

Manufacturing Data Science 製造數據科學

Assignment 2

Due Date: 5pm, Nov. 12, 2021

Please solve the following questions and justify your answer. **Show all your analysis result including equation/calculation or Python code in your report.** Upload your “zip” file including (1) MSWord/PDF report and Python code; or (2) notebook (including answer and code), with file name: “MDS_Assignment2_ID_Name.zip” to NTU COOL by due. The late submission is not allowed.

1. (60%) Stepwise Regression and Regularization- Ridge, Lasso, and Elastic Net

Data Source: a flotation plant in a mining process

<https://www.kaggle.com/edumagalhaes/quality-prediction-in-a-mining-process> and are attached in the file **MDS_Assignment2_Mining.zip**.

Dataset provided by EduardoMagalhãesOliveira. Data collection methodology is shown as follows.

Hardware sensors, like temperature, pH, flow, density and all continuous process variables, where data were collected every 20s with no transformation (the dataset here shows raw data). Quality variables, like % of silica content, % of iron ore content and so on are quality measurements made by laboratory analysis. A sample of the iron ore pulp is collected in the field/shop floor, every 15 minutes. Those samples are sent to lab for analysis. So, on every two hours, lab give a feedback of quality analysis, in other words, only every two hours you have a lab/quality measurement of the product stream (iron ore concentrate), which gives you a sense of the quality of the product (iron ore pulp concentrate).

The main goal is to use this data to predict how much impurity is in the ore concentrate. As this impurity is measured every hour, if we can predict how much silica (impurity) is in the ore concentrate, we can help the engineers, giving them early information to take actions (empowering!). Hence, they will be able to take corrective actions in advance (reduce impurity, if it is the case) and also help the environment (reducing the amount of ore that goes to tailings as you reduce silica in the ore concentrate).

Content

The first column shows time and date range (from march of 2017 until september of 2017). Some columns were sampled every 20 second. Others were sampled on a hourly base.

The second and third columns are quality measures of the iron ore pulp right before it is fed into the flotation plant. Column 4 until column 8 are the most important variables that impact in the ore quality in the end of the process. From column 9 until column 22, we can see process data

(level and air flow inside the flotation columns, which also impact in ore quality. The last two columns are the final iron ore pulp quality measurement from the lab.

Target is to predict the last column, which is the % of silica in the iron ore concentrate.

Inspiration

Is it possible to predict % Silica Concentrate every minute?

How many steps (hours) ahead can we predict % Silica in Concentrate? This would help engineers to act in predictive and optimized way, mitigating the % of iron that could have gone to tailings.

Also, for the Amina Flow, Ore Pulp Flow, and Flotation Column, etc. data are the “commas” in those cells supposed to be “decimals” as shown in the csv file. For example, given the column “% Silica Concentrate”=“1,31”, it means that the concentrate is 1.31%.

According to the description mentioned above, if the factor “% Silica Concentrate” is regarded as the response variable (y) and all factors (**except the date and % Iron Concentrate**) are independent variables, how to identify the importance variable which significantly affects the “% Silica Concentrate” (y)?

- (a) **(5%)** Identify the important variable by **linear regression** with ordinary least squares (OLS) (i.e. ranked by p-value).
- (b) **(5%)** Identify the important variable by **stepwise regression**. (hint: you can select forward selection, backward elimination, or both)
- (c) **(5%)** Give a comparison between (a) and (b). The results are consistent?
- (d) **(5%)** From a methodology aspect, what’s the difference between Ridge regression and Lasso? Why does Lasso support the variables selection rather than ridge? (hint: answer with description or formulation. No computation needed.)
- (e) **(5%)** What’s the benefit to use the Elastic Net? (hint: answer with description or formulation. No computation needed.)
- (f) **(5%)** Identify the important variable by **ridge regression**.
- (g) **(5%)** Identify the important variable by **lasso**.
- (h) **(5%)** Identify the important variable by **elastic net**.
- (i) **(5%)** Give a comparison between (f), (g) and (h). The results are consistent?
- (j) **(5%)** What is “adaptive elastic net”? Why we need it? How to build it? Please simply describe or formulate it. (No computation needed.)
- (k) **(5%)** Which columns are highly-correlated? Show the table of the coefficient estimation by using linear regression. Any multicollinearity problem?
- (l) **(5%)** Is it possible to predict % Silica in Concentrate without using % Iron Concentrate column (as they are highly correlated)? Why? What’s the potential issue? How to address it?

2. (40%) Principal Component Analysis (PCA)- Faulty Steel Plates

Data Source: <https://www.kaggle.com/uciml/faulty-steel-plates>

Dataset provided by Semeion, Research Center of Sciences of Communication, Via Sersale 117, 00128, Rome, Italy. www.semeion.it

This dataset comes from research by Semeion, Research Center of Sciences of Communication. The original aim of the research was to correctly classify the type of surface defects in stainless steel plates, with six types of possible defects (plus "other"). The Input vector was made up of 27 indicators that approximately describe the geometric shape of the defect and its outline.

There are 1941 plates with 34 variables. The first 27 columns (i.e. independent variables) describe some kind of steel plate faults seen in images, i.e., X1-X27, as

{X_Minimum, X_Maximum, Y_Minimum, Y_Maximum, Pixels_Areas, X_Perimeter, Y_Perimeter

SumofLuminosity, MinimumofLuminosity, MaximumofLuminosity, LengthofConveyer, TypeOfSteel_A300, TypeOfSteel_A400, SteelPlateThickness, Edges_Index, Empty_Index, Square_Index, OutsideXIndex, EdgesXIndex, EdgesYIndex, OutsideGlobalIndex, LogOfAreas, LogXIndex, LogYIndex, Orientation_Index, Luminosity_Index, SigmoidOfAreas}

The last seven columns (i.e. dependent variables) are one hot encoded classes, i.e. if the plate fault is classified as "Stains" there will be a 1 in that column and 0's in the other columns.

{Pastry, Z_Scratch, K_Scratch, Stains, Dirtiness, Bumps, Other_Faults}

These data can be found in <http://archive.ics.uci.edu/ml/datasets/steel+plates+faults>, and are attached in the file **MDS_Assignment2_Steelplates.xlsx**.

Perform a principal component analysis (PCA) on these data.

- (a) (10%) Use the Logistic Regression to predict the defect type **"Bumps"** in the testing dataset. That is, **only consider the Bumps column** in the 7 defect types. What's the accuracy and confusion matrix?
- (b) (5%) For PCA, which variables should not be put into the PCA analysis (eg. binary variable).
- (c) (5%) What's eigenvalues and eigenvectors of covariance matrix in PCA? How can you interpret from these eigenvalues?
- (d) (5%) Plot a **scree plot** and decide the most appropriate number of principal components (PCs) to use.
- (e) (10%) Use the selected PCs and perform Logistic Regression to predict the defect type **"Bumps"** again. What's the accuracy and confusion matrix?
- (f) (5%) What conclusions can you draw from above analysis? Give a comparison between (a) and (e). The results are consistent?

Note

1. Show all your work in detail. **Innovative idea is encouraged.**
2. If your answer refers to any external source, please “must” give an academic citation. Any “plagiarism” is not allowed.