Disaster Relief Project: Part 2

DS 6030 | Spring 2023 | University of Virginia

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.

Objective

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

Submission Format

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

1. **Rmarkdown (.Rmd)** file which contains the code
2. **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 30 pages!**

We will look at both documents.

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 (Microsoft 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). 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.

Grading

Document format (5 pts)

Compiled document well structured and easy to read:

- organized well
- tables/plots fit on the page
- plots are labeled correctly
- etc.

Coding (5 pts)

We look for well organized code. All code is shown and executes without errors. The R code in the code chunks should be visible and easy to follow. Use `echo = TRUE` for all chunks that were actually used (e.g., personal notes to yourself or preliminary coding attempts shouldn't be shown).

Data Wrangling and EDA (10 pts)

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

Analyze and discuss EDA for both the training and the holdout set.

- How similar are the two datasets?
- What do you expect based on this result?

Model Fitting, Tuning Parameter Selection, and Evaluation (15 pts)

Use 10-fold cross-validation to tune and evaluate the performance of 7 models:

- Logistic Regression
- LDA (Linear Discriminant Analysis)
- QDA (Quadratic Discriminant Analysis)
- KNN (K-nearest neighbor)
- Penalized Logistic Regression (elastic net penalty)
- Random forest
- Support vector machine

Train your models using the `HaitiPixels.csv` data (provided in Module 3). Validate your models using the data in the `Hold Out Data.zip` file (provided in Module 9)

Describe how you approach the training, the tuning and the threshold evaluation.

- overall model building process well defined and *explained*.

- describe and justify parameter tuning and model selection (if applicable)
- describe and justify model validation
- describe and justify threshold selection
  - It should be clear how the threshold was applied
- describe and justify metrics used for model performance evaluation
- It should be clear what features were used for each model family.
  - E.g., by using a formula or explicit calculation of model matrix.
- It should be clear if any pre-processing was used (e.g., scaling).

For each of the models,

- 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
  - use ROC curves and relevant metrics; how are they derived

Results (15 pts)

We expect the following results:

- Cross validation performance Table (5 pts)
- Hold-out Test Performance Table (5 pts)
- ROC curves for training and holdout (5 pts)

Describe how the metrics were calculated under the cross-validation framework. - E.g., is it an average, a sum, a max, etc.

Compare and contrast the results.

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

Conclusions (50 pts)

Report **at least six** conclusions;

- four of them must be:
  1. A discussion of the best performing algorithm(s) in the cross-validation and hold-out data. (they may not be the same)
  2. A discussion or analysis justifying why your findings above are compatible or reconcilable.
  3. A recommendation and rationale regarding which algorithm to use for detection of blue tarps.
  4. A discussion of the relevance of the metrics calculated in the tables to this application context.
- And two more of your choice.
- **Separate your conclusions into different sections**

Holdout dataset

Additional data derived from images like this:

- Several files:
  - orthovnir078_ROI_NON_Blue_Tarps.txt
  - orthovnir078_ROI_Blue_Tarps.txt
  - orthovnir069_ROI_NOT_Blue_Tarps.txt
  - orthovnir069_ROI_Blue_Tarps.txt
  - orthovnir067_ROI_NOT_Blue_Tarps.txt
  - orthovnir067_ROI_Blue_Tarps.txt
  - orthovnir067_ROI_Blue_Tarps_data.txt
  - orthovnir057_ROI_NON_Blue_Tarps.txt
- You will need to process the data and convert them into a data frame suitable for your analysis
- The hold-out set has about **2 million data points**.
- Here is the format:

```
; ENVI Output of ROIs (5.5.1) [Tue Nov 05 11:56:48 2019]
; Number of ROIs: 1
; File Dimension: 4583 x 6796
;
; ROI name: Region #1
; ROI rgb value: {255, 0, 0}
; ROI npts: 295510
;      ID     X     Y      Map X       Map Y        Lat         Lon    B1   B2   B3
        1  2296   222  772981.96  2050019.86  18.523111  -72.414329   116  138  120
        2  2296   223  772981.96  2050019.78  18.523110  -72.414329   104  121  109
        3  2295   223  772981.88  2050019.78  18.523110  -72.414330   106  125  109
        4  2290   224  772981.47  2050019.70  18.523110  -72.414333   114  135  109
```

```
5  2291   224  772981.55  2050019.70  18.523110  -72.414333  111  133  107
```

- You will make several decisions about the data.
  - Which of the columns map to Red, Green, and Blue?
  - How do you assign class labels?
  - Are all files relevant?