EE 102 PROJECT Fall 2020

Regression Analysis to Predict Electrical Power Output

For this project, you should work with at most two partners(i.e., **at most three** students in one group, groups can be made of 2 to 3 students). The purpose of the project is for you to gain experience in applying statistical methods to a real data set and getting introduced to machine learning. The project is composed of two parts:

A. Problem Definition, Data Collection, Data Division, and Data Understanding

B. Modeling and Prediction

For statistical analysis, modeling and prediction, you are free to choose any of the following programming languages: Matlab, R, Python, C, Excel, Octave etc. If you choose Matlab, know that you can access it for free through the COE Virtual Computer Lab. Let me know if this is what you would like to do so I can provide additional information.

Part A.

Step1. Problem Definition, Data Collection, and Data Division

In this project, we are going to analyze the effect of various input variables on an output variable.

You will be analyzing and predicting the electrical power output of a power plant that operates with 2 gas turbines and 1 steam turbine. (called a combined cycle power plant).

There are 4 input variables: the "Ambient Temperature (AT)", the "Atmospheric Pressure (AP)", the "Relative Humidity (RH)" and the 'Vacuum (V)" which represents the Exhaust Steam Pressure

The output variable is the " hourly full load electrical power output (PE)" of the combined cycle power plant.

The real data set is available online at

https://archive.ics.uci.edu/ml/index.php

https://archive.ics.uci.edu/ml/datasets/combined+cycle+power+plant in the Data Folder.

Please pick sheet 3.

Note that each row in the Excel file represents one instance. All together there are 9568 rows of data points collected when the power plant was set to operate over 2 years.

Your first task is to Divide the data set into two groups

a. **Training Data**: save 80% of the original dataset into an Excel sheet

b. **Test Data**: take the remaining 20% of the dataset and save into an Excel sheet

Note that the test data is **NOT** included in the training data. They do not overlap!

For example, you could take the first 2000 instances and save into the Test Data set and the remaining data (around 8000 instances) would be saved into the Training Data set.

Create a form (in Excel) including the 2 sets of data.

Step 2. Understanding the Data

Before deriving prediction models in the next Part, we have to understand the statistics of the data sets.

1. For each data set (training and test), find the summary statistics of the dataset (mean, median, mode, minimum, maximum, variance, standard deviation).

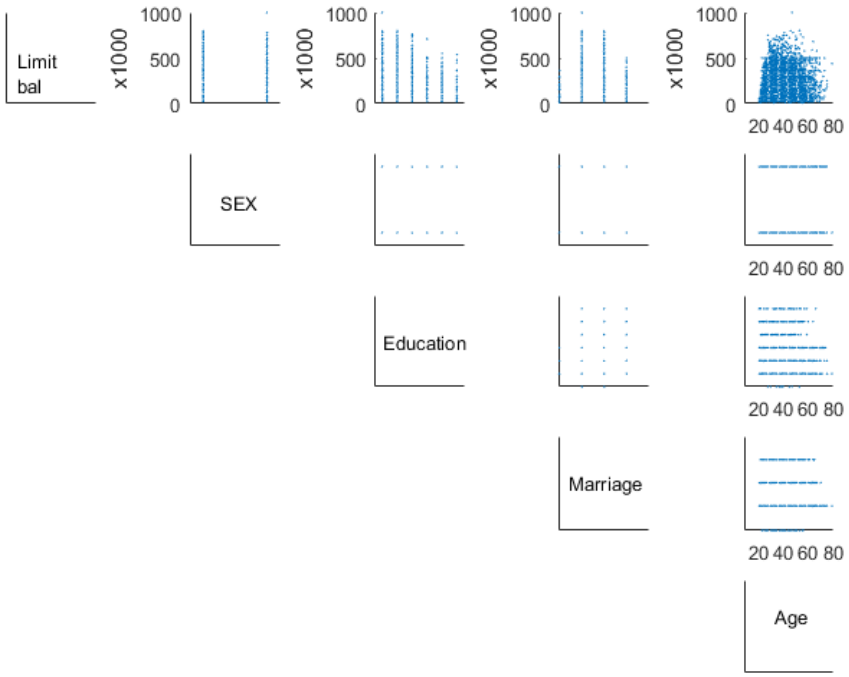
   Remember there are 4 input variables and there is one output variable in each group of data. In your report, create a table so that each row is one variable, and each column is one of the summary statistics. Do that for the training data set and the test data set, so you should have 2 tables.

   Make observations and write your comments in the report.

