**ISEN 613-Fall 2019 Course Project**

**Objective**

Develop models to predict the object in an image.

**Data**

The dataset contains 50,000 color images coded in RGB (red, green, and blue). The images are only 32x32. Each row in the dataset corresponds to an image. There are 3x32x32 = 3072 columns (features) representing the color values in each column. The first 1024 columns contain the red values. The second 1024 columns contain the green values. The third 1024 columns contain the blue values. The image is stored in row-major order which means that the first 32 columns correspond to the first row in the image, the next 32 columns correspond to the second row, and so on. The entry value can range from 0 to 255.

There 10 classes corresponding to different objects that an image from the dataset can contain. They are coded by 0-9. Some images may contain multiple objects. There is an extract.R script that allows you to visualize the images.

**Remark:** The number of features (3072) is comparable to the total number of images (50000). You may use them all, try to extract some higher level features that are interpretable, or add more features. For example, by taking cross-product of the vectors you can create 3072x3072 more features and use them in training.

**Project Instructions**

Download the data folder. You can run the extract.R code to get the training data in list format. Each element of the list contains the RGB codes of an image in a data frame. extract.R includes a drawImage function to plot an image.

You are welcome to try as many models as possible ranging from basic models (e.g. logistic regression, LDA, QDA, linear SVM) to more complex models (e.g. considering transformation of inputs). After trying different models, pick your best three.

For each of these models, report your technical analysis, the training error, and an approximation of the test error using a technique you see fit (e.g. cross-validation) in **2-pages including the figures and tables**. In total, you will have **6-pages** of technical analysis for all the three models you worked on. After that, pick your best model and explain why it is your best candidate in **1-page**.By Dec 2nd 5:00pm, you need to submit the technical analysis of your three models (total **7-pages**), as well as the code for your best model. Then, we provide the test data and **you run the exact same code you submitted to us** on the test data to calculate your test error. Briefly report that result (you will not need more than **1 page**). If you want to modify your code to achieve a better test error, you have the option to do so, in which case you need to report the changes as well as the improved rate in **2 pages**. By Dec 6th, 5:00pm, you need to add your test results, your optional modification, and your executive summary to the previous report, and submit the final report on e-campus. If you have a modification upon your best model, you should submit the code for that as well.

The final version of your report should have the following structure:

1. Executive summary (at most **1 page**, due Dec6th) page 1
2. Technical report of model 1 (at most **2 pages**, due Dec 2nd ) page 2-3
3. Technical report of model 2 (at most **2 pages**, due Dec 2nd) page 4-5
4. Technical report of model 3 (at most **2 pages**, due Dec 2nd) page 6-7
5. Comparison of the 3 models and specifying the best model you want to use for testing (at most **1 page**, due Dec 2nd) page 8
6. Evaluating test points only on your **best** model and reporting the test error (at most **1 page**, due Dec6th) page 9
7. (Optional) Steps to improve your best model **or** introducing a better model (at most **2** **pages**, due Dec6th) page 10-11

**Evaluation of the project:** The project has 20 points + 1 bonus:

* Executive summary (3points)
* Technical analysis of model 1-3 (12 points)
* Reasoning for choosing the best model (1 point)
* Evaluating your test result (2 points)
* Competition: the group with lowest test error will receive (2 points). Other groups will either receive 1 point or 0 point.
* (Bonus) the improvement upon your best method (1 point)

**Rules:**

1- The results **must be reproducible.** If we run your best model on test points and get a different test error from what is reported, you will not get points from the competition and lose the points from the technical analysis of the model. The same rule applies to modified code (if submitted).

2- Your code should be a .R file with clear comments on how to run.

3- The report should be concise and to-the-point. The font size should be at least 11, and exceeding the suggested page limits will result in point deduction.

4- Late submissions are penalized at the rate of 0.5 point per hour.

5- Make sure to write a concise executive summary. Remember from HW 2 that this is a **non-technical** summary of your analysis for a broad audience. Do not include any graphs or statistical concepts in the executive summary.

6- In the last paragraph of the executive summary, the contribution of each individual in the project should be clarified. Evidently, if a group member does not contribute to the project, s/he will not receive any points from the project.