STAT 154: Project 2 Cloud Data

Release date: **Wednesday, April 10**

Due by: **11 PM, Wednesday, May 1**

Please read carefully!

- It is a good idea to revisit your notes, slides and reading; and synthesize their main points BEFORE doing the project.
- *For this project, we adapt a zero tolerance policy with incorrect/late submissions (no emails please) to Gradescope.*
- The recommended work of this project is at least 20 hours (at least 10 hours / person). Plan ahead and start early.
- We need two things:
  (a) A main pdf report **(font size at least 11 pt, less or equal to 12 pages)** generated by Latex, Rnw or Word is required to be submitted to Gradescope.
    - Provide top class (research-paper level) writing, useful well-labeled figures and no code in this pdf. Arrange text and figures compactly (.Rnw may not be very useful for this).
    - You can choose a title for the report and a team name as per your liking (*get creative!*). Do provide the names and student ID of your teammates below the title.
    - Your report should conclude with an acknowledgment section, where you provide brief discussion about the contributions of each member, **and** the resources you used, credit all the help you took and briefly outline the way you proceeded with the project.

  (b) A link to your GitHub Repo at the end of your write-up that contains all your code (see Section 5 for more details).
- **Be visual *and* quantitative:** Remember projects are graded differently when compared to homework—one line answer without explanation is usually not enough. Make your findings succinct and try to convince us with good arguments supported by numbers and figures. Putting yourself in reader's shoes and reading the report out loud usually helps. The standards for grading are *very high* this time. We will be very picky with figures: Lack of proper titles and axis labels will lead to loss of several points.

Overview of the project

The goal of this project is the exploration and modeling of cloud detection in the polar regions based on radiance recorded automatically by the MISR sensor abroad the NASA satellite Terra. You will attempt to build a classification model to distinguish the presence of cloud from the absence of clouds in the images using the available signals/features. Your dataset has "expert labels" that can be used to train your models. When you evaluate your results, imagine that your models will be used to distinguish clouds from non-clouds on a large number of images that won't have these "expert" labels.

On Piazza, you will find a zip archive with three files: **image1.txt**, **image2.txt**, **image3.txt**. Each contains one picture from the satellite. Each of these files contains several rows each with 11 columns described in the Table below. All five radiance angles are raw features, while NDAI, SD, and CORR are features that are computed based on subject matter knowledge. More information about the features is in the article **yu2008.pdf**. The sensor data is multi-angle and recorded in the red-band. For more information about MISR, see **http://www-misr.jpl.nasa.gov/**.

01	y coordinate
02	x coordinate
03	expert label (+1 = cloud, -1 = not cloud, 0 unlabeled)
04	NDAI
05	SD
06	CORR
07	Radiance angle DF
08	Radiance angle CF
09	Radiance angle BF
10	Radiance angle AF
11	Radiance angle AN

Table 1: Features in the cloud data.

1 Data Collection and Exploration (30 pts)

(a) **Write a half-page summary** of the paper, including at least the purpose of the study, the data, the collection method, its conclusions and potential impact.

(b) **Summarize** the data, i.e., % of pixels for the different classes. **Plot well-labeled beautiful maps** using x, y coordinates the expert labels with color of the region based on the expert labels. **Do you observe some trend/pattern? Is an i.i.d. assumption for the samples justified for this dataset?**

(c) **Perform a visual and quantitative EDA** of the dataset, e.g., summarizing (i) pairwise relationship between the features themselves and (ii) the relationship between the expert labels with the individual features. **Do you notice differences** between the

two classes (cloud, no cloud) based on the radiance or other features (CORR, NDAI, SD)?

2 Preparation (40 pts)

Now that we have done EDA with the data, we now prepare to train our model.

(a) (Data Split) **Split the entire data** (image1.txt, image2.txt, image3.txt) into three sets: training, validation and test. Think carefully about how to split the data. **Suggest at least two non-trivial different ways** of splitting the data which takes into account that the data is not i.i.d.

(b) (Baseline) **Report the accuracy of a trivial classifier** which sets all labels to -1 (cloud-free) on the validation set and on the test set. In what scenarios will such a classifier have high average accuracy? *Hint: Such a step provides a baseline to ensure that the classification problems at hand is not trivial.*

(c) (First order importance) Assuming the expert labels as the truth, and without using fancy classification methods, suggest three of the "best" features, **using quantitative and visual justification**. Define your "best" feature criteria clearly. Only the relevant plots are necessary. Be sure to give this careful consideration, as it relates to subsequent problems.

(d) Write a generic cross validation (CV) function **CVgeneric** in R that takes a generic classifier, training features, training labels, number of folds K and a loss function (at least classification accuracy should be there) as inputs and outputs the K-fold CV loss on the training set. Please remember to put it in your github folder in Section 5.

3 Modeling (40 pts)

We now try to fit different classification models and assess the fitted models using different criterion. For the next three parts, we expect you to try *logistic regression and at least three other methods*.

(a) **Try several classification methods and assess their fit using cross-validation (CV). Provide a commentary on the assumptions for the methods you tried and if they are satisfied in this case.** Since CV does not have a validation set, you can merge your training and validation set to fit your CV model. **Report** the accuracies across folds (and not just the average across folds) and the test accuracy. CV-results for both the ways of creating folds (as answered in part 2(a)) should be reported. Provide a brief commentary on the results. Make sure you honestly mention all the classification methods you have tried.

(b) **Use ROC curves to compare the different methods.** Choose a cutoff value and highlight it on the ROC curve. Explain your choice of the cutoff value.

(c) (Bonus) Assess the fit using other relevant metrics.

4 Diagnostics (50 pts)

Disclaimer: The questions in this section are open-ended. Be visual and quantitative! The gold standard arguments would be able to convince National Aeronautics and Space Administration (NASA) to use your classification method—in which case Bonus points will be awarded.

(a) Do an in-depth analysis of a good classification model of your choice by showing some diagnostic plots or information related to convergence or parameter estimation.

(b) For your best classification model(s), do you notice any patterns in the misclassification errors? Again, use quantitative and visual methods of analysis. Do you notice problems in particular regions, or in specific ranges of feature values?

(c) Based on parts 4(a) and 4(b), can you think of a better classifier? How well do you think your model will work on future data without expert labels?

(d) Do your results in parts 4(a) and 4(b) change as you modify the way of splitting the data?

(e) Write a paragraph for your conclusion.

5 Reproducibility (10 pts)

In addition to a writeup of the above results, please provide a one-line link to a public GitHub repository containing everything necessary to reproduce your writeup. Specifically, imagine that at some point an error is discovered in the three image files, and a future researcher wants to check whether your results hold up with the new, corrected image files. This researcher should be able to easily re-run all your code and produce all your figures and tables. This repository should contain:

(i) The pdf of the report,

(ii) the raw Latex, Rnw or Word used to generate your report,

(iii) your R code (with CVgeneric function in a separate R file),

(iv) a README file describing, in detail, how to reproduce your paper from scratch (assume researcher has access to the images).

You might want to take a look at the GitHub's tutorials https://guides.github.com/.

Final remarks

- Make sure to read the instructions for the submission on Page 1.
- Note that we will enforce a **zero tolerance policy for last minute / late requests (no emails please) this time.** Start early and plan ahead. If something is falling apart or not working, see us in office hours.