STA141C Spring 2020 Final project

V. Budilov

6/5,hna HGVCYDDQHJMMMMMMMMMMMMMMMMM-
MMSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSJMSYYYYYYYYY JV
/2020

STA141C Spring 2020 Final project

Planned activities toward completing this project.

The activietes are suggested and may be changed/dropped/added during the project.

1. Implement parallelization. In the current implementation, only one CPU is used in the algorithm. Make it possible to use more than one CPUs. Note that you should let users to decide if they want to use parallelization.
2. Data loading functionality implementation. Allow users to specify file names to run the model rather than loading the whole data in the main process.
3. Implementing C++ substitute functions or creating new functionalities using C++. Functions are written in pure R, it is possible, for example, to convert the function lm1 to c++ code. Your might need look at how RcppArmadillo's fastLm.R and fastLm.cpp. (Spoiler, it is not easy, but if you insist, here is a some slides about it: https://scholar.princeton.edu/sites/default/files/q-aps/files/slides_day4_am.pdf)
4. Creating documentation and implementing tests. Write tests and documentations
5. Implementing more models More models? Logistic regression? GLM?

Implemented tasks.

Parallelization

1. Parallelization has been implemented. Function blblm1() is an analog of the original function blblm(). The difference of bkblm1() from blblm() is that the new function uses parallelization. bkblm1() was used for the development and testing. The final implementation of parallelization has been included into blblm().

Generalized Linear Model

2. Generalized linear model has been added to the package. Original function signature

```
blblm(formula, data, m = 10, B = 5000)
```

has been changed to

```
blblm(formula, data, m = 10, B = 5000, ifgeneral=FALSE,family = "gaussian", parallel=FALSE).
```

Logical ifgeneral parameter tells if we need to use generalized linear model, the default value is FALSE. If the generalized linear model required, we need to change ifgeneral to TRUE. The default family is ste to be "gaussian". When parallelization is required, the value of the parameter *parallel* has to be set to TRUE.

Test cases blablabla

balbla

Documentation blablabla

balbla

Results and Discussion

```
{r}
#library(nycflights13)
#flights1<-flights[1:2000,]
fit <- blblm(mpg ~ wt * hp, data = mtcars, m = 3, B = 100)
coef(fit)
fit <- blblm(mpg ~ wt * hp, data = mtcars, m = 3, B = 100, parallel = TRUE)
coef(fit)
```

Comparision between non-parallel and parallel calculations has been done using various data lengths, boostraping counts, data subdivisions, numbers of workers.

```
{r message=FALSE, warning=FALSE}
library(nycflights13)
flights1<-flights[1:100000,]
f1<-function(){blblm(arr_delay ~ distance + dep_delay, data = flights, m = 24, B = 1000)
  return (0)}
f2<-function(){blblm(arr_delay ~ distance + dep_delay, data = flights, m = 24, B = 1000,parallel=TRUE)
  return (0)}
result<-bench::mark(
f1(),
f2()
)
result
```

For six workers, m=4, and B=1000, the results are

```
## # A tibble: 2 x 6
##   expression      min   median `itr/sec` mem_alloc `gc/sec`
##   <bch:expr> <bch:tm> <bch:tm>     <dbl> <bch:byt>    <dbl>
## 1 f1()          4.76m    4.76m   0.00350     159GB  4.91
## 2 f2()          2.75m    2.75m   0.00606     213MB  0.00606
```

zorro

```
# content of the chunk above

entire fligths B=700
<S3: bench_expr>    6.48m   6.48m   0.002572499 112GB
<S3: bench_expr>    3.79m   3.79m   0.004402274 209MB
```

Some examples of usage

Default usage (corresponds to the original blbml() function)

Default usage with simple linear model, and no prallelization

```
fit <- blblm(mpg ~ wt * hp, data = mtcars, m = 3, B = 100)
coef(fit)
```

```
## (Intercept)          wt          hp       wt:hp
## 53.23084716 -8.54900595 -0.18478001  0.04183136
```

Parallelization

Usage with simple linear model, and prallelization

```
fit2 <- blblm(mpg ~ wt * hp, data = mtcars, m = 3, B = 100, parallel=TRUE)
coef(fit2)
```

```
## (Intercept)          wt          hp       wt:hp
## 48.03149289 -7.31439938 -0.12235599  0.02641407
```

Generalized linear model

Usage with generalized linear model without parallelziation

```
fit <- blblm(mpg ~ wt * hp, data = mtcars, m = 4, B = 100, ifgeneral=TRUE)
coef(fit)
```

```
## (Intercept)          wt          hp       wt:hp
## 48.43533226 -7.50246842 -0.12386731  0.02710822
```

Generalized linear model with parallelization

Usage with generalized linear model with parallelziation

```
fit <- blblm(mpg ~ wt * hp, data = mtcars, m = 5, B = 100, ifgeneral=TRUE, family='quasi', parallel=TRU
coef(fit)
```

```
## (Intercept)          wt          hp       wt:hp
##  58.7850315 -11.5058320  -0.1968608   0.0549630
```

```
#fit2 <- blbglm(mpg ~ wt * hp, data = mtcars, m = 3, B = 100)
#fit2$coef
#coef(fit2)
```