Lab 4: Cloud Data
Stat 215A, Fall 2021

Due: Friday November 12, 11:59 PM

The goal of this lab is the exploration and modeling of cloud detection in the polar regions based on radiances recorded automatically by the MISR sensor aboard the NASA satellite Terra. You will attempt to build a prediction model to distinguish cloud from non-cloud using the available signals. Your dataset has "expert labels" that you can use 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".

The data can be found in image_data.zip which contains three files: `image1.txt`, `image2.txt`, and `image3.txt`. Each of these files contains one "picture" from the satellite. Each of these files contains 11 columns described below. NDAI, SD and CORR are features based on subject matter knowledge. They are described in the article `yu2008.pdf` in the lab4 folder. The sensor data is multi-angle and recorded in the red-band. More information on MISR is available at `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

Please push a "lab4" folder to your repository by the deadline. Although this is a group project, **each member of the group must push a copy of the project to their own repository**. A suggestion is to make a separate (private) GitHub repository for collaboration on this project and then copy the final report and results to your individual repositories. At a minimum, your "lab4" folder should contain the files below:

- **lab4.Rmd** or lab4.Rnw: the raw report + code with your group members' names
- **lab4.pdf**: the output of lab4.Rnw/lab4.Rmd. This output should not contain any code.
- **R**/: a folder containing .R scripts (e.g. load.R and clean.R) that will be sourced in lab4.Rmd

You may create additional folders if you wish, but please keep your project directory organized. Push everything necessary to reproduce your report, but nothing else (e.g., do not push your **data**/ or **documents**/ folders). Note also that you do not have to create blinded files.

1 EDA

1. Plot the expert labels for the presence or absence of clouds, according to a map (i.e. use the X, Y coordinates).
2. Explore the relationships between the radiances of different angles, both visually and quantitatively. Do you notice differences between the two classes (cloud, no cloud) based on the radiances? Are there differences based on the features (CORR, NDAI, SD)?

2 Modeling

1. Some of the features might be better predictors of the presence of clouds than others. Assuming the expert labels are the truth, suggest three of the best features, using quantitative and visual justification. Be sure to give this careful consideration, as it relates to subsequent problems.
2. Develop several (i.e., at least three) 0-1 classifiers for the presence of clouds, using your best features, or others, as necessary. Provide a brief description of the classifier, state the assumptions of the classification models if any, and test if the model assumptions are reasonable.
3. Assess the fit for different classification models e.g. using cross-validation, AIC, and/or the ROC curve. Think carefully about how to choose folds in cross validation and/or your training/test splits.
4. Pick a good classifier. Show some diagnostic plots or information related to convergence, parameter estimation, and/or feature importance.
5. For your best classification model(s), perform some post-hoc EDA. 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?
6. How well do you think your model will work on future data without expert labels?

As usual, please carefully document your analysis pipeline, justify the choices you make, and place your work within the domain context.

3 Collaborative Work

Collaborative work is at the heart of applied statistics, an important part of which is fair credit for the work among the group members. Therefore we ask for the following:

1. Select a research lead for your group project (notify Bin and Omer by e-mail).
2. Outline a working plan for your group, building on the individual strengths of each group member (email the plan to Omer by **November 5th**).
3. Write explicitly the individual contributions in the final report. These contributions will be used to calculate an 'effort' term for your grade. Your individual grade would be:

$$0.3 \cdot \text{Effort} + 0.7 \cdot \text{Report} \tag{3.1}$$

The report part would be the same for all group members.

4 Reproducibility

I will use the scf cluster to test the reproducibility of your code, by cloning into your repository and attempting to compile your reports. For your convenience, I provide you with the script I will be using to test your code. If you wish, you can use the script to test your code in the following way:

1. Go into your lab4 directory, for example using my directory:

   ```
   cd /Users/omerronen/Documents/Phd/215a2021/stat-215-a-gsi/lab4
   ```

2. Run

   ```
   bash test.sh .
   ```

To perform testing on scf, you may need to transfer the data file there which can be done using *SCP* (see detailed instructions on this link).

Lastly if you wish to use any R packages that require installation, please ask for approval here - Link. I will try to install it on scf, and approve if possible.