10 Academy July 2020 Training - Weekly Challenge: Week 3

This case will be used for the week 3 of training for Batch 3.

Change point analysis to quantify the impact of African government policy interventions to slow the spread of COVID-19

Business objective

The main business objective of this challenge is to quantify the statistical significance of a public health policy introduced by African governments to slow down the spread of COVID-19.

Situational Overview (Business Need)

The African Union has hired Batch3 LLC, a data science consultancy, to gather insights from across the continent on which public health and social measures are most effective at reducing the spread of Covid-19. The contract states that they are looking for evidence-based insights and they want to learn from existing initiatives launched a competition for all Young African Data Scientists to provide evidence-based insight on the effectiveness of recent public health government policies in the continent as it relates to Covid-19.

The contract wants Batch 3 LLC to investigate the following countries to identify which interventions had the most effect on the country under study. The African Union is looking to the analysis results from Batch3 LLC to both drive policy recommendations to countries in Africa on how to deal with Covid-19 as well as to help countries prepare for future pandemics.

Batch 3 LLC has assigned the following teams to study the following countries.

Project Country	**Project Team Members**
Egypt	Natanan Meleta Iyanuloluwa Osuolale Adnan Adetunji Karen Ngugi Victor Anisi Emmanuel Patrick
Ethiopia	Biniyam Tiruye Dawit Yilma Nahom Negussie Nabil Seid Yassin Rahel Weldegebriel Muluwork Geremew Busayo Olushola
Ghana	Samuel Negash Hailu Patrick Ojunde Augustine Anankum Ayebilla Avoka Johnson Obeng Judy Muriithi Evander Eghan
Kenya	Adah Kibet Sharleen Muoki Brian Odhiambo Kevin Karobia Gaitho Claire Munyole Gerald Okioma
Nigeria	Lawal Ogunfowora Victoria Akintomide Tijesunimi Olashore Olusola Timothy Ogundepo Temilade Adelakun Aminu Bello Abdullahi Memory Apiyo
Rwanda	Jeannette Uwizeyimana Bessy Mukaria Moyinoluwa Sobowale John Lotome Rofiah Adeshina Stephany Wanjiru
Senegal	Abubakar Alaro David Elvis Abdulazeez Shittu Ridwan Amure

	Idowu ilekura Wamuyu Wanjohi
South Africa	Chala Getu Anastasia Kiiru Oluwasegun Ajikobi Glory Odeyemi Gbenga Jayeola Ken Mbaya

Data

The data that will be used comes from the COVID-19 project by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The python package that comes with the key code (Priesmann et al.) already contains a python class to download the relevant data from John Hopkins database.

Alternatively, if you want to use a code you have already used before in your week0 challenge, you can find a Python code to connect and fetch data for a particular country in the jupyter notebook[1].

Learning Outcomes:

Skills:

- Statistical Modelling
- Using PyMC3 - a standard Bayseian modelling package in Python
- Working with remote data
- Running and modifying a moderately large python package

Knowledge:

- Probability distributions and choosing the relevant one for a given task
- Bayesian inference
- Monte Carlo Markov Chain
- Model comparison
- Policy analysis

Communication:

- Reporting to government bodies

[1] https://github.com/10acad/QuickStart2020/blob/master/week3/notebook/tenx_covid19_analysis.ipynb

Team

Instructor: Yabebal Fantaye with Jean-Henock
Tutors: Sebastian, Abla, Usman
Special External Support: Jonas Dehning

Key Dates:

- Discussion on the case - 1130 Rwanda time on Monday 3 August 2020. Use #all-week3 to pre-ask questions.
- Interim Solution - 2000 Rwanda time on Tuesday 4 August 2020.
- Final Submission - 2000 Rwanda time on Saturday 8 August 2020

Grading for the week

There are 100 points available for the week.

15 points - community growth and peer support. This includes supporting other learners by answering questions (Slack), asking good questions (Slack), participating (not only attending) daily standups (GMeet) and sharing links and other learning resources with other learners.

30 points - presentation and reporting.

10 points - interim submission

25 points for the final submission. This is measured through:

- Clarity of writing (10 points)
- Clarity of structure and message including appropriate usage of graphs (5 points)
- Professionalism/production value (free of spelling errors, use of same font, well produced) (5 points)
- Balance between being 'full of information' and 'easy to understand' (5 points)

45 points - final submission

Validity of model and recommendations (25 points)

Quality of code (20 points)

Badges

Each week, one user will be awarded one of the badges below for the best performance in the category below.

In addition to being the badge holder for that badge, each badge winner will get +20 points to the overall score.

