
Reconhecimento de Padrões Inteligência Geoespacial Aprendizagem Computacional em Biologia

2020/2021

Project Assignment Covid-19 prediction

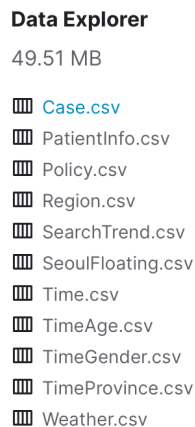
1 Background

Covid-19 has been for more than a year now a pandemic that has affected everyone globally. Several dataset exists with Covid-19 data and pattern recognition techniques are actively being researched to shed some light on data relations.

Your job in this assignment is to predict the outcome of the disease using patient data along with background data, as the weather.

2 Dataset Description

Consider the dataset available at <https://www.kaggle.com/kimjihoo/coronavirusdataset?select=Case.csv>. This dataset contains several csv files:



Data Explorer
49.51 MB

- Case.csv
- PatientInfo.csv
- Policy.csv
- Region.csv
- SearchTrend.csv
- SeoulFloating.csv
- Time.csv
- TimeAge.csv
- TimeGender.csv
- TimeProvince.csv
- Weather.csv

Figure 1: Available files (from [1]).

Data from 5165 patients including sex, age, country, province, city, date, and the outcome of the case is available (state). The outcome takes one of the possibilities:

1. **Released**, 2929 cases, 56.71%
2. **Isolated**, 2158 cases, 41.78%
3. **Deceased**, 78 cases, 1.51%

The rest of the information includes:

1) Case Data

- **Case**: Data of COVID-19 infection cases in South Korea

2) Patient Data

- **PatientInfo**: Epidemiological data of COVID-19 patients in South Korea
- **PatientRoute**: Route data of COVID-19 patients in South Korea (currently unavailable)

3) Time Series Data

- **Time**: Time series data of COVID-19 status in South Korea
- **TimeAge**: Time series data of COVID-19 status in terms of the age in South Korea
- **TimeGender**: Time series data of COVID-19 status in terms of gender in South Korea
- **TimeProvince**: Time series data of COVID-19 status in terms of the Province in South Korea

4) Additional Data

- **Region**: Location and statistical data of the regions in South Korea
- **Weather**: Data of the weather in the regions of South Korea
- **SearchTrend**: Trend data of the keywords searched in NAVER which is one of the largest portals in South Korea
- **SeoulFloating**: Data of floating population in Seoul, South Korea (from SK Telecom Big Data Hub)
- **Policy**: Data of the government policy for COVID-19 in South Korea

Figure 2: Information in the dataset (from [1]).

This information is stored in different cvs files, nevertheless it is possible to find the links between for instance the date of a case and the weather on the specified date and province.

Note: There are some missing values, that you may choose to filter or replace.

3 Objective

Your task is to develop classifiers for Covid-19. Consider three scenarios:

- **Scenario A (Binary Classifier)**: where one wants to distinguish the **Released** outcome from the others using just patient information (PatientInfo.csv);
- **Scenario B (Binary Classifier)**: where one wants to distinguish the **Deceased** outcome from the others using additional information as the weather;
- **Scenario C (Three-class Classifier)**: where one wants to distinguish **Released**, **Isolated**, and **Deceased** outcomes.

4 Practical Assignment

4.1 Data import

Develop scripts for feature data import. Organize data into sub-sets, relating to each source type you intend to test, e.g.: features from patient data; features weather; features from trends; features from regions; features from policies.

4.2 Feature Selection and Reduction

Some of the supplied features may be useless, redundant or highly correlated with others. In this phase, you should consider to use feature selection and dimensionality reduction techniques, and see how they affect the performance of the pattern recognition algorithms. Analyse the distribution of the values of your features and compute the correlation between them. Make sure you know your features! Do not forget to present your findings in the final report.

4.3 Experimental Analysis

You should be able to design experiences in order to run the pattern recognition algorithms in the given data and evaluate their results. Define the appropriate performance metrics and justify your choices!

Run the experiments multiple times! To be able to present average results and standard deviations (of the metrics used) you should split the training set in parts and use cross-validation. At the end you should be able to choose the best classifier and evaluate them in a testing set.

Do not forget that manually inspecting the predictions of your algorithms can give you precious insights of where they might be failing (and why), and what you can do to improve them (e.g. what makes the algorithm fail in this particular case? what special characteristic does it have that makes it so hard? how can I make the algorithm better deal with those cases?). Go back and forward to the Pre-processing, Feature reduction and Feature Selection phases until you are satisfied with the results. It is a good idea to keep track of evolution of the performance of your algorithm during this process. Try to show these trends in your final report, to be able to justify all the issues involved (choosing parameters, model fit, etc.)

4.4 Pattern Recognition Methods

You can write your own code in your language of choice or use the functions and methods available in Matlab and in the Statistical Pattern Recognition STPRTool used in the classes (since you are already familiarized with it). The methods used in your work should be described as well as discussion of the parameters used. Try out different pattern recognition algorithms. You should try to understand how they perform differently in your data.

4.5 Results and Discussion

Present and discuss final results obtained in your Project assignment. This problem was already studied by other authors. Compare your results with the results from other sources. In this problem one important aspect is to evaluate among the data available the more appropriate for the different scenarios.

4.6 Code & Graphical User Interface (GUI)

You should deliver your software code in MATLAB, or in any other programming language you used during the project.

For your project you should write code for a graphical user interface (GUI). The GUI should improve the interaction of the user with the code by providing options for data-loading, feature selection/dimensionality

reduction, classification, post-processing, validation and visualization.

Remember to comment your code. Write also a help section to your code that tells the purpose of the function, usage, and explanation of parameters.

5 Documentation

Write documentation (in Portuguese or in English) about your project. The documentation should include a cover page where course name, project title, date, names and student numbers of the authors are mentioned.

Describe the methods used for classification in such detail that the reader would be able to implement the same kind of functions for feature extraction and classification just based on your documentation and some basic background in pattern recognition. Always justify your choices, even when their are based on intuition. Do not forget to verify your assumptions! Include classification results with the given data to your documentation. At the end of your documentation you should have a list of all references used.

5.1 Requirements

Practical assignment is meant to be done in groups of two persons, but up to three are allowed. If someone wants to work alone, this is also possible. Larger groups are in principle not allowed.

5.2 Project Submission & Deadlines

1. Project First Milestone (**Deadline: 23rd April 2021!**)

Deliverables:

- Data Preprocessing (Scaling, Feature Reduction (PCA & LDA), Feature Selection, etc.);
- Minimum Distance classifier, Fisher LDA for Scenario A.
- Code + short report.

2. Project Final Goal (**Deadline: 21st May 2021!**)

Deliverables:

- Data Preprocessing (Scaling, Feature Reduction (PCA & LDA), Feature Selection, etc.);
- Several classifiers;
- Final Report
- Matlab code + GUI.

3. Presentation and Discussion (**from 24th to 28th May 2021!**)

References

[1] Jimi Kim, Seojin Jang, Woncheol Lee, Joong Kun Lee, Dong-Hwan Jang, "DS4C Patient Policy Province Dataset: a Comprehensive COVID-19 Dataset for Causal and Epidemiological Analysis", 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada. <https://www.cmu.edu/dietrich/causality/CameraReady-accepted%20papers/55%5CCameraReady%5Cpaper.pdf>