MiniProject 1: Analyzing COVID-19 Search Trends and Hospitalization

COMP 551, Fall 2020, McGill University

October 1, 2020

Please read this entire document before beginning the assignment

Preamble

- **Quiz TA's**; Arna Ghosh and Howard Huang
- This mini-project is due on **October 16th** at 11:59pm EST. Late work will be automatically subject to a 20% penalty, and can be submitted up to 5 days after the deadline. No submissions will accepted after this 5 day period.
- This mini-project is to be completed in groups of three. There are three tasks outlined below which offer one possible division of labour, but how you split the work is up to you. All members of a group will receive the same grade. It is not expected that all team members will contribute equally to all components. However every team member should make integral contributions to the project.
- You will submit your assignment on MyCourses as a group. You must register your group on MyCourses and any group member can submit. See MyCourses for details.
- You are free to use libraries with general utilities, such as matplotlib, numpy, scipy, pandas and sklearn for Python.

Background

In this miniproject, you will be exploring two COVID19-related datasets. The goal is to gain experience in deploying unsupervised and supervised machine learning techniques to tackle a real-world data science problem. You are encouraged to explore techniques you have learned in class to visualize the data and thereafter form a hypothesis about possible patterns in the data.

Task 1: Acquire, preprocess, and analyze the data

Your first task is to acquire the data, analyze it, and clean it (if necessary). You will use two publicly available datasets provided by Google Research in this project, outlined below.

- **Dataset 1 (Search Trends dataset)**: This aggregated, anonymized dataset shows trends in search patterns for symptoms and is intended to help researchers to better understand the impact of COVID-19. Read the dataset details and how it was generated here. We suggest using the weekly resolution dataset that can be downloaded from the associated github repo.
- **Dataset 2 (COVID hospitalization cases dataset)**: This is an open source dataset that aggregates public COVID-19 data sources into a single dataset. The dataset includes time series data for COVID-19 cases, deaths, tests, hospitalizations, discharges among other attributes. For more information about the dataset, refer to the associated github repo. We suggest using the hospitalizations from this dataset for the United States regions that are also present in the Search Trends dataset. You can get the data (provided under the CC-BY license) from here. https://github.com/google-research/open-covid--19-data/tree/master/data/exports/cc_by

The essential subtasks for this part of the project include:

1. Download the datasets. *Hint: If you are working locally, you could try cloning the repository and then using the csv files from your cloned version. Do not forget to mention the date (or version) of the dataset that you ended up using in your report.*

2. Load the datasets into Pandas dataframes or NumPy objects (i.e., arrays or matrices) in Python.

3. Clean the data. Are there any symptoms that have no search data available? Do all regions have valid hospitalization data (you can assume regions to have valid hospitalization data if they have sufficient non-zero entries)? **You should remove regions and features that have too many missing or invalid data entries.**

4. Merge the two datasets. Note that the time resolution is different for the two datasets, the search symptoms is weekly whereas the hospitalization cases are at the daily resolution. Your task is to bring both the datasets at the weekly resolution and thereafter merge them into one array (Numpy or Pandas).

Task 2: Visualize and cluster the data

Your next task is to leverage dimensionality reduction techniques and visualize the data. The subtasks for this part include:

1. Visualize the evolution of popularity of various symptoms across different regions over time. Specifically, you need to visualize how the distribution of search frequency of each symptom aggregated across different regions changes over time. You can only do these plots for some of the most popular symptoms only. *Hint: checkout some visualizations here for inspiration.*

2. Visualize the search trends dataset in a lower dimensional space. Use Principal Component Analysis (PCA) to reduce the data dimensionality. *Hint: You may treat each time point as an independent data point.*

3. Explore using a clustering method – e.g., *k-means* – to evaluate possible groups in the search trends dataset. Do the clusters remain consistent for raw as well as PCA-reduced data?

Task 3: Supervised Learning

In this part, you will compare two supervised learning frameworks, namely *K-nearest neighbours* (KNN) and *decision trees*, to predict the hospitalization cases given the search trends data. The specific subtasks for this part include:

1. Split the data into train and validation sets using two strategies – based on regions and based on time. Specifically, in the first case, you need to keep all data from some regions in the validation set and train on the rest (keep 80% regions in training set and 20% in validation set, doing this multiple times to estimate cross-validation results). In the second case, you need to keep data for the last couple of timepoints (keep data after '2020-08-10') from all regions in the validation set and train on the rest of the data

2. Compare the regression performance of KNNs and decision trees for each of the train-validation split strategies. Note that you can report a 5-fold cross-validation performance for region-based train-validation split, wherein you vary which regions are kept in the validation set for each fold. Please clearly report your validation error in both cases.