Visualization - quality of visualizations, understandability, skimmability, choice of visualization
Quality of code - reliability, maintainability, efficiency, commenting - in future this will be CICD
Innovative approach to analysis -using latest algorithms, adding in research paper content and other innovative approaches
Writing and presentation - clarity of written outputs, clarity of slides, overall production value
Most supportive in the community - helping others, adding links, tutoring those struggling

The goal of this approach is to support and reward expertise in different parts of the Data Scientist toolbox.

Group Work Policy

You are expected to complete the project with your assigned group. The conception of the model and developing the code and modelling is to be done within groups; all members of the group can submit the same code. We recommend that everyone keeps a copy of this code in their own GitHub repository.

Reporting, both for the interim report as well as for the final report, must be done individually.

We expect all group members to contribute equally. We leave the assignment of roles within groups to the group members.

Late Policy

Our goal is to prepare successful learners for the work and submitting late, when given enough notice, shouldn't be necessary.

For interim submissions, those submitted 1-6 hours late will receive a maximum of 50% of the total possible grade. Those submitted >6 hours late may receive feedback, but will not receive a grade.

For final submissions, those submitted 1-24 hours late, will receive a maximum of 50% of the total possible grade. Those submitted >24 hours late may receive feedback, but will not receive a grade.

When calculating the leaderboard score:

- From week 4 onwards, your lowest week's score will not be considered.
- From week 8 onwards, your two lowest weeks' scores will not be considered.

Instructions

Objectives:

The global (business) objective is divided into 4 sub-objectives

- Defining the data analysis workflow
- Understanding the model and data
- Extending the model for African countries
- Extracting statistically valid insights in relation to the business objective

Task 1

The business objectives are a mission statement for a data analysis project. It has to be translated into a set of data analysis objectives - a workflow. A well articulated data analysis workflow contains the following elements

- A clear understanding of the data to be used - how it is generated, sampled, and compiled.
- A clear understanding of the model - the inputs to the model, the parameters that make up the model, the output of the model
- A clear understanding of the assumptions - assumptions on the data that couldn't be easily verified, critical assumptions of the model that has to be stated with any result obtained using the model.
- A clear understanding of the limitations of the analysis - reality is complex and models are always an approximation. As the well known statistician George Box summarised it "All models are wrong, but some are useful". In this context the data analysis workflow has to outline the expected limitations of the work being done, the context the results will be interpreted, and some sensitivity test that needs to be carried out to determine the impact of some of the assumptions and limitations.
- The main media channels and formats the result will be shared and communicated with the stakeholders (e.g. the public).

In the following two sub-tasks you are required to address the first two sub-objectives. Neither of these tasks include coding, but require a good understanding of the main paper of the project. Read the references given below, do your own research, and discuss with your colleagues as necessary. Doing well on these tasks will give you a strong foundation to complete the data analysis part of the project.

Task 1.1

To ensure your project will be completed on time, you must plan in detail your data analysis end-to-end workflow.

After reading the main paper of the challenge, you should make sure you understand

the following key point: **what are the main analysis steps you will have to carry out to achieve the objective of this challenge?** This is equivalent to defining data science objectives. Given the complex nature of the challenge, there may be a number of new concepts that you should familiarise yourself with.

The first task in outlining a detailed workflow is to find out the key concepts of the project and ensure you have a working level understanding. In order to do that, your task is to write down an explanation in your words for the following points:

- Explain in your words the purpose of the SIR/SEIR model in the current project?
- We will use the COVID19 cases data for this challenge, comment on the output of an SIR/SEIR model in relation to this data.
- List the processes that affect the generation of the COVID19 cases data in your country, and which part of this process is modelled by the SIR/SEIR model.
- Explain the difference between the SIR and SEIR model.
- Explain the distinct characteristics of an exponential function.
- Explain the similarities and differences of an exponential growth, exponential decay, geometric progression, and logistic growth.
- Explain what approximation leads the SIR/SEIR model to take an exponential form.
- Explain how the rate parameters in the SIR/SEIR model are estimated from a COVID19 cases data.
- Explain which probability distributions will be used to model the SIR/SEIR rate parameters. Why?
- Describe the expected outputs of the modelling phase of this challenge.
- How are predictions to future dates, for example one week from the date of the last date of the training data, is done?
- Describe how the effectiveness of government COVID19 non-pharmaceutical interventions policies are evaluated in this challenge.

This reference might be a good help for understanding the mathematics behind the SIR model.

Task 1.2

If one has a clear understanding of the whole process that goes in a statistical modelling of a given business objective, applying the concept to a similar data and business objective should be straightforward.

In this task you are required to use your understanding from **task 1.1.** and write what you think needs to be changed to adapt it to the following scenario.