2. In principle, at this point you would clean your data, i.e., remove outliers. In machine learning, it is usually tricky to remove outliers because what is really an outlier? We will discuss about that in class. Luckily for this dataset you do not need to do any cleaning, it has already been done for you. So leave as is.

3. Just for the training dataset, obtain scatter plots USE a scatter plot matrix to display multiple scatter plots in the same figure. An example for a credit data set is given below (in this example, limit balance is the output variable, and there are 4 input variables : sex, education, marriage, age).

   Make observations and write your comments using the scatter plots.



4. Find the correlation coefficients (see class notes) between your input variables and your output variable.

   What are the input variables that show the strongest correlation to the output?.

   Pick 3 input variables out of the set of 4 that show the most promise in predicting the output.

   For the rest of the project you will be working with those 3 chosen input variables

Step 3. Feature Scaling

The idea here is to make sure the input variables and output variable are on a similar scale.

The simplest method is rescaling the range of each input variable to the interval [0, 1]. The general formula is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

For example, suppose you have m instances of temperature x, min(x) is the smallest value of the m instances . Note you have already computed the max and min in Part A. You can find the transformed value x' of x (the original raw data) from the equation above.

From now on, you will be working with the normalized training data and test data. And this is what is

meant when those data sets are invoked.

Part B.

Deriving linear regression models (using training data) and measuring their performances (on test data)

Step 4. Derive linear regression models from the Training Data

We will use the linear least squares curve fitting method to derive models for prediction.

We want to express the output variable as a linear combination of input variables, such as

$$\text{Output Variable} = \sum_{k=1}^{n} constant_k * input\ variable_k + constant_{n+1} \quad \text{(Equation 1)}$$

In the training phase, the values taken on by the output variable and the values taken on by the input variables are known from the training dataset. The objective is to find the (n+1) constants in Equation (1)

Read the Help file for the Matlab command "mldivide", same as symbolic matrix left division "https://www.mathworks.com/help/matlab/data_analysis/linear-regression.html"

Remember to consider only the training data set in this step

There will be 7 models to derive because of **7 feature subsets considerations** i.e.:

- Models with only **one** input variable (n=1; there are 3 models since you have selected 3 input variables, one model for each input variable). Find the corresponding two constant values in each model.
- Models with **two** input variables (n=2; there are 3 models, choose 2 input variables from 3 input variables, you will have 3 combinations). Find the corresponding three constant values in each model.
- Models with **three** input variables (n=3; there is only one model). Find the corresponding four constant values in the model.

Step 5 . Test the models on the test data – Prediction Accuracy

For each instance in the test data, you will use the input variables of that instance to calculate the **predicted output** value using any of the 7 models you obtained in the previous step. The **actual output** value of the power is known and available to you in the test data.

Therefore you can compare the predicted output (that is modelled-based) to the actual output by performing an error analysis. To measure the goodness of the fit, we will consider two different error metrics;

One of your error metrics is the **mean squared error (MSE)**

$$MSE = \frac{\sum_{j=1}^{m}(actual\ value - predicted\ value)^2}{m}$$

where m is the number of data samples in your test data.

Another metric is **the R^2 value** (https://www.mathworks.com/help/matlab/data_analysis/linear-regression.html). R^2 is a statistical measure of how close the data are to a fitted regression line.

Discuss your observations based on the error metrics. Which model performs best (recall that you have 7 models to investigate from Step 4)? Why?

WHAT TO SUBMIT?

FINAL REPORT DUE: **DEC 7th** (Submit ONLINE a pdf file on Canvas)

FINAL CODE/DATA-SET DUE: **DEC 7th** (Submit all your code (if applicable such as Matlab code or Python code) and dataset (as an Excel file) on Canvas)

Each group submits one report.

The report should contain:

1. Title page: Project title + Group Member Names + etc.,
2. Abstract: Brief Project Description
3. List of Figures and Tables
4. Glossary: Description of Important Terms and Abbreviations used in Your Project
5. Introduction and Background: Importance of the project
6. Discussion of Data
7. Analysis: includes choice of the linear regression.
8. Results. Are they satisfactory? Would a nonlinear regression yielded better results?
9. Conclusions (difficulties faced, any other observations/comments about the project)
10. Discussion of contribution of each team member
11. References (with proper citation in the report)

Make sure you answer all the questions asked from you in the project . Comment on your findings and provide sufficient support, including appropriate plots. Make sure all the figures are incorporated in the report properly.

Reading Assignment:

https://en.wikipedia.org/wiki/Predictive_analytics