Disaster Relief Project: Part 1

DS-6030

Contents

In this project, you will use classification methods covered in this course to solve a real historical data-mining problem: locating displaced persons living in makeshift shelters following the destruction of the earthquake in Haiti in 2010.

Following that earthquake, rescue workers, mostly from the United States military, needed to get food and water to the displaced persons. But with destroyed communications, impassable roads, and thousands of square miles, actually locating the people who needed help was challenging.

As part of the rescue effort, a team from the Rochester Institute of Technology were flying an aircraft to collect high resolution geo-referenced imagery. It was known that the people whose homes had been destroyed by the earthquake were creating temporary shelters using blue tarps, and these blue tarps would be good indicators of where the displaced persons were – if only they could be located in time, out of the thousands of images that would be collected every day. The problem was that there was no way for aid workers to search the thousands of images in time to find the tarps and communicate the locations back to the rescue workers on the ground in time. The solution would be provided by data-mining algorithms, which could search the images far faster and more thoroughly (and accurately?) then humanly possible. The goal was to find an algorithm that could effectively search the images in order to locate displaced persons and communicate those locations rescue workers so they could help those who needed it in time.

This disaster relief project is the subject matter for your project in this course, which you will submit in two parts. You will use data from the actual data collection process was carried out over Haiti. Your goal is to test each of the algorithms you learn in this course on the imagery data collected during the relief efforts made Haiti in response to the 2010 earthquake, and determine which method you will use to as accurately as possible, and in as timely a manner as possible, locate as many of the displaced persons identified in the imagery data so that they can be provided food and water before their situations become unsurvivable.

You will document the performance of several models using cross-validation (Part I) and a hold-out testing set (Part II). In **Module 6** you will submit the results for Part I that includes performance of the models we have covered in Modules 1-5. In **Module 12** you will submit the results for Part II that includes performance of a few other models, overall conclusions, and recommendations on the preferred model for this application.

1 Submission Format

For Part I you will submit **two** deliverables:

1. **PDF document** which contains the results in a report format. You can use Word or any other text processing software to prepare this document. The emphasis of the report is to show results and discuss your findings. You are completely free in how you organize your report. However, the report must contain the minimum requirements listed below. **The PDF must not have more than 20 pages!**
2. **Rmarkdown (.Rmd)** file which contains the code

We will look at both documents.

2 Collaboration and Help

- While all work must be your own, you are permitted to discuss this project with classmates and post questions and answers on the discussion boards (e.g., teams).
  - However, you are **not** permitted to work collaboratively.
- You are not permitted to copy code. You will no doubt come across examples on the internet. You can consult them to help understand the concept or process, but *code in your own words.*
- It is a scholarly responsibility to attribute all your work. This includes figures, code, ideas, etc. Think of it this way: will someone who reads your submission think that it is your original idea, figure, code, etc? Add a link and/or reference to all sources you used to solve a problem. It is really of no value to you when you just copy someone else's solutions (other then preserve a grade that you didn't earn).
- If you use generative AI, list it as a reference and describe what you used it for.

It is not always easy to tell what qualifies as an honor code violation, so do not be afraid to talk to me about it. Such discussions do not imply guilt of any kind.

3 Project Part I DUE in Module 6

Use 10-fold cross-validation to evaluate the performance of 5 models:

- Logistic Regression
- LDA (Linear Discriminant Analysis)
- QDA (Quadratic Discriminant Analysis)
- KNN (K-nearest neighbor)
- Penalized Logistic Regression (elastic net penalty)

Download the `HaitiPixels.csv` data from https://gedeck.github.io/DS-6030/project/HaitiPixels.csv.

3.1 Document Format

Document is well structured and readable. **No code** shown in writeup. Maximum size for writeup: 20 pages (excluding Code that is moved to the end of the document excluded)

Expected sections:

- Introduction
- Data
- Description of Methodology
- Results
- Conclusions

Graphs suitably sized and clearly readable; axis labels and legends descriptive; captions describe the content of the graph.

3.2 Coding

Code **must** not be shown in the final PDF document. If you knit your document, the code should either be hidden. Add the following to the start of your document:

````
```{r hide-code, include=FALSE}
knitr::opts_chunk$set(echo=FALSE)
```
````

You can include the code in an Appendix at the end of the document. It will not count to the total page count. Add the following to the end of your document to display all code at once.

````
# Appendix {-}
```{r ref.label=knitr::all_labels(), echo=TRUE, eval=FALSE}
```
````

All code is well organized, and executes without errors. The R code should be easy to follow.

We will mainly look at the PDF report. This means, the report must contain **all** information required to understand and repeat the study without consulting the code.

We will **only** inspect your R code if there are problems in compiling your document so ensure we can understand what you implemented.

3.3 Data (loading, wrangling, and EDA)

Data loaded correctly and exploratory data anlysis (EDA) is performed to better understand the data

3.4 Method section

Model training, tuning, and validation described in details and suitable to allow reader to repeat the process without looking at code.

Overall model building process well defined and *explained*.

- describe used software
- describe and justify parameter tuning and model selection (if applicable)
- describe and justify model validation
- describe and justify threshold selection
- describe and justify metrics used for model performance evaluation

3.5 Results: Model Fitting, Tuning Parameter Selection, and Evaluation

Train five models:

- use random seeds for reproducibility
- describe and show parameter tuning and discuss results (use tables and/or plots)
- describe and show results of threshold selection
- describe and discuss model performance

This can be done individually or in summary.

3.6 Results: Summarize and discuss ROC curves and performance metrics

Model performance summarized in one or more tables or figures. Expected information shown:

- ROC curves and AUC
- Optimal model tuning parameters
- Selected threshold
- Accuracy, TPR, FPR, Precision calculated at selected threshold

3.7 Conclusions

Three or more clearly identifiable conclusions. This section is more important than the previous sections (as reflected in the points). Give sufficient explanation and justification for each conclusion.

One conclusion must be:

- determination and justification of which algorithm works best.

Additional conclusions should be observations you've made based on your work on this project, such as:

- What additional recommend actions can be taken to improve results?
- Were there multiple adequately performing methods, or just one clear best method? What is your level of confidence in the results?
- What is it about this data formulation that allows us to address it with predictive modeling tools?
- How effective do you think your work here could actually be in terms of helping to save human life?
- Do these data seem particularly well-suited to one class of prediction methods, and if so, why?

These are only suggestions, pursue your own interests. Your *best two additional* conclusions will be graded. **Make sure that the 3 conclusions are clearly separated.**