STAT 154: Project 2 Cloud (Christopher Lau 26364374, Dat duc vu, 3031823757)

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.

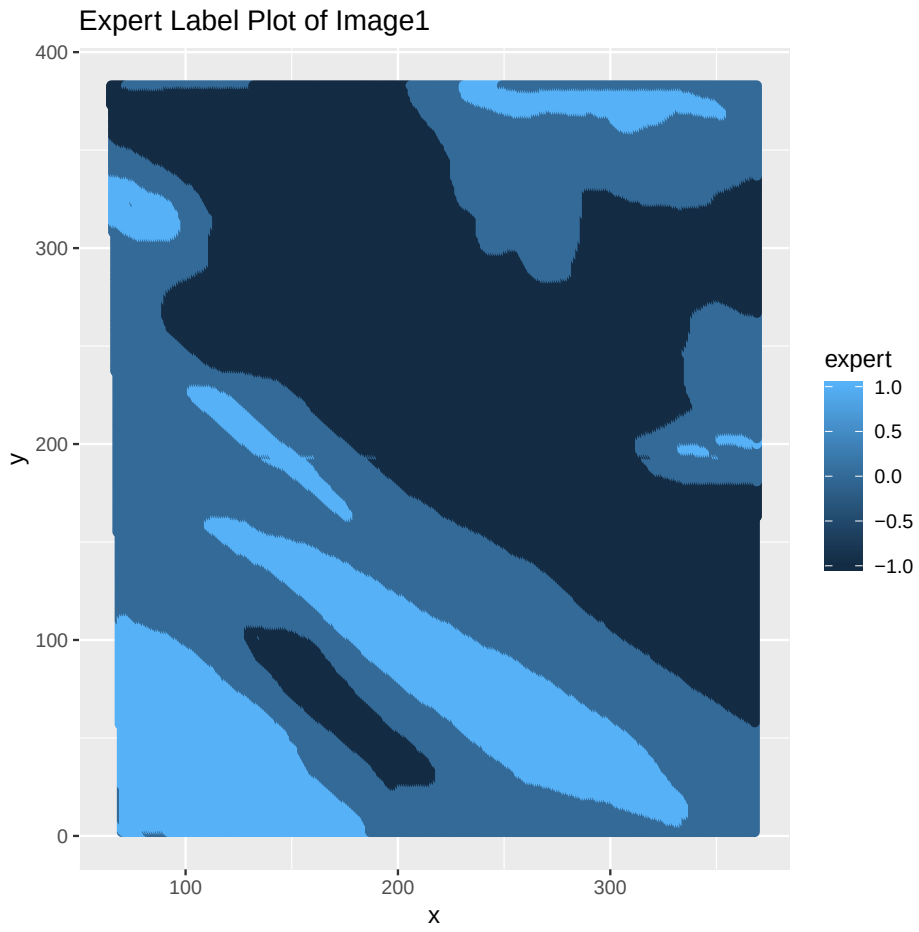
Global climate models are important to scientific and public interest, and a key component is the Earth's increasing amount of atmospheric carbon dioxide. The clouds in the Arctic play an important role in modeling sensitivity to increasing air temperatures, thus leads to the attempt to measure the different properties of clouds. Unfortunately, the properties of clouds can be difficult to isolate, because ice/snow surfaces scatter light in a similar way, leading to cloud detection issues. NASA's satellite launched the MISR, capable of measuring 360-km of Earth's surface in four spectral bands at nine different angles. This leads to an impossibly large dataset that is impossible to classify except

