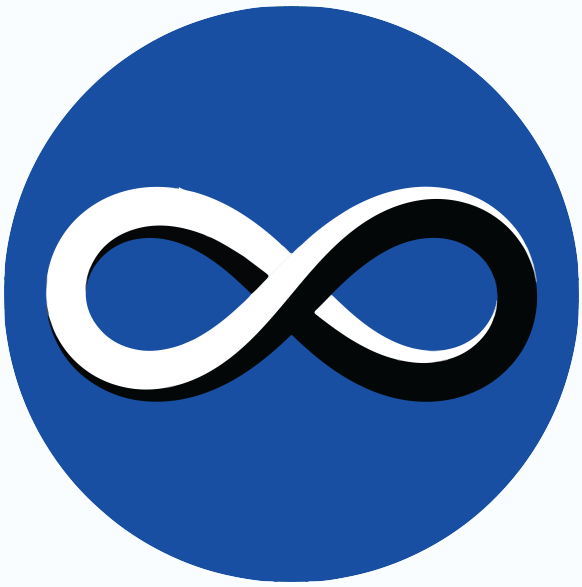
Project Member Application

SUBLIME

2025-26



Mathematics Club

Centre for Innovation
IIT Madras

Instructions

General Instructions

- Mention the following details at the start of your application.

Name:	Insti Nickname:
Roll No:	CGPA:
Phone (WA):	Email:

- Join the Aspiring Team Members WhatsApp group for further updates: Whatsapp Group
- The recommended font is a standard font size 11-13.
- The applications have to be submitted in PDF format, named as:
  `<First_name>_<Roll_Number>_Mathematics_Club_PM_SUBLIME.pdf`

  For example, `Aditi_ME23B248_Mathematics_Club_PM_SUBLIME.pdf`.
- You can upload the finished applications in this Google Form
- You may submit the completed application on or before **11:59 PM, 30/05/2025.**

Note:

- Do look up the resources mentioned at the end.
- It is fine even if you don't answer all the questions.
- Focus on the compulsory questions before attempting the bonus questions.
- If you have any queries, you can reach out to the Project Leads anytime you want:
  - Aditi Vaidya (ME23B248): 9545149055
  - Deenabandhan N (EE23B021): 7904387884

Contents

HR

0. Tell us a bit about yourself. Assume you are a master of flexing and proceed.
1. Why do you want to join this project? What skills or qualities do you possess that make you suitable for the work involved?
2. Mention all PoRs/activities you are planning to take this year. Weekly, how much time do you think you will be able to commit to this project? How much time will you commit to other PoRs and academics?
3. What would keep you motivated throughout the period of two semesters? How would you ensure that you are consistent in contributing to the project throughout the tenure?

§1 Statistical Learning: Making Sense of Your Messy Data

Statistical learning refers to a set of tools for modeling and understanding complex datasets. It helps us make sense of known data and derive insights from it which can be used to predict some insights for unknown data. Let us represent an input dataset (predictors) that has p features (variables) and $n$observations as a $n \times p$ matrix and output to this input (response) as a $n \times 1$ column vector.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

We assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$, which can be referred as $Y = f(X) + \varepsilon$ where f is a fixed unknown function and ε is a random error term with average as 0 and variance as σ^2.

In essence, Statistical learning refers to a set of approaches for estimating the function f. In this project **SUBLIME**, we will try to explore few of these approaches which will work when the number of observation (n) and number of variables (p) are very high.

§2 Tug of War: Bias vs. Variance

Estimation of function f can be done by two methods , parametric and non-parametric.

- Parmetric: This method involves assumption of some initial function and finding the parameters. We can infer that this approach is less flexible i.e., more restrictive. Eg: Linear regression where we assume the function follows a linear trend over all features.
- Non-parametric: This method does not explicitly assume any functional format but tries to estimate f that gets as close to the data points as possible. This method is more flexible as there is no assumption that is made while taking the function. Eg: Spline estimation.

To quantify the extent to which predicted response is close to the original one, we use Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i) \right)^2$$

Our goal is to estimate function $\hat{f}$ with the training data for which MSE is less with one of the above methods. As we have seen, the Non-parametric methods are more flexible which will result in less training MSE. But in reality we mostly use parametric methods for datasets with large features. This is because MSE of training data is an underestimate of MSE of test data.

Questions

1. Can you explain the reason for the above sentence with the terms bias and variance?

2. *Bonus:* Simulate linear regression and spline regression across a dataset and show why parametric (less flexible) methods are better.

