MEM T380 – Applied Machine Learning in Mechanical Engineering

Case Studies Assignment 3
Supervised Regression

Dimitrios Fafalis, PhD
Dimitrios.Fafalis@drexel.edu

due Thursday, June 1, 2023

Students' names & ID:

1. ________________________ ____________
2. ________________________ ____________

Submit your files (.ipynb) and a report (.pdf) on Blackboard by due date.

Regression Prediction of Bead Geometry for GMAW-based Rapid Manufacturing!

CASE STUDY 1. points 100 – **Bead geometry regression prediction**
This mini case study is related to predicting the geometry of beads during the fabrication of metallic parts with *laser welding* and **gas metal arc welding (GMAW)**. The predictions will be made using the regression supervised learning techniques of least-squares regression and regression trees you learned during weeks 5 and 6. This case study is based on the

Research article [Xiong et al., 2014]: ”Jun Xiong, Guangjun Zhang, Jianwen Hu, and Lin Wu, 2014, **Bead geometry prediction for robotic GMAW-based rapid manufacturing through a neural network and a second-order regression analysis**, *Journal of Intelligent Manufacturing* 25, pages 157-163 (2014), DOI 10.1007/s10845-012-0682-1”.

This article is available in the assignment 4 post. Although the authors of the paper used an artificial neural network (ANN) machine learning approach and a second-order regression analysis, for the purposes of this mini case study we will explore this problem with regression analyses and regression trees. In a later lecture, we will revisit this case study and practice with ANN. Refer to this article to familiarize yourselves with the problem.

The raw data for this case study are available on table 2 of the paper [Xiong et al., 2014]. You have to transfer them into an appropriate format and file before you load them in `python`.

The data consist of four (4) predictors (aka regressors, independent variables):

- wire feed rate F [$\mathrm{m\,mm^{-1}}$]
- welding speed S [$\mathrm{cm\,min^{-1}}$]
- arc voltage V [V]
- nozzle-to-plate distance D [mm]

There ara two (2) response variables (aka dependent variables, responses):

- width of bead W [mm]
- height of bead H [mm]

The goal of this case study is to develop least-squares and regression trees ML models that predict the geometry (width and height) of the beads during the GMAW process.

1 Data Exploration Tasks:

20 points. The very first steps in developing a `Machine Learning` model are to load, explore, and preprocess the data. The goal of this task is to explore the data and try to listen to what they want to tell us!

1. reading from **table 2** of the research paper [Xiong et al., 2014], load the data into an excel worksheet or a `.csv` file; give appropriate names to the features (columns); save the excel worksheet with the name `bead_geometry_gmaw_train.xlsx`.

2. using the `pandas` command `read_excel`, load the dataset and store it into a `pandas` dataframe. Refer to pandas.read_excel documentation for examples and additional options.

3. print a summary of the information included in the dataframes using the functions `df.info`, `df.dtypes` and `df.describe`.

4. print the properties of the dataframe, e.g. `df.info` and/or `df.describe` .

5. use the `seaborn` command `pairplot` to visualize `bivariate` relationships between the `numerical` fields grouping them by the `categorical` fields (if applicable).

6. Create a heatmap of the correlation matrix of the features.

7. identify from the figures in item 5 and the correlation matrix from item 6 which `numerical` features (fields) are the **strongest** to be used in predicting the width and height of the beads created during the GMAW process.

2 Regression with Ordinary Least-Squares (OLS):

2.1 Simple Linear Regression:

20 points.
For this part of the case study you may refer to the lecture codes of week 7a.

1. Based on the data exploration you performed in section 1, choose **one** strong feature from the set of `F, S, V, D` as predictor (regressor) and **one** response variable (width `W` or height `H`). Write down the mathematical expression of the linear univariate model.

2. Create a Simple Linear Regression `SLR` model using the `LinearRegression` class of the `sklearn.linear_model` package. Documentation: `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html`

3. Comment on the results: the model coefficients, the intercept, the `R-squared` and determine whether the regression coefficients are significant. In other words, does this model fit the data well? Explain.

4. Plot the regression line against the raw data (only the feature and response you selected in item 1).

5. Select a different predictor variable from the set of `F, S, V, D` (other than the one you used in item 1) and the other response variable (width `W` or height `H`). Repeat the tasks 1 to 4.

6. Do you observe any common behaviors between the two models? Do you think that the two simple models you created are adequate to fit the data and that they can be used to make correct predictions for unseen data?

2.2 Multivariable Polynomial Regression:

20 points.

For this part of the case study you may refer to the lecture codes of weeks 7 and 8.

In the previous section you modeled a subset of the available data with simple linear regression models. However, the dataset consists of more features that the researchers collected under the assumption that these features may play a significant role in the final geometry of the beads. Let's explore here whether their assumption (hypothesis) is correct or not.

