Imperial College London

Department of Mathematics

MATH96007 - MATH97019 - MATH97097
Methods for Data Science
Years 3/4/5

Coursework 1 – Data exploration, regression, classification

Deadline: 4 November 2019, 5 pm

General instructions

The goal of this project is to analyse a data set using some of the tools introduced in the lectures, but also following your own initiative. Coursework tasks are different from exams: they can be more open-ended and may require going beyond what we have covered explicitly in lectures. Initiative and creativity are important as is the ability to pull together the course content, draw new links between subjects and back up your analysis with relevant computations. The quality of presentation and communication are very important, so use good combinations of tables and figures to present your results.

Submission instructions can be found at the end of this document. You must submit a Jupyter notebook, execute it, and submit the html file. You can produce a notebook in Google Colab, or you are free to use your local Jupyter notebook through the Anaconda environment downloaded on your computer. Before you produce your html, make sure all cells are executed, so their output appears in your html file.

To gain the marks for each part you are required to: (1) complete the task as described, (2) comment on your code so that we can understand each step, and (3) provide a brief written introduction to the task explaining and discussing what you did. *You are allowed to use high-level Python commands and packages unless indicated (e.g., not allowed in Task 1.4)*

This coursework is worth 15% of the total mark, and contains *no mastery component* for MSci and MSc students.

Overview

In this first coursework, you work with a data set that consists of weather and climbing data of Mount Rainier. Mount Rainier is a 4,392 meters high stratovolcano in Washington, USA, and is considered difficult to summit. In Task 1, you explore the data set with some descriptive statistics and visualisations. In Task 2, you implement and compare different regression and classification algorithms we have introduced in the lectures.

Start: The data set can be found as CSV files on BlackBoard. Download the data set. You will find two files: one gives weather data, the other climbing data. Your aim is to find relationships between both. There is also a test data set that will be used to check your results.

> *Hint: It is helpful to import both CSV files with pandas and show their first five lines to see what their columns describe.*

> *Hint: In Google Colab, you can upload the CSV files from your local drive with "from google.colab import files", and then "files.upload()".*

Task 1: Exploring data (30 marks)

1.1. Cleaning data **(10 marks)**

1.1.1. In the file climbing_statistics.csv, find the rows where both the date and route are identical, and write commands that aggregate their information as follows: (i) obtain the total number of Attempts and Successes for the same route and date; (ii) compute the percentage of successful attempts to summit for the same route and date.

Note: In most datasets, there will be faulty or incomplete data. If there are any rows with inconsistent data based on your computations above, ***eliminate them*** *from your data set.*

1.1.2. In the file Rainier_Weather.csv, delete all columns that do not contain any information about weather conditions.

1.1.3. Merge the two CSV files into one file containing all the above data by matching the dates such that each row contains a date and route and all the corresponding columns with the weather variables and the climbing statistics. The resulting merged file constitutes your data set.

1.2. Visualising data **(10 marks)**

1.2.1. Plot the Success Percentage, Attempts and Successes as a function of time in three separate subplots. You will need to transform dates into a linear time variable for the x-axis of the plot.

1.2.2. Plot all weather variables (except Wind Direction) over time in *one diagram* with two y-axes. The y-axis on the left of the diagram should contain data for Temperature and Wind Speed Daily, the y-axis on the right of the diagram should contain Relative Humidity and Solar Radiation. Plot the time-series in different colours and place a legend to the right of the diagram.

1.2.3. Plot all five weather variables as histograms (50 bins) in *one diagram,* assigning a different colour to each distribution. Specify two x-axes: the x-axis at the bottom should correspond to Temperature, Relative Humidity, and Wind Speed Daily (range from -5 to 100); the x-axis at the top should correspond to Wind Direction and Solar Radiation (range from 0 to 375).

1.3. Visualising relationships between descriptors **(5 marks)**

Write a function that shows 5x5 scatter plots of the five weather variables against each other. Comment on any observable relationships between variables and the possible impact on any regression task.