§3 When in Doubt, Fit a Line

Let us assume that our model is linear with all features i.e,

$$Y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

here $\beta_0, \beta_1, \cdots, \beta_p$ are parameters that determine the function f. To estimate these parameters, we try to minimize a parameter called Residual Sum of Squares (RSS) and this method is called 'least squares approach' .

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 = (\mathbf{y} - X\boldsymbol{\beta})^\top (\mathbf{y} - X\boldsymbol{\beta})$$

On finding the minima of RSS for all range of β_i values we get optimal values as

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{y}$$

In most of the cases, few features won't have much significance in the response. So the β value for such features should ideally be zero. But due to the irreducible error ε and least square optimization, it would be a non-zero value. So to predict if a feature has significance in the response, we compute t-statistics on the parameter.

Questions

3. Explain what t-statistic is and take this dataset containing 5 features and tabulate each feature's t-value.

§4 How Close Did We Get, Really?

§4.1 T-Test: The Truth Teller or Just a Show

The t-statistic is useful for assessing the significance of individual features in a dataset. However, when multiple features are insignificant simultaneously, relying solely on individual t-tests can increase the risk of error. In this case, we will use F-statistic to predict if there is more than one feature which is insignificant in the response.

Questions

4. Explain the reason why we can't rely solely on t-statistics to assess the significance of multiple insignificant features, and how does the F-statistic address this limitation?

§4.2 The Art of Counting Errors

After training a model on a dataset, we often assess its accuracy using measures such as Residual Standard Error (RSE) and Correlation $(Y,\hat{Y})$ which is also referred as R^2.

Questions

5. What does each of these terms represent, and how do they estimate the model's error? What do the extreme values of the R^2 term indicate?
6. *Bonus:* Give a pictorial description for the extreme values of R^2

§5 Subsetting the Data

While estimating the parameters for linear regression we used least squares approach. Although this method works perfectly for cases where $n >> p$, it will have a high variance in the case of $n \approx p$ or $p > n$. So to overcome this issue, we select a subset from the set of variables and use them to train our data. If we brute force all possible combinations, we will have 2^p models to estimate which is computationally expensive.

Algorithm 5.1 Subset Selection

1. Start with a null model, M_0 : $(\beta_1 = \beta_2 = \cdots = \beta_p = 0)$.
2. For $k = 0, \ldots, p-1$:
   a) Consider all $p-k$ models that augment the predictors in M_k with one additional predictor.
   b) Choose the best among these $p-k$ models, and call it M_{k+1}.
   Here, **best** is defined as having the smallest Residual Sum of Squares (RSS) or the highest R^2.

The above algorithm reduces the number of comparisons from 2^p to just $p+1$. Now we need to choose the best from these $p+1$ models. We cannot use RSS or R^2 as measures because the number of features are different for these $p+1$ models.

Questions

7. What are the various metrics used to evaluate the performance of these $p+1$ models, and how do they differ from one another?

§6 Thinking Outside the (Least) Square

We have used least squares approach to estimate the parameters however, it has two major disadvantages.

- It does not work when number of features is comparable to that of observations.
- It does not shrink any parameter exactly to zero even if the feature does not have any significance in the response.

There are two other methods that we use to overcome this issue, Ridge regression and the Lasso. In Ridge regression we try to minimize ,

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2$$

and while using Lasso , we try to minimise ,

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|$$

where λ is a tunable parameter.

Questions

8. What are the key differences between these two methods, how do they improve upon the least squares approach?
9. What role does the parameter λ play in the optimisation process?

§7 The Handyman and His Tools

Now that you have gained a conceptual understanding of statistics, lets familiarize ourselves with the tools required to execute this efficiently. Every sub-part of this section will introduce you stages in the pipeline of the project.

§7.1 Getting started with R

To begin with, we must install R. Depending on the OS in use you can look this up, it will provide you with the step by step instructions required to install R.

- You can check if R has been successfully installed by running `help.start()` from the terminal. This should lead you to a web-page consisting of detailed documentation of the language, feel free to play around with it!
- Find a simple data set of your choice. Copy the data in a spread sheet in google sheets (remember to give access to the mathematics club account!)
- Load this data in R, write a simple script to perform linear regression on the data. Keep all that you have learnt from the previous section in mind whilst doing so!

Questions

10. Can you justify the choice of R for a project based in statistics? Mention the advantages R brings in terms of functions and libraries that we can explore as a part of this project. If you're not convinced, recommend alternatives.

11. Perform a simple linear regression on a data set of your choice (in R). Make sure to give us access to the script and data set.

