obtained. This plot further confirmed the information obtained from the pairplot.

Correlation plot

Figure 2: Correlation plot Gegori1/DL_Specialization ### Benchmark models

Seven linear models were trained to check their predictive power.

- Linear regression, with and without standardization:

A linear regression model was fitted with the predictors and no transformation. It was found that the model was slower to train, than the one with standardization transformation, but the predictive power remained unchanged. Therefore, this transformation was applied to the rest of the models.

Linear model. Real vs Predicted

Figure 3: Linear model. It can be observed that the large skewness present in the output variable leads to large errors for large values of the variable.

- Linear regression, with variable selection:

A variable selection by highest correlation between pairs, with a threshold of 0.7, was applied to the data. Which decreased the number of features by a factor of 3. The selected variables were: `RH_2, RH_5, T8, RH_9, Press_mm_hg, RH_out, Windspeed, Visibility, Tdewpoint`. As can be seen in the correlation_plot, the majority of these variables, had low linear correlation with respect to the dependent variable, indicating a possible drop of predictive power from the remaining variables.

- Partial Least Squares (PLS) Regression:

The data was fitted using PLS regression. An iterative process was employed to determine the optimal number of components. It was determined that 12 components provided the best results, indicating the "sweet spot" for this analysis.

PLS score vs number of components

Figure 4. Score vs number of components for the PLS regression

- Transformation to target variable:

Due to high skewness of the target variable, a logarithmic and a square transformation to this variable was applied following the next procedure. The output variables was transformed after partition, the linear model was applied and the resulting prediction was transformed inversely. The obtained variable was measured against the real values.

- Lasso Regression:

A lasso regression, with default α parameter, was used to fit the data.

- Results

Model	Test score	Train score	Test RMSE	Train RMSE
Regression original data	0.1544	0.1537	93.8604	96.5442
Regression normalized data	0.1544	0.1537	93.8604	96.5442
Regression with selected features	0.0325	0.0298	100.3980	103.3671
PLS regression	0.1546	0.1521	93.8516	96.6372
Regression log transformed target	0.0790	0.0968	97.9600	99.7386
Regression root square transformed target	0.1427	0.1383	94.5070	97.4197
Lasso regression	0.1533	0.1523	93.9259	96.6244

As can be seen from the previous table, the regression with the original data, the normalized data and the PLS regression have the highest accuracy, being the PLS regression the one with the highest interpretability thanks to the reduce number of components and the possibily of interpreting the coefficients of the components.

Neural Network models:

A neural network network with 400 hidden neurons, hyperbolic tangent activation function, Adam optimizer and mean squared error as loss function was trained. The model was trained for 400 epochs, with a learning rate of 0.01. To avoid saturation of the activation function, the data was normalized using the standardization method.

Table with number of parameters and architecture