STAT 639 Project

Group Sign-UP Due: 2/9 11:59pm Central Time
Project Submission Due: 4/18 11:59pm Central Time

The project consists of three parts: (1) supervised learning task, (2) unsupervised learning task, and (3) presentation. You are allowed to work in groups up to 3 people from the same section (600 or 700). You need to write a **report as one pdf** summarizing your work. It should show clearly what you have done and how you obtain the results. The report should be no more than 4 pages (including title, figures, tables, and references) on 8 1/2 x 11-inch paper. Please use at least 12 point font size, double-spacing, and 1 inch margin all sides. State in your report clearly the name and UIN of each group member. You also need to submit **two R code** files. One R code exactly reproduces all results (including both supervised and unsupervised learning tasks) in your report (you may want to set the RNG seed using `set.seed()` function). The second R code exactly reproduces the best result in the classification problem. In addition, for the classification problem, you need to submit an **RData** file; details see below. Turn in one zip file containing the pdf, R, RData files, and presentation slides by the due date listed above. Each group should turn in one zip file only. All files should be named using the group number, e.g. 37.zip, 37_all.R and 37_best.R if your group number is 37.

1 Supervised Learning [40 pts]

We consider a classification problem. You will find `class_data.RData` in the module `Project` on Canvas. It has three variables:

- `x`: training data covariates
- `y`: training data response
- `xnew`: testing data covariates

Train various classifiers to this dataset. You can also use classifiers that are not covered in this course (as long as you provide description of the method and readable R code). Summarize all the findings including what classifiers you tried, how the paramters were tuned and how the testing error was estiamted etc. in your report. Out of all the classifiers you have tried, report in detail the best classifier such as a desciption, the best set of tuning parameters, the estimated testing classification error, etc. In addition, save that best prediction result and estimated testing error using the command (replace `###` by your group number)

```
save(ynew,test_error,file="###.RData")
```

where `ynew` stores the (best) predicted lables and `test_error` stores the estimated testing error. **Please strictly follow the rules for the RData file; 5 points will be deducted for any violations**. I will use `ynew` to compute the actual testing classification error. Obviously, smaller testing error receives better score. But your score will also depend on how close your estimated testing error to the actual testing error. The group with the smallest testing error receives 20 bonus points on this project.

2 Unsupervised Learning [40 pts]

You will find another dataset `cluster_data.RData` in the module `Project`. It has 1000 observations and 784 variables. The task for you is to cluster the observations into an unknown number K of groups. We did not discuss the issue of choosing K in class. You would need to search the literature and find a meaningful way to determine K. In your report, describe the method(s) you have tried for clustering and choosing K, and state how many clusters you find. Again, you are not restricted to using clustering methods covered in this course. You do not need to submit an RData file for this task.

3 Group Presentation [20 pts]

Each group will have TBA min to present your project live (Section 600) or via recorded videos (Section 700). The order of the live presentations will be assigned and announced later in the semester. Both supervised and unsupervised tasks have to be covered. Your presentation will be judged based on clarity and timeliness. Section 700 students must submit recorded videos by the last day of presentation to the class google drive folder "DATA MINING & ANALYSIS 2023 SPRING/Project_Recorded_Videos/". Like other submitted files, please use your group number for the video file name.