Lab 4: Cloud Data
Stat 215A, Fall 2014

Due: Wednesday, November 12, 4:00 PM
Note – November 11th is a school holiday, so we will not have a regular lab section on November 11th. The lab is due online the following day. There is no need to turn in a printed copy of the lab.

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".

On bSpace you will find a zip archive with three files: `image1.txt`, `image2.txt`, and `image3.txt`. Each 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` available on bSpace. 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

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 quan- titatively. 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.

Only the relevant plots are necessary. Be sure to give this careful consideration, as it relates to subsequent problems.

2. Develop several 0-1 classifiers for the presence of clouds, using your best features, or others, as necessary. Be sure to 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 using cross-validation, AIC, and/or the ROC curve. Think carefully about how to choose folds in cross validation.

4. Pick a good classification model. Show some diagnostic plots or information related to convergence or parameter estimation.

5. 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?

6. How well do you think your model will work on future data without expert labels?

3 Reproducibility

1. In addition to a writeup of the above results, please provide a 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.

2. This repository should contain:

   (a) The raw LaTeX or knitr used to generate your report.

   (b) The R code used.

   (c) The final pdf.

   (d) A README file describing, in detail, how to reproduce your paper from scratch.

3. For the purpose of this lab, the repository does not need to contain the raw image files. You may assume that the researcher has access to the three `image?.txt` files in a local directory on his / her computer.