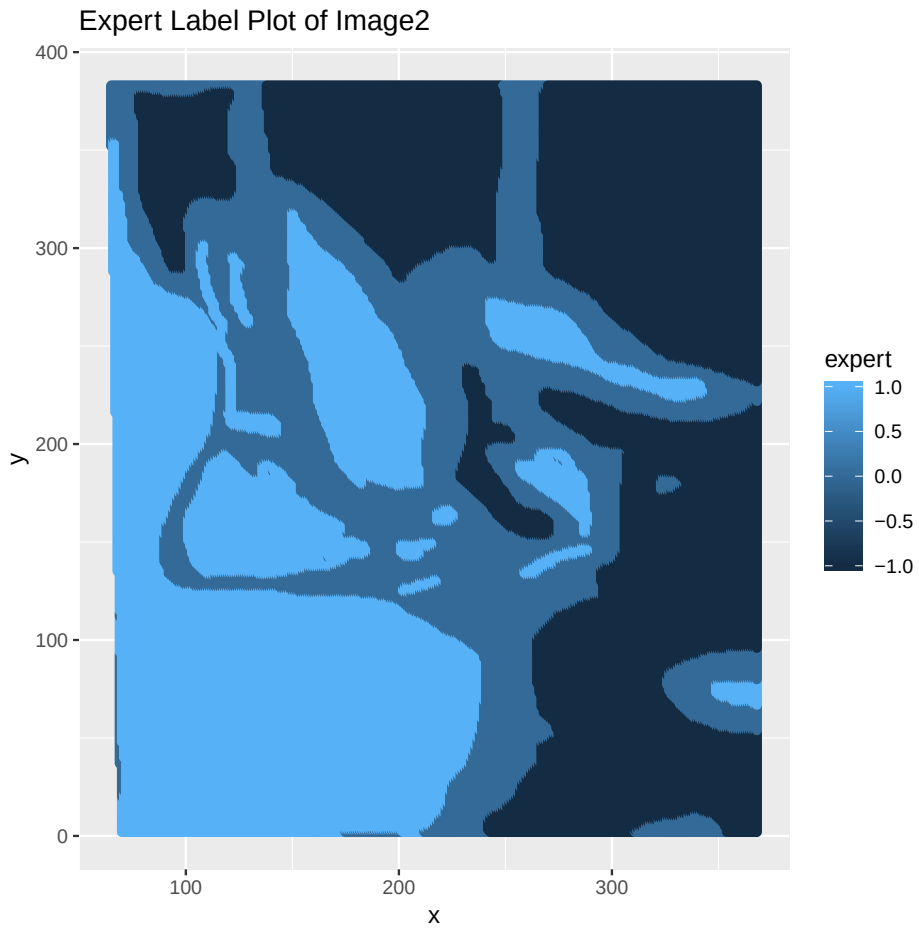
by some data analysis algorithm.

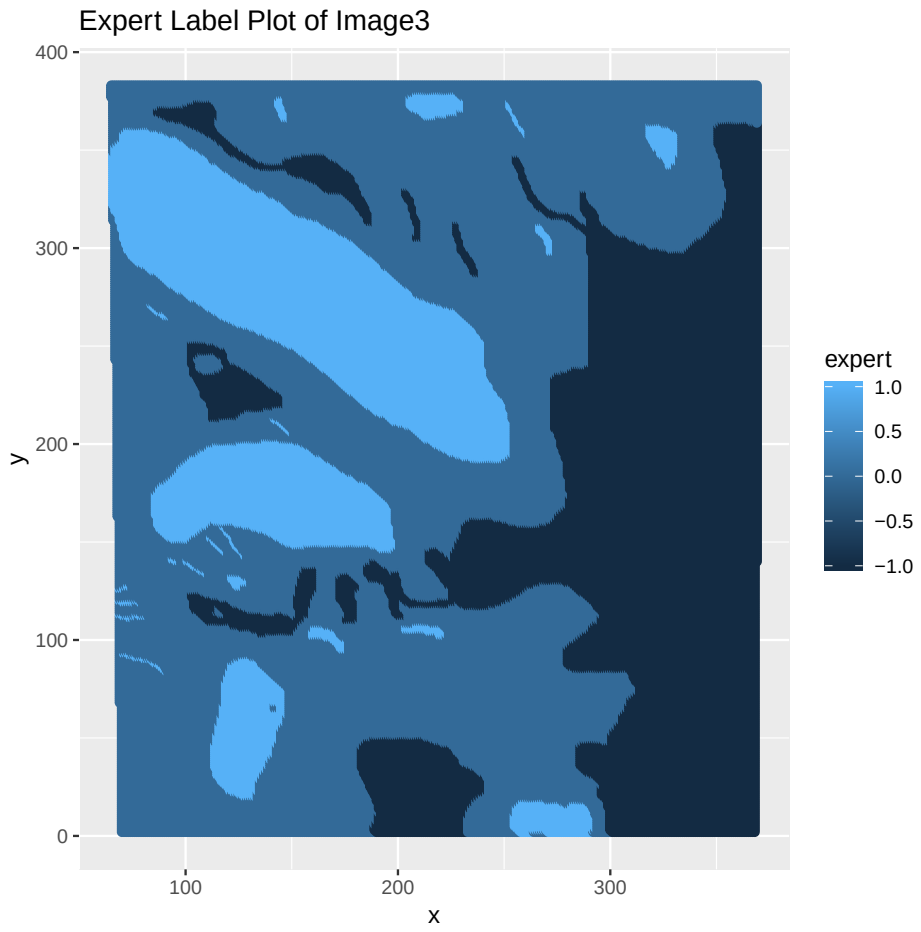
The goal is to build cloud detection algorithms that can process the MISR data set: searching for cloud-free conditions and using its correlations to find the reverse. The surface is modeled, and we find the (CORR) correlation of MISR images of the same scenery from different viewing directions, the standard deviation (SD) of MISR nadir camera pixel values across a scene, and a normalized difference angular index (NDAI) that characterizes the change in a scene with changes in the MISR view angle/direction. The data used in this study is collected from 10 MISR orbits of orbit path 26, which each contained six data units. The study concentrates on repeated visits so the experts could gain familiarity, improving the (EXPERT LABEL) process, which is the best method for producing validation data to compare against the model built by the algorithms and features.

