proposal

March 15, 2025

1 Bayesian Regression & MLE for Student Performance Analysis

Team member: Hung Dinh - 19774520

Theme: Bayesian vs Frequentist Approach on Regression - Student Performance

Github Repo: https://github.com/hd54/stat447c

All commits are done by me and me only.

Introduction:

Performance have always been a concern for a lot of students regardless of education level, whether it be high school, university, or college. Good performance can mean greater opportunities for higher education, awards, and even jobs, so students want to be successful in their courses. However, there's always a disparity in students' performance, which can be seen in grade distributions of exams, homework, etc. It's possible that students' background or how they treat the class affect their performance. This project seeks to see how different factors contribute to students' performance (in particular, final exam score or course grade letter). The main goal would to be able to predict performance based on most influential predictors.

Dataset:

Student performance may vary throughout the years due to societal changes, introduction of new technologies (such as ChatGPT), etc. This leads me to choose some of the more recent datasets as possible candidates:

https://www.kaggle.com/datasets/lainguyn123/student-performance-factors

https://www.kaggle.com/datasets/joebeachcapital/students-performance

[1]:
```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse
2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- Conflicts ------------------------------------------
tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
```

```
x dplyr::lag()    masks stats::lag()
i Use the conflicted package
(<http://conflicted.r-lib.org/>) to force all conflicts to
become errors
```

[2]:
```
dataset_1 <- read.csv("StudentPerformanceFactors.csv")
head(dataset_1)
```

A data.frame: 6 x 20

	Hours_Studied <int>	Attendance <int>	Parental_Involvement <chr>	Access_to_Resources <chr>	Extrac <chr>
1	23	84	Low	High	No
2	19	64	Low	Medium	No
3	24	98	Medium	Medium	Yes
4	29	89	Low	Medium	Yes
5	19	92	Medium	Medium	Yes
6	19	88	Medium	Medium	Yes

[3]:
```
dataset_2 <- read.csv("StudentsPerformance_with_headers.csv")
head(dataset_2)
```

A data.frame: 6 x 33

	STUDENT.ID <chr>	Student.Age <int>	Sex <int>	Graduated.high.school.type <int>	Scholarship.type <int>
1	STUDENT1	2	2	3	3
2	STUDENT2	2	2	3	3
3	STUDENT3	2	2	2	3
4	STUDENT4	1	1	1	3
5	STUDENT5	2	2	1	3
6	STUDENT6	2	2	2	3

Approaches:

Overall, the dataset contains a lot of discrete variables. For example, in the 2nd dataset, student age tends to be mostly 2, but this describes the range of 22-25 instead. It's possible to use some categorical variables to capture this or some type of encoding.

I can start simple with a comparative study of different types of regression using frequentist and Bayesian approach. I can do a comparative study on traditional (without penalty) and regularized regression for each approach. For frequentist approach, I can start by optimizing the number of variables used for regression through removing collinearity (i.e. we can use forward selection along with VIF analysis). I would expect to see regularized model to perform at least as well on predicting data due to the extra penalty added.

For Bayesian approach, we can use hierarchical model. It happens that we can treat the regularizer term as a part of the prior, since it controls how much information we can learn from data, whereas frequentist approach will include an extra term as penalty value. I can also experiment with using different priors for discrete and continuous variables, whereas dummy variables can be used for frequentist paradigm as mentioned above.

References show that I can use a Laplace prior in response to frequentist LASSO regression, and a Normal prior for ridge regression. This is good news - I would expect the performance between frequentist and Bayesian approach to be somewhat similar, and I don't have to come up with a

prior distribution. The only thing left is to find a good variance for the chosen distribution to simulate the regularizer term. Note that the implementation for both methods are very similar with the difference being in the chosen prior for the regularizer term.

In short, the project will be carried out in 4 different tasks:

- Perform EDA: optimize the number of variables used for modeling, some visualisations, etc.
- Perform frequentist approach with Lasso/ridge regression
- Perform Bayesian approach using Laplace/Normal priors
- Finally, compare the result between each approach (should produce similar result)

References: https://haines-lab.com/post/on-the-equivalency-between-the-lasso-ridge-regression-and-specific-bayesian-priors/