1. Create two multivariable second-degree regression models using all the predictors (independent variables); create one model for the response width of bead `W` and a second model for the response height of bead `H`. Your models should have the following mathematical description, as given by equations (2) and (3) of the paper,

$$
\begin{aligned}
Y = {} & \beta_0 + \beta_1 F + \beta_2 S + \beta_3 V + \beta_4 D \\
& + \beta_{11} F^2 + \beta_{22} S^2 + \beta_{33} V^2 + \beta_{44} D^2 + \\
& + \beta_{12} FS + \beta_{13} FV + \beta_{14} FD + \\
& + \beta_{23} SV + \beta_{24} SD + \beta_{34} VD
\end{aligned} \tag{1}
$$

   where Y can be either of the response variables W or H. As you can see, the model contains a full second-degree polynomial, consisting of all the linear terms, all the pure quadratic terms (i.e. X^2) and the linear interaction terms (i.e. $X_i \cdot X_j$).
   In eq. (1) β_0 is the constant intersect, $\beta_1, \beta_2, \beta_3, \beta_4$ are the linear coefficients, $\beta_{11}, \beta_{22}, \beta_{33}, \beta_{44}$ are the quadratic coefficients, and $\beta_{12}, \beta_{13}, \beta_{14}, \beta_{23}, \beta_{24}, \beta_{34}$ are the interaction coefficients.

2. Create two quadratic (polynomial) regression models based on the mathematical equation (1) in item 1. Create one model to predict the response height of bead `H` and a second one for the width of bead `W`. You will first have to preprocess and transform the data with the `sklearn.preprocessing.PolynomialFeatures` class. Then, use the `LinearRegression` class of the `sklearn.linear_model` package, as usual.

3. For each quadratic model comment on the results you obtained: the model coefficients, the `R-squared` and determine whether the regression coefficients are significant. In other words, does this model fit the data well? How does the performance of these models compare to the results from part 1? Explain.

2.3 Predicting with Multivariable Polynomial Regression:

20 points.
This part requires that you have completed the tasks of the previous section 2.2.

1. Read from paper [Xiong et al., 2014] the sections *Second-order regression modeling* and *Selecting the most accurate model*. The paper provides a short dataset to test the predictability power of the regression models developed. These testing data are available on table 3 of the paper.

2. reading from **table 3** of the research paper [Xiong et al., 2014], load the data into an excel worksheet; give appropriate names to the features (columns); save the excel worksheet with the name `bead_geometry_gmaw_test.xlsx`.

3. using the `pandas` command `read_excel`, load the dataset from the file `bead_geometry_gmaw_test.xlsx` and store them into a `pandas` dataframe.

4. based on the multivariable quadratic regression models you developed in items 1 and 2 of section 2.2, use the command `predict` to predict both response variables bead width `W` and bead height `H` on the testing dataset.

5. from the predicted results calculate the **root mean square error (RMSE)**:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - y_i\right)^2} \tag{2}$$

   where n is the total number of testing data, y_i is the original values of the response variables, $f(x_i)$ are the predicted values. Alternatively, you may import and use the corresponding function from `sklearn`.

6. from the predicted results calculate the **mean absolute percentage error (MAPE)**:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{f(x_i) - y_i}{y_i}\right| \tag{3}$$

7. looking at the $R-$squared, the `MAPE` and `RMSE`, conclude whether the predictive accuracy of the models is satisfactory. Alternatively, you may import and use the corresponding function from `sklearn`.

3 Regression with Decision Trees (RDT):

20 points.
For this task you are required to use the training dataset `bead_geometry_gmaw_train.xlsx` to train the regression trees and the testing dataset `bead_geometry_gmaw_test.xlsx` to test the predictive accuracy of the regression trees.

1. Create a `RDT` model using `DecisionTreeRegressor` class of the `sklearn.tree` package, the training dataset and the default options (this will generate the most deep tree). Experiment with different values of `max_depth`, and `criterion`. Refer to `sklearn` documentation sklearn.tree.DecisionTreeRegressor for explanations and examples.

2. Visualize the `RDT` model you created in item 1 using the method `plot_tree`. Display both the text description and a graph.

3. For the `RDT` model you created in item 1, make predictions on the testing dataset using the `predict` method and the calculate the `mean_absolute_error`, the `mean_squared_error` and the `r2_score`, from the `sklearn.metrics` package.

4. Create a random forest regressor model (`RFR`) model using the `sklearn.ensemble` class `RandomForestRegressor`, the training dataset, and the following arguments: `n_estimators=200, bootstrap = True, max_features = 'sqrt'`. Make predictions on the testing dataset using the `predict` method and the calculate the `mean_absolute_error`, the `mean_squared_error` and the `r2_score`, from the `sklearn.metrics` package.

5. Find the minimum optimal number of trees in a random forest regressor model using a `for` loop and iterating over 1000 `n_estimators`. Create a plot ov `n_estimators` vs `r2_score` and identify the optimal number of trees.

6. Create a random forest regressor model using the optimal number of trees you found in item 6 to predict the bead width `W` and bead height `H` for the testing dataset.

7. For this particular case study to predict the bead geometry for GMAW-based rapid manufacturing processes, which supervised `Machine Learning` approach would you adapt, the least squares models or the regression decision trees? Please elaborate on your choice.

References

[Xiong et al., 2014] Xiong, J., Zhang, G., Hu, J., and Wu, L. (2014). Bead geometry prediction for robotic gmaw-based rapid manufacturing through a neural network and a second-order regression analysis. *Journal of Intelligent Manufacturing*, 25:157–163.