LELEC2870 Project

Industrial Process Modeling

Cyril de Bodt `cyril.debodt@uclouvain.be` Maxwell a.138
Dimitri de Smet `dimitri.desmet@uclouvain.be` Maxwell a.138
Luc Rocher `luc.rocher@uclouvain.be` Euler SC03B

November 13, 2018

Introduction

Machine learning methods can be used to solve many practical problems in a wide range of applications such as weather forecast, customer clustering, medical diagnostics, spam blocking, financial time series prediction or signal de-noising, ... It is also used to model industrial process for which the physical model is too complex to be accurately simulated. It can then be useful to predict the characteristics of the output of a manufacturing process without the need of running it many times with different parameters.

In the scope of the courses final project, you will try to predict a concrete's compressive strength based on its ingredients and its age.

Instructions

Model

You will build regression models that predict concrete compressive strenght based on its ingredient and its age. You can use any of the methods seen during the lectures. We expect you to, at least, implement linear regression, KNN[1] and one other non-linear method. Features selection and model selection shall also be part of your work. Pay attention to the fact that the model

[1]K-Nearest Neighbour : Regression model with metaparameter K that predicts the output of a sample as the mean of output of the K nearest neighbours in the features space.

selection can require a lot of computation time. You are advised to explore the metaparameters space according to the time available.

Prediction

Once your model is properly selected and validated, you are asked to produce predictions Y2 on the data X2 for which we have kept secret the corresponding targets. Your prediction quality criteria shall be the root mean squared (rms) error. In addition, you will provide in your report, an estimate of the rms error that you expect on your prediction.

Report

You will produce a report documenting your technical choices and experimental results. We do not need a course on the methods you use. We are more interested in what you did and why. Try to illustrate your results by graphics (with legends) and comment them. Be critical about what you observe and try to give a possible justification of the obtained results. Summarize your results and observations in a conclusion. A strict maximum of 7 pages (font of size 11 or larger) will be observed. Annexes might be included in the digital version only. Plastic covers and bindings are not required (a paper clip or a staple will do).

The report will be the basis of a discussion which will take place during the exam session. The main objective of your report will be to convince the reader of the follwing:

1. that you selected the best model with the best hyper-parameters for the predictions on the second dataset
2. that you applied a proper methodology to assess the accuracy level of your predictions
3. that you discussed the features impact on the output.

You will work by group of two or alone on the provided dataset. All your figures and computation need to be reproducible by us running your implementation code on the provided data.

Data

On the course website, you will find two csv files called X1.csv and X2.csv. The first one is the labeled dataset and the second will be used to make your

prediction. Paste these two files in your working directory. The data will be loaded in your workspace by running the following commands.

```
X1=csvread('X1_t1.csv',1,0); (Matlab)
X1<-read.table("X1_t1.csv",sep=",",header=TRUE) (R)
X1=pandas.read_csv("X1_t1.csv") (Python)
```

Each row in the data corresponds to an instance (a piece of concrete). Each column represents one of the features listed here :

1. Cement (component 1) [kg/m3] Input Variable
2. Blast Furnace Slag (component 2)[kg/m3] Input Variable
3. Fly Ash (component 3)[kg/m3] Input Variable
4. Water (component 4) [kg/m3] Input Variable
5. Superplasticizer (component 5) [kg/m3] Input Variable
6. Coarse Aggregate (component 6)[kg/m3] Input Variable
7. Fine Aggregate (component 7) [kg/m3] Input Variable
8. Age [Day] (1 365) Input Variable
9. Concrete compressive strength [MPa] Output Variable (only in X1)

Programming languages

The programming language you will use is up to you but we strongly recommend Matlab, Python or R. The best language to use here is the one you already master the most. As we may run your code, please specify where your toolboxes and packages come from.

For matlab users, the Netlab toolbox that is on the moodle website can be used for MLP regression.

R users can use all the packages they can get with "`install.packages`" commands.

Python users can use all the packages they can get with "`pip`" commands.

Agenda

- As soon as possible: Register your group (maximum two people) on course website.
- Thursday November the 29th and Thursday December the 13th at 4:15pm: We will be available in the lecture room to answer your questions.
- Thursday December the 20st, 1 p.m
  - submit your work as an archive (.zip) containing the following items
    - Your report (pdf)
    - A csv file called "Y2.csv" no header line: one line per prediction values.
    - A folder containing all scripts you wrote for the project.
  - In addition, you will provide three paper copies of your report (no cover) to be left in front of the assistants' office (Maxwell a.138).

Evaluation Criteria

- Respect of the instructions and deadlines
- Quality of the report and its defense (discussion)
- Proper validation of your model(s)
- Consistency between the report, your implementation and your prediction
- reproducibility of your results

Tips

Here is a list of advices for the project.

Before any analysis:

- A fast univariate analysis of the data is always useful.

- Normalise your data if necessary.
- Some outliers might be excluded from the learning set (if you decide to remove some observations, explain why you removed them).

While coding:

- Pay attention to avoid unnecessary loops: scripting languages work better when an operation is expressed in matrix form. You also should remove all operation which could be done outside of the loops.

Before you send us the report:

- Make sure that each choice is justified, each result is commented and interpreted; each graph should have a legend.
- Comment your model performances.
- You can discuss the project with other students, in fact, it is a great idea ! You could compare your results and those obtained by other groups, but remember that it is not allowed to copy what other did...
- Make sure that you provided a prediction vector in a csv file named "Y2.csv" that contains one line per prediction (515 lines) and no header.

We will be happy to answer during the Q/A sessions or on appointment. Good luck !