1.4. Splitting the data into training and validation sets **(5 marks)**

Although in this coursework we will **not** use T-fold cross-validation in the standard form, it is important to know how to split the data into training and validation sets.

Write a function that randomly divides the data set into a ‘training’ set with 80% of the rows, and a ‘validation’ set with the remaining 20%. Make sure that the training and validation sets are *well sampled, e.g.,* with respect to the Success Percentage.

Do this by writing your own set of commands (i.e. do not use the sklearn or any other predefined train test split function).

Task 2: Compare regression and classification algorithms (70 marks)

In Task 2, you will use the Mount Rainier data set to compare a couple of regression and classification algorithms. With the regression algorithms, you will explore relationships between weather conditions and climbing statistics. Using the classifiers, you should try and identify which weather conditions are optimal to maximise the probability of reaching the summit, according to the historic records of attempts in the data set.

Note: Both the regression tasks and the classification tasks below should be carried out on the 'training' set only, and then tested on the 'validation' set that you have obtained in Task 1.4

2.1 Regression **(40 marks)**

2.1.1. Linear regression **(10 marks)**

- Using the cleaned and merged data prepared in Task 1, obtain a linear regression model to predict the 'Success Percentage' using the four weather features 'Temperature', 'Relative Humidity', 'Wind Speed Daily' and 'Solar Radiation' as predictors.
- Report the parameters of the model and the in-sample error from the training set.
  Apply the model to the validation data obtained in Task 1 to predict the 'Success Percentage' and compute the out-of-sample error (MSE) from this vallidation set and compare with the in-sample error.
- Test the model on out-of-sample data by predicting the 'Success Percentage' for the weather conditions from 28-31 December 2018 available as a CSV file on Blackboard (this is the test set).

2.1.2. Ridge regression **(20 marks)**

Repeat the above task for Ridge Regression models. You will need to scan the penalty parameter of the Ridge models and establish the optimal value of the penalty for this dataset by examining the error on the validation set.

Explain your results using plots, code, and the effect of the parameter of in-sample and out-of-sample errors.

2.1.3. Discussion **(10 marks)**

Briefly discuss the results of both regression algorithms and identify the source of the differences in performance, if they exist, based on your exploration of the descriptors in the data set. You can support your answers with additional computations or methods.

2.2 Classification **(30 marks)**

2.2.1. Preparation of the data as categorical variables **(10 marks)**

To solve the classification tasks you need to unfold your data so that each row contains a sucess variable that is binary, i.e., 'success=1' or 'failure=0' for each attempt. Use the functionalities in Python to create from the data provided an expanded table of climbs with date and weather descriptors as in the original table, but a binary outcome variable of success.

2.2.2. Logistic regression **(10 marks)**

Train a logistic regression classifier on the training data with the same four weather conditions as features and the binary success of summiting Mount Rainier as the output.

After the classifier is trained, estimate the probability to successfully summit on the validation data. Compare the errors in the training data and validation data using different metrics.

2.2.3. Naive Bayes **(10 marks)**

Same as for the logistic regression classifier in 2.2.2, but training a Naive Bayes classifier on the training data with the same four weather conditions as features and the binary success of summiting Mount Rainier as the output.

After the classifier is trained, estimate the probability of summiting on the validation data. Compare and discuss the errors in the training data and validation data.

Submission instructions

You will upload two documents to Blackboard, wrapped into a single zip file:

1) Your notebook as an ipynb file.
2) Export your notebook as an html file.

You are also required to comply with these specific requirements:

- Name your files as 'SurnameCID.zip', e.g. Smith1234567.zip. Do not submit multiple files.
- Your ipynb file must produce all plots that appear in your html file, i.e. make sure you have run all cells in the notebook before exporting the html.
- Use clear headings: your html must make it clear where the answers to each question are, e.g. 'Task 1.1.1', etc.

Needless to say, projects must be your own work.

You may discuss the analysis with your colleagues but the code, writing, figures and analysis must be your own. The Department may use code profiling and tools such as Turnitin to check for plagiarism, and plagiarism cannot be tolerated.

Copying and plagiarism, if they occur, may force the Department to stop offering project-based courses such as this one.