§7.2 Need For Speed

At its core, we are looking at an optimisation problem. The mathematical aspect of optimisation is looked into in the previous section, here we will look into code optimisation. That is to write code that will run fast and efficiently. Let's switch gears to C++ for a while. Take a look at this function:

```cpp
#include <vector>

std::vector<std::vector<int>> multiply(const std::vector<std::vector<int>>& A,
    const std::vector<std::vector<int>>& B) {
    int r1 = A.size();
    int c1 = A[0].size();
    int c2 = B[0].size();
    std::vector<std::vector<int>> result(r1, std::vector<int>(c2, 0));

    for (int i = 0; i < r1; ++i) {
        for (int j = 0; j < c2; ++j) {
            for (int k = 0; k < c1; ++k) {
                result[i][j] += A[i][k] * B[k][j];
            }
        }
    }
    return result;
}
```

- While the logic for this looks intact, the cost of such computation is quite high.
- A function such as matrix multiplication will be called multiple times, large-scale simulations require millions of matrix multiplications. Even a 1-second delay per multiplication adds up to days or weeks of total runtime.
- This task involves optimising the process as much as possible. Here are a few hints: It is useful to think about two things: namely, how loops are structured and how memory is accessed. (Some techniques to look up: loop unrolling, cache friendly code and loop blocking)
- The idea behind parallel computation: breaking down a large task into smaller sub tasks that can be implemented simultaneously.

- The aim of this task is to conceptually understand the pipeline of code, and introduce you to the possibility of writing code that runs efficiently rather than just doing the bare minimum job. This exercise should be an exploration of how code really works. It is completely alright if you face difficulty in execution of the same.

Questions

12. Find ways in which the above code can be optimised. In addition to parallel computing write about your understanding for any two techniques (you are always free to explore more than required). How much computation time is saved with each optimisation method?
13. Implement these three techniques (in C++) and analyse their performance (of all three combined as well as each technique individually).

§7.3 Go, get the data first

Being a project based on statistics, there is one more area of interest and that is data loading. Here's how to go about it. You can find a dataset of your choice, do note that it should have a minimum of 10^6 data points. Remember to quote the source of the data set in your application.

Questions

Steps for data loading

14. Open the contents of the file using `ifstream`. Print the contents of line number 2025 and line 3106.
15. Split each line into the required fields.
16. Find an appropriate data structure for storing these values.
17. *Bonus:* Parallelise this process if possible!

§7.4 The Simulation Situation

Now that you have learn how to load the dataset, lets work on a simulation in C++.

Questions

18. Use the dataset provided here to simulate linear regression with the Lasso method in C++.
19. You are encouraged to use the Eigen library for efficient matrix operations and linear algebra support.

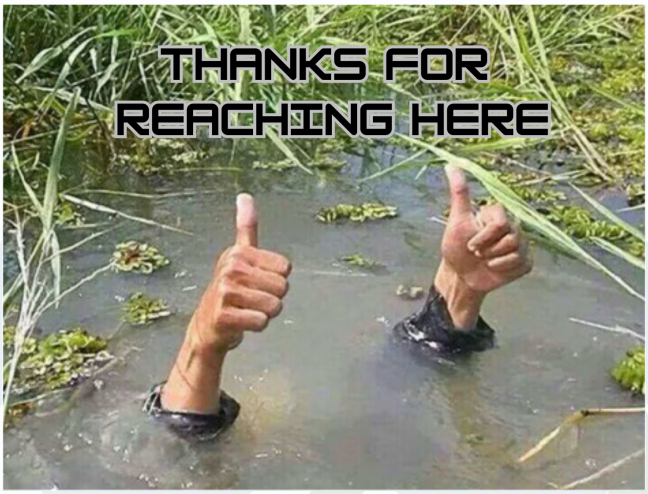
§7.5 It's a Wrap!

Time has come to link R and C++, we are using C++ for speed, but building the package in R. The next obvious step is to make the functions written in C++ callable in R. This is achieved through something called **wrappers**. Packages such as `Rcpp` allow you to do the same.

Questions

End of beginning:

20. Call the matrix multiplication function that you have optimised in C++ in R.



Resources

1. Statistical Learning with R - Stanford School of Humanities and Sciences
2. The Elements of Statistical Learning - Trevor Hastie, Robert Tibshirani, Jerome Friedman
3. An Introduction to Statistical Learning - Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani
4. Loop Unrolling, Cache friendly code and loop blocking - Description of a few optimisation methods
5. Wrapper for C++ in R - Help with `Rcpp`