Disaster Relief Project: Part 1

DS-6030

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.

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

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 (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). 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.

Project Part I (100 points) 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)

Use the `HaitiPixels.csv` data (provided in Module 3).

Document Format (5pts)

Compiled document well structured and easy to read:

- the compiled document must not have more than **25 pages**
- organized well
  - the provided template should satisfy the minimal requirements
  - (optional) make it better (e.g., used tabbed sections, custom css, different themes)
- tables/plots fit on the page
- there are not extraneous outputs (e.g., printing a matrix that fills 2 pages)
- plots are labeled correctly
- etc.

Coding (5 pts)

All code is shown, well organized, 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).

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

Data Wrangling and EDA (10 pts)

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

Model Fitting, Tuning Parameter Selection, and Evaluation (30 pts)

- 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 (e.g., to the estimated probabilities or logit of the probabilities).
  - describe and justify metrics used for model performance evaluation
    * use ROC curves to compare models
  - 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).
  - It should be clear how cross-validation was implemented.
- For each of the 5 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

Performance Table (20 pts)

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

- Optimal model tuning parameters
- AUC
- Selected threshold
- Accuracy, TPR, FPR, Precision calculated at selected threshold
- Describe how the metrics were calculated under the cross-validation framework.
  - E.g., is it an average, a sum, a max, etc.

Conclusions (30 pts)

Report on at least 3 conclusions. This section is more important than the previous sections (as reflected in the points). Give sufficient explanation and justification for each conclusion.

- One must be your **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:
  - Examples:
    * 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.