Homework #03

데이터사이언스를 위한 컴퓨팅 1 (2022년도 2학기, M3239.005500)

Due: 2022년 11월 1일 (화) 23시 59분

1 Read file, calculate correlation and run simple linear regression [100pts]

Implement two classes: `DataFrame` and `Matrix`. `DataFrame` reads a `csv` file and can generate a `Matrix` object using specific columns. `Matrix` class stores a vector ($n \times 1$ matrix) or a matrix data and provide interface of matrix operations.

Instruction:

1. `DataFrame` class should have the following member functions:
   a. `int ReadData(std::string FileName, char sep, char comment, bool IsHeader)`

      `sep` is a separator of the file. Lines start with comment should be skipped. If `IsHeader` is `true`, the first line (after skipping) should be the header.
   b. `Matrix GetMatrix(int index[], int nColumn)`

      Generate a `Matrix` using the columns in `DataFrame`. `index[]` is the array of the column index. Note that `0` means the first column.
2. `Matrix` class should have the following member functions and operators:
   a. `+`, `-` : matrix addition and subtraction
   b. `*` : matrix multiplication
   c. `Matrix Transpose()` : return transpose matrix
   d. `Matrix GetSubVectorbyColumn(int column)` : return a specific column as a $n \times 1$ matrix.
   e. `void Print()` : print the values of the matrix
   f. `int GetNumRow()` : return the number of rows
   g. `int GetNumColumn()` : return the number of columns
   h. `double GetVal(int x, int y)` : return the value in `(x, y)` of the matrix. Note that `x` and `y` start from zero.
3. Implement the following function:
   a. `Matrix Cor(Matrix &mat, int method = 1)` : calculate the Pearson correlation (`method = 1`) or Kendall's tau (`method = 2`) of the all pairs of the columns. Returned matrix should be $m \times m$ matrix if the input matrix is $n \times m$. Diagonal elements of the returned matrix should be equal to 1. When calculating the Kendall's tau, treat ties as concordant pairs. (i.e., treat (i, j) as a concordant pair when $x_i = x_j$ or $y_i = y_j$)
   b. `Matrix SimpleLinearRegression(Matrix &X, Matrix &Y)` : fit the simple linear regression of `X` and `Y`. `X` and `Y` should be $n \times 1$ matrices. Note that simple linear regression includes the intercept. So the returned value should be 2×1 `Matrix` of the coefficients.

Implementation:

Add all code related to class and function definition to `Matrix.h` and all implementation to `Matrix.cpp`. `main.cpp` have a test code.

`Makefile` is provided to help compilation and job scheduling. Please compile and run it at GSDS server. You can run test code by running "`make test`". You can check your result using data in `coris.txt`. Below is the results.

```
(base) cfdsta@login0:~/HW3$ make test
g++    -c -o main.o main.cpp
g++    -c -o Matrix.o Matrix.cpp
g++ -o matrix main.o Matrix.o
salloc --nodes=1 --ntasks-per-node=1 --time=5 --cpus-per-task=1 --mem=1G ./matrix
salloc: Granted job allocation 33752
Matrix M^T M:
9.03372e+06 306037 1.64988e+06
306037 12358.5 58913.1
1.64988e+06 58913.1 326131
Pearson:
1 0.158296 0.3565
0.158296 1 0.440432
0.3565 0.440432 1
Kendall:
1 0.161732 0.277009
0.161732 1 0.334535
0.277009 0.334535 1
Pearson + Kendall:
2 0.320028 0.633509
0.320028 2 0.774967
0.633509 0.774967 2
Pearson - Kendall:
0 -0.00343565 0.0794908
-0.00343565 0 0.105896
0.0794908 0.105896 0
Simple Linear Regression Output:
130.9
1.5667
salloc: Relinquishing job allocation 33752
```

2 Submission Instruction

- Compress `Matrix.h` and `Matrix.cpp` as a single file and report it to ETL.
- You cannot change `main.cpp` and `Makefile`. You don't need to submit them.
- The file name you submit should be `YOUR_ACCOUNT_HW03.zip`. (ex: `cfds123_HW03.zip`)
- **Make sure your code works well on the GSDS server.** Your code will be scored automatically by the program on the GSDS server. If you don't follow the submission instruction, a penalty may occur.
- If you want to use your grace day, you must notify the TA by e-mail (`kisung.nam@snu.ac.kr`) when submitting the homework. If you don't notify, we will judge that you want to save your grace day for the next homeworks, so your homework is considered unsubmitted. Even if you use your grace day, your homework should be submitted through ETL.