STAT40830 - Adv Data Prog with R (online)

Assignment 1

Due on Wednesday 1st July 2020 11:59pm (IST).

This assignment is worth 20% of your overall grade for this module. You should carry out your analysis in `R` and submit your `R` script (in .R file format) to Brightspace before the assignment deadline. You will be assessed on all material covered from Weeks 1 to 4.

This assignment consists of two sets of questions concerned with two different datasets, both available in `R`.

Instructions

1. Download the Assignment1.R script from Brightspace. This script contains some code necessary for you to complete the assignment, and has space for you to write your own code. Save this file in the format XXXXXXXX.R where XXXXXXX is your student number.

2. Run the first two lines of code in the script. The first is to empty your environment and the second is to set the seed. This is so that we can mark your script. **You must run these lines of code.**

3. You should extend the provided `R` scipt with the instructions outlined below in Questions 1 and 2.

4. Remember that the code should be easy to read and to work with. You should add appropriate comments, use sensible notation and ensure your code is concise and efficient.

5. Submission: upload on Brightspace the modified .R file and .pdf files for your plots. Note that a new submission will replace your previous files.

Answer the following two questions in your `R` script.

Question 1

This question is based on the materials of Weeks 1 and 2. You should prepare your solution using only functions that have been introduced in the first two weeks.

Implement a function called `regression_fit` with arguments:

- `x_g`: a vector of scaled predictors
- `x`: a vector of scaled covariates
- `y`: a vector of scaled responses
- `p`: an integer ≥ 0, indicates the order of the polynomial regression (defaults to 1)
- `method`: the algorithm used to optimise the parameters of the Gaussian process regression (defaults to 'BFGS').

The function should fit a polynomial regression with no intercept term and powers of the covariate up to p; also, it should fit a Gaussian process regression with parameters optimised using the method indicated. The starting point of the optimisation should be c(0,0,0) in the log scale.

The function returns a list of two vectors that contain the predicted values using the polynomial regression and Gaussian process regression, respectively.

Run your function on the trees dataset to predict the height (in feet) of Black Cheery Trees from the measurement of the tree diameter (in inches) - note that this is labelled Girth in the data set and is measured 4ft 6in above the ground. Use an order of polynomial 4 and using the BFGS optimisation method for the GP regression model. Create a scatterplot of the data along with your fitted models and ensure this is appropriately labelled and includes a legend. Save this as 'regression.pdf'.

Question 2

This question is based on the materials of Weeks 3 and 4. You should use `magrittr` format whenever this can provide a more readable and concise style, as illustrated in Week 4. You should use `ggplot2` to create the plots for this question and `ggsave` to export them. Remember to redefine plot titles and axes to give them meaning.

We will be using the `flights` dataset in the `nycflights13` package. This dataset contains airline data for all flights that departed NYC (i.e. JFK, LGA or EWR) in 2013.

(a) Create a new dataset '`flights_2`' that contains only flights from 'EWR' to 'LAX'. Recast the '`carrier`' variable as a factor, with levels in the following order: 'UA', 'VX', 'AA'.

(b) Create a barplot where the bars show the number of flights from 'EWR' to 'LAX' for each of the carriers.
Save the plot as 'plot_1.pdf'.

(c) Calculate the average air time for each carrier for flights from 'EWR' to 'LAX'.
Plot the estimated densities for each of the underlying empirical distributions (i.e. 1 figure with 3 continuous lines, each corresponding to a different carrier).
Save the plot as 'plot_2.pdf'.

(d) When producing the plot for Q2.c) the following warning message appears:
`Removed 45 rows containing non-finite values (stat_density).`
Why did we get this warning message?
What could be done to avoid this message?

(e) Using the `magrittr` format, define a function called '`speed`' that takes a flights data.frame and adds a new column with value equal to the average speed in miles per hour.
Plot boxplots for the speed by month, for all flights from 'EWR' to 'LAX'.
Save the plot as 'plot_3.pdf'.

(f) Create multiple scatterplots to visually explore how delay at departure affects delay at arrival by carriers ('EWR' to 'LAX' only).
The scatterplots share the same y-axis but have different x-axes and different points colours.
Save the plot as 'plot_4.pdf'.

Please ensure that all documents are uploaded to Brightspace **clearly and on time**. It is students' responsibility in taking care of uploading all assignments. Pending or incomplete submissions after the deadline will be considered late and be penalised according to UCD guidelines, unless proven and valid justification (e.g. doctor's certificate) is provided. **Plagiarism is prohibited**. Please refer to the UCD Plagiarism Policy.

In addition, you must not discuss your answers or attempts on any of the module forums or with your classmates. Questions relating to the assignment will not be answered as the work is assessed.

Good luck and enjoy the assignment.