Logistic Regression and Monte Carlo Simulation

Fall 2022, MATH8050: Homework 5
Your Name, Section XXX

Due October 12, 12:00 PM

General instructions for homeworks: Please follow the uploading file instructions according to the syllabus. Each answer must be supported by written statements as well as any code used. Your code must be completely reproducible and must compile. For writing mathematical expressions in R Markdown, refer to the homework template posted on Canvas, a 30-minute tutorial, or LaTeX/Mathematics.

Advice: Start early on the homeworks and it is advised that you not wait until the last day. While the professor and the TA's check emails, they will be answered in the order they are received and last minute help will not be given.

No late homeworks will be accepted.

R Working Environment

Please load all the packages used in the following R chunk before the function `sessionInfo()`

```
# load packages


sessionInfo()
```

Total points on assignment: 10 (reproducibility) + 45 (Q1) + 45 (Q2)

Reproducibility component: 10 points.

1. (45pts total) Load the `Smarket` data from the `ISLR2` R package. This dataset consists of percentage returns for the S&P 500 stock index over 1,250 days, from beginning 2001 until the end of 2005. For each date, we have recorded the percentage returns for each of the five previous trading days, `Lag1` through `Lag5`. We have also recorded `Volume` (the number of shares traded on the previous day, in billions), `Today` (the percentage return on the date in question), and `Direction` (whether the market was `Up` or `Down` on this date). The goal is to predict `Direction` (a qualitative response) using other predictors.

a. (10pts) Write down a logistic regression model with all the assumptions and fit the model using the R function `glm()`. List all the parameter estimates and what did you find based on the p-values? What does that mean in a practical sense?

b. (20pts) The MLE for logistic regression can only be obtained numerically. Write your own function called `my.glm()` to implement the Newton-Raphson algorithm to the get the MLE of the logistic regression model in Part (a), and verify your numerical solutions with the those obtained in Part (a) via `glm()`.

c. (15pts) Divide the data into testing set by randomly sampling 15% of the data and treat the resting datasets as the training set. Write your own prediction function called `my.predict()` to get the prediction. Report the false positive and false negative mis-classification errors. For random sampling, use `sammple()` in R. To ensure that your numerical results are reproducible, you need to set the random seed (e.g., `set.seed(12345)`) before using `sample()`.

2. (45pts, equally weighted) Estimate

$$I = \int_{-\infty}^{\infty} \exp(-x^4)\, dx$$

a. Evaluate I using substitution and the gamma function. This is not an approximation given that we can actually compute the integral exactly. We will use this exact expression when comparing to Monte Carlo and Importance sampling in the other tasks below.

b. Estimate I using Monte Carlo. How does this compare to the exact value obtain in a?

c. Estimate I using importance sampling. How does this compare to the exact value obtain in a?