We conclude that the three physical features - the linear correlation of MISR radiation measurements, standard deviation of MISR nadir red radiation measurements, and a normalized difference angular index, contain sufficient information to separate clouds from ice/snow surfaces. The ELCM algorithm based on the three features, which combines classification and clustering frameworks, is suitable for real-time MISR data processing and is more accurate than existing algorithms. Firstly, statisticians worked on the engineering of real-time processing, rather than simply performing ex post facto analysis on finished scientific work. Secondly, statistical thinking is important and has the ability to create solutions for modern scientific problems.

(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.


Expert Label Plot of Image1
400
300
200
100
0
y
100
200
300
x
expert
1.0
0.5
0.0
−0.5
−1.0


Expert Label Plot of Image2
400
300
200
100
0
y
100
200
300
x
expert
1.0
0.5
0.0
−0.5
−1.0



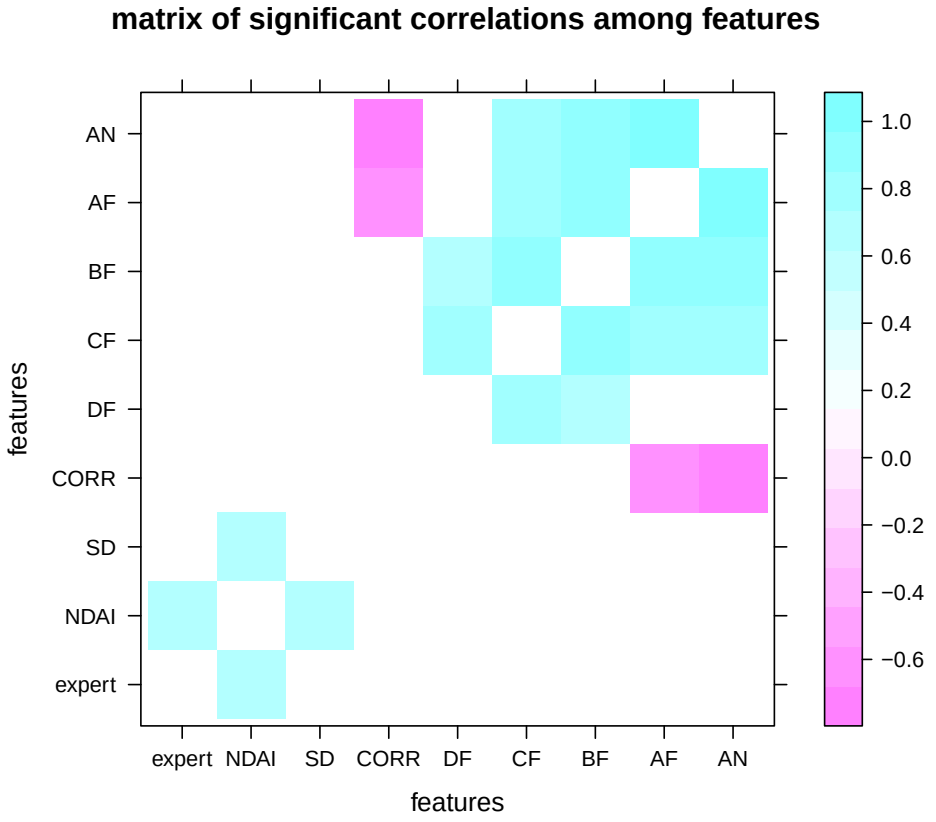
Do you observe some trend/pattern? Is an i.i.d. assumption for the samples justified for this dataset?