1. Scenario: public interest in corruption
    a. **Data**: number of daily tweets containing the word corruption from a given country for a year

b. **Business Objective**: disentangle the relative interest of people in corruption issues because of the following triggers:
    i. Reports emerged at the beginning of February, that an important politician was taking bribes
    ii. This politician was convicted on the 15th of August.
    iii. The slow rise during this year of every-day bribery (police, clerks, etc.)

Task 2

In this Task you will use Bayesian inference technique to estimate the SIR model parameters, and apply Bayesian model comparison to select models that best fit the observed data. The major analysis steps you will follow are:

1. Git fork the following repository project python package
2. Follow the readme to pip install the repository you forked and make sure you are able to run the notebook located in the root directory in the following path: scripts/interactive/example_one_bundesland.ipynb
3. Adopt the notebook to run a similar analysis for your country
4. Interpret the result and write up your finding, paying particular attention to the limitations.

Task 2.1

Based on the COVID-19 daily cases, you are expected to do the following

a. Download the COVID19 case data for your assigned country
b. Pre-process the downloaded data such that the starting date of the data is when the number of covid19 cases in your country reaches 100 and dominated by a community transmission.
c. It is unlikely but, if there are dates that have NaN values, perform linear regression to fill these missing values. Make sure the final data has a continuous date - ensure no date is missing. Zero number of cases for a given date is ok.
d. Split the data into one part used for inference (training set), and an other used for to validate a forecast (validation set):
    i. **Training set** includes all dates from the time the community transmission reaches 100 to July 25 2020.
    ii. **Validation set** includes dates from 25 July 2020 to one final date in the covid19 cases data.
e. Plot the training data together with the model that is sampled from the posterior of the SIR model. The posterior of the SIR model means distributions on the Lambda, Mu, and other parameters. A single model curve means a single sample from the posterior distribution.
f. Use the validation data set to evaluate the forecasting power of the model you generated using the training set. If you are happy with your model, you can run it to make predictions until the 10th of August.

g. Find the dates where your country introduced the following policies. You can find the policies different countries introduced in the Oxford COVID tracker[2] & IMF policy tracker[3].
    i. Banning major gatherings
    ii. School closures
    iii. Required social distancing
    iv. Mask wearing mandatory
    v. Relaxing the previous rules
h. Find those policies that are announced after the number of cases reach 100. Make sure you have at least one policy that falls in your data date range.
i. Introduce the dates of the policies you found in step (g) as change points in your model with prior centered at the times the country introduced these policies. In the example paper, Germany had three specific points which were mild social distancing, strong social distancing and contact ban on 2020-3-9, 2020-3-16, and 2020-3-23.
j. You are expected to get the following results after performing the parameter estimation and model comparison steps
    i. Posterior distribution of SEIR model parameters
    ii. The number of change points required in the model to best fit the data

Task 2.2:

Interpret the result and write a report to showcase your work and help explain how data and your model can help to predict case progression in the country that you are working on. Pay attention to the following details

- Your target audience should be those with a general scientific knowledge, such that they can understand your posting without needing very specific domain knowledge.
- You must explain why this statistical modelling technique is suited to the data you are using
- Give an overview of how your model works and how it builds on the Science publication
- Discuss the change points' and what the relative effect of each one was (and when they were observed, as per your model).
- What are the predicted number of cases for the country for the week of 10 August 2020?
- Explain the limitations of your analysis given the data available.
- Highlight any lessons learned that countries could adopt for future pandemics

[2] https://covidtracker.bsg.ox.ac.uk/stringency-scatter
https://github.com/OxCGRT/covid-policy-tracker
[3] https://www.imf.org/en/Topics/imf-and-covid19/Policy-Responses-to-COVID-19

Interim Submission (Due Tuesday 4 August 2020 20hr Rwanda time)

- Share a report that addresses the points from task 1 (answer all questions in task 1.1. & task 1.2). Maximum of 3 pages - PDF format please.

Feedback

You may not receive detailed comments on your interim submission, but will receive a grade.

Final Submission (Due Saturday 8 August 2020 20hr Rwanda time)

- Link to your code in GitHub
- A complete report with details addressing both task 1 & task 2. PDF format please, max 5 pages.

Feedback

You will receive comments/feedback in addition to a grade.

References

1. **Key paper** Jonas Dehning et al. (2020): Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions
2. **Key Code**: https://github.com/Priesemann-Group/covid19_inference_forecast
3. As a introduction to Bayesian statistics and the python package (PyMC3)
4. Why is it difficult to accurately predict the COVID-19 epidemic?

On SIR/SEIR Model

5. Intuition on SIR Model
6. On SEIR Model
7. COVID-19 dynamics with SIR model and the references therein

On Bayesian Change point detection

8. Bayesian Changepoint Detection with PyMC3
9. Model Comparison using PyMC3