STA 521 Project 2 Cloud Data Report

Yicheng Shen (yicheng.shen@duke.edu) & Yunhong Bao (yunhong.bao@duke.edu)

24 October, 2022

1 Data Collection and Exploration

(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) pair- wise 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

(a) (Data Split) Split the entire data (imagem1.txt, imagem2.txt, imagem3.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 CVmaster 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

(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. Use quantitative measures and show clean and interpretable figures!

4 Diagnostics

(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

In addition to a writeup of the above results, please submit a zip file containing everything necessary to reproduce your writeup to Gradescope "PROJ2 code". 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 zip file should contain:

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

(ii) your R code (with CVmaster function in a separate R file),

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