3. **[Optional]** Explore other prediction strategies. For example, one strategy could be to learn separate models for predicting hospitalization in each region or cluster from Task 2.

Deliverables

You must submit two separate files to MyCourses **(using the exact filenames and file types outlined below)**:

1. **code.zip**: Your data processing, classification and evaluation code (as some combination of .py and .ipynb files).

2. **writeup.pdf**: Your (max 5-page) project write-up as a pdf (details below).

Project write-up

Your team must submit a project write-up that is a maximum of five pages (single-spaced, 11pt font or larger; minimum 0.5 inch margins, an extra page for references/bibliographical content can be used). We highly recommend that students use LaTeX to complete their write-ups. You have some flexibility in how you report your results, but you must adhere to the following structure and minimum requirements:

Abstract (100-250 words)

Summarize the project task and your most important findings. For example, include sentences like "In this project we investigated the performance of two regression models, namely k-nearest neighbours and decision trees, on predicting COVID-19 hospitalization cases from related symptoms search", "We found that the k-nearest neighbour regression approach achieved worse/better accuracy than decision trees and was significantly faster/slower to train."

Introduction (5+ sentences)

Summarize the project task, the two datasets, and your most important findings. This should be similar to the abstract but more detailed. You should include background information and potential citations to relevant work (e.g., other papers analyzing these datasets).

Datasets (5+ sentences)

Very briefly describe the datasets and how you processed them. If you have come up with new new features to get better results, you should explain it here. Present the exploratory analysis you have done to understand the data, e.g. visualization plots and data filtering.

Results (7+ sentences, possibly with figures or tables)

Describe the results of all the experiments mentioned in Task 2 and 3 (at a minimum) as well as any other interesting results you find. At a minimum you must report:

1. A visualization of the search trends data in lower dimensions
2. Same plot as above but with cluster labels for each data point to illustrate the clustering results
3. A comparison of regression performance (mean squared error or mean absolute error) between KNN and decision trees on the aforementioned cross-validation schemes

Discussion and Conclusion (5+ sentences)

Summarize the key takeaways from the project and possibly directions for future investigation.

Statement of Contributions (1-3 sentences)

State the breakdown of the workload across the team members.

Evaluation

The mini-project is out of 100 points, and the evaluation breakdown is as follows:

- Completeness (20 points)
  - Did you submit all the materials?
  - Did you run all the required experiments?
  - Did you follow the guidelines for the project write-up?
- Correctness (40 points)
  - Are you cross-validation schemes implemented correctly?
  - Are your models used/implemented correctly?
  - Are you visualizations informative and visually appealing?
  - Are your reported accuracy close to (our internal) reference solutions?
  - If you proposed any features, did your proposed features actually improve performance, or do you adequately demonstrate that it was not possible to improve performance?
- Writing quality (25 points)
  - Is your report clear and free of grammatical errors and typos?

  - Did you go beyond the bare minimum requirements for the write-up (e.g., by including a discussion of related work in the introduction)?
  - Do you effectively present numerical results (e.g., via tables or figures)?

- Originality / creativity (15 points)
  - Did you go beyond the bare minimum requirements for the experiments? For example, did you investigate which features are the most useful (e.g., by correlating them with your predictions or removing them from your data)?
  - Did you use other publicly available data to run more interesting experiments (e.g., using neighbourhood information, or weather conditions for different states). This could potentially give you better performance on the validation set.
  - within the context of producing the required results did you propose a creative idea?
  - **Note:** Simply adding in a random new experiment will not guarantee a high grade on this section! You should be thoughtful and organized in your report in explaining why you performed an additional experiment and how it helped in evaluating your hypothesis.

Final Remarks

You are expected to display initiative, creativity, scientific rigour, critical thinking, and good communication skills. You don't need to restrict yourself to the requirements listed above - feel free to go beyond, and explore further

You can discuss methods and technical issues with members of other teams, but **you cannot share any code or data with other teams**.