Cloud and cloud-free areas appear to have trends, or grouping. The appearance of each in a pixel, affect the possibility of cloud or cloud-free surrounding pixels, due to grouping or clustering. Therefore, the i.i.d. assumption is unjustifiable, as there are correlations and relationships within the data, based on the x, y coordinates.

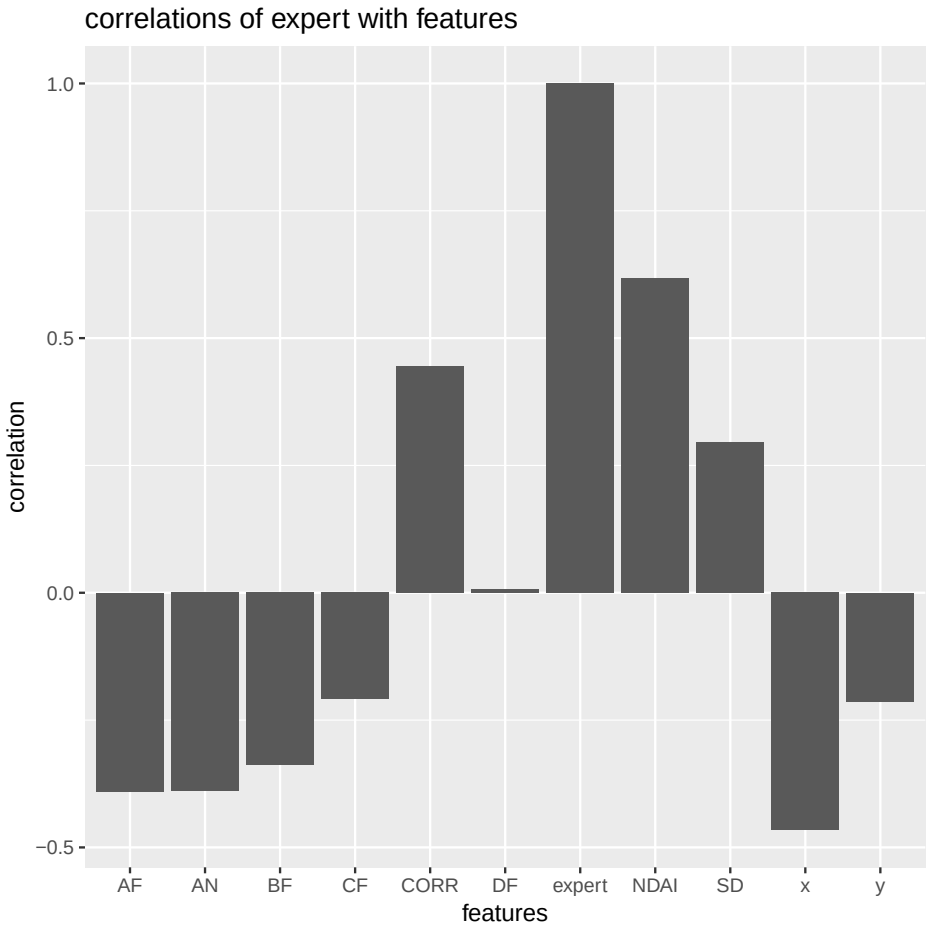
(c) **Perform a visual and quantitative EDA** of the dataset, e.g., summarizing (i) pairwise relationship between the features themselves and (ii) the relationship between the expert labels with the individual features.

	expert	NDAI	SD	CORR	DF	CF	BF	AF	AN
expert	NA	0.6169	NA	NA	NA	NA	NA	NA	NA
NDAI	0.6169	NA	0.6311	NA	NA	NA	NA	NA	NA
SD	NA	0.6311	NA	NA	NA	NA	NA	NA	NA
CORR	NA	NA	NA	NA	NA	NA	NA	-0.6039	-0.6820
DF	NA	NA	NA	NA	NA	0.8496	0.6991	NA	NA
CF	NA	NA	NA	NA	0.8496	NA	0.9119	0.8216	0.7727
BF	NA	NA	NA	NA	0.6991	0.9119	NA	0.9530	0.9043

```
AF          NA      NA      NA -0.6039      NA 0.8216 0.9530       NA  0.9706
AN          NA      NA      NA -0.6820      NA 0.7727 0.9043   0.9706      NA
```




```
     var1   var2  correlation
1  expert      y -0.213372321
2  expert      x -0.465822566
3  expert expert  1.000000000
4  expert   NDAI  0.616934624
5  expert     SD  0.295447745
6  expert   CORR  0.444059231
7  expert     DF  0.006550085
8  expert     CF -0.208279170
9  expert     BF -0.337948500
10 expert     AF -0.389741017
11 expert     AN -0.389358825
```




```
# A tibble: 3 x 11
  expert        y        x       NDAI        SD      CORR       DF       CF
   <dbl>    <dbl>    <dbl>      <dbl>     <dbl>     <dbl>    <dbl>    <dbl>
1     -1 217.1906 266.1333 -0.2627047  2.978540 0.1400588 271.2948 256.8658
2      0 193.7892 207.3163  1.8206888 11.713588 0.1831792 270.9059 244.8691
3      1 154.1517 161.1891  1.9495617  9.844765 0.2629963 272.2473 232.4672
# ... with 3 more variables: BF <dbl>, AF <dbl>, AN <dbl>
```

Do you notice differences between the two classes (cloud, no cloud) based on the radiance or other features (CORR, NDAI, SD)?
The correlation graph tells me that cloud/cloud-free is positively related to the calculated features (NDAI, SD, CORR), and negatively to radiance, and there are differencesin the means of the features based on cloud/cloud-free labeling.

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.

One way to split the data would be to use the expert label grouping, and ensure the approximate percentages of cloud, cloud-free, and ambiguous data from each split, is the same as the entire data set. You can do this by splitting every grouped value of expert label separetely. Another way to split the data is to directly use the expert label's percentages of cloud, cloud-free, and ambiguous data to sample into each split, but this is much less accurate than using the groups made by experts.

(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.*

```
[1] 0.3989857
```

The accuracy of such a classifier would have high average accuracy on data sets with little to no clouds in the validation and test sets. A good example of this would be if the data was split poorly, and very little cloud covered data was split into the validation and test sets.

(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.
Based on the graph plot I generated at the end of 1c, it tells me that the best correlations with the expert labeling are NDAI, and the 2 radiance angles AF, AN. Therefore, the three best features I would assume to be those three.

(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.

Method 1: Logic Regression Model

For Logistic Regression Model, the dependent variable should be in the range of 0 to 1 and that's why we rescaled the expert labels from -1,0,1 to 0,0.5,1 and we assume that it has the same interpretation.

```
glm.pred    -1     0     1
      -1 25156  6359   519
      0   1348 10452  6363
      1   1166  8181  9568
```

Logistic regression accuracies across 10 folds are:

```
 [1] 0.6533708 0.6522749 0.6544735 0.6538732 0.6524076 0.6539670 0.6529689
 [8] 0.6565072 0.6527733 0.6540420
```

Average logistic regression accuracy is:

```
[1] 0.6536658
```

Testing accuracy is:

```
[1] 0.6536636
```

Method 2: Linear Discriminant Analysis

```
lda.class    -1     0     1
       -1 24716  5290   208
       0   2130 15447  7155
       1    824  4255  9087
```

LDA accuracies across 10 folds are:

```
 [1] 0.7157304 0.7116037 0.7121182 0.7116387 0.7144453 0.7131551 0.7132062
 [8] 0.7128939 0.7131913 0.7132344
```

Average LDA regression accuracy is:

```
[1] 0.7131217
```

Testing accuracy is:

```
[1] 0.7126114
```

Method 3: QDA QDA accuracies across 10 folds are:

```
[1] 0.7144402 0.7178497 0.7157516 0.7085037 0.7126447 0.7126166 0.7115222
[8] 0.7141278 0.7149236 0.7129932
```

Average QDA regression accuracy is:

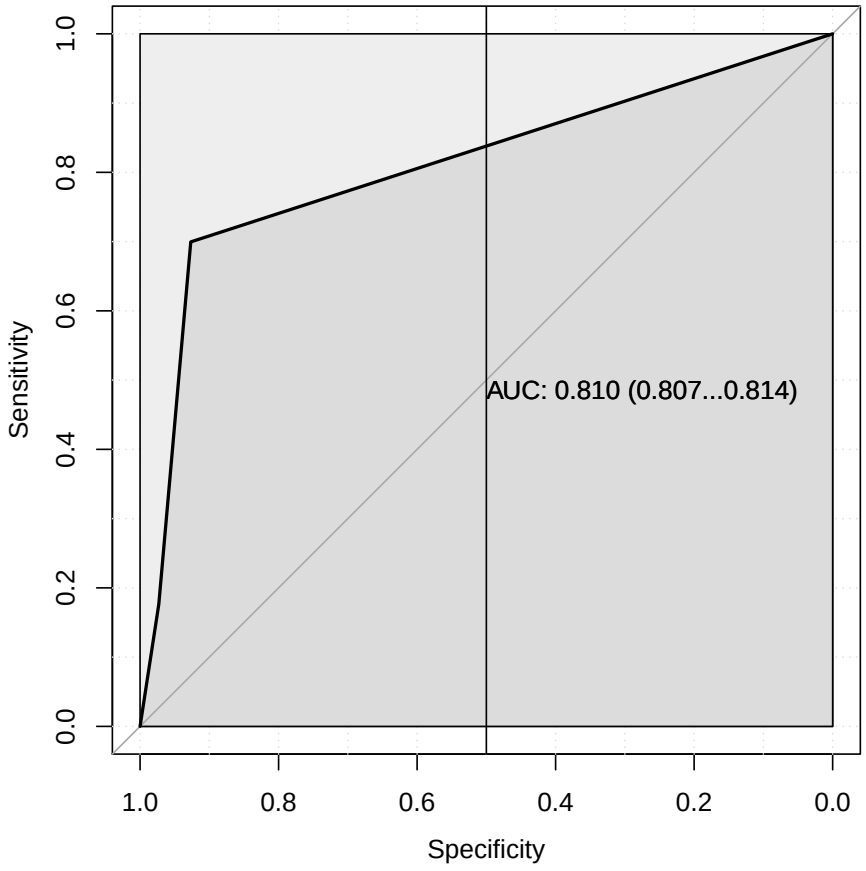
```
[1] 0.7135373
```

Testing accuracy is:

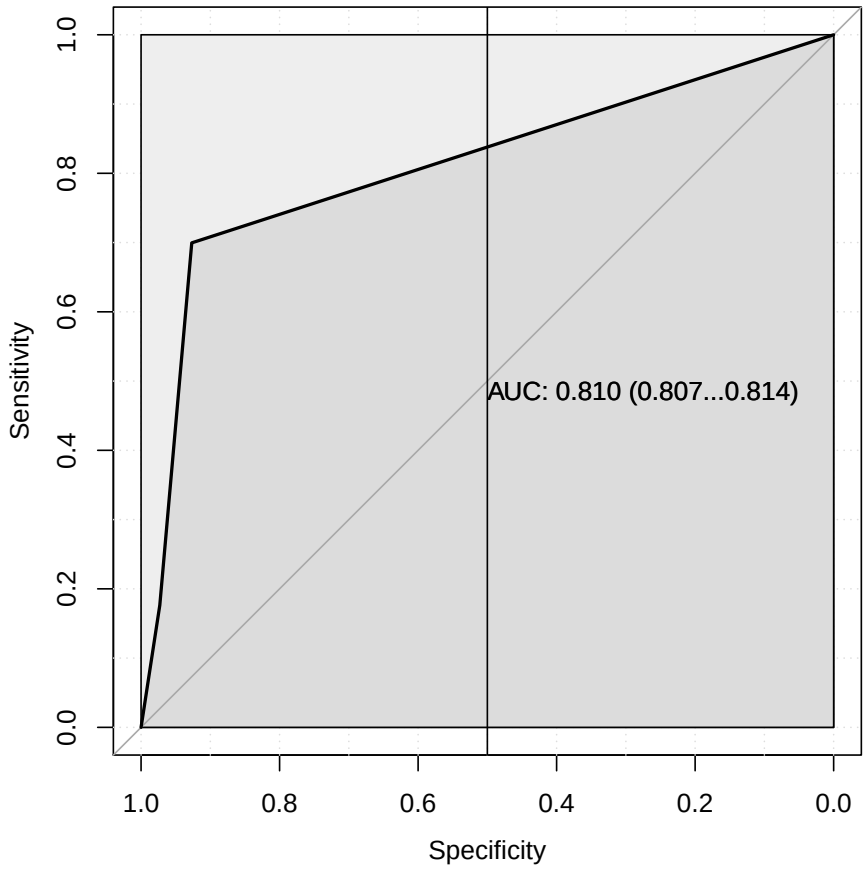
```
[1] 0.7113236
```

(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.

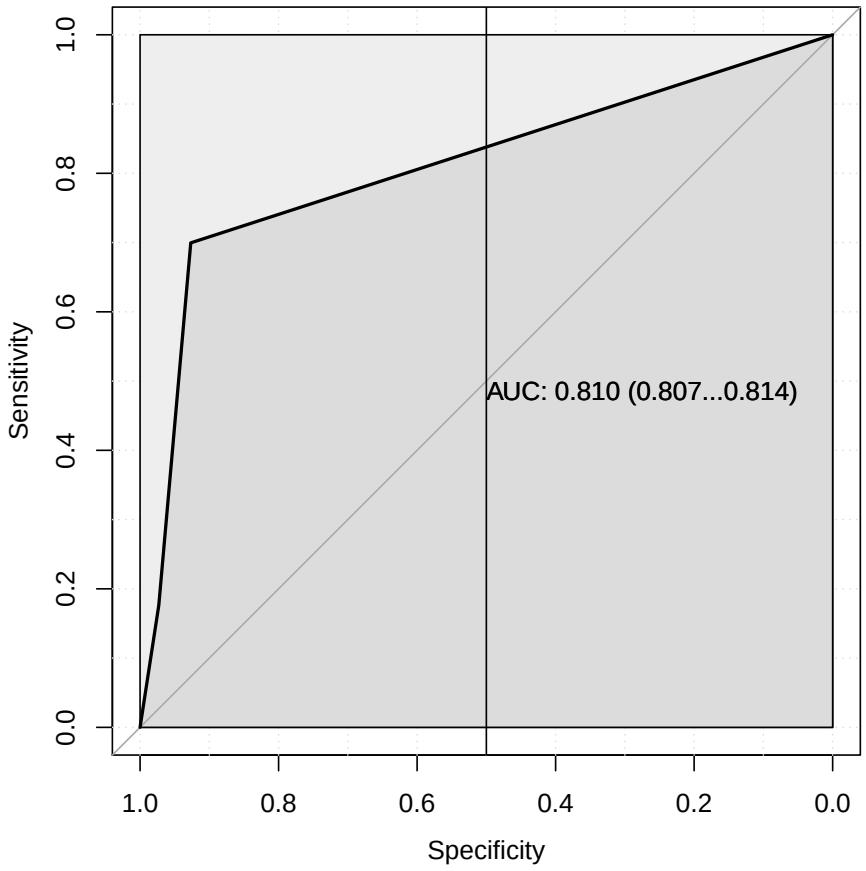
ROC Curve for Logistic Regression:



ROC Curve for LDA:



ROC Curve for QDA:



(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).

https://github.com/chrislaustopher/stat154proj2
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.