University of Science
Introduction to DS

Final project

Collect, pre-process, explore data, create model from data

Nguyen Bao Long
baolongnguyen.mac@gmail.com

Abstract

In this project, you are going to handle a reality problem using data. Specifically, you will perform a full data process which contains data collection, data pre-processing, data exploration, and data modeling. Select an interesting topic, crawl some data and have fun!

1 Evaluation & Submission guidelines

1.1 Evaluation

- You have to create a 4-member group. Your team will have 15 minutes for presenting the project.
- Data collection: 2p
- Data pre-processing and exploration: 3.5p
- Data modeling: 3.5d
- Presentation: 1p

1.2 Submission guidelines

- Make a folder named `ID1_ID2_ID3_ID4` (e.g. `123_456_789_910`), where `ID1` is the smallest ID in the group. The contents of the folder is as follows
  - `slide.pdf`: A presentation of the project (less than or equal to 30 pages).
  - `./src`: Folder that contains all source code of the project.
  - `./data`: Folder that contains data. If the data size is larger than 10MB, upload your data to cloud and put an URL in `./data`.
- After that, compress `ID1_ID2_ID3_ID4` in form of a `*.zip` file and submit on Moodle.
- **Please strictly follow the instructions**.

2 Requirements

In this section, you will go through all the process of a data project. Firstly, you have to select a topic that you are interested in (traffic, population, finance, Covid19,...). Next, you are going to collect the data. Then pre-process and explore data to gain insights about them. Finally, you will create models from data to solve problems such as regression, classification, clustering,...

2.1 General requirements

- Use GitHub to manage source code and collaborate with team. Your repository must have at least 20 commits when the project is done.
- Create a plan with specific steps and deadline for each of them for each member of team. Make sure that every member has full knowledge about what team is doing.
- It is recommended that you use `Jupyter Notebook` to do this assignment.
- You can use Trello, a really good tool for managing tasks (assign, mark as done, to-do).

2.2 Data collection

- You have to prepare data by your own self in order to analyze and explore them.
- The data can be collected from the Internet by parsing `HTML` code or using API.
- Note that **you are not allowed to use available datasets**. Your dataset must have **at least 5 fields and 1000 observations**.

2.3 Data pre-processing & Exploration

These processes can be done in parallel since they are involved in each other.

2.3.1 Data pre-processing

- In order to analyze the data well, you have to pre-process them.
- At each pre-process step, you have to explain the reason why you decide to do that.

2.3.2 Data exploration

- Basic level: How many rows and columns are there in your data? What is the meaning of each row/column? What is the datatype of each column? Is this suitable datatype for the column? What is the distribution of the data in each column?
- Make at least 5 questions. For each question:
  - What is the purpose of answering this question?
  - How do you find the answer for this question in the data?
  - The answer should be visualized so reader can easily capture the idea.

2.4 Data modeling

First, you have to specify the problem (regression, classification, clustering) you are going to solve. Second, you might want to prepare your data so you can train and test your models. Third, design machine learning models, train and test them with the prepared data. Final, evaluate your models using metrics such as accuracy, mean square error,...

2.4.1 Problem statement

- State your problem.
- What are the purposes/benefits of solving this problem?

2.4.2 Data preparation

- Pre-process (handle missing data, normalize data,...) your data so you can input to the models.
- Split your data into training, validating and testing set.

2.4.3 Create, train & Test models

- Create more than one model so you can compare them to each other to decide whether one is good or not.
- Choose metrics to evaluate models.
- Process: Train your models on training set, then fine-tune the model on validating set in order to get the best hyper-parameters. After that, re-train your models on (training + validating set). Finally, test your models on testing set and evaluate them.
- Visualize